Politechnika Wrocławska

Wydział Informatyki i Zarządzania

Katedra Inteligencji Obliczeniowej

Wrocław, Wybrzeże Wyspiańskiego 27

Rozprawa doktorska

Metoda pogłębianej analizy semantycznej w zadaniu wykrywania relacji semantycznych między fragmentami tekstu w języku polskim

mgr inż. Paweł Kędzia

Promotor: dr hab. inż. Urszula Markowska-Kaczmar, prof. nadzw.

Promotor pomocniczy: dr inż. Maciej Piasecki

Wrocław, 2018

Spis treści

Streszczenie

Tworzenie abstraktów dokumentów, maszynowe tłumaczenia, odpowiedzi na pytania to tylko kilka przykładów zastosowań przetwarzania języka naturalnego (ang. *Natural Language Processing – NLP*), które wiążą się z analizą zdania np. w celu określenia jego struktury oraz ekstrakcją zawartych w nim informacji. Najbardziej efektywnym rozwiązaniem byłoby zastosowanie głębokiego parsera, który daje pełną reprezentację semantyczną zdania. Cechuje się dużą dokładnością, ale wolnym przetwarzaniem. Biorąc pod uwagę fakt, że w języku polskim występuje swobodny szyk zdania oraz bogata fleksja, opracowanie głębokiego parsera, który posiadałby wysoką kompletność analizowanych zdań, jest jeszcze trudniejsze i bardziej długotrwałe niż w językach niefleksyjnych. Dlatego często stosuje się jedynie płytki parser, który oznacza jedynie elementy zdania, ale nie wyszczególnia wewnętrznej struktury ani ich roli w zdaniu. W rozprawie zaproponowano podejście alternatywne, które polega na przyrostowym wydobywaniu coraz dokładniejszych informacji dotyczących zdań w przetwarzanym tekście. Informacje pozyskiwane są na podstawie modelowania zależności semantycznych wewnątrz tekstu oraz dołączania do tekstu wiedzy pozatekstowej, to znaczy takiej, która nie jest bezpośrednio reprezentowana w przetwarzanym tekście. Zakłada też łączenie tak pozyskanej wiedzy w celu otrzymania dokładniejszej reprezentacji badanego tekstu. Tę ideę zaimplementowano w metodzie *pogłębianej analizy semantycznej PAS*. Obejmuje ona przetwarzanie wstępne tekstu, formalizację tekstu do postaci grafów, dołączanie informacji pozyskanej z zewnętrznych źródeł wiedzy, takich jak Słowosieć i ontologia wysokiego poziomu SUMO, do zbioru już istniejących informacji. Wszystkie te operacje realizowane są na strukturach grafowych przyjętych jako formalna reprezentacja wiedzy w metodzie *PAS*.

Przy projektowaniu metody starano się o niezależność jej działania od języka analizowanych tekstów, jednak w rozprawie skoncentrowano się jedynie na tekstach napisanych w języku polskim. Metoda została wykorzystana w zadaniu wykrywania relacji semantycznych między fragmentami tekstów. W rozprawie zaproponowano nowe cechy opisujące relacje między fragmentami tekstów. Wykorzystują one podobieństwo grafów reprezentujących informacje z tekstów, pomiędzy którymi zachodzi dana relacja. Zadanie rozpoznawania relacji między fragmentami tekstów jest rozumiane jako zadanie klasyfikacji. Każdy fragment tekstu reprezentowany jest za pomocą wektora cech. Jest on podawany na wejście klasyfikatora, którego zadaniem jest rozpoznanie jednej z 17 relacji semantycznych odnoszących się do porównywanych fragmentów tekstów. Skuteczność proponowanego rozwiązania została zbadana na drodze eksperymentalnej i porównana z efektami klasyfikacji uzyskanymi przy użyciu cech dotychczas stosowanych i opisanych w literaturze. Przeprowadzona analiza statystyczna wyników eksperymentalnych pokazała, że nowe cechy, uzyskane dzięki wykorzystaniu metody *PAS*, w sposób statystycznie istotny poprawiają jakość klasyfikacji w stosunku do cech powszechnie stosowanych w literaturze. Poszczególne etapy przetwarzania użyte w metodzie, zostały opracowane jako narzędzia do przetwarzania języka, dostępne na zasadach otwartej licencji GPL. Etapy te, były realizowane w ramach projektu CLARIN-PL[1] oraz w ramach zadania *A13* projektu *SyNaT*[2].

[1] CLARIN (Common Language Resources & Technology Infrastructure) – Ogólnoeuropejska infrastruktura naukowa – umożliwia badaczom z dziedziny nauk humanistycznych i społecznych wygodną pracę z bardzo dużymi zbiorami tekstów.

[2] **S**ystem **Na**uki i **T**echniki – `http://synat.pl/`

Większość opracowanych narzędzi dostępna jest również w postaci usług sieciowych (ang. *web-service*).

Spis rysunków

Spis tablic

Rozdział 1

Wstęp

W czasach mediów społecznościowych coraz więcej informacji gromadzonych jest w postaci elektronicznej. Informacje te mogą być w postaci obrazowej, dźwiękowej lub tekstowej a największa ilość opublikowanych, zgromadzonych i wyszukiwanych informacji znajduje się w internecie. Ręczne przeszukiwanie i przetwarzanie takich dużych zasobów staje się niemożliwe. Dlatego naturalnym dążeniem wydaje się automatyzacja przetwarzania i wyszukiwania informacji. Niniejsza praca koncentruje się na przetwarzaniu i analizie informacji tekstowej, wyrażonej w języku naturalnym, którym zajmuje się przetwarzanie języka naturalnego (NLP). Dziedzina ta stanowi połączenie metod sztucznej inteligencji i językoznawstwa. Ma ona na celu między innymi automatyczne wyszukiwanie, przetwarzanie i analizę informacji wyrażonej w języku naturalnym. Dążymy do tego, aby komputery były zdolne do automatycznego generowania pytań i odpowiedzi na nie w języku naturalnym, bez udziału człowieka. W tym kontekście zadanie to wymaga rozumienia języka i automatycznego wnioskowania. Wymaga też odpowiedniej dla komputera reprezentacji tekstu wyrażonego w języku naturalnym, gdyż komputerowe przetwarzanie nieustrukturalizowanych danych nie jest możliwe.

Człowiek w ciągu tysięcy lat w drodze nabytego doświadczenia nauczył się rozumieć język naturalny. Czuje emocje czytając tekst w języku naturalnym i często potrafi wizualizować jak dana rzecz mogłaby wyglądać w rzeczywistości. Komputery są w stanie przetwarzać dane o pewnej strukturze znacznie szybciej niż robią to ludzie, jednak niejednoznaczność i nieprecyzyjność języka naturalnego powoduje, że NLP jest trudne do implementacji. Aby to było możliwe należy zidentyfikować i wyekstrahować reguły językowe, by można było przełożyć nieustrukturalizowaną formę języka na postać zrozumiałą dla komputera. Stosowane metody często są oparte o algorytmy uczenia maszynowego, dzięki czemu uczą się reguł językowych poprzez analizowanie zbiorów przykładów. Jednak wymagają one dużego zbioru tekstów, takich jak książki, czy kolekcje zdań, których wiele przykładów można znaleźć w internecie. Im więcej danych jest analizowanych, tym dokładniejszy jest model budowany przez aplikację.

Przetwarzanie języka naturalnego służy do rozwiązywania takich zadań, jak na przykład: automatyczne tłumaczenia, automatyczne streszczanie tekstów, generowanie słów kluczowych, analiza nastawienia wypowiedzi czy też automatyczne generowanie odpowiedzi na pytania. W procesie wyszukiwania odpowiedzi na pytania mogą okazać

się przydatne zależności między fragmentami tekstu. Mogą mówić o tym, że z jednego fragmentu tekstu wynika inny (ang. *Textual entailment*) lub że pomiędzy tymi fragmentami tekstów zachodzi relacja semantyczna, w której jeden fragment jest *uszczegółowieniem* innego albo że dwa fragmenty tekstów opisują ten sam temat, lecz podają sprzeczne informacje (*sprzeczność*).

Zadanie rozpoznawania relacji semantycznych między dwoma porównywanymi tekstami jest problemem, na którym koncentruje się niniejsza praca w kontekście NLP. Rozpoznawanie tych relacji zostało w pracy sprowadzone do problemu klasyfikacji, w którym jakość rozpoznawania zależy od zdolności dyskryminacyjnych wektora cech. W pracy zaproponowano nowe cechy, których opracowanie było możliwe dzięki autorskiej metodzie *PAS – **P**ogłębianej **A**nalizy **S**emantycznej*, służącej do stopniowego wydobywania coraz dokładniejszych informacji z analizowanego tekstu. Informacje te mogą być opisywane na różnych poziomach abstrakcji, np. składniowym lub semantycznym. Oryginalnym rozwiązaniem zaimplementowanym w tej metodzie jest również opracowanie zapisu formalnego analizowanego tekstu i sposobu dołączania do tej reprezentacji ekstrahowanych z tekstu informacji. Do elementów nowatorskich należy również zaliczyć zaproponowaną miarę określającą podobieństwo porównywanych tekstów. Metoda PAS oraz zaproponowana metryka zostały zaimplementowane w systemie `grafon`, który posłużył do przeprowadzenia badań przedstawionych w części eksperymentalnej.

1.1. Przetwarzanie języka naturalnego

Proces przetwarzania języka naturalnego jest wieloetapowym procesem, który w zależności od zastosowania może być różnie realizowany, poczynając od wyróżnienia w tekście poszczególnych znaków, wyrazów, zdań, od analizy morfologicznej form wyrazowych, poprzez analizę struktur wyrażeń i większych całostek, aż po kontekstową analizę ich znaczeń. Można wyróżnić kilka etapów przetwarzania języka naturalnego wspólnych dla wielu zastosowań, które w większości odpowiadają tradycyjnym dla językoznawstwa poziomom opisu języka. Można tu wymienić etap:

— *segmentacji*, który obejmuje podstawowe czynności związane z podziałem tekstu na poszczególne elementy takie jak tokeny – w tym wyrazy, zdania czy akapity.

— *morfologiczny* – na którym formy wyrazowe są odróżniane od niewyrazowych tokenów, a następnie są poddawane analizie w zakresie informacji gramatycznych i słowotwórczych wyrażanych poprzez nie. Określane są własności morfo-syntaktyczne, takie jak: *lemat* (inaczej podstawowa forma morfologiczna[1]), części mowy (lub dokładniej, klasy gramatyczne), powiązane z nimi zespoły kategorii gramatycznych (np. przypadek, liczba czy też rodzaj) oraz ich możliwe wartości. Często do tego poziomu włącza się również ujednoznacznienie tych własności dla konkretnych wystąpień form wyrazowych w tekście, tj. ujednoznacznianie morfo-syntaktyczne[2], (np. następuje podjęcie decyzji czy słowo „piec" odnosi się do czynności „piecze-

[1] Arbitralnie wyznaczona forma wyrazowa o określonych wartościach kategorii gramatycznych, np. mianownik liczby pojedynczej dla rzeczowników, która reprezentuje zbiór form wyrazowych różniących się co do wartości kategorii, ale o tym samym znaczeniu leksykalnym.

[2] Można to również wyodrębnić jako osobny etap analizy morfo-syntaktycznej, który jest pośredni pomiędzy morfologicznym a syntaktycznym.

nia" – a więc jest to czasownik, czy też dotyczy bytu jakim jest „piec" – czyli rzeczownika).

— *syntaktyczny* – w ramach którego określane są zależności składniowe na poziomie wyodrębnionych wyrazów, fraz, całostek zdań itd. Kwestia rodzaju i wielkości wyodrębnianych elementów jest również określana w modelu przyjętym na potrzeby opisu. Rozpoznawana jest struktura składniowa wyrażeń i zdań w ideale w sposób abstrahujący od znaczenia wyrazów i większych struktur.

— *semantyczny* obejmuje analizę semantyczną wypowiedzi w języku naturalnym na różnych poziomach ich wewnętrznej budowy. Bardzo często analiza semantyczna jest silnie powiązana z analizą syntaktyczną, a co za tym idzie również rozpoznawane struktury budowy semantycznej są powiązane ze strukturami syntaktycznymi – zasada kompozycyjności. Etap ten obejmuje również analizę znaczeń słów (wystąpień lematów) w tekście , wyrażeń językowych i zdań (np. relacje semantyczne, forma logiczna, struktura predykatowa, itp.), jak również struktury semantycznej wypowiedzi (np. zjawiska anafory i koreferencji). Na poziomie leksykalnym kluczowe jest ujednoznacznienie znaczeń leksykalnych słów (tj. wystąpień lematów) poprzez ich odniesienie do przyjętego repozytorium znaczeń lub przestrzeni znaczeń (np. ciągłej, wektorowej w modelu nienadzorowanej indukcji znaczeń) (np. rozróżnianie między „zamkiem" w kurtce a „zamkiem" w drzwiach). Specyficznym osobnym zagadnieniem jest rozpoznawanie i klasyfikacja semantyczna wystąpień *nazw własnych* (szerzej jednostek identyfikujących), (np. *Zakład Ubezpieczeń Społecznych* jako nazwa własna organizacji oraz potencjalnie powiązany z reprezentacją obiektu w bazie wiedzy). Na najwyższym poziomie analiza semantyczna obejmuje również rozpoznawanie relacji pomiędzy zdaniami oraz fragmentami tekstu, np. relacji wynikania semantycznego lub sprzeczności.

— *pragmatyczny* odnosi się do analizy wszystkich zjawisk związanych z użyciem wyrażeń językowych jako wypowiedzi w określonym kontekście (nadawca, odbiorca wraz z ich cechami, wiedzą, celami działania, wpływem chwilowego kontekstu, kultury, społeczeństwa itd.). Analiza na etapie pragmatyki może obejmować kompleksowe zjawisko *analizy dyskursu* (często wyróżnianej jako osobny etap), gdzie rozpoznawane są wypowiedzi różnych nadawców jako składowe całości. Analiza na etapie pragmatycznym często jest również sprowadzana do rozwiązywania szeregu prostszych zagadnień częściowych, które umożliwiają częściowe wydobywanie informacji (i później wiedzy) z wypowiedzi językowych, jak rozwiązywanie niejednoznaczności w odniesieniach koreferencyjnych i anaforycznych (częściowo problem z obszaru semantyki, ale możliwy do pełnego rozstrzygnięcia jedynie z wykorzystaniem wiedzy o kontekście interpretacji), rozpoznanie odniesień do sytuacji (zdarzeń, procesów, stanów) wraz z ich powiązaniem z kontekstem interpretacji, czy rozpoznaniem odniesień do czasu i przestrzeni. Wszystkie te zjawiska mogą być w dużej mierze rozpoznawane już na etapie analizy semantycznej, ale ich pełna interpretacja w relacji do bazy wiedzy o świecie wchodzi już w zakres etapu analizy pragmatycznej.

Dla każdego tradycyjnego etapu przetwarzania zaproponowano wiele *narzędzi językowych* (programów komputerowych), które realizują poszczególne zadania szczegółowe. Niektóre typy narzędzi językowych na tyle przyjęły się w procesach przetwarzania, że możemy mówić o *podstawowych narzędziach językowych*, podobnie często stosowane typy *zasobów językowych* (sformalizowane opisy wiedzy o języku naturalnym na róż-

Tabela 1.1. Wynik działania taggera WCRFT dla zdania „Kazał kurze ścierać kurze" (Radziszewski, 2013)

Kazał	kurze	ścierać	kurze
kazać	kura	ścierać	kurz
praet:sg:m1:perf	subst:sg:dat:f	inf:imperf	subst:pl:acc:m3

nych poziomach jego opisu), wykorzystywanych na różnych etapach przetwarzania, określane są mianem *podstawowych zasobów językowych*. Ograniczając się w dalszych rozważaniach w tym wprowadzeniu do tekstu oraz zakładając, że dysponujemy już tekstem pozyskanym z pierwotnego źródła (np. wydobytym z dokumentu webowego lub rozpoznanym z postaci drukowanej), *programy do segmentacji tekstu* są stosowane na początku procesu przetwarzania, aby podzielić tekst na segmenty różnego poziomu granulacji: tokeny (poziomu wyrazowego, np. wyrazy, symbole, liczby, daty itp.), zdania (dokładniej segmenty poziomu zdań, np. również zwroty w dialogu) oraz ewentualnie inne o większej granulacji (np. tytuły, śródtytuły, akapity itp.). Narzędzia do segmentacji mogą być z dużym powodzeniem oparte na wyrażeniach regularnych, np. dla języka polskiego *Toki* (A configurable tokeniser) (Radziszewski i Śniatowski, 2011), który wykorzystuje *reguły SRX Marcina Miłkowskiego* (Miłkowski i Lipski, 2009). Oczywiście siła ekspresji wyrażeń regularnych oraz bardzo ograniczona wiedza o tekście dostępna na tym etapie przetwarzania powoduje, że w niektórych przypadkach nie można podjąć właściwej decyzji, np. w przypadku użycia kropki jako znaku o dwóch rolach w tekście: elementu skrótu i znaku interpunkcyjnego, tzw. haplologii kropki (Przepiórkowski, 2004), ale stanowią one bardzo małą część wszystkich decyzji jakie ma podjąć tzw. segmenter.

Na etapie analizy morfologicznej, zwykle najpierw jest stosowany *analizator morfologiczny*, który przypisuje do rozpoznanych form wyrazowych wszystkie możliwe analizy morfologiczne, obejmujące lemat, klasę gramatyczną oraz wartości kategorii gramatycznych. Następnie, *tager* – program do ujednoznaczniania morfo-syntaktycznego – wybiera analizy właściwe dla kontekstu danego wystąpienia formy wyrazowej, najlepiej jedną analizę (w niektórych przypadkach nie jest to możliwe). Dla języka polskiego bardzo często stosowanym analizatorem jest *Morfeusz* (Woliński, 2006). W tabeli 1.1 przedstawiony został wynik działania narzędzia (tagera) ujednoznaczniającego morfo-syntaktycznie teksty języka polskiego, w tym przypadku *WCRFT* (Wrocław CRF Tagger) (Radziszewski, 2013), dla zdania „Kazał kurze ścierać kurze". Pierwszy wiersz pokazuje słowo w formie występującej w tekście, wiersz drugi, to forma podstawowa słowa, wiersz trzeci, to określenie części mowy z dodatkowymi informacjami morfologicznymi dla słowa. Słowo „Kazał" posiada formę podstawową „kazać" oraz jest to pseudoimiesłów (`praet`) o formie podstawowej bezokolicznikowej, liczbie pojedynczej (`sg`) mówiący o rodzaju męskim (`m1`) i aspekcie dokonanym (`perf`). Każde słowo posiada przypisaną formę podstawową oraz tak zwany tag w określonym tagsecie (zbiorze dostępnych tagów możliwych do przypisania). W przytoczonym przykładzie wykorzystany został tagset *Narodowego Korpusu Języka Polskiego* (*nkjp*)[3]. WCRFT został zastosowany w większości eksperymentów opisanych w pracy ze względu na

[3] Opis wszystkich tagów dostępny jest na stronie `http://nkjp.pl/poliqarp/help/ense2.html`. Ostatni dostęp dnia 16.08.2018

jego wysoką efektywność oraz łatwość integracji z innymi wykorzystywanymi narzędziami. Warto jednak dodać, że dla języka polskiego zbudowano kilka bardzo dobrych tagerów, z których wiele wykazuje się większą dokładnością działania, np. PANTERA (Acedański i Gołuchowski, 2009, Acedański, 2010) bazujący na mechanizmie uczenia się systemów prostych reguł zaproponowanym przez Brilla (Brill, 1992, Rychlikowski i inni, 2017) wykorzystujący sieci głębokie i modele znakowe, (Krasnowska-Kieraś, 2017, Wróbel, 2017) oparte na dwukierunkowych sieciach LSTM (dające najlepszy wynik w momencie składania niniejszej pracy, ale też o największej złożoności pamięciowej i obliczeniowej, trudne do zastosowania na czystym tekście w połączeniu z Morfeuszem) oraz (Kobyliński i inni, 2018) łączący kilka tagerów w zespół (przez to również bardzo wymagający pod względem złożoności).

Wyniki etapu analizy morfo-syntaktycznej są często wejściem do analizy syntaktycznej, w ramach której wyrażeniom językowym zostaje przypisana sformalizowana reprezentacja ich struktury. Dla języka polskiego jednym z narzędzi łączącym dobrą skuteczność z wysoką efektywnością jest parser Aliny Wróblewskiej[4] (Wróblewska, 2014) generujący reprezentację w postaci grafu zależności składniowych (rodzaju relacji składniowych). Podejmowane są również obiecujące próby łączenia analizy syntaktycznej z morfo-syntaktyczną w ramach jednego narzędzia opartego na kaskadzie głębokich sieci neuronowych – COMBO (Rybak i Wróblewska, 2018).

Jednym z fundamentalnych problemów dla etapu *semantycznego* jest przypisanie reprezentacji semantycznej dla słów (dokładniej wystąpień lematów) w tekście. Jednak, aby to zrobić najpierw trzeba ujednoznacznić (ang. *Word Sense Disambiguation* – ujednoznacznianie sensów słów) w jakim znaczeniu dany lemat pojawia się w określonym miejscu, np. „zamek" występujący w tekście może być użyty w znaczeniu *zamka w drzwiach*, zamka w *broni*, zamka – *budowli*, czy też zamka – *układu* podczas gry w hokeju. Przypisanie znaczeń do wystąpień lematów zwykle odbywa się w odniesieniu do przyjętego repozytorium znaczeń leksykalnych[5], w którym są one przynajmniej wymienione, często opisane w mniej lub bardziej sformalizowany sposób. Dla języka polskiego takim repozytorium może to być *Słowosieć* (ang. *plWordNet*)[6] (Piasecki i inni, 2009a, Maziarz i inni, 2012)), czyli bardzo duża leksykalna sieć semantyczna, która również może być postrzegana jako wielki relacyjny słownik semantyczny języka polskiego. Słowosieć została zbudowana całkowicie ręcznie przez zespół leksykografów. Proces budowy opierał się na analizie bardzo dużych korpusów języka polskiego i był wspomagany przez szereg narzędzi językowych do eksploracji danych językowych, m.in. poprzez wykrywanie potencjalnej liczby znaczeń dla danego lematu na podstawie dużego zbioru tekstów (Broda i Kędzia, 2011). Słowosieć opisuje znaczenia leksykalne poprzez bogaty zbiór relacji leksykalno-semantycznych (ponad 50 typów i blisko 120 wraz z podtypami) oraz w wersji 3.1 zapewnia opis dla 259 000 znaczeń wyróżnionych dla ponad 178 000 różnych lematów.

Dla języka polskiego były podejmowane próby opracowania narzędzi do ujednoznaczniania znaczeń słów, np. (Bas i inni, 2008) czy też (Broda i Piasecki, 2011).

[4] `http://zil.ipipan.waw.pl/PDB/PDBparser`

[5] W podejściach opartych na indukcji znaczeń leksykalnych, zbiór znaczeń jest automatycznie ustalany na podstawie dużego korpusu tekstów, ale są one zwykle nieintuicyjne i niezrozumiałe dla człowieka.

[6] `http://plwordnet.pwr.edu.pl`

Metody te jednak oparte są na maszynowym uczeniu, które wymaga wielu ręcznie oznaczonych wzorcowych tekstów z przypisanymi znaczeniami do wystąpień poszczególnych lematów. Systemy takie cechują się większą dokładnością działania w stosunku do systemów uczonych metodami słabonadzorowanymi wykazując przy tym niską kompletność. Co gorsza, proces ręcznej anotacji znaczeniami korpusu jest niezwykle pracochłonny (dziesiątki osobolat dla kilkudziesięciu tysięcy lematów) i przez to niemożliwy do zastosowania na praktyczną skalę. Dlatego też, na potrzeby niniejszej pracy zaadaptowane zostało słabonadzorowane podejście do ujednoznaczniania znaczeń leksykalnych z pracy (Agirre i Soroa, 2009), które pierwotnie zostało zaproponowane i rozwinięte dla języka angielskiego. Wykorzystuje ono jako podstawę algorytm PageRank (Altman i Tennenholtz, 2005) zastosowany do WordNetu dla języka angielskiego jako grafu powiązań pomiędzy zbiorami synonimicznych znaczeń. Algorytm PageRank został dostosowany do problemu ujednoznaczniania znaczeń leksykalnych poprzez uwzględnienie kontekstu występowania znaczeń, tzw. spersonalizowany PageRank (ang. *personalised*) (Agirre i Soroa, 2009, Gutiérrez i inni, 2012, Agirre i inni, 2013). Atutem tego podejścia jest uzyskanie wysokiej kompletności, która rozumiana jest jako liczba słów, dla których metoda może przypisać znaczenie, jednak ze słabszą precyzją w porównaniu do systemów uczonych maszynowo. W przypadku zastosowania dla języka polskiego w oparciu o Słowosieć jako bazę wiedzy, podejście takie umożliwia przypisanie znaczenia dla każdego lematu, którego znaczenie jest opisane w Słowosieci. Zaproponowano adaptację tej metody, tj. *personalizowanego PageRanku* (PPR), która została dostosowana do języka polskiego, m.in. poprzez rozszerzenia bazy wiedzy o nowe źródła wiedzy, spoza wordnetu (Kędzia i inni, 2014a, 2015).

Nazwy własne funkcjonują poza leksykonem języka naturalnego i dlatego problem wykrywania ich wystąpień w tekście oraz ich kategoryzacji jest odrębny w stosunku do ujednoznaczniania znaczeń leksykalnych. Dla języka polskiego istnieją już metody oraz narzędzia do wykrywania i kategoryzacji wystąpień nazw własnych: (Walas i Jassem, 2010, Marcinczuk, 2015, Smywiński-Pohl, 2013, Waszczuk i inni, 2013) o dobrej jakości działania. Jednakże, ze względu na łatwość integracji, dużą efektywność obliczeniową oraz dokładność działania dla wielu kategorii nazw, w pracy wykorzystano podejście zaprezentowane w (Marcinczuk, 2015).

Problemem z pogranicza etapów przetwarzania *semantycznego* oraz *pragmatycznego* jest streszczanie tekstów. Powszechnie wykorzystywane są dwa podejścia w procesie automatycznego streszczania tekstów: opisowe oraz ekstrakcyjne. Podejście opisowe polega na generowaniu zdań opisujących w sposób skondensowany zawartość analizowanego tekstu. Podejście ekstrakcyjne polega na wybieraniu z analizowanego tekstu tych zdań, które niosą ze sobą największą wartość informacyjną. Wymaga to określenia w pewnym stopniu relacji semantycznej pomiędzy zdaniem a całością tekstu. Dla języka polskiego zaproponowano jedynie kilka podejść w tej dziedzinie, np. w pracy (Jassem i Pawluczuk, 2015) przedstawiono metodę służącą do ekstrakcyjnego streszczania tekstów newsowych napisanych w języku polskim.

Większość narzędzi językowych jest konstruowana w oparciu o metody maszynowego uczenia. Wymaga to uprzedniego zbudowania zasobów językowych – korpusów anotowanych ręcznie. W przypadku etapów semantycznego i pragmatycznego konstrukcja takich korpusów staje się szczególnym wyzwaniem pod względem koncepcji i nakładów pracy. Budowę zasobów i narzędzi łączy się często w kompleksowe prace

badawcze. Opracowane zasoby mogą posłużyć do uczenia reguł lub opracowywania miar wykrywających terminologię, np. w pracach (Pazienza i inni, 2005, Marciniak i Mykowiecka, 2014) przedstawiono proces budowy narzędzi do ekstrakcji terminologii z tekstów medycznych.

W procesie przetwarzania języka naturalnego, szczególnie na etapie semantycznym, ale też pragmatycznym, bardzo ważną rolę mogą pełnić *parsery semantyczne*. Celem jego działania jest wygenerowanie dla wyrażeń językowych ich szczegółowej, sformalizowanej reprezentacji semantycznej względem przyjętego modelu opisu i często też bazy wiedzy o kontekście interpretacji (aspekt pragmatyczny). Narzędzia takie rozwiązują między innymi problemy relacji semantycznych między słowami i wyrażeniami w analizowanych zdaniach. W większości przypadków parsery semantyczne umożliwiają przedstawienie znaczeń wyrażeń językowych, w tym zdań, za pomocą sformalizowanej reprezentacji semantycznej opartej często na wyrażeniach logicznych w ramach podzbioru logiki predykatów pierwszego rzędu. Dla języka polskiego zaproponowany został w ostatnich latach parser ENIAM [7] (Jaworski i Kozakoszczak, 2016), który generuje częściową reprezentację semantyczną tekstu oraz jest w stanie przypisać do części słów odpowiadające im znaczenia ze Słowosieci. W procesie analizy ENIAM wykorzystuje wiedzę o strukturach predykatowych języka polskiego wyrażoną w słowniku ram walencyjnych – *Walentego* (Przepiórkowski i inni, a,b, 2014). Istotnym ograniczeniem parsera jest niska kompletność względem możliwości przetworzenia zdań w praktycznym tekście – dla wielu zdań nie wygeneruje wyników ze względu na występujące w nich błędy językowe lub też nierozpoznane konstrukcje syntaktyczne. ENIAM charakteryzuje się również wysoką złożonością obliczeniową, co utrudnia jego zastosowania.

1.2. Motywacje do podjęcia tematu

Jednym z narzędzi językowych użytecznych w procesie analizy języka naturalnego jest głęboki parser semantyczny. Jak wspomniano w poprzednim rozdziale, taki parser powinien umożliwiać:

— określenie zależności semantycznych pomiędzy słowami występującymi w tekście,
— uzupełnienie informacji wyciągniętej z tekstu o dodatkową wiedzę pozatekstową (*wiedzę pozyskaną z ontologii*, w celu wzbogacenia i uzupełnienia informacji wydobytych bezpośrednio z tekstu,
— utworzenie formalnej reprezentacji wiedzy użytecznej w automatycznej analizie tekstu.

Parser głęboki tworzy pełną reprezentację logiczną, na przykład za pomocą logiki predykatów pierwszego rzędu, co umożliwia pełny ciąg wnioskowania. Biorąc jednak pod uwagę cechy języka polskiego, takie jak swobodny szyk zdania, bogatą fleksję i niejednoznaczności semantyczne słów, opracowanie takiego parsera jest procesem bardzo trudnym, żmudnym i długotrwałym.

Opracowana metoda **PAS** służąca do analizy semantycznej, ma w pewnym stopniu naśladować efekt działania głębokiego parsera i realizować przedstawione wyżej wymagania. Wzbogaca ona informacje wydobyte z tekstu o wiedzę pozatekstową pozyskaną z

[7] `https://clarin-pl.eu/dspace/handle/11321/264`

innych źródeł wiedzy. Daje też możliwość modelowania częściowej informacji wydobytej z tekstu.

1.3. Cel pracy i zadania badawcze

Uwzględniając przedstawioną motywację do podjęcia tematu, cel pracy można zdefiniować jako:

Opracowanie metody *pogłębianej analizy semantycznej*, która w sposób przyrostowy wydobywa wiedzę z tekstu, użyteczną w wykrywaniu relacji semantycznych między fragmentami tekstów (ang. *chunks*) w języku polskim. Zakłada się, że użycie metody powinno poprawić jakość wykrywania relacji w stosunku do wyników metod opisywanych w literaturze.

Opracowanie metody *pogłębianej analizy semantycznej* wymagało realizacji następujących zadań:

1. **Opracowanie płytkiego parsera semantycznego wykrywającego role semantyczne w obrębie frazy nominalnej.**
   Zadanie to wiąże się z rozszerzeniem anotacji nadawanych przez płytki parser składniowy IOBBER, zaprezentowany w pracach (Radziszewski i inni, 2012, Radziszewski i Pawlaczek, 2013), o zależności składniowe między słowami i z przypisaniem do zależności składniowych informacji o semantycznych zależnościach słów. Opis metody przedstawiony został w rozdziale 4.2.2.
2. **Adaptacja do języka polskiego metody do ujednoznacznia znaczeń słów.**
   To zadanie wymagało rozpoznania literaturowego podejść do wykrywania znaczeń leksykalnych słów i wyboru jednego z nich. Zaadaptowano do języka polskiego metodę opartą o nienadzorowane ujednoznacznianie znaczeń leksykalnych. Jako zasób znaczeń przyjęto Słowosieć. Opis zaadaptowanego podejścia znajduje się w rozdziale 4.2.2
3. **Opracowanie metody oraz narzędzia do rzutowania Słowosieci na ontologię SUMO.**
   Wykorzystanie do analizy tekstu opisu z użyciem pojęć ontologii wysokiego poziomu, jaką jest SUMO (Niles i Pease, 2004), daje możliwość uogólniania wiedzy wywiedzionej z tekstu. Dlatego kolejnym zadaniem było opracowanie metody umożliwiającej rzutowanie Słowosieci na pojęcia ontologii SUMO. Podczas rzutowania wykorzystano istniejące już rzutowania zbudowane ręcznie: Princeton WordNet na SUMO oraz Słowosieci na Princeton WordNet. Proces rzutowania oraz przykładowe reguły rzutujące znaczenia ze Słowosieci na pojęcia SUMO przedstawiony został w rozdziale 4.2.3.
4. **Opracowanie metody budowania grafów z analizowanego tekstu.**
   Metoda ***P****ogłębianej* ***A****nalizy* ***S****emantycznej PAS* reprezentuje informacje wydobyte z tekstu za pomocą grafów. Dlatego pierwszym krokiem było opracowanie metody przekształcającej dowolny tekst w jego reprezentację grafową. Graf taki może być dalej porównywany z innym, reprezentującym inny tekst (fragment tekstu) lub można do niego dołączać inny graf reprezentujący wiedzę pozyskaną z innych źródeł.

5. **Opracowanie metody wycinania spójnych podgrafów.**
   Metoda *PAS* umożliwia dołączanie wiedzy pozatekstowej do wiedzy wydobytej z tekstu. Zarówno wiedza tekstowa, jak i pozatekstowa reprezentowane są w postaci grafów. W celu dołączenia wiedzy pozatekstowej z dostępnych zasobów wiedzy pozatekstowej, należy odpowiedni fragment grafu reprezentujący elementy wiedzy z tekstu wydobyć w jak najkrótszym czasie. Problem wycinania minimalnego spójnego podgrafu jest problemem NP-trudnym i złożonym obliczeniowo. Dlatego w pracy ograniczono się do opracowania metody umożliwiającej wycinanie spójnych podgrafów, dzięki czemu możliwe jest ich znalezienie w krótszym czasie.
6. **Opracowanie metody łączenia wiedzy pozatekstowej z wiedzą tekstową**. Po wycięciu grafów spójnych z dostępnych źródeł wiedzy, kolejnym krokiem było połączenie ich z wiedzą wydobytą z tekstu. Zadanie sprowadzało sie do opracowania metody łączenia wiedzy z zewnętrznych źródeł wiedzy z wiedzą wydobytą z tekstu, zapisaną w grafie.

Następujące zadania badawcze pozwoliły na weryfikację osiągnięcia celu pracy:

1. Określenie jakości rozpoznawania relacji semantycznych między fragmentami tekstów przy użyciu cech zaproponowanych w literaturze. Osiągnięte wyniki stanowią **poziom odniesienia** dla oceny wyników z użyciem opracowanej metody *PAS*.
2. Wykorzystanie metody pogłębianej analizy semantycznej do **reprezentacji wiedzy z tekstu**, **pozyskania nowej wiedzy** oraz zbudowanie nowych cech opartych o reprezentację grafową. Określenie jakości rozpoznawania relacji semantycznych między fragmentami tekstów przy użyciu takiego wektora cech.
3. **Porównanie wyników** i przeprowadzenie analizy statystycznej,

1.4. Struktura pracy

Niniejsza praca składa się z siedmiu rozdziałów. Rozdział 1 to wstęp do pracy, w którym przedstawiono problematykę poruszaną w rozprawie. Tutaj też zdefiniowano cel pracy i zadania badawcze. W rozdziale 2 przedstawiono podstawowe pojęcia, które dotyczą różnych sposobów reprezentacji wiedzy oraz scharakteryzowano zasoby wiedzy – Słowosieć oraz ontologię SUMO. Rozdział 3 stanowi przegląd literaturowy w zakresie wykrywania relacji semantycznych między fragmentami tekstów oraz różnych miar ich podobieństwa. Warto podkreślić, że tutaj też przedstawiono nową miarę $CTXBowSim$, za pomocą której możliwe jest określanie podobieństwa dwóch grafów, przy użyciu podobieństw otoczeń węzłów posiadających takie same identyfikatory w obu grafach. W rozdziale 4 opisano autorską metodę pogłębianej analizy semantycznej – *PAS*. Opracowana metoda umożliwia modelowanie wiedzy wyciągniętej z dowolnego tekstu za pomocą grafów. Rozdział 5 zawiera opis problemu wykrywania relacji semantycznych jako zadania klasyfikacji. Znajduje się tu również opis zbioru danych, krótki opis wybranych klasyfikatorów wykorzystanych w badaniach oraz opis cech dla tych klasyfikatorów. Rozdział 6 to badania przeprowadzone w celu oceny proponowanego rozwiązania. Badania przeprowadzone zostały dla wszystkich zbiorów cech z uwzględnieniem wszystkich klasyfikatorów. Uzyskana jakość klasyfikacji została porównana z poziomem odniesienia. Sprawdzono także istotność statystyczną otrzymanych wyników. Rozdział 7, to podsumowanie realizacji całej pracy. Zawiera opis najważniejszych osiągnięć oraz

perspektywę rozwoju metody. Całość uzupełnia wykaz symboli oraz oznaczeń we wzorach. Przedstawiono również definicje najistotniejszych terminów używanych w pracy.

Rozdział 2

Reprezentacja wiedzy

Kluczowym zagadnieniem dla opracowanej metody był wybór *reprezentacji wiedzy* wydobytej z tekstu. W niniejszym rozdziale przedstawiono przegląd różnych reprezentacji wiedzy oraz uzasadniono wybór grafów na potrzeby metody *pogłębianej analizy semantycznej PAS*.

Zgodnie z definicją w „Nowej encyklopedii powszechnej" PWN (enc, 2017), *wiedza* to:

> „w najogólniejszym sensie rezultat wszelkich możliwych aktów poznania; w węższym znaczeniu – ogół wiarygodnych informacji o rzeczywistości wraz z umiejętnością ich wykorzystywania;",

informacja zaś, według tego samego źródła (enc, 2017) definiowana jest jako:

> „obiekt abstrakcyjny, który może być w postaci zakodowanej zapisywany (na nośnikach informacji), przesyłany, przetwarzany za pomocą programów komputerowych i używany do sterowania urządzeniami. "

Upraszczając nieco przytoczone definicje, przyjmiemy dalej, że *informacja* to fakty o czymś/kimś, a *wiedza* to umiejętność wykorzystania tej informacji, np. poprzez abstrahowanie, uogólnianie, czy też konstrukcję reguł. Tekst jako zapis wypowiedzi w języku naturalnym może być zarówno nośnikiem informacji jak i wiedzy. Informacja w tekście może być wyrażona na różnych poziomach analizy języka naturalnego: morfologicznym (np. część mowy, przypadek, rodzaj lub użycie wielkiej/małej litery), składniowym, semantycznym (np. znaczeń słów) oraz pragmatycznym (np. akty mowy i efekty ich użycia – zamierzone i faktyczne). Informacja taka częściowo jest wyrażona bezpośrednimi środkami, np. użycie wielkiej litery, ale w znaczącej większości jest wyrażona nie wprost i wymaga daleko idącej kontekstowej interpretacji wyrażeń językowych, np. jedynie na podstawie kontekstu danego słowa możemy wnioskować o jego znaczeniu, które z kolei jest możliwe do reprezentacji jedynie względem przyjętego zbioru wyróżnianych znaczeń. Opierając się na przyjętym modelu struktur reprezentacji informacji, możemy łączyć elementy informacji reprezentowanej w tekście w większe agregaty, np. mając informację o *kocie* w jednym zdaniu, a informację o *ssakach* w innym, wiedząc, na podstawie leksykonu semantycznego że *kot* jest *ssakiem*, można odnieść informacje dotyczące *ssaków* również do *kotów*, ale zablokować takie automatyczne powiązanie w

drugą stronę, jako że informacje dotyczące *kotów* niekoniecznie muszą dotyczyć *ssaków*. W ten sposób, poprzez daleko idącą interpretację informacji opartą na przyjętych modelach przechodzimy od informacji wydobytej z tekstu do wiedzy. Efektywne wykorzystanie wiedzy zawartej w tekście, obrazie czy dźwięku w aplikacji komputerowej, wiąże się z wyborem konkretnego sposobu jej reprezentacji odpowiednio dobranego do rozważanego problemu (Gärdenfors, 2004), z uwzględnieniem poziomu szczegółowości, jaki dostarcza nam konkretna metoda reprezentacji wiedzy. Reprezentacja wiedzy związana jest nie tylko z formalnym sposobem jej zapisu, ale również wydobywaniem wiedzy z tekstu i rozszerzaniem na jej podstawie bazy wiedzy.

2.1. Sposoby reprezentacji wiedzy zawartej w tekście

Załóżmy, że naszym celem jest budowa wyszukiwarki, która ma służyć do odnalezienia dokumentów, w których znajduje się potencjalna odpowiedź na pytanie zadane w skróconej, hasłowej postaci. W takim przypadku, jedną z najprostszych reprezentacji wiedzy jest **binarny model wektorowy** (Manning i inni, 2008). W modelu takim każdy z dokumentów $\mathcal{D}$ przynależących do kolekcji $\mathcal{C}$ ($\mathcal{D} \in \mathcal{C}$), w której wyszukiwane będą dokumenty, reprezentowany jest za pomocą wektora $\vec{v}(\mathcal{D})$. Elementy wektora, to informacja o występowaniu słowa w_i w dokumencie $\mathcal{D}$; $w_i \in \{0, 1\}$:

$$\vec{v}(\mathcal{D}) = \begin{bmatrix} w_1 & w_2 & w_3 & ... & w_l \end{bmatrix} \tag{2.1}$$

Wymiar l wektora $\vec{v}(\mathcal{D})$ równy jest liczbie unikalnych słów w przetwarzanej kolekcji dokumentów $\mathcal{D}$. Jeżeli budowana jest reprezentacja dla wielu dokumentów, przykładowo dla całej kolekcji $\mathcal{C}$, wtedy każdy dokument z tej kolekcji jest transformowany do wektora, którego liczba wymiarów równa jest liczbie wszystkich unikalnych słów we wszystkich dokumentach tej kolekcji. Efektem końcowym jest macierz $\mathbf{M}$, w której kolumny reprezentują słowa występujące w kolekcji dokumentów, zaś wiersze reprezentują dokumenty z tej kolekcji. Niech

— dokument $\mathcal{D}_1$ zawiera zdanie „Kazał kurze ścierać kurze.”,
— $\mathcal{D}_2$ zawiera zdanie „Koń, jaki jest, każdy widzi.”,
— a $\mathcal{D}_3$: „Każdy kazał ścierać kurze.”

Po sprowadzeniu wszystkich liter do małych oraz opuszczeniu znaków interpunkcyjnych, na wektor unikalnych słów składają się słowa:

$$\begin{bmatrix} \text{jaki} & \text{jest} & \text{kazał} & \text{każdy} & \text{koń} & \text{kurze} & \text{ścierać} & \text{widzi} \end{bmatrix} \tag{2.2}$$

Kolejne wiersze w macierzy $\mathbf{M}$ zawierają informacje o występowaniu konkretnych słów w dokumentach: kolejno $\mathcal{D}_1$, $\mathcal{D}_2$, $\mathcal{D}_3$:

$$\mathbf{M} = \begin{bmatrix} 0 & 0 & 1 & 0 & 0 & 1 & 1 & 0 \\ 1 & 1 & 0 & 1 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 & 1 & 1 & 0 \end{bmatrix} \tag{2.3}$$

Standardowe podejście (Manning i inni, 2008) do wyszukiwania odpowiedzi na zadane pytanie opiera się na transformacji zapytania, podobnie jak dokumentów w kolekcji, do wektorów i na tej podstawie odnalezieniu najbardziej podobnego dokumentu do zapytania. Podobieństwo wektora zbudowanego dla zapytania $\vec{z}$ do wektora dla dokumentu

$\mathcal{D}$ określane jest przy wykorzystaniu wybranych miar podobieństwa lub odległości wektorów zapytania $\vec{z}$ i dokumentu $\mathcal{D}$ – oznaczony jako $\vec{v}(\mathcal{D})$. Często używana jest miara kosinusowa opisana wzorem 2.4.

$$\cos(\vec{z}, \vec{v}(\mathcal{D})) = \frac{\vec{z} \cdot \vec{v}(\mathcal{D})}{\|\vec{z}\|\|\vec{v}(\mathcal{D})\|} = \frac{\sum_{i=1}^{n} \vec{z}_i \vec{v}(\mathcal{D}_i)}{\sqrt{\sum_{i=1}^{n} \vec{z}_i^2} \sqrt{\sum_{i=1}^{n} \vec{v}(\mathcal{D}_i)^2}} \tag{2.4}$$

Jest ona zdefiniowana jako suma iloczynów kolejnych elementów i, wektora zapytania $\vec{z}$ oraz wektora reprezentującego dokument $\vec{v}(\mathcal{D})$. Wartość ta normalizowana jest iloczynem długości wektora zapytania $\vec{z}$ oraz wektora $\vec{v}(\mathcal{D}_i)$ reprezentującego dokument. Inna popularną miarą jest odległość euklidesowa opisana wzorem 2.5:

$$d(\vec{z}, \vec{v}(\mathcal{D})) = \sqrt{\sum_{i=1}^{n} (\vec{z}_i - \vec{v}(\mathcal{D})_i)^2} \tag{2.5}$$

Jest ona zdefiniowana jako pierwiastek z sumy kwadratów różnicy i-tych elementów wektora zapytania $\vec{z}$ oraz wektora $\vec{v}(\mathcal{D}_i)$ reprezentującego dokument.

Standardowe podejście oparte o wektory binarne zastąpione może być poprzez określenie częstości występowania danego słowa w danym dokumencie (Manning i inni, 2008). Można obliczyć macierz $\mathbf{M}_{freq}$ opisującą częstotliwość występowania słów, która dla rozważanego przykładu przyjmuje następującą postać (wzór 2.6):

$$\mathbf{M}_{freq} = \begin{bmatrix} 0 & 0 & 1 & 0 & 0 & 2 & 1 & 0 \\ 1 & 1 & 0 & 1 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 & 1 & 1 & 0 \end{bmatrix} \tag{2.6}$$

Tak jak poprzednio, kolejne wiersze odnoszą się do dokumentów, kolumny do słów. Macierz częstości $\mathbf{M}_{freq}$ można normalizować względem najczęściej występującego słowa w dokumencie. Normalizacja przedstawiona została za pomocą wzoru 2.7, gdzie $tf_{w_i,\mathcal{D}_j}$ oznacza częstotliwość słowa w_i w dokumencie $\mathcal{D}_j$.

$$tf_{w_i,\mathcal{D}_j} = \frac{tf_{w_i,\mathcal{D}_j}}{max_{k=1}^{|\mathcal{D}_j|}\{tf_{w_k,\mathcal{D}_j}\}} \tag{2.7}$$

Często spotykanym sposobem ważenia wartości w wektorze opisującym dokument jest przekształcenie za pomocą miary *tf.idf*, czyli iloczynu wartości *tf* oraz *idf*. Miara ta jest złożeniem liczności występowania słowa w_i w dokumencie i liczby dokumentów zawierających to słowo. *idf* to odwrotna częstość słowa w_i w kolekcji dokumentów $\mathcal{C}$ (wzór 2.8), nazwanej jako idf_{w_i}.

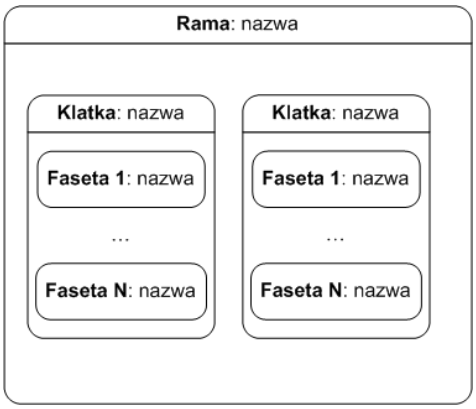
$$idf_{w_i} = \log \frac{|\mathcal{C}|}{|\{D \in C : w_i \in D\}|} \tag{2.8}$$

Definiuje się ją jako stosunek liczby dokumentów w korpusie $\mathcal{C}$, do liczby dokumentów zawierających przynajmniej jedno wystąpienie danego słowa w_i. Logarytm pomaga w przypadku dużych wartości na traktowanie ich w podobny sposób – wartości logarytmu

będą bliższe sobie niż dwie duże liczby. Słowa częste będą dostawały niskie wartości *tf.idf*, zaś słowa rzadkie wartości wysokie. Wartości te, po uszeregowaniu będą tworzyły ranking słów, który można odpowiednio skrócić – usuwając częste (poniżej zadanej przez użytkownika wartości *idf*) i rzadkie słowa (powyżej zadanej przez użytkownika wartości *idf*).

Omawiana reprezentacja wektorowa posiada istotną wadę wynikającą z braku możliwości wyrażenia zależności między słowami w tekście. Niemożliwe jest (w prosty sposób) określenie funkcji słów w tekście (np. wprowadzenie informacji, że dane słowo w tekście pełni rolę podmiotu), gromadzenie informacji o słowach, czy też uogólnienie informacji reprezentowanej przez poszczególne słowa na podstawie zewnętrznych źródeł wiedzy. Nie jest możliwa reprezentacja wiedzy, która nie jest zapisana bezpośrednio w tekście, a o której wiemy „z własnego doświadczenia" – np. to, że kura to ptak, a nie ssak.

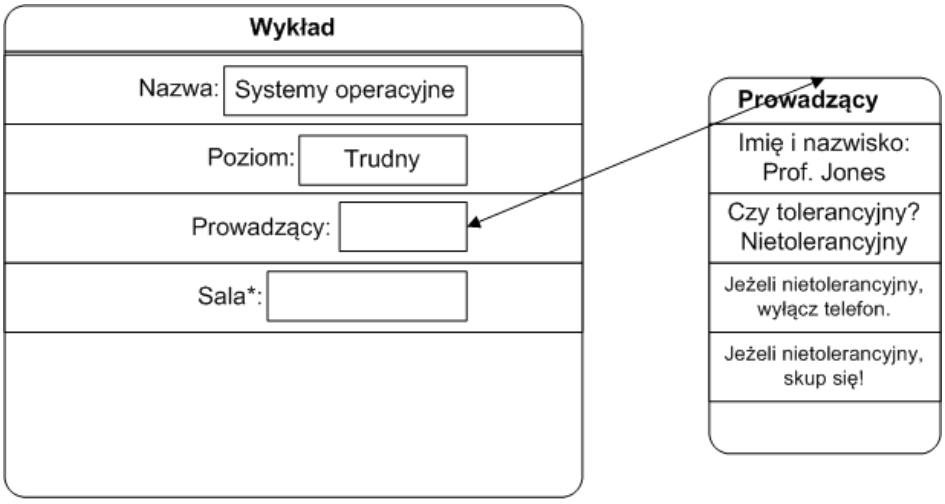
Innym sposobem reprezentacji wiedzy w tekście, posiadającym większą siłę ekspresji są **ramy** (Minsky, 1981). Teoria ram opiera się między innymi na ludzkiej percepcji, strukturach pamięciowych oraz sposobie wnioskowania. Jest to abstrakcyjne przedstawienie wyobrażeń o świecie, modelowane za pomocą abstrakcyjnych pojęć reprezentowanych w postaci formalnych struktur nazywanych *ramami* (ang. *frames*) zawierających *klatki* (ang. *slots*). Rysunek 2.1 przedstawia ogólny schemat *ramy*. Za pomocą



Rys. 2.1. Ogólna postać ramy

ram można nie tylko przedstawić wiedzę zapisaną w tekście, ale przede wszystkim zamodelować stan wiedzy, jaki posiadamy (Bobrow i inni, 1986). Każda budowana rama oraz każda klatka posiadają swoją nazwę, która służy do ich identyfikacji. W skład ramy, która może reprezentować właściwości określonego bytu, wchodzą klatki, które przechowują poszczególne cechy lub informację o takich bytach, np. rama *Ssak* posiada klatkę *Odżywia się*. Wewnątrz klatek zawarte są *fasety*, które odzwierciedlają właściwości klatek. Dodatkowo, rama może zawierać specjalne klatki, które wskazują na dziedziczenie, czy też domyślne ich wartościowanie. Jeżeli dwie ramy posiadają klatkę o tej samej nazwie, wtedy są one w relacji ze sobą. Na rysunku 2.2 przedstawiona została przykładowa rama obrazująca ramy nazwane jako *Wykład* i *Prowadzący*, wraz z relacją mówiącą o tym, że *wykład* posiada *prowadzącego*.

Często zamiast ram, preferuje się modelowanie za pomocą **sieci semantycznych** (Collins i Quillian, 1995, Lehmann, 1992). Idea Quilliana zakładała, że pamięć ludzką



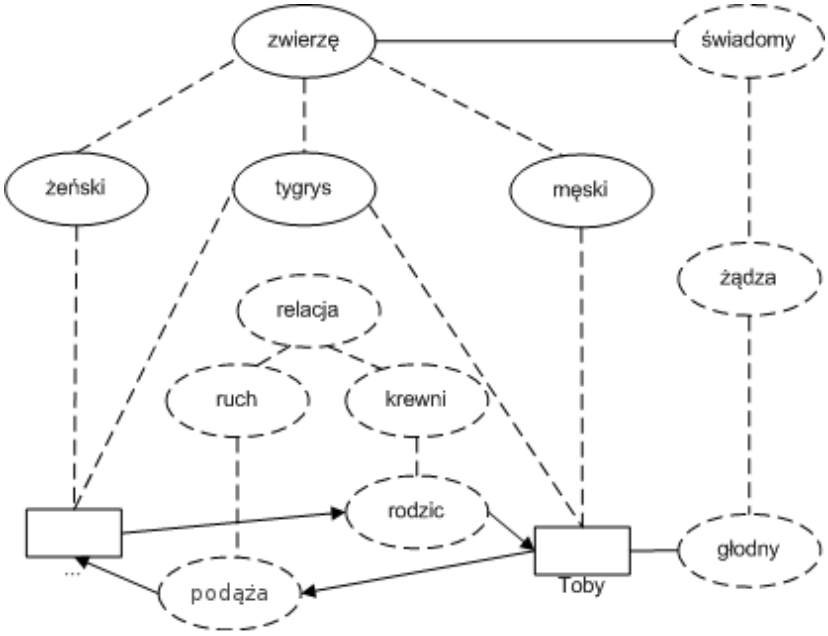
Rys. 2.2. Przykład wypełnionej ramy, na podstawie `http://www.doc.ic.ac.uk/~sgc/teaching/pre2012/v231/lecture4.html`, ostatni dostęp 16.08.2018

najlepiej opisuje model asocjacyjny, czyli cykliczne opisywanie pojęć innymi pojęciami, połączonych ze sobą różnymi relacjami. Sieci semantyczne reprezentowane są przez grafy, które mogą być skierowane oraz mogą posiadać przypisane etykiety do krawędzi oraz węzłów. Przykład takiej sieci przedstawiony został na rysunku 2.3. Prostokąty reprezentują obiekty: nazwany obiekt *Toby* – tygrys oraz nienazwany obiekt tygrysicy – rodzaju *żeńskiego*. Przerywaną linią oznaczona została relacja `is-a`, ciągłe linie, to połączenia między konkretnymi bytami/jednostkami. Na rysunku 2.3 pokazano, że *tygrys* jest *zwierzęciem*, *zwierzę* może być płci *żeńskiej* lub *męskiej*, do czynienia mamy z konkretnym tygrysem, który nazywa się *Toby*, posiada *rodziców* i jest *głodny*. Przeważnie węzły reprezentują klasy, zaś łuki to relacje między nimi. Przyjmijmy, że L_V to skończony zbiór etykiet węzłów, oraz L_E to skończony zbiór etykiet krawędzi. Formalnie, według zapisu z pracy (Champin i Solnon, 2003) graf etykietowany $\mathbf{G}$ definiowany jest jako trójka: $\mathbf{G} = (V, r_V, r_E)$, gdzie:

— V oznacza skończony zbiór węzłów,
— $r_V \subseteq V \times L_V$ to relacja przypisująca węzłom etykiety, to znaczy składa się ze zbioru par (n_i, l), takich, że węzeł $n_i \in V$ posiada przypisaną etykietę $l \in L_V$,
— $r_E \subseteq V \times V \times L_E$ jest relacją, która przypisuje etykiety do krawędzi, r_E to trójka (n_i, n_j, l), takich, że krawędź pomiędzy węzłami $n_i \in V$ oraz $n_j \in V$ posiada przypisaną etykietę $l \in L_E$; kierunek w przypadku grafów skierowanych ustalany jest od węzła n_i do węzła n_j.

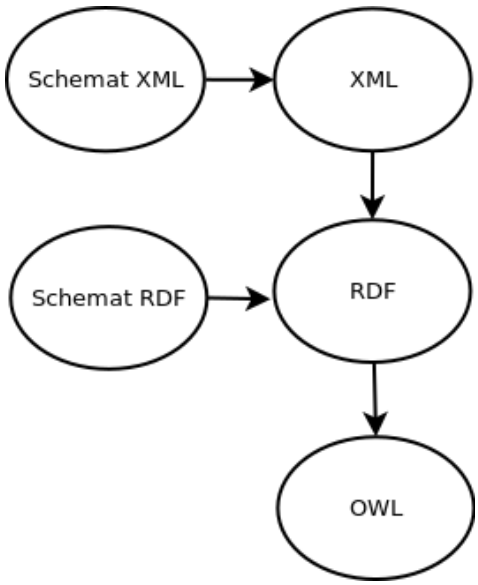
Idea sieci semantycznych wykorzystana została w internetowej sieci semantycznej (ang. *Semantic Web*) szczegółowo przedstawiona w (Kashyap i inni, 2008)[1]). Opracowany został standard opisu zasobów – RDF (ang. *Resource Description Framework*) przedstawiony w pracy (Candan i inni, 2001) [2] oraz zapisu wiedzy – OWL (ang. *Web Ontology Language*) opisany w (Antoniou i van Harmelen, 2004) [3]. Na rysunku 2.4 przedstawiony został podział na standardowy opis za pomocą *XML*, *RDF* oraz *OWL*. *XML* to

[1] https://www.w3.org/standards/semanticweb/
[2] https://www.w3.org/RDF/
[3] https://www.w3.org/OWL/



Rys. 2.3. Przykład sieci semantycznej, na podstawie (Lehmann, 1992)

meta-język, za pomocą którego możliwe jest zapisanie dowolnej informacji posiadającej nazwę z przypisaną jej wartością. *RDF* bazuje na składni języka *XML* i jego założeniem jest opis zasobu za pomocą trójelementowego wyrażenia: podmiotu, orzeczenia/predykatu (własności) i dopełnienia/obiektu (wartości). Zaś *OWL* to rozszerzenie *RDF* i służy do opisywania zasobów sieciowych przy pomocy ontologii. W pracy (Widdows,



Rys. 2.4. Podział na zasoby semantyczne oraz warstwę wiedzy, na podstawie `http://instructionaldesign.com.au/content/semantic-web`, ostatni dostęp 09.08.2016

2004) reprezentacja grafowa została wykorzystana do przedstawienia zależności mię-

dzy słowami z kolekcji tekstów. Krawędzią łączone są ze sobą wyrazy podobne, bądź powiązane semantycznie.

Zaletami wykorzystania sieci semantycznych jako sposobu reprezentacji informacji są: możliwość zapisu relacji między pojęciami, możliwość rozbudowy sieci semantycznej w dowolnym momencie oraz prostota wizualizacji.

Jednak sama sieć semantyczna nie daje możliwości wnioskowania na podstawie zapisanej w niej wiedzy i informacji, gdyż jest jedynie reprezentacją. Wnioskowanie logiczne z sieci semantycznej odbywa się po transformacji tej sieci do wyrażeń logicznych (Deliyanni i Kowalski, 1979) lub nadania sieci interpretacji logicznej, np. logiki deskryptywnej jak jest w przypadku OWL poziomu trzeciego.

Najlepszym podejściem z silnie ugruntowanymi podstawami teoretycznymi, które umożliwiają wnioskowanie i dowodzenie prawdziwości zdań, jest zapis wiedzy bezpośrednio w języku logiki predykatów pierwszego rzędu (Hobbs i inni, 1993, Hobbs, 2008, Barker i inni, 2007) lub logiki predykatów wyższych rzędów. Oczywiście posługując się jednym typem logiki nie byłoby możliwe dokładne zamodelowanie semantyki tekstu, dlatego do wyrażenia *możliwości* stosuje się logikę modalną, do wyrażenia *czasu* logikę temporalną lub, jak to zostało zaproponowane w pracy (Hobbs, 2008), logikę predykatów pierwszego rzędu, w której wyróżniono predykaty określające czas, następstwo występowania zdarzeń, czy też możliwość zajścia pewnego zdarzenia przy pomocy odpowiedniej interpretacji. Rozważmy dla przykładu zdanie:

„John called the Boston office" (*John zadzwonił do biura w Bostonie*)
(Hobbs i inni, 1993).

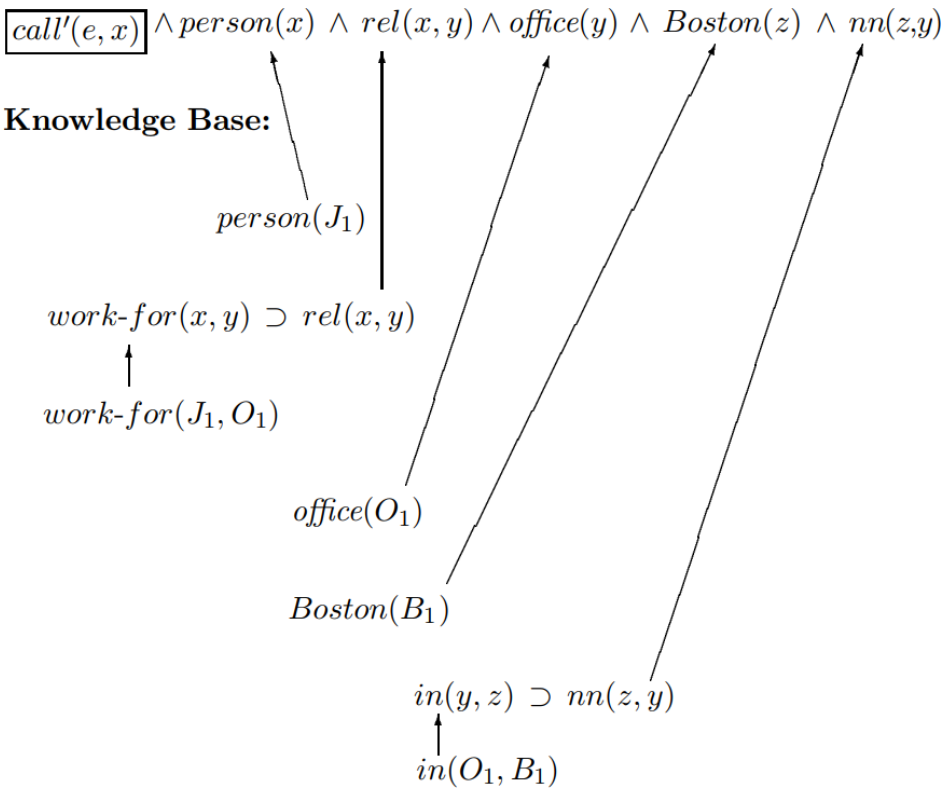
Jego forma logiczna może być przedstawiona jako:

$$(\exists x, y, z, e) call'(e, x) \wedge person(x) \wedge rel(x, y) \wedge office(y) \wedge Boston(z) \wedge nn(z, y) \quad (2.9)$$

gdzie:
- e przyjmuje wartości z wydzielonego podzbioru uniwersum reprezentującego zdarzenia i użycie e oznacza, że dany predykat odnosi się do zdarzenia e,
- x jest osobą,
- $rel(x, y)$ oznacza, że między x i y zachodzi relacja rel,
- y to biuro,
- $nn(z, y)$ to niedospecyfikowana relacja zachodząca między *Bostonem*, który został wartością zmiennej z, a biurem y.

Na rysunku 2.5 przedstawiono przykładową postać logiczną dla zdania „John called the Boston office." (pominięty został aspekt reprezentacji momentów czasu) z podziałem na formę logiczną przedstawioną w pierwszej linii rysunku oraz bazę wiedzy (bazę faktów), których prawdziwość można wywnioskować z rozważanego zdania przy założeniu prawdziwości zdania. Są one ujęte na rysunku jako **Knowledge Base**.
Zdanie to może być interpretowane jako zestaw faktów:

1. `Boston(`B_1`)`, co oznacza, że B_1 to miasto *Boston*.
2. `office(`O_1`)` $\wedge$ `in(`O_1`,` B_2`))` mówi o tym, że O_1 znajduje się w *Bostonie*.
3. `person(`J_1`)` wskazuje, że J_1 jest zmienną wskazującą na bycie osobą.
4. `work-for(`J_1`,`O_1`)` informuje o tym, że John J_1 pracuje dla biura O_1.
5. `in(y,z)` $\supset$ `nn(z,y)`, mówi o tym, że jeżeli y jest w z, wtedy między z oraz y zachodzi niedospecyfikowana relacja nn oznaczająca kompozycję.



Rys. 2.5. Przykład logicznej reprezentacji dla zdania: „John called The Boston office." (Hobbs i inni, 1993) z wydzieloną bazą wiedzy (ang. *Knowledge Base*) zawierającą pojedyncze predykaty całego wyrażenia logicznego budującego zdanie

6. `work-for(x,y)` $\supset$ `rel(x,y)`, zaś wskazuje na to, że jeżeli x pracuje dla y, to między nimi zachodzi relacja, w której x jest przez y niejako zmuszany do pracy (x ma obowiązek pracować dla y).

Niestety w rzeczywistości pojawiają się również problemy związane z niejednoznacznościami wypowiedzi. Weźmy dla przykładu zdanie:

Na stole w pudełku leży koc.

W momencie interpretacji tego zdania natykamy się na niejednoznaczność struktury składniowo-semantycznej *na stole w pudełku*. Nie jest wiadome, czy stół jest w pudełku, czy też pudełko leży na stole. Ta niejednoznaczność powoduje, że jeżeli budowalibyśmy reprezentację semantyczną tego tekstu, należałoby rozróżnić dwa przypadki:

1. Koc leży na stole, a stół jest w pudełku.
2. Pudełko jest na stole, a w pudełku znajduje się koc.

Na koniec, warto zauważyć, że niezależnie od sposobu reprezentacji, problem niejednoznaczności interpretacji będzie pojawiał się w przypadku bardzo wielu wypowiedzi w języku naturalnym.

2.2. Wybór sposobu reprezentacji wiedzy

Niewątpliwie logika jest najlepszym rozwiązaniem umożliwiającym wnioskowanie na podstawie zgromadzonej wiedzy. Posiada ona jednak dużą wadę – niemożliwość wnioskowania z niepełnej wiedzy. Wnioskowanie na podstawie faktów nieistniejących w bazie wiedzy, będzie skutkowało otrzymaniem odpowiedzi *fałszu* na zadane pytanie do bazy wiedzy. Nie byłoby to problemem, gdybyśmy posiadali pełną wiedzę o analizowanym tekście, znali dokładnie pojęcia występujące w tekście, potrafili odnieść je do zewnętrznego źródła wiedzy, w którym byłyby one opisane formalnie, potrafili stwierdzić zakres kwantyfikatorów czy też negacji.

Aby otrzymać pełny opis formalny tekstu, należałoby zbudować głęboki parser semantyczny. Jednak warto podkreślić, że zadanie budowy parsera głębokiego, który umożliwiałby rozpoznanie dokładnej struktury semantycznej zdania/tekstu jest dużym wyzwaniem. Dla języka polskiego, który posiada swobodny szyk zdania, bogatą fleksję oraz skomplikowane zasady gramatyczne problem ten staje się jeszcze większy.

Dlatego metoda opracowana w ramach niniejszej pracy zakłada selektywne pogłębianie wiedzy wydobywanej z analizowanego tekstu. Brak głębokiego parsera dla języka polskiego, efektywnie i skutecznie działającego na dowolnych tekstach, wyklucza niejako wykorzystanie logiki jako docelowego sposobu reprezentacji wiedzy. Sposobem reprezentacji wiedzy najbardziej zbliżonym pod względem siły ekspresji do logiki są sieci semantyczne, które w wielu przypadkach mogą być formalnie interpretowane jako wyrażenia logiczne. Umożliwiają one zapisanie informacji niepełnej, rozpoznanej na różnych poziomach opisu języka. Ich transformacja do reprezentacji grafowej również jest dużą zaletą, ponieważ posiadając hierarchię pojęć wraz z określeniem zależności semantycznych między słowami, można uzupełnić opis reprezentowany jako sieć semantyczna za pomocą języka regułowego, np. OWL (Antoniou i van Harmelen, 2004), operującego na tej sieci i wnioskować tylko z wybranych fragmentów formalnie zbudowanego opisu. Biorąc pod uwagę przytoczone argumenty, w opracowanej w pracy metodzie pogłębianej analizy semantycznej, przyjęto grafowy sposób reprezentacji wiedzy.

2.3. Zasoby wiedzy

W zaproponowanej metodzie pogłębianej analizy semantycznej, wykorzystywane są dwa główne zasoby wiedzy: ontologia wysokiego poziomu SUMO (Pease, 2011) oraz Słowosieć (Piasecki i inni, 2009a). Są one dostępne na otwartej licencji.

2.3.1. Słowosieć

Słowosieć (ang. plWordNet)[4] jest to *leksykalna sieć semantyczna*, na którą składają się *jednostki leksykalne* oraz *relacje leksykalno-semantyczne* między nimi. Jednostka leksykalna jest rozumiana w tym przypadku w wąskim, technicznym sensie jako trójka: lemat (podstawowa forma morfologiczna – forma hasłowa), część mowy oraz identyfikator konkretnego znaczenia leksykalnego. Przykładowo wyraz „zamek” posiada siedem

[4] Strona internetowa"`http://plwordnet.pwr.wroc.pl` (Piasecki i inni, 2009a, Dziob i Piasecki, 2018, Piasecki i inni, 2016b, Maziarz i inni, 2016, 2014), do pobrania z `http://nlp.pwr.wroc.pl/plwordnet/download`

znaczeń, z których każde ma przypisany unikalny, jednoznacznie identyfikujący je numer:

— *zamek.1* to budowla,
— *zamek.2* – rodzaj zamknięcia w drzwiach,
— *zamek.3*, według definicji Słowosieci jest to *Urządzenie do łączenia lub zabezpieczania w ustalonym położeniu elementów maszyny*,
— *zamek.4* wskazuje na mechanizm broni palnej, który służy do zamykania i otwierania części lufy,
— *zamek.5* jest to blokada w informatyce,
— *zamek.6* – rodzaj zapięcia,
— *zamek.7* – zagranie taktyczne w hokeju.

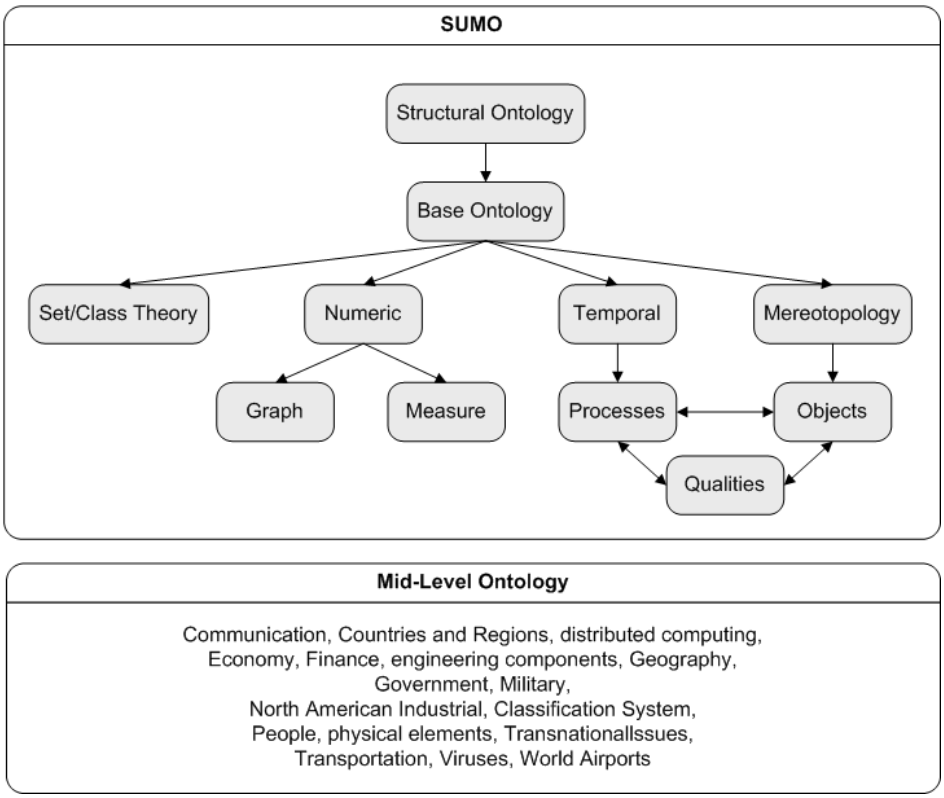
Każda jednostka leksykalna posiada przypisaną część mowy. W Słowosieci do każdej jednostki leksykalnej przypisana jest również jej dziedzina semantyczna. Jednostki leksykalne są ze sobą połączone za pomocą *relacji jednostek leksykalnych*. Relacje takie, to np. *deminutywność* (zdrobnienie): *zamek.2* → *zameczek.2*, która mówi o tym, że jednostka leksykalna *zameczek.2* jest zdrobnioną formą od *zamek.2*, a przede wszystkim wyraża semantykę bytu mniejszego, przyjemniejszego, itd. Innym przykładem relacji między jednostkami leksykalnymi jest *synonimia międzyparadygmatyczna*, która łączy ze sobą jednostki, które mają bardzo zbliżone znaczenie, jednak wyrażone są za pomocą innych części mowy np. *zamek.1* → *zamkowy.1*. Jedną z najważniejszych relacji jednostek leksykalnych jest *synonimia* (Maziarz i inni, 2013). Wiąże ona ze sobą te jednostki leksykalne, które oznaczają dokładnie to samo oraz mają tę samą część mowy. Na przykład *zamek.6* jest w relacji *synonimii* do *suwak.1*, *ekler.2* oraz *zamek błyskawiczny.1*. Te jednostki leksykalne, które połączone są ze sobą relacją synonimii, tworzą *synsety* – czyli zbiory synonimów. Pomiędzy synsetami zachodzą również relacje, są to relacje synsetów, które wynikają ze współdzielenia pewnych relacji pomiędzy jednostkami wchodzącymi w skład tych synsetów. Przykładem takich relacji jest *hiperonimia* – relacja mówiąca o tym, że tworzące synset jednostki są ogólniejsze od jednostek tworzących synset z nimi połączonych. Przykładem *hiperonimii* jest relacja między *zamkiem.1*, a *budynkiem.1*, mówimy wtedy, że *budynek.1* jest *hiperonimem zamku.1*, a *zamek.1* jest *hiponimem budynku.1*. Relacja *meronimii* – jednostki tworzące synset są w relacji część-całość z jednostkami tworzącymi drugi synset, na przykład *zamek.4* jest częścią *broni palnej.1*. Słowosieć może zostać przedstawiona na dwóch poziomach:

— jako graf, w którym w węzłach znajdują się synsety, zaś krawędzie grafu są to relacje między synsetami,
— bądź też graf powiązań jednostek leksykalnych, gdzie w węzłach znajdują się jednostki leksykalne, a krawędzie to relacje łączące jednostki.

Znając semantykę poszczególnych relacji można na przykład wnioskować, że *kot*, to rodzaj *zwierzęcia* (*hiponim* zwierzęcia).

2.3.2. SUMO

SUMO (ang. ***S**uggested **U**pper **M**erged **O**ntology*) (Pease, 2011) to ontologia wysokiego poziomu, która zawiera ogólne pojęcia odnoszące się do świata rzeczywistego. SUMO jest definiowane przy pomocy aksjomatów pogrupowanych w jedenaście modułów opisanych poniżej. Moduły te są ułożone w hierarchię widoczną na rysunku 2.6.



Rys. 2.6. Struktura SUMO, na podstawie (Pease, 2011)

Poniższy opis modułów zawiera ich odpowiednik w wersji angielskiej oraz tłumaczenie na język polski. Podział na moduły jest następujący:

Ontologia strukturalna (ang. *Structural Ontology*) – zawiera podstawowe definicje do określenia, że:
- coś jest instancją klasy (relacja *instance*),
- jedna klasa jest podklasą innej (*subclass*),
- jedna relacja jest podrelacją innej (*subrelation*),
- określenie typu ograniczeń dla relacji lub funkcji (*domain*, *domainSubclass*, *range*, *rangeSubclass*).

Ontologia podstawowa (ang. *Base Ontology*) – zawiera pojęcia najwyższego poziomu: *Entity*, *Physical*, *Abstract*, *Object*, *Substance* oraz kilka poziomów niższych. Oprócz tego, zawiera podstawowe typy liczbowe, funkcje arytmetyczne oraz nierówności. Podstawowe różne typy relacji również są definiowane na tym poziomie: *TotalValueRelation*, *PartialValueRelation*. Zdefiniowane zostały również podstawowe role (*CaseRole*): *agent*, *patient*, *origin*, *destination*. Dodatkowo, oprócz pojęć, zawiera relacje: *part*, *property*, *manner*, kilka podstawowych relacji semiotycznych (Pease, 2011, str. 147) oraz wyrażenia modalności: *wants*, *desires*, *believes*, *capability*, *hasPurpose*.

Teoria zbioru/klasy (ang. *Set/Class Theory*) – podstawowy zbiór notacji za pomocą zbiorów, takie jak: suma, różnica, przecięcie. Definicja funkcji KappaFn - umożliwia definiowanie klas “dynamicznie”, jako nienazwane klasy.

Funkcje numeryczne (ang. *Numeric*) – dodawanie *AdditionFn*, odejmowanie (*subtraction*), pierwiastek kwadratowy (*square root*) etc.

Określenia czasowe (ang. *Temporal*) – do określeń czasu zostały włączone również relacje pomiędzy zdarzeniami.

Mereotopologia (ang. *Mereotopology*) – mereotopologia, połączenie merologii – dziedziny filozofii i logiki „o kolekcjach i składaniu się" oraz topologii – o przestrzeniach. Definiuje różne abstrakcyjne operacje na obiektach, takie jak: *bottom* czy też relacje pomiędzy obiektami na przykład *connected*.

Graf (ang. *Graph*) – definiuje podstawowe notacje z teorii grafów oraz wiele rzeczywistych struktur grafowych jak na przykład *sieć komputerowa*, *system transportowy*, *system elektryczny*. Zdefiniowane zostały między innymi takie pojęcia jak: węzły, krawędzie, ścieżki, wagi, najkrótsze ścieżki.

Miara (ang. *Measure*) – Definiuje różne klasy miar np. *AreaMeasure*, *VolumeMeasure*, jak również bardziej szczegółowe *Acre*, *Litre*. Oprócz miar zdefinowane są również funkcje jak: *MiliFn* (np. do konwersji), *PerFn* (do określania stosunku) oraz funkcje operujące na obiektach fizykalnych: *length*, *width*, *height*.

Procesy (ang. *Processes*) – jedna z największych części SUMO. Najwyższa część dotyczy *DualObjectProcess*, *NaturalProcess*, *IntentionalProcess*, *Motion* oraz *InternalChange*.

Obiekty (ang. *Objects*) – definicje pojęć dotyczącymi obszarów, elementów „świata", np. *Roadway*. Podstawowa taksonomia biologiczna (*Organisms*), części mowy (np. *Verb*), kilka rodzajów tekstów (*Texts*) oraz relacja (między nimi i nie tylko) np. *SeriesVolumeFn*, *authors*. Relacje rodzinne, typy organizacji (*Organization*) oraz inne: *Residence*, *Vehicle*, *Weapon*, *Clothing*.

Własności (ang. *Qualities*) – obejmuje duży zakres abstrakcyjnych pojęć (*Abstract*), włączając w to *Plans* oraz *Arguments*, liczne relacje, atrybuty relacyjne (*Attributes*) oraz różne pojęcia odnoszące się do zobowiązań i norm, w tym pojęcia prawne. W module tym występują również *PhysicalAttributes* oraz pojęcia odnoszące się do wyglądu obiektów.

Hierarchia pojęć SUMO definiowana jest za pomocą relacji: *subclass* (jedno pojęcie jest bardziej specyficzne od innego) oraz *instance* (rozumiana jako instancja klasy). Przykładem *subclass* jest: *subclass(Human,Mammal)*, co należy interpretować jako: „Ludzie są ssakami". Przykładem zapisu relacji *instance* jest: *instance(John, Human)*, które należy interpretować jako: „John jest instancją człowieka".

Rozdział 3

Przegląd literatury

Problem modelowania zależności między fragmentami tekstów został po raz pierwszy opisany w (Trigg, 1983) i w nowszej wersji w (Trigg i Weiser, 1986). Powstał jako efekt prac nad narzędziem do strukturyzacji dużego zbioru artykułów naukowych. Miały być one ze sobą połączone tworząc hierarchię, na przykład w formie drzewa z zależnościami między nimi. Między tekstami modelowane były połączenia *fizyczne* oraz *semantyczne*. Fizyczne – wskazywały na odwoływanie się do innej pozycji literaturowej, połączenia semantyczne wynikały z relacji semantycznej zachodzącej między fragmentami tekstów. Przykładem połączeń może być *krytyka*. Jeżeli fragment tekstu $T1$ odwoływał się do $T2$ *krytykując* go, to połączenie fizyczne skierowane było: $T1 \rightarrow T2$, zaś połączenie semantyczne $T1 \leftarrow T2$ (ponieważ, aby wiedzieć, że $T1$ krytykuje $T2$, należało najpierw przeczytać $T1$).

Podobnie jak w ostatnim wymienionym wyżej artykule, również w (Allan, 1996) zaproponowano zestaw typów połączeń między dokumentami. Połączenia te były rozumiane jako semantyczne, takie jak *okoliczność* (ang. *circumstance*), *przyczyna* (ang. *cause*), *cel* (ang. *purpose*), *sprzeczność* (ang. *contrast*), czy też *równoważność* (ang. *equivalence*). Połączenia klasyfikowane były do jednej z trzech grup: *dopasowanie wzorca*, *ręczne* oraz *automatyczne*. Grupa wskazywała na sposób podejścia do określenia połączenia między fragmentami tekstów. Rozważano więc wykorzystanie wzorców regularnych, ręczne wykrywanie lub wykrywanie automatyczne.

W (McKeown i Radev, 1995) autorzy uzupełnili omówiony wyżej zestaw relacji. Opracowany przez nich zbiór relacji został wykorzystany do streszczania wielodokumentowego (ang. *multi-document summarization*) i był ukierunkowany na właściwości tekstowe, takie jak *podobieństwo*, *sprzeczność*, *różnica*, *komplementarność*.

Model relacji *Cross-document Structure Theory* (dalej oznaczany *CST*) został zaproponowany w (Radev, 2000). Jest to rozwinięcie podejścia *Rhetorical Structure Theory* (Mann i Thompson, 1987) oraz idei relacji między różnymi dokumentami (Allan, 1996). Model *CST* pierwotnie zaaplikowany został do zadania wielodokumentowego streszczania. Obejmuje on zestaw 24 relacji semantycznych, które mogą być symetryczne lub niesymetryczne. Relacja symetryczna oznacza, że relacja zachodzi między tekstem $T1 \rightarrow T2$ oraz $T2 \rightarrow T1$. Niesymetryczne relacje zachodzą jedynie w jedną stronę, tzn. $T1 \rightarrow T2$ albo $T2 \rightarrow T1$; $T1$ oraz $T2$ to teksty, między którymi zachodzi

relacja. Szczegółowy przegląd prac literaturowych w zakresie określania relacji (połączeń) semantycznych między dokumentami/fragmentami tekstów oraz historia powstania modelu CST opisano w (Maziero i inni, 2014). W pracy tej autorzy zaproponowali również zmiany mające na celu lepsze doprecyzowanie relacji.

3.1. CST – Model relacji semantycznych między fragmentami tekstów

Wykrywanie relacji semantycznych zachodzących między dwoma fragmentami tekstów przy użyciu modelu *CST* jest problemem dość często opisywanym w literaturze. O ile problematyka zachodzenia relacji między tekstami jest dość mocno wyeksplorowana, o tyle automatyczne wykrywanie relacji jest rzadziej opisywane. W znacznej części prac opisujących automatyczne wykrywanie relacji problem ten jest traktowany jako problem klasyfikacji. Sprowadza się to do uczenia klasyfikatora na wektorach cech powstałych z wcześniej zaanotowanego zbioru dokumentów. Takie podejście może być również łączone z podejściem regułowym.

3.1.1. Opis relacji semantycznych ujętych w modelu CST

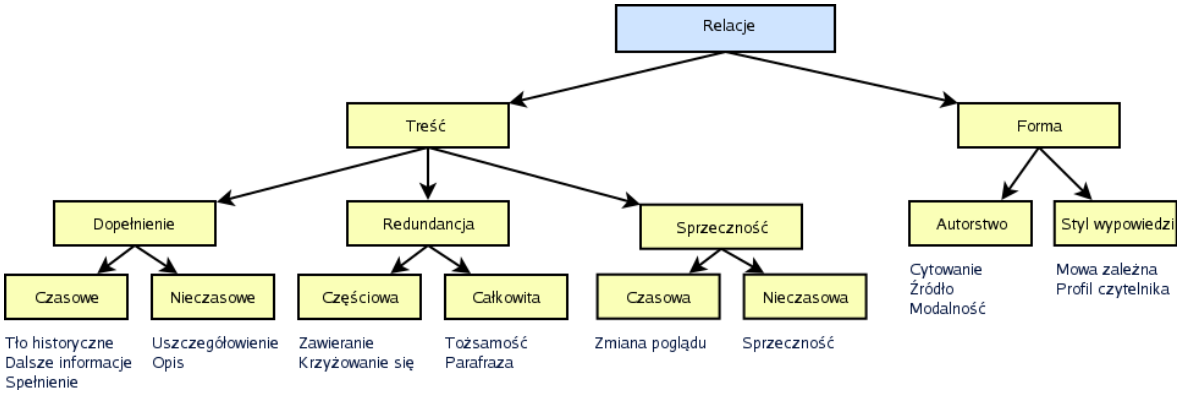
Rozważane w pracy relacje semantyczne pochodzą z modelu CST (Maziero i inni, 2010), który w sposób formalny definiuje poszczególne relacje. Teoria CST została opracowana pierwotnie dla języka angielskiego. Używana jest do odwzorowania powiązań pomiędzy fragmentami tematycznie podobnych do siebie dokumentów (jak również pomiędzy fragmentami tego samego dokumentu) w relacje semantyczne. CST, w odróżnieniu od (RST) (Mann i Thompson, 1988), nie zakłada, że dwa fragmenty tekstu pochodzą z tego samego dokumentu.

W ramach projektu CLARIN-PL zestaw relacji z modelu CST został zaadaptowany do języka polskiego. Każda z relacji posiada szczegółowy opis, znacząco rozszerzony w stosunku do wytycznych CST (Radev i inni, 2003) [1].

Na rysunku 3.1 przedstawiony został podział relacji CST według (Maziero i inni, 2010), ze względu na ich cechy semantyczne dotyczące treści i formy. Opisując relacje semantyczne między fragmentami tekstów skoncentrujemy się na relacjach zachodzących między dwoma zdaniami - zdaniem $\mathcal{S}_1$, a $\mathcal{S}_2$. Definicje poszczególnych relacji zaczerpnięte zostały z wytycznych dotyczących znakowania relacji semantycznych między fragmentami tekstów. Wytyczne opracowane zostały w grupie naukowej G4.19 z Politechniki Wrocławskiej. Niestety, do tej pory nie ma publikacji dokładnie opisującej relacje z modelu CST dostosowane do języka polskiego. Poniżej przedstawione są skrócone definicje poszczególnych relacji wraz z przykładowymi parami zdań, które obrazują ich semantykę.

— Mowa zależna – zdanie $\mathcal{S}_1$ pośrednio cytuje coś, co zostało bezpośrednio przytoczone w zdaniu $\mathcal{S}_2$. Mowa zależna może polegać na parafrazie pewnych słów przy zachowaniu wszakże pełnej tożsamości znaczeniowej. Przykład: $\mathcal{S}_1$: *Pan Cuban zagwarantował tłumowi darmowe cukierki.*, $\mathcal{S}_2$: *„Gwarantuję darmowe cukierki" – Pan Cuban powiedział do tłumu.*

[1] `http://clair.si.umich.edu/clair/CSTBank/annotation_guide.pdf`



Rys. 3.1. Podział relacji z modelu *CST*, ze względu na treść i formę, na podstawie (Maziero i inni, 2010)

— Streszczanie – zdanie $\mathcal{S}_1$ streszcza zdanie $\mathcal{S}_2$ (tj. $\mathcal{S}_1$ zawiera kluczowe informacje z $\mathcal{S}_2$), np. $\mathcal{S}_1$: *Mets wygrali tytuł w siedmiu grach.* $\mathcal{S}_2$: *Po wyczerpujących sześciu grach, Mets przybyli dziś wieczorem by wziąć sobie tytuł.*
— Spełnienie – zdanie $\mathcal{S}_1$ potwierdza wystąpienie (spełnienie się) zdarzenia przewidzianego w (zapowiedzi z) zdaniu $\mathcal{S}_2$ np. $\mathcal{S}_1$: *Po podróży do Austrii, Pan Green wrócił do domu w Nowym Jorku.* $\mathcal{S}_2$: *Pan Green pojedzie do Austrii.*
— Uszczegółowienie – zdanie $\mathcal{S}_1$ dostarcza dodatkowych szczegółów opisanych ogólniej w zdaniu $\mathcal{S}_2$. $\mathcal{S}_1$ zawiera specyficzne informacje, które nie zostały opisane w $\mathcal{S}_2$, lecz oba zdania przekazują tę samą ideę, np. $\mathcal{S}_1$: *Publiczna telewizja RAI donosi, że pilot powiedział, iż miał problem z silnikiem.* $\mathcal{S}_2$: *Włoska telewizja podała, że sygnał SOS nadany przez pilota spowodowany był problemem technicznym.*
— Krzyżowanie się – $\mathcal{S}_1$ przedstawia informacje X i Y, $\mathcal{S}_2$ przedstawia informacje X i Z, np. $\mathcal{S}_2$: *Samolot rozbił się, uderzając w 25 piętro budynku Pirelli znajdującego się w centrum miasta Milan.* $\mathcal{S}_2$: *Mały turystyczny samolot zderzył się z najwyższym budynkiem w Milanie.*
— Zawieranie – $\mathcal{S}_1$ zawiera wszystkie informacje z $\mathcal{S}_2$, oraz dodatkowe informacje niewystępujące w $\mathcal{S}_2$, np. $\mathcal{S}_1$: *Z trzema zwycięstwami w tym roku Green Bay ma najlepszy wynik w lidze NFL.* $\mathcal{S}_2$: *Green Bay trzy razy osiągnął zwycięstwo w tym roku.*
— Zmiana poglądu – zdanie $\mathcal{S}_1$ opublikowane zostało później niż zdanie $\mathcal{S}_2$, jednak ta sama osoba przedstawia dwie odmienne opinie dotyczącego tego samego bytu, tej samej sytuacji lub prezentuje ten sam fakt w innym świetle, np. $\mathcal{S}_1$: *Giuliani skrytykował Związek Oficerów jako "zbyt wymagający" podczas rozmów o umowach.* $\mathcal{S}_2$: *Giuliani pochwalił Związek Oficerów, który dostarcza prawnego wsparcia i pomocy swoim członkom.*
— Tożsamość – zdanie $\mathcal{S}_1$ oraz zdanie $\mathcal{S}_2$ to dokładnie takie same zdania.
— Opis – w zdaniu $\mathcal{S}_1$ znajduje się opis bytu wspomnianego w zdaniu $\mathcal{S}_2$, np. $\mathcal{S}_1$: *Greenfield, emerytowany generał i ojciec dwojga dzieci, odmówił komentarza.* $\mathcal{S}_2$: *Pan Greenfield pojawił się wczoraj w sądzie.*
— Źródło – zdanie $\mathcal{S}_1$ zawiera źródło informacji występującej w zdaniu $\mathcal{S}_2$, np. $\mathcal{S}_1$:

Durczok powiedział, że Ameryka wypowie wojnę Korei. $\mathcal{S}_2$: *Ameryka wypowie wojnę Korei.*

— Cytowanie – zdanie $\mathcal{S}_1$ bezpośrednio cytuje fragment zdania $\mathcal{S}_2$ pochodzącego z innego dokumentu, np. $\mathcal{S}_1$: *Wcześniejszy artykuł cytuje księcia Alberta mówiącego: „Nigdy nie będę uprawiał hazardu".* $\mathcal{S}_2$: *Książę Albert kontynuował mówiąc: „Nigdy nie będę uprawiał hazardu"*

— Sprzeczność – zdanie $\mathcal{S}_1$ i zdanie $\mathcal{S}_2$, pochodzące z tego samego momentu czasu, zawierają konfliktujące ze sobą informacje, np. $\mathcal{S}_1$: *Na pokładzie zestrzelonego samolotu były 122 osoby.* $\mathcal{S}_2$: *126 osób było na pokładzie samolotu.*

— Parafraza – zdanie $\mathcal{S}_1$ oraz zdanie $\mathcal{S}_2$ zawierają dokładnie takie same informacje wyrażone jednak różnymi słowami, np. $\mathcal{S}_1$: *Wałęsa był agentem SB.* $\mathcal{S}_2$: *Lechu współpracował jako agent z SB.*

— Modalność – zdanie $\mathcal{S}_1$ z dodatkiem ramy modalnej przedstawia wersję informacji przedstawionych w zdaniu $\mathcal{S}_2$, np. $\mathcal{S}_1$: *Uważa się, że Sean Combs posiada kilka posiadłości wartych wiele milionów.* $\mathcal{S}_2$: *Puffy posiada cztery wielomilionowe domy w rejonie Nowego Jorku.*

— Tło historyczne – zdanie $\mathcal{S}_1$ opisuje kontekst historyczny przedstawiony w zdaniu $\mathcal{S}_2$, np. $\mathcal{S}_1$: *To był czwarty raz jak członek rodziny królewskiej rozwiódł się.* $\mathcal{S}_2$: *Duke Windsor wczoraj rozwiódł się z Duchess Windsor*

— Dalsze informacje – zdanie $\mathcal{S}_1$ zawiera dodatkowe informacje w stosunku do zdania $\mathcal{S}_2$, jednak $\mathcal{S}_1$ opublikowane zostało po $\mathcal{S}_2$, np. $\mathcal{S}_1$: *102 ofiary zostały odnotowane w rejonie trzęsienia ziemi.* $\mathcal{S}_2$: *Do tej pory nie potwierdzono ofiar trzęsienia ziemi.*

Analizując poszczególne relacje można zauważyć, że możliwe jest dokonanie ich podziału ze względu na redundancję informacji występującej w analizowanych tekstach. Redundancja ta może być pełna (jak w relacjach *Tożsamość* oraz *Parafraza*) lub częściowa, jak w przypadku relacji *Krzyżowanie się* czy też *Zawieranie*. Relacje w modelu *CST* tworzą między sobą hierarchię. Na przykład relacja *Modalność* jest podrelacją *Zawierania*, a relacje takie jak *Cytowanie*, *Źródło*, *Streszczanie* dziedziczą cechy zarówno z *Krzyżowania się*, jak i *Zawierania*.

3.1.2. Wykrywanie relacji CST

Praca (Zhang i inni, 2003) jako jedna z pierwszych przedstawia problem automatycznego wykrywania relacji. Zastosowano podejście oparte o uczenie maszynowe z wykorzystaniem boostingu (Schapire i Singer, 2000). Uczono klasyfikator na podstawie cech ekstrahowanych z par zdań, pomiędzy którymi zachodziła relacja. Ocena skuteczności metody odbyła się w dwóch krokach. W pierwszym zastosowano *klasyfikację binarną* – czyli rozpoznanie czy dla zadanej pary zachodzi jakakolwiek relacja. W kroku drugim przeprowadzano *klasyfikację n-arną*, czyli rozpoznanie konkretnej relacji między parą zdań. Wektory wejściowe były budowane uwzględniając cechy z różnych poziomów analizy języka:

— leksykalnej – liczba tokenów w zdaniu $\mathcal{S}_1$, liczba tokenów w zdaniu $\mathcal{S}_2$, liczba wspólnych tokenów;

— składniowej – liczba tokenów z zadaną częścią mowy *pos* w zdaniu $\mathcal{S}_1/\mathcal{S}_2$, liczba wspólnych tokenów mających część mowy *pos* (*pos* to jedna z sześciu rozpatrywanych części mowy);

— semantycznej – odległość semantyczna pomiędzy głowami kluczowych fraz ze zdania $\mathcal{S}_1$ oraz zdania $\mathcal{S}_2$, podobnie jak w (Jiang i Conrath, 1997);

Stwierdzenie czy pomiędzy daną parą zdań $\mathcal{S}_1$ a $\mathcal{S}_2$ zachodzi jakakolwiek relacja, możliwe było na poziomie $0,7267$ *miary-f* (miara zdefiniowana wzorem 6.2) dla pełnego wektora cech. Wykorzystanie jedynie cech leksykalnych i składniowych umożliwiło rozpoznanie występowania relacji na poziomie $0,7278$ *miary-f*. Same zaś cechy leksykalne dawały wynik na poziomie $0,6$ *miary-f*. W klasyfikacji n-arnej rozpoznanie głównych relacji, wyrażone za pomocą *miary-f* dało następujące wyniki: *Tożsamość* (ang. *Equivalence*) $0,3902$, *Opis* (ang. *Description*) – $0,1622$, *Zawieranie* (ang. *Subsumption*) to $0,0588$. Średnia wartość *miary-f* jaką udało się uzyskać na zbiorze danych CSTBank (Radev i inni, 2004), to $0,3502$.

Rozwinięciem omówionego podejścia jest praca (Zhang i Radev, 2005), gdzie także zastosowano podejście dwuetapowej klasyfikacji oraz uczenie maszynowe z wykorzystaniem boostingu. Jednak w procesie klasyfikacji wykorzystano dane z CSTBank oraz dane nieoznaczone, dzięki czemu dla binarnej klasyfikacji udało się uzyskać wartość *miary-f* równa $0,8839$. Rozpoznawanie poszczególnych relacji, nadal posiadało niską wartość *miary-f*.

Model relacji *CST* wykorzystany został również w (Aleixo i Pardo, 2008) do znajdowania w zbiorze dokumentów tych, które są do siebie podobne pod względem tematyki. Badano jak cechy, które opierają się na wartościach liczonych za pomocą kosinusowej miary podobieństwa (równanie 2.4) oraz miary określającej współczynnik pokrywania się słów w dwóch zdaniach $\mathcal{S}_1$ i $\mathcal{S}_2$ (równanie 3.1), wpływają na wykrywanie relacji dzięki progowi odcięcia poniżej/powyżej danej wartości.

$$WordOverlap(\mathcal{S}_1, \mathcal{S}_2) = \frac{Liczba\_wspolnych\_slow(\mathcal{S}_1, \mathcal{S}_2)}{Liczba\_slow(\mathcal{S}_1) + Liczba\_slow(\mathcal{S}_2)} \tag{3.1}$$

Autorzy przeprowadzili analizę wpływu odcięcia wartości podobieństwa uwzględnianych w wektorze cech. Jest to jedna z niewielu prac poruszających tematykę rozpoznawania relacji CST dla języka innego niż angielski. W badaniach wykorzystano korpus tekstów dla języka portugalskiego.

W artykule (Zahri i Fukumoto, 2011) przedstawiono rozpoznawanie relacji *CST* przy wykorzystaniu *maszyny wektorów nośnych* – SVM (ang. *Support Vector Machine*). Warto zauważyć, że zbiór rozpoznawanych relacji ograniczono do pięciu: *Tożsamość*, *Parafraza*, *Zawieranie*, *Krzyżowanie*, *Uszczegółowienie*. Dla przykładów uczących zaczerpniętych z CSTBanku wektory wejściowe zawierały powiększony zestaw cech w stosunku do (Aleixo i Pardo, 2008) o następujące cechy:

— podobieństwo kosinusowe zdań $\mathcal{S}_1$ i $\mathcal{S}_2$, pomiędzy którymi zachodzi relacja z modelu *CST* (równanie 2.4);
— współczynnik pokrywania się słów w zdaniach $\mathcal{S}_1$ i $\mathcal{S}_2$;
— informacja o tym, które zdanie jest dłuższe – dłuższe zdanie przyjmowało w wektorze cech wartość 1, krótsze 0;
— współczynnik pokrywania słów ze zdania $\mathcal{S}_1$ w zdaniu $\mathcal{S}_2$ oraz słów zdania $\mathcal{S}_2$ w zdaniu $\mathcal{S}_1$.

Podobne podejście opisane jest w (Kumar i inni, 2012). Autorzy zastosowali SVM oraz sieci neuronowe. Ograniczyli się jednak do czterech relacji *Tożsamość*, *Krzyżowanie*, *Zawieranie* oraz *Opis*. Zestaw cech zawierał:

— wartość miary kosinusowej obliczonej dla wektorów reprezentujących dwa zdania, obliczanej na wartościach wag *tf.idf* dla słów występujących w tych zdaniach,
— współczynnik pokrywania się słów,
— różnicę długości zdań $\mathcal{S}_1$ i $\mathcal{S}_2$
— „typ długości zdania $\mathcal{S}_1$" – cecha ta przyjmowała wartość: 1, jeżeli $\mathcal{S}_1$ było dłuższe niż $\mathcal{S}_2$, ,−1, jeżeli $\mathcal{S}_1$ było krótsze niż $\mathcal{S}_2$ oraz 0, jeżeli długość obu zdań była taka sama.

Uzyskana wartość *miary-f* dla wykrywania relacji *Tożsamość* wynosiła $0,91$, dla *Zawierania* $0,59$, $0,54$ dla *Opisu*, a $0,62$ dla *Krzyżowania*. Sieć neuronowa *MLP* została zastosowana do wykrywania relacji: *Tożsamość*, *Zawieranie*, *Opis* oraz *Krzyżowanie się* w (Yogan i Naomie, 2013). Wektory cech bazowały na tych zaproponowanych w (Zahri i Fukumoto, 2011). Uzupełniono je jednak o następujące cechy:

— podobieństwo fraz rzeczownikowych – dla fraz rzeczownikowych ze zdań $\mathcal{S}_1$ oraz $\mathcal{S}_2$ obliczane było ich podobieństwo za pomocą miary Jaccarda (równanie 3.15);
— podobieństwo fraz czasownikowych – dla fraz czasownikowych ze zdań $\mathcal{S}_1$ oraz $\mathcal{S}_2$ obliczane było ich podobieństwo za pomocą miary Jaccarda (równanie 3.15).

Wyniki dla tak zbudowanego wektora cech przedstawiały się następująco: *miara-f* dla relacji *Tożsamość* wyniosła $0,98$, dla *Zawierania* $0,77$, *Opisu* $0,77$, a dla *Krzyżowania* $0,71$.

Ze względu na niejednoznaczność w interpretacji niektórych definicji relacji modelu *CST* podanych w (Radev, 2000), zaproponowano zmiany (Maziero i inni, 2014). Obejmowały one taksonomię relacji, która wprowadzała podział na formę oraz zawartość informacyjną. Dopracowane zostały również definicje poszczególnych relacji oraz wprowadzone zostały dodatkowe wymagania odnośnie współwystępowania niektórych typów relacji. Dla zmodyfikowanego modelu opracowana została metoda do automatycznego wykrywania nowych relacji. Zastosowano klasyfikatory probabilistyczne (naiwny klasyfikator Bayesowski), maszynę wektorów nośnych (SVM) oraz klasyfikatory symboliczne (J48). Połączono również podejście oparte o uczenie maszynowe z wykrywaniem relacji za pomocą reguł. Zestaw cech zawierał cechy z różnych poziomów przetwarzania języka naturalnego, które odnosiły się do obu zdań $\mathcal{S}_1$ i $\mathcal{S}_2$, dla któwych określano relację modelu *CST*. Były to:

1. różnica w liczbie słów zdań $\mathcal{S}_1$ i $\mathcal{S}_2$,
2. procent wspólnych słów w zdaniu $\mathcal{S}_1$ w stosunku do $\mathcal{S}_2$,
3. procent wspólnych słów w zdaniu $\mathcal{S}_2$ w stosunku do $\mathcal{S}_1$,
4. pozycja $\mathcal{S}_1$ w tekście (0 – jedno z pierwszych trzech zdań; 1 – środek; 2 – jedno z ostatnich trzech zdań w dokumencie),
5. pozycja $\mathcal{S}_2$ w tekście (0 – jedno z pierwszych trzech zdań; 1 – środek; 2 – jedno z ostatnich trzech zdań w dokumencie),
6. liczba słów najdłuższego wspólnego podciągu zdań $\mathcal{S}_1$ i $\mathcal{S}_2$,
7. różnica w liczbie rzeczowników między zdaniem $\mathcal{S}_1$ i $\mathcal{S}_2$ (z wyłączeniem nazw własnych),
8. różnica w liczbie przysłówków między zdaniem $\mathcal{S}_1$ i $\mathcal{S}_2$,
9. różnica w liczbie przymiotników między zdaniem $\mathcal{S}_1$ i $\mathcal{S}_2$,
10. różnica w liczbie czasowników między zdaniem $\mathcal{S}_1$ i $\mathcal{S}_2$,
11. różnica w liczbie nazw własnych między zdaniem $\mathcal{S}_1$ i $\mathcal{S}_2$,
12. różnica w liczbie liczb między zdaniem $\mathcal{S}_1$ i $\mathcal{S}_2$,

13. liczba takich samych synonimów w $\mathcal{S}_1$ i $\mathcal{S}_2$.

W badaniach uwzględniono trzy typy klasyfikatorów: wieloklasowy, binarny, hierarchiczny. Dla klasyfikatora wieloklasowego otrzymano średnią wartość *miary-f* równą $0,40$. Średnia wartość *miary-f* dla klasyfikatora binarnego, to $0,67$ (jednak ocena ta nie uwzględnia decyzji końcowej, a jedynie ocenę rozpoznawania poszczególnych relacji). Klasyfikator hierarchiczny uzyskiwał średnią wartość *miary-f* równą $0,72$.

Jak już wspomniano, w ramach projektu CLARIN-PL[2] (Piasecki, 2014) opracowana została pierwsza wersja systemu do automatycznego wykrywania relacji między fragmentami tekstów. Była to adaptacja do języka polskiego podejścia wykorzystującego cechy z (Maziero i inni, 2014) przedstawiona w pracy magisterskiej (Janz, 2016). Wykrywanie relacji odbywa się na zasadzie klasyfikacji dwóch fragmentów tekstów i automatycznego oznaczenia relacji zachodzącej między nimi na podstawie wektorów cech. Liczba klas równa jest liczbie relacji semantycznych zawartych w modelu *CST* i opisanych w rozdziale poprzednim z dodatkową klasą „brak relacji". *Brak relacji*, oznacza, że pomiędzy dwoma analizowanymi tekstami nie zachodzi żadna relacja. Do rozwiązania problemu użyto zespół klasyfikatorów binarnych. Każdy z nich rozpoznaje jedną relację. Uczenie odbywało się na zasadzie *jeden przeciw wszystkim*, to znaczy przykłady pozytywne, to przykłady z danej klasy, negatywne zaś to przykłady z pozostałych klas relacji. Jako klasyfikator wykorzystano *SVM*, ze względu na wysoką zdolność uogólniania, umiejętność uczenia się próbek małolicznych oraz możliwość separacji obiektów z dużym marginesem separującym. Przynależność obiektu do danej klasy określana była na podstawie najwyższego stopnia pewności klasyfikatora. Uzyskane wyniki przedstawiono w Tabeli 3.1. Warto zauważyć, że wartość *precyzji* dla relacji *Krzyżowanie się* jest

Tabela 3.1. Wyniki jakości klasyfikacji relacji semantycznych, mierzone za pomocą *precyzji*, *kompletności* oraz *miary-f* dla 5 relacji z modelu CST z wykorzystaniem zespołu klasyfikatorów binarnych, źródło: (Janz, 2016)

Relacja	**Precyzja**	**Kompletność**	**Miara F**
Krzyżowanie się	0,78	0,37	0,47
Zawieranie	0,32	0,43	0,36
Parafraza	0,10	0,67	0,17
Tożsamość	0,86	0,87	0,86
Brak relacji	0,55	0,78	0,64

znacząco wyższa od kompletności, co oznacza, że system rzadziej podawał odpowiedzi, jednak trafnie. W przypadku *Zawierania*, zarówno *precyzja* jak i *kompletność* są na niskim poziomie. Relacja *Parafrazy* poprawnie wykrywana była bardzo rzadko. Zaś dla *Tożsamości*, uzyskane wyniki są na wysokim poziomie zarówno dla *precyzji* oraz *kompletności*. *Brak relacji* poprawnie wykrywany był w 55% przypadków.

Podsumowując, można zauważyć, że uzyskane wyniki dla języka polskiego odbiegają od wyników uzyskanych dla języka angielskiego. Warto się zatem zastanowić, czy cechy wykorzystywane w badaniach nie są zbyt prostymi cechami dla języka polskiego. Szukanie odpowiedzi na tak postawione pytanie przedstawione jest w rozdziale 6, w którym opisano przeprowadzone eksperymenty z wykorzystaniem metody *PAS*.

[2] `http://clarin-pl.eu/`

3.2. Podobieństwo tekstów a wykrywanie relacji między fragmentami tekstów

Jednym z zadań rozwiązywanych w dziedzinie przetwarzania języka naturalnego jest określanie podobieństwa tekstów. Jest ono wykorzystywane na przykład w systemach antyplagiatowych w celu sprawdzenia czy analizowane teksty są do siebie podobne i w jakim stopniu. Rozwiązanie tego zadania wymaga zdefiniowania miary określającej podobieństwo dwóch tekstów/fragmentów tekstów. Możemy ogólnie powiedzieć, że miara ta, za pomocą wartości liczbowej, określa stopień podobieństwa (lub niepodobieństwa) dwóch obiektów.

Sposób pomiaru podobieństwa jest ściśle powiązany ze sposobem reprezentacji obiektów, między którymi chcemy to podobieństwo policzyć. Istotne jest również jakimi atrybutami obiekty są opisywane. Dla atrybutów ilościowych, które przyjmują wartości liczbowe podobieństwo *Sim* często bazuje na obliczaniu odległości $d_{i,j}$ pomiędzy obiektami i oraz j. Dla cech o wartościach binarnych, czy nominalnych (przyjmujących wartości będące etykietami, dla których nie istnieje wynikające z natury danego zjawiska uporządkowanie) nie można wprowadzić pojęcia odległości i wymagane jest zdefiniowanie innych miar.

Aby możliwe było stwierdzenie, że dany obiekt jest bardziej podobny do danego, a do innego mniej, wartość generowana przez miarę, powinna być skalowana do określonego przedziału. Przyjmując, że o_1 oraz o_2 to obiekty, między którymi określane jest podobieństwo za pomocą miary s, warunki jakie powinna spełniać ta miara są następujące (Schenker i inni, 2005):

$$0 \geqslant s(o_1, o_2) \geqslant 1 \tag{3.2}$$

$$s(o_1, o_2) = s(o_2, o_1) \tag{3.3}$$

$$s(o_1, o_2) = 1 \rightarrow o_1 \cong o_2 \tag{3.4}$$

Punkt 3.2 mówi o tym, że wartości generowane przez miarę powinny być z przedziału $\langle 0, .., 1 \rangle$. Wzór 3.3 oznacza, że miara powinna być symetryczna, tzn. wartość podobieństwa obiektu o_1 do obiektu o_2, powinna być taka sama jak wartość podobieństwa o_2 do o_1. Zależność 3.4 informuje o tym, że jeżeli o_1 jest identyczne z o_2, wtedy miara powinna generować wartość 1. W przypadku niespełnienia jednej z właściwości powinno się mówić o odległości dwóch obiektów, nie o ich podobieństwie.

W dalszej części przyjmiemy, że miara spełniająca założenia 3.2, 3.3 oraz 3.4, będzie oznaczana jako s. Jeżeli miara definiowana jest na podstawie odległości, wówczas zapis przyjmuje oznaczenie d, co niekoniecznie musi oznaczać spełnienie założeń 3.2, 3.3 oraz 3.4. Dolny indeks przypisany do tych oznaczeń, na przykład d_{XYZ} lub s_{XYZ}, oznacza nazwę miary. Na przykład zapis d_{WGU} wskazuje, że miara o nazwie WGU została zdefiniowana na podstawie odległości. Jak wspomniano, sposób obliczania podobieństwa między tekstami/ fragmentami tekstów (ogólnie obiektów) jest zależny od reprezentacji tych obiektów. W dalszej części skoncentrowano się na dwóch z nich. W podrozdziale 3.2.1 przedstawione zostały sposoby mierzenia podobieństwa dwóch obiektów przy wykorzystaniu ich reprezentacji wektorowej. W podrozdziale 3.2.2 przedstawiono opisane w literaturze miary podobieństwa służące do określania podobieństwa grafów.

3.2.1. Miary podobieństwa tekstów dla reprezentacji wektorowej

Najczęściej wykorzystywaną reprezentacją tekstu jest reprezentacja wektorowa. Poszczególne współrzędne wektora są są cechami opisującymi tekst. Wykorzystujemy wówczas miary, które operują na wektorach cech. Podstawowymi miarami wykorzystywanymi do mierzenia podobieństwa wektorów są:

1. *Miara kosinusowa* (równanie (2.4) oblicza kąt między dwoma wektorami (Dehak i inni, 2010).
2. *Odległość euklidesowa* (równanie 2.5) (Gower, 1982).

3.2.2. Miary podobieństwa dla reprezentacji grafowej tekstów

Metoda PAS do reprezentacji wiedzy wykorzystuje grafy. Określanie podobieństwa dwóch tekstów sprowadza się w tym wypadku do określenia podobieństwa między dwoma grafami reprezentującymi te teksty. Bardziej formalnie możemy powiedzieć, że podobieństwo dwóch tekstów $\mathcal{T}_1$ oraz $\mathcal{T}_2$ wyrażane jest jako podobieństwo grafu $\mathbf{G}_1$ reprezentującego $\mathcal{T}_1$ do grafu $\mathbf{G}_2$ reprezentującego tekst $\mathcal{T}_2$ i wyrażane jest za pomocą funkcji $Sim(\mathbf{G}_1, \mathbf{G}_2)$.

Jedną z najbardziej znanych miar do określania podobieństwa grafów jest *Graph Edit Distance* (GED). Idea tej miary opiera się na odległości edycyjnej zaproponowanej w pracy (Wagner i Fischer, 1974), służącej do określania liczby operacji potrzebnych do przekształcenia jednego ciągu znaków w drugi. W (Sanfeliu i Fu, 1983) zaadaptowano ideę odległości edycyjnej do grafów. Wykorzystywane są operacje: wstawiania, usuwania oraz zmiany etykiety (zarówno węzłów jak i krawędzi). Każda z operacji posiada swój koszt wykonania. Formalnie jest to funkcja rzutująca $M : \mathbf{G}_x \rightarrow \mathbf{G}_y$, gdzie $\mathbf{G}_x \subseteq \mathbf{G}_1$, a $\mathbf{G}_y \subseteq \mathbf{G}_2$, $\mathbf{G}_1 = (V_1, E_1)$, a $\mathbf{G}_2 = (V_2, E_2)$; V to zbiór węzłów w grafie, E to zbiór krawędzi. Sposób obliczania miary GED w przekształcającej graf $\mathbf{G}_1$ w graf $\mathbf{G}_2$ za pomocą akcji atomowych można przedstawić jako sekwencję kolejnych kroków do wykonania:

1. Jeżeli węzeł $n \in V_1$, ale $n \notin V_x$, wtedy usuwany jest węzeł n z kosztem c_{nd}.
2. Jeżeli $n \in V_2$, ale $n \notin V_y$, wtedy dodawany jest węzeł n z kosztem c_{ni}.
3. Jeżeli $M(n_i) = n_j$, dla $n_i \in V_x$ oraz $n_j \in V_y$, wtedy zastępowany jest węzeł v_i z v_j z kosztem c_{ns}.
4. Jeżeli krawędź $e \in E_1$, ale $e \notin E_x$, wtedy usuwana jest krawędź e z kosztem c_{ed}.
5. Jeżeli $e \in E_2$, ale $e \notin E_y$, wtedy wstawiana jest krawędź e z kosztem c_{ei}.
6. Jeżeli $M(e_i) = e_j$, dla $e_i \in E_x$, oraz $e_j \in E_y$, wtedy zastępowana jest krawędź e_i z e_j z kosztem c_{es}.

Warto zaznaczyć, że dla danej pary grafów $(\mathbf{G}_1, \mathbf{G}_2)$ istnieje wiele rzutowań M. Koszt pojedynczego rzutowania oznaczany jest jako $\gamma(M)$ i jest sumą kosztów c, operacji atomowych. Jako wartość miary s_{GED}, wybierane jest rzutowanie z najmniejszym kosztem $\gamma(M)$:

$$\begin{aligned} Sim(\mathbf{G}_1, \mathbf{G}_2) &= s_{GED}(\mathbf{G}_1, \mathbf{G}_2) \\ &= \min_{\forall M}(\gamma(M)) \end{aligned} \tag{3.5}$$

Przystępując do opisu kolejnych miar, należy wprowadzić pojęcie *maksymalnego wspólnego podgrafu* (oznaczanego jako *mcs*) oraz *minimalnego wspólnego supergrafu*

(oznaczanego jako MCS). Definicje mcs oraz MCS zaczerpnięte zostały z pracy (Bunke, 1997).

Maksymalny wspólny graf g $mcs(\mathbf{G}_1, \mathbf{G}_2)$ rozumiany jest jako graf, zawierający węzły i krawędzie występujące zarówno w $\mathbf{G}_1$ jak i $\mathbf{G}_2$ i niezawierający węzłów ani krawędzi występujących tylko w jednym z grafów $\mathbf{G}_1$ lub $\mathbf{G}_2$. Formalnie możemy to wyrazić w postaci trzech warunków do spełnienia:

1. $\mathbf{g} \subseteq \mathbf{G}_1$,
2. $\mathbf{g} \subseteq \mathbf{G}_2$,
3. nie ma innych podgrafów $\mathbf{g}'$ ($\mathbf{g}' \subseteq \mathbf{G}_1$,$\mathbf{g}' \subseteq \mathbf{G}_2$), takich, że $|\mathbf{g}'| > |\mathbf{g}|$. W tym zapisie, oznaczenie $|\mathbf{G}|$, to wielkość grafu $\mathbf{G}$, mierzona za pomocą liczby węzłów w tym grafie.

Jeżeli graf $\mathbf{G}_i$ jest podgrafem grafu $\mathbf{G}_j$, wtedy graf $\mathbf{G}_j$ jest supergrafem grafu $\mathbf{G}_i$. Definicja **minimalnego wspólnego supergrafu g**, $MCS(\mathbf{G}_1, \mathbf{G}_2)$ przekłada się na trzy warunki do spełnienia:

1. $\mathbf{G}_1 \subseteq \mathbf{g}$,
2. $\mathbf{G}_2 \subseteq \mathbf{g}$
3. nie ma innych supergrafów $\mathbf{g}'$ ($\mathbf{G}_1 \subseteq \mathbf{g}'$, $\mathbf{G}_2 \subseteq \mathbf{g}'$), takich, że $|\mathbf{g}'| < |\mathbf{g}|$.

Sposób wyznaczania mcs oraz MCS przedstawiony jest w (Levi, 1973).

Równanie 3.6 przedstawia odległość (ang. *distance*) – d, zaproponowaną przez Bunke (Bunke i Shearer, 1998), w której wykorzystuje się maksymalny wspólny podgraf $\mathbf{g}$, znormalizowany przez rozmiar grafu większego. Ze względu na to, że nie jest spełniona właściwość 3.4 jest ona oznaczana jako d (czyli odległość), a nie miara podobieństwa s. Jeśli przyjmiemy, że $\mathbf{G}_1 = \mathbf{G}_2$ (grafy są identyczne), wtedy $mcs(\mathbf{G}_1, \mathbf{G}_2) = \mathbf{G}_1 = \mathbf{G}_2$ a $max\{|\mathbf{G}_1|, |\mathbf{G}_2|\} = |\mathbf{G}_1| = |\mathbf{G}_2|$, zatem wartość podobieństwa $Sim(\mathbf{G}_1, \mathbf{G}_2)$ jest równa $1 - \frac{|\mathbf{G}_1|}{|\mathbf{G}_1|} = 0$. Jeśli zaś $mcs(\mathbf{G}_1, \mathbf{G}_2) = 0$, wtedy wartość podobieństwa $Sim(\mathbf{G}_1, \mathbf{G}_2)$ będzie równa 1.

Aby przekształcić tę odległość na miarę, wystarczy uwzględnić jedynie odjemnik w równaniu 3.6. Wtedy wielkość tę można interpretować jako miarę podobieństwa s, wysoką w pierwszym przypadku oraz niską w przypadku drugim. Równanie 3.7 przedstawia odległość d_{MCS} przekształconą na miarę podobieństwa s_{MCS}. Zgodnie z założeniami dotyczącymi miary, obiekty identyczne posiadają wartość podobieństwa równą 1.

$$\begin{aligned} Sim(\mathbf{G}_1, \mathbf{G}_2) &= d_{MCS}(\mathbf{G}_1, \mathbf{G}_2) \\ &= 1 - \frac{|mcs(\mathbf{G}_1, \mathbf{G}_2)|}{max\{|\mathbf{G}_1|, |\mathbf{G}_2|\}} \end{aligned} \tag{3.6}$$

$$\begin{aligned} Sim(\mathbf{G}_1, \mathbf{G}_2) &= s_{MCS}(\mathbf{G}_1, \mathbf{G}_2) \\ &= \frac{|mcs(\mathbf{G}_1, \mathbf{G}_2)|}{max\{|\mathbf{G}_1|, |\mathbf{G}_2|\}} \end{aligned} \tag{3.7}$$

Alternatywna do 3.6 odległość została zaproponowana w (Wallis i inni, 2001), (wzór 3.8). Podobnie jak 3.6, oblicza ona odległość między dwoma porównywanymi grafami, uwzględniając *mcs*.

$$\begin{aligned} Sim(\mathbf{G}_1, \mathbf{G}_2) &= d_{WGU}(\mathbf{G}_1, \mathbf{G}_2) \\ &= 1 - \frac{|mcs(\mathbf{G}_1, \mathbf{G}_2)|}{|\mathbf{G}_1| + |\mathbf{G}_2| - |mcs(\mathbf{G}_1, \mathbf{G}_2)|} \end{aligned} \tag{3.8}$$

$$\begin{aligned} Sim(\mathbf{G}_1, \mathbf{G}_2) &= s_{WGU}(\mathbf{G}_1, \mathbf{G}_2) \\ &= \frac{|mcs(\mathbf{G}_1, \mathbf{G}_2)|}{|\mathbf{G}_1| + |\mathbf{G}_2| - |mcs(\mathbf{G}_1, \mathbf{G}_2)|} \end{aligned} \tag{3.9}$$

Produkowane wartości są z przedziału $\langle 0, 1\rangle$. Wartość 1 oznacza całkowite niepodobieństwo (maksymalną odległość między grafami), a 0 minimalną odległość między grafami (identyczność grafów). Różnica względem odległości opisanej równaniem 3.6 dotyczy sposobu normalizacji wyliczonej wielkości grafu *mcs*. Uwzględnia ona zarówno sumę wielkości obu grafów jak i wielkość grafu *mcs*. Jeśli $mcs(\mathbf{G}_1, \mathbf{G}_2) = 0$, to odległość przyjmuje wartość maksymalnej odległości dwóch grafów. Jeśli zaś $\mathbf{G}_1 = \mathbf{G}_2$, to $mcs(\mathbf{G}_1, \mathbf{G}_2) = \mathbf{G}_1 = \mathbf{G}_2$, a $d_{WGU}(\mathbf{G}_1, \mathbf{G}_2) = 0$, czyli oznacza identyczność grafów. Przejście z odległości d_{WGU} do miary podobieństwa s_{WGU} odbywa się analogicznie jak dla wyrażenia 3.7 i polega na uwzględnieniu odjemnika z wyrażenia 3.8. Miara opisana jest równaniem 3.9.

Odległość przedstawiona równaniem 3.10 zaproponowana przez Bunke w pracy (Bunke, 1997), w odróżnieniu od 3.7 oraz 3.9, nie uwzględnia współczynnika normalizującego podobieństwo grafów do przedziału $\langle 0, 1\rangle$ oraz uniemożliwia przejście w prosty sposób z odległości d_{UGU} na miarę s.

$$\begin{aligned} Sim(\mathbf{G}_1, \mathbf{G}_2) &= d_{UGU}(\mathbf{G}_1, \mathbf{G}_2) \\ &= |\mathbf{G}_1| + |\mathbf{G}_2| - 2 \cdot |mcs(\mathbf{G}_1, \mathbf{G}_2)| \end{aligned} \tag{3.10}$$

Zachowana została jednak idea dotycząca podobnych grafów. Jeżeli $\mathbf{G}_1 = \mathbf{G}_2$, wtedy $d_{UGU}(\mathbf{G}_1, \mathbf{G}_2) = 0$. Oznacza to, że najniższa wartość miary – równa 0, wskazuje na identyczność grafów. Wartości wyższe określają moc niepodobieństwa, która nie posiada skończonej górnej granicy. Uogólniając, odległość d_{UGU} przyjmuje wartości z przedziału $\langle 0, |\mathbf{G}_1 \cup \mathbf{G}_2|\rangle; |\mathbf{G}_1 \cup \mathbf{G}_2| \in \mathbb{N}_+$.

W odróżnieniu od przedstawionych wcześniej miar podobieństwa wykorzystujących *mcs*, miara wyrażona równaniem 3.11 uwzględnia również wielkość *supergrafu MCS*. Jest definiowana jako różnica między wielkością supergrafu MCS a wielkością maksymalnego wspólnego podgrafu *mcs*.

$$\begin{aligned} Sim(\mathbf{G}_1, \mathbf{G}_2) &= d_{MMCS}(\mathbf{G}_1, \mathbf{G}_2) \\ &= |MCS(\mathbf{G}_1, \mathbf{G}_2)| - |mcs(\mathbf{G}_1, \mathbf{G}_2)| \end{aligned} \tag{3.11}$$

Zachowana jest tutaj zasada niepodobieństwa dwóch grafów. Im wyższa wartość produkowana przez miarę, tym mniej podobne są do siebie dwa grafy. Jeśli $\mathbf{G}_1 = \mathbf{G}_2$, to $mcs(\mathbf{G}_1, \mathbf{G}_2) = MCS(\mathbf{G}_1, \mathbf{G}_2)$, co odpowiada wartości tej miary równej 0.

Alternatywą do miary podobieństwa opisanej w 3.11, uwzględniającą *mcs* oraz MCS jest odległość d_{MMCSN} zaproponowana w pracy (Fernández i Valiente, 2001). Jest ona przedstawiona wzorem 3.12. Podobnie do 3.6 czy też 3.8, odległość ta skalowana do przedziału $\langle 0, 1\rangle$, gdzie 0 oznacza minimalną odległość (maksymalne podobieństwo) a 1 maksymalną odległość (minimalne podobieństwo).

$$\begin{aligned} Sim(\mathbf{G}_1, \mathbf{G}_2) &= d_{MMCSN}(\mathbf{G}_1, \mathbf{G}_2) \\ &= 1 - \frac{|mcs(\mathbf{G}_1, \mathbf{G}_2)|}{|MCS(\mathbf{G}_1, \mathbf{G}_2)|} \end{aligned} \tag{3.12}$$

$$\begin{aligned} Sim(\mathbf{G}_1, \mathbf{G}_2) &= s_{MMCSN}(\mathbf{G}_1, \mathbf{G}_2) \\ &= \frac{|mcs(\mathbf{G}_1, \mathbf{G}_2)|}{|MCS(\mathbf{G}_1, \mathbf{G}_2)|} \end{aligned} \tag{3.13}$$

Przekształcenie odległości d_{MMCSN} do miary s_{MMCSN} przebiega analogiczne do poprzednich i polega na uwzględnieniu z równania 3.12 jedynie odjemnika. Miara przyjmuje postać przedstawioną w równaniu 3.13. Jeśli przyjmiemy, że $\mathbf{G}_1 = \mathbf{G}_2$, to $mcs(\mathbf{G}_1, \mathbf{G}_2) = MCS(\mathbf{G}_1, \mathbf{G}_2)$, a to odpowiada wartości $s_{MMCSN}(\mathbf{G}_1, \mathbf{G}_2) = 1$.

Do obliczania podobieństwa grafów, poza miarami wykorzystującymi *maksymalny wspólny podgraf* oraz *minimalny wspólny supergraf*, można zaadaptować miary operujące na zbiorach. Jedną z takich miar jest podobieństwo Jaccarda (Jaccard, 1912) przedstawione w równaniu 3.14. Wzór ten można prosto przekształcić do obliczenia podobieństwa grafów. Wystarczy potraktować zbiór A jako zbiór węzłów grafu $\mathbf{G}_1$, a zbiór B jako zbiór węzłów grafu $\mathbf{G}_2$.

$$J = \frac{|A \cap B|}{|A| + |B| - |A \cap B|} \tag{3.14}$$

Zatem podobieństwo Jaccarda dostosowane do obliczania podobieństwa grafów będzie posiadało postać przedstawioną w równaniu 3.15.

$$\begin{aligned} Sim(\mathbf{G}_1, \mathbf{G}_2) &= s_{JACCARD}(\mathbf{G}_1, \mathbf{G}_2) \\ &= \frac{|V_1 \cap V_2|}{|V_1| + |V_2| - |V_1 \cap V_2|} \end{aligned} \tag{3.15}$$

gdzie V_1 oznacza zbiór węzłów grafu $\mathbf{G}_1$, V_2 to zbiór węzłów grafu $\mathbf{G}_2$, $|V_1|$ to liczba wierzchołków grafu $\mathbf{G}_1$, a $|V_2|$ jest liczbą wierzchołków grafu $\mathbf{G}_2$.

Przedstawione wcześniej miary wykorzystujące *mcs* oraz MCS, nie uwzględniają sytuacji, w której porównywane grafy będą posiadały taki sam zestaw węzłów, jednak będą połączone innymi krawędziami. Z tego względu, w niniejszej pracy zaproponowana została miara s_{CTXBow} opisana równaniem 3.21. Jako składowa miary s_{CTXBow} występuje s_{SN} bazująca na mierze Jaccarda, opisana wzorem 3.16. Określa ona liczbę tych samych węzłów występujących w grafach $\mathbf{G}_1$ i $\mathbf{G}_2$, ale normalizowana jest przez sumę różnych węzłów obu grafów. Jeżeli przecięcie zbiorów $V_1 \cap V_2$ będzie puste (grafy

nie posiadają ani jednego takiego samego węzła), wtedy miara produkuje wartość 0, jeśli $\mathbf{G}_1 = \mathbf{G}_2$, to $s_{SN}(\mathbf{G}_1, \mathbf{G}_2) = 1$.

$$\begin{aligned} Sim(\mathbf{G}_1, \mathbf{G}_2) &= s_{SN}(\mathbf{G}_1, \mathbf{G}_2) \\ &= \frac{|V_1 \cap V_2|}{|V_1 \cup V_2|} \end{aligned} \tag{3.16}$$

Miara s_{CTXBow} uwzględnia nie tylko przecięcie się grafów ze sobą, ale także sąsiedztwo NBH takich samych węzłów n w obu grafach. Sprawdza podobieństwo sąsiedztwa węzła n w grafie $\mathbf{G}_1$ i grafie $\mathbf{G}_2$, gdzie jako podobieństwo rozumiana jest część wspólna obu tych sąsiedztw. Dla grafów skierowanych, sąsiedztwo definiowane jest jako węzły wejściowe n_{in}, czyli węzły, od których krawędź skierowana jest do węzła n, oraz wyjściowe n_{out}, czyli węzły, do których istnieje krawędź skierowana od węzła n.

$$NBH(\mathbf{G}_1(n)) = \{\mathbf{G}_1(n_{in}) \cup \mathbf{G}_1(n_{out})\} \tag{3.17}$$

$$NBH(\mathbf{G}_2(n)) = \{\mathbf{G}_2(n_{in}) \cup \mathbf{G}_2(n_{out})\} \tag{3.18}$$

$$s_{NBH}(\mathbf{G}_1, \mathbf{G}_2, n) = \frac{|NBH(\mathbf{G}_1(n)) \cap NBH(\mathbf{G}_2(n)))|}{|NBH(\mathbf{G}_1(n)) \cup NBH(\mathbf{G}_2(n))|} \tag{3.19}$$

$$\begin{aligned} \mathbf{G}_{min} = \mathbf{G}_1 &\iff |\mathbf{G}_1| \leqslant |\mathbf{G}_2| \\ \mathbf{G}_{min} = \mathbf{G}_2 &\iff |\mathbf{G}_2| < |\mathbf{G}_1| \end{aligned} \tag{3.20}$$

W powyższych wzorach $\mathbf{G}(n_{in})$ to węzły wejściowe do węzła n w grafie $\mathbf{G}$, a $\mathbf{G}(n_{out})$ to węzły wyjściowe z węzła n w tym grafie. W równaniu 3.17 zdefiniowane zostało otoczenie węzła n w grafie $\mathbf{G}_1$, a za pomocą równania 3.18 otoczenie tego węzła w grafie $\mathbf{G}_2$. Tak zdefiniowane otoczenia traktowane są jako zbiory węzłów, dla których można zastosować miarę opisaną równaniem 3.16. Podstawiając za $V_1 = NBH(\mathbf{G}_1(n)))$, $V_2 = NBH(\mathbf{G}_2(n)))$, można obliczyć podobieństwo sąsiedztw węzła n w grafach $\mathbf{G}_1$ oraz $\mathbf{G}_2$. Po podstawieniu równań 3.17 oraz 3.18 do 3.16 otrzymujemy miarę podobieństwa $s_{NBH}(\mathbf{G}_1, \mathbf{G}_2, n)$, opisaną wzorem 3.19.

Aby obliczyć podobieństwo między grafami $\mathbf{G}_1$ i $\mathbf{G}_2$, dla każdego węzła w grafie mniejszym $\mathbf{G}_{min}$ (warunki wyboru grafu mniejszego opisane są w równaniu 3.20) obliczana jest wartość miary podobieństwa określona równaniem 3.19. Graf $\mathbf{G}_{min}$ wybierany jest ze względów optymalizacyjnych, ponieważ obliczenie podobieństw sąsiedztw węzła n możliwe jest tylko w przypadku, kiedy zarówno $\mathbf{G}_1$, jak i $\mathbf{G}_2$ zawierają ten węzeł. Dla węzłów w grafie $\mathbf{G}_{min}$ obliczana jest suma wartości podobieństw ze wzoru 3.19, przedstawiona we wzorze 3.21.

$$\begin{aligned} Sim(\mathbf{G}_1, \mathbf{G}_2) &= s_{CTXBow}(\mathbf{G}_1, \mathbf{G}_2) \\ &= \frac{\sum_{n \in \mathbf{G}_{min}} s_{NBH}(\mathbf{G}_1, \mathbf{G}_2, n)}{|\mathbf{G}_{min}|} \end{aligned} \tag{3.21}$$

Aby miara produkowała wartości w przedziale $\langle 0, 1 \rangle$, suma podobieństw otoczeń dla węzłów z grafu $\mathbf{G}_{min}$ normalizowana jest przez wielkość tego grafu – liczbę węzłów w tym grafie. Łatwo zauważyć, że jeżeli otoczenie każdego węzła n będzie takie samo, wartość produkowana przez miarę będzie równa 1. Jeżeli zaś grafy nie będą posiadały wspólnych węzłów, wtedy dla każdego węzła n w $\mathbf{G}_{min}$: $|s_{NBH}(\mathbf{G}_1, \mathbf{G}_2, n)| = 0$, a miara wyprodukuje wartość równą 0. Wartości bliższe 1 wskazują na większe podobieństwo grafów do siebie, wartości bliższe do 0 wskazują na większe niepodobieństwo obu grafów.

Rozdział 4

Autorska metoda pogłębianej analizy semantycznej – PAS

Podczas projektowania metody ***P**ogłębianej **A**nalizy **S**emantycznej* (*PAS*), uwzględnione zostały następujące założenia i ograniczenia:

- wejściem do metody jest tekst napisany w języku naturalnym,
- metoda nie powinna być zależna od języka,
- metoda powinna pozwalać użytkownikowi na kontrolowanie poziomu pogłębiania informacji wydobywanych z tekstu w kontekście:
  - słów występujących w tekście,
  - relacji między tymi słowami,
- informacje wydobyte z tekstu mogą być niekompletne,
- metoda powinna umożliwiać dołączanie wiedzy pozatekstowej,
- metoda powinna mieć modularną strukturę,
- wiedza wydobyta z tekstu oraz wiedza pozatekstowa będą reprezentowane za pomocą grafów (sformalizowana wiedza).

Opis metody rozpoczyna się od przedstawienia jej idei a w kolejnym podrozdziale szczegółowo omówione są wszystkie jej moduły.

4.1. Ogólna idea metody

Na rysunku 4.1 przedstawiono ogólną wizję sposobu pogłębiania wiedzy wydobywanej z tekstu, który został zaimplementowany w metodzie *PAS*. Wejściem jest tekst napisany w języku naturalnym oraz wymagania użytkownika opisane w *scenariuszu realizacji*. Wyjściem jest sformalizowana wiedza. Jest ona pozyskiwana w sposób przyrostowy na podstawie stopniowego przetwarzania tekstu wejściowego. Można do tej wiedzy dołączyć również wiedzę wydobytą z innych źródeł. Taki przyrostowy proces powiększania wiedzy o przetwarzanym tekście został nazwany *pogłębianiem*. Szare strzałki na rysunku symbolizują iteracyjne wykonywanie modułów.

Metoda rozpoczyna działanie od *przetwarzania wstępnego*, obejmującego głównie wstępną obróbkę tekstu oraz przypisywanie do niego informacji składniowych i semantycznych.

W module *Analiza wiedzy* do istniejącej bazy wiedzy dołączane są coraz dokładniejsze informacje wydobywane z tekstu. W tym celu metoda wykorzystuje między innymi opracowane w ramach niniejszej pracy płytkie parsowanie semantyczne (Kędzia i Maziarz, 2013), ujednoznacznianie znaczeń słów (zaadaptowane z języka angielskiego i odpowiednio dopasowane do języka polskiego (Kędzia i inni, 2015, 2014b)) oraz wydobywanie nazw własnych (Marcinczuk, 2015). Ten moduł odpowiada za formalizację wydobywanej wiedzy. Opcjonalnie może tu być wykonana generalizacja wiedzy lub dołączanie coraz to dokładniejszych informacji wydobywanych z tekstu lub zewnętrznych źródeł wiedzy.

W etapie *Ekstrakcja wiedzy o świecie* do analizowanego tekstu można przypisać wiedzę pozatekstową. Zastosowanie tego etapu zależy od potrzeb użytkownika. Dołączanie do wiedzy tekstowej wiedzy pozatekstowej znajdującej się w ontologiach stało się możliwe dzięki opracowaniu w ramach pracy doktorskiej metody rzutowania polskiej Słowosieci na ontologię SUMO (Kędzia i Piasecki, 2014). Dołączanie nowej wiedzy wymaga jej wcześniejszej formalizacji i odbywa się w module *Analiza wiedzy*.

Dzięki temu, że struktura metody jest zaprojektowana w sposób modularny, możliwe jest jej dostosowanie do konkretnych potrzeb użytkownika. Jeżeli nie zachodzi potrzeba dołączania wiedzy pozatekstowej (np. zawartej w Słowosieci), istnieje możliwość formalizacji i pogłębiania wiedzy z *przetwarzania wstępnego*. Jeśli interesująca jest wiedza pozatekstowa, bez wiedzy wyciągniętej z analizowanego tekstu (dalej nazywaną skrótowo wiedzą tekstową), to możliwe jest wykorzystanie modułu *Ekstrakcja wiedzy o świecie* i formalizacja tej wiedzy w module *Analiza wiedzy*. Jeżeli jednak użytkownik zainteresowany jest wiedzą pełnotekstową oraz pozatekstową, wówczas w module *Analiza wiedzy* dokonywane jest połączenie wydobytej wiedzy.

Podczas kolejnych kroków, moduł *przetwarzania wstępnego* wykorzystuje *scenariusz użytkownika*, aby odczytać sposób pogłębiania oraz poziomy podstawowego przetwarzania. Informacja może być wydobywana z tekstu zarówno na poziomie pojedynczych słów (np. znaczenie słowa, część mowy danego słowa), słów wielowyrazowych (np. wielowyrazowe nazwy własne – *Zakład Ubezpieczeń Społecznych*), jak i relacji między słowami (np. określenie relacji semantycznej zachodzącej między dwoma słowami – *Jan zabił Pawła* ***nożem*** *– nóż* jest *instrumentem* wykorzystanym do zabicia).

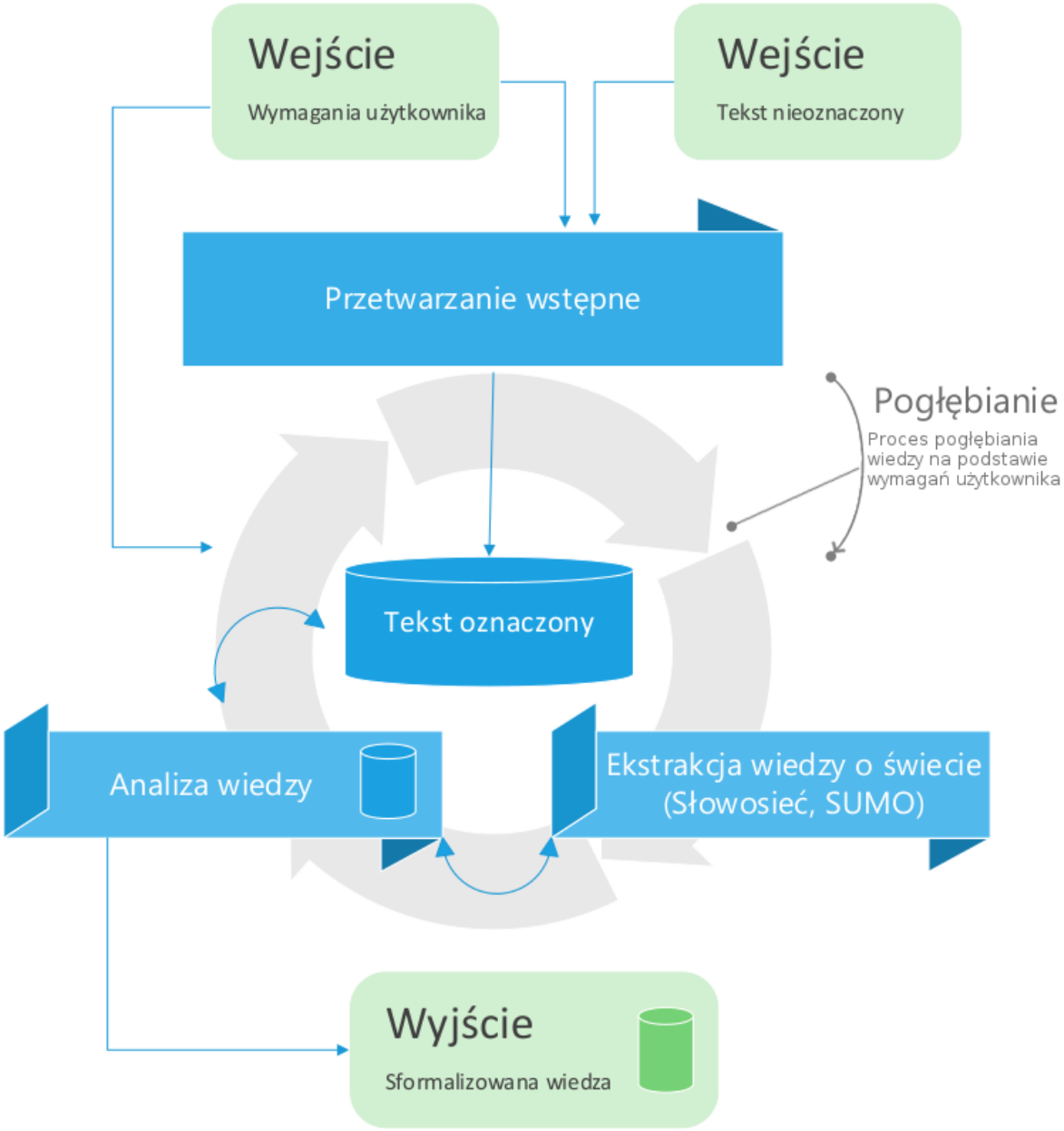
Słowo występujące w tekście może zostać przedstawione oraz zinterpretowane w różny sposób, w zależności od poziomu przetwarzania, na jakim operujemy. Dla przykładu słowo *ogrodu* w zdaniu:

Z ogrodu zoologicznego we Wrocławiu uciekł wąż Boa Dusiciel i przemieszcza się w stronę Ostrowa Tumskiego.

może zostać zinterpretowane na poziomie:

1. *segmentacji* jako ciąg znaków $[o, g, r, o, d, u]$;
2. *morfosyntaktycznym* jako słowo, którego forma podstawowa to *ogród*, liczby pojedynczej w dopełniaczu, rodzaju męskiego;
3. *semantycznym* jest to znaczenie `‘ogród 1(msc)‘` ze Słowosieci, czy też określenie, że słowo jest elementem nazwy własnej *Ogród Zoologiczny*.

Ważnym założeniem w metodzie *PAS* jest możliwość dopasowania poziomu reprezentacji słowa (który to poziom, powiązany jest z poziomem przetwarzania języka) do



Rys. 4.1. Ogólna wizja zaproponowanej metody *Pogłębianej Analizy Semantycznej – PAS*

problemu rozwiązywanego z wykorzystaniem metody *PAS*. Szczegółowo sposoby reprezentacji słowa przedstawione są w podrozdziale 4.2.1.

4.2. Szczegółowy opis elementów składowych metody

Na rysunku 4.2 pokazano elementy składające się na metodę *PAS*. Wśród nich wyróżniono:

— *wymagania użytkownika* opisujące konkretną realizację przebiegu metody,
— *przetwarzanie wstępne*, podczas którego do tekstu przypisywane są podstawowe informacje, takie jak znaczenia słów czy też role semantyczne zachodzące między słowami z tekstu,
— *ekstrakcję wiedzy o świecie* z ontologii wysokiego poziomu (jest to możliwe jest dzięki rzutowaniu Słowosieci na ontologię SUMO),
— *analizę wiedzy*, gdzie następuje uaktualnianie i łączenie zapisanej w sposób formalny wiedzy. Przyjęto reprezentację grafową, gdyż umożliwia ona zapis oraz wnioskowanie z wiedzy niekompletnej.

Warto w tym miejscu zwrócić uwagę na fakt, że przyjęta w pracy reprezentacja wiedzy w postaci grafów odbiega od powszechnie przyjętego podejścia, w którym słowo z tekstu staje się węzłem w grafie, a etykieta węzła jest tożsama z formą bazową słowa (Abdulsahib i Kamaruddin, 2015).

W *PAS* każde słowo z tekstu odzwierciedlone jest za pomocą węzła w budowanym grafie, jednak metoda umożliwia określenie sposobu reprezentacji słowa (w *scenariuszu realizacji*), co w efekcie prowadzi do tego, że etykieta węzła nie musi posiadać takiej samej formy jak słowo w tekście. W zależności od wykorzystania metody może to być znaczenie, pojęcie, czy też forma bazowa słowa z tekstu. Relacje między słowami reprezentowane są przez skierowane krawędzie grafu z przypisanymi etykietami zgodnymi z typem relacji. Kierunek oraz nazwa krawędzi tożsame są z relacją zachodzącą między dwoma słowami.

Dalsza część rozdziału zawiera opis każdego modułu składowego metody *PAS*.

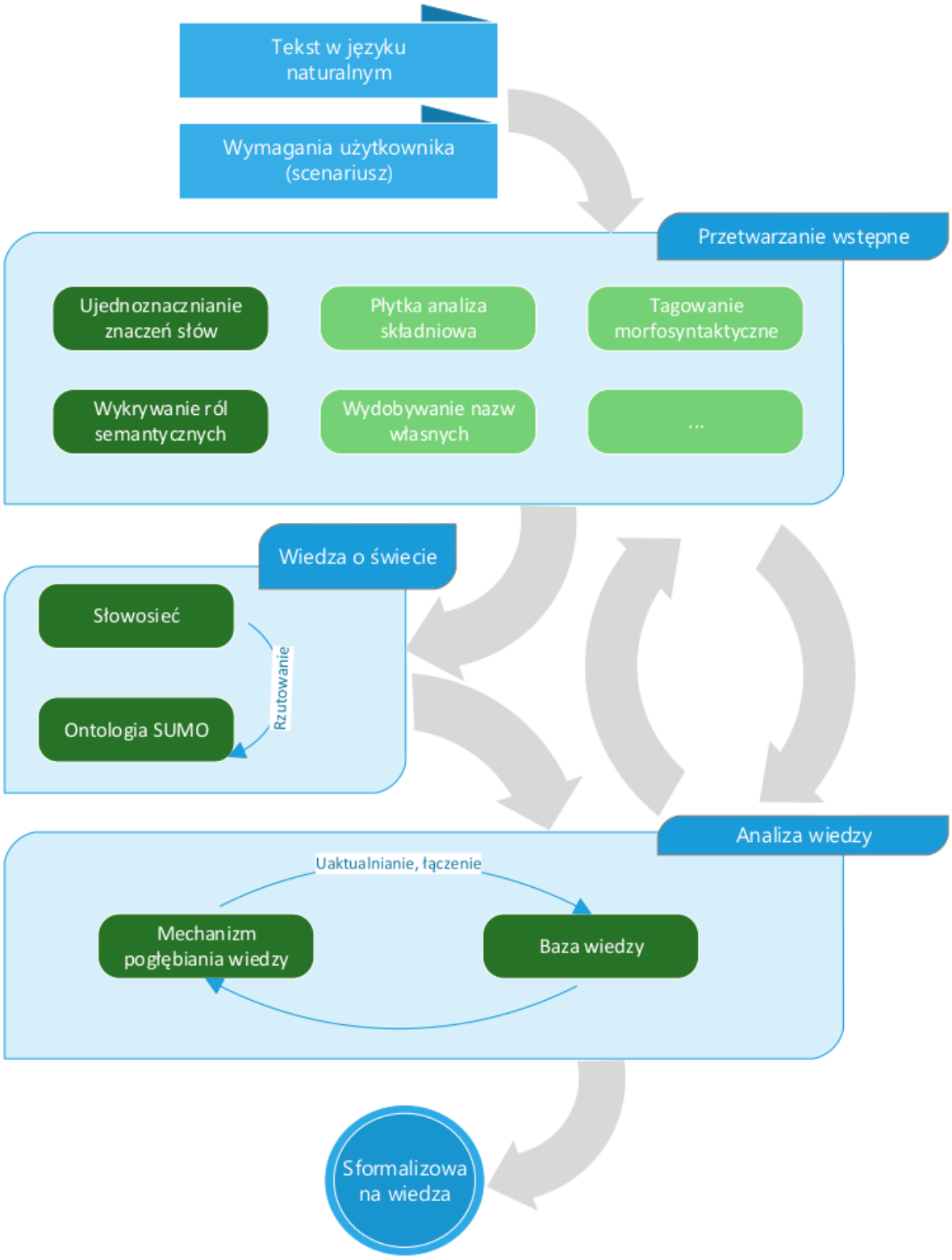
4.2.1. Wymagania użytkownika, scenariusz realizacji metody

Formalny opis konkretnego przebiegu metody *PAS* znajduje się w scenariuszu realizacji metody i wynika z wymagań użytkownika odnośnie modelowania relacji semantycznych. Scenariusz zapisywany jest w pliku konfiguracyjnym i podawany jako wejście do metody.

Wymagania użytkownika

Zanim przystąpimy do określenia scenariusza realizacji metody, skoncentrujemy się na określeniu wymagań odnośnie modelowania relacji semantycznych między fragmentami tekstów. Można to zilustrować następującym przykładem. Załóżmy, że postawiony przez użytkownika problem dotyczy modelowania zależności semantycznych wewnątrz tekstu oraz dołączania do tekstu reprezentacji wiedzy pozatekstowej (takiej, która nie jest bezpośrednio reprezentowana w tekście). Należy wówczas w module przetwarzania wstępnego wykonać następujące kroki:

1. podzielić tekst na tokeny, frazy, zdania,



Rys. 4.2. Szczegółowa wizja metody PAS

2. ujednoznacznić tekst morfosyntaktycznie,
3. wewnątrz zdań wyszukać i rozpoznać relacje składniowe,
4. określić role semantyczne poszczególnych słów w zdaniu,
5. rozpoznać wystąpienia nazw własnych, jak również wyrażenia będące odniesieniami do wystąpień nazw własnych,
6. ujednoznacznić tekst przypisując wyrazom konkretne znaczenia,
7. dołączyć do informacji tekstowych informacje pozatekstowe,
8. uogólnić informacje wydobyte z tekstu oraz informacje pozatekstowe.

Powyższa lista przedstawia jedną z możliwych realizacji metody, której efektem będzie reprezentacja semantyczna oraz składniowa. Lista kroków, które mogą być wykonane na tekście wejściowym wraz z odpowiadającymi im nazwami używanymi w pliku konfiguracyjnym scenariusza użycia, jest następująca (na końcu każdego kroku podana jest nazwa konkretnej aplikacji, która ten krok realizuje):

1. rozpoznanie i klasyfikacja wystąpień nazw własnych (Marcińczuk i inni, 2013) – `Liner`,
2. określenie relacji semantycznych pomiędzy wystąpieniami nazwami własnymi – `Serel`,
3. dzielenie zdania na frazy (frazy rzeczownikowe *NP*, przymiotnikowe *AdjP*, czasownikowe *VP* oraz uzgodnione *AgP*) (Radziszewski i Pawlaczek, 2012) – `Iobber`,
4. określenie zależności składniowych wewnątrz fraz *NP* oraz *AdjP* – `Defender`,
5. wykrywanie ról semantycznych wewnątrz fraz *NP* oraz *AdjP* – `Npsemrel` opisany w opracy (Kędzia i Maziarz, 2013),
6. ujednoznacznienie słów znaczeniami ze Słowosieci – `Wsd` (Kędzia i inni, 2015, Piasecki i inni, 2016a),
7. przypisanie do słów pojęć z ontologii SUMO – `Sumo`,
8. przypisanie do zdań reprezentacji za pomocą drzew zależności składniowych – `Malt` (Wróblewska, 2014).

Wymagania użytkownika względem wykonania metody obejmują również sposób reprezentacji w grafie pojedynczego słowa z tekstu. Węzeł w grafie, który reprezentuje słowo może opisywać je jako:

1. formę ortograficzną słowa, w jakiej wystąpiło ono w tekście,
2. formę podstawową słowa,
3. część mowy,
4. formę podstawową słowa połączoną z częścią mowy,
5. znaczenie słowa,
6. pojęcie z ontologii wysokiego poziomu, które jest uogólnieniem znaczenia tego słowa.

Wybór sposobu reprezentacji słowa wynika z celu, jaki stawia użytkownik metody. Jeżeli w końcowym etapie istotne jest rozróżnienie słowa po formie ortograficznej (na przykład odróżnienie słowa *ogrodzie* od *ogród*), to jako sposób reprezentacji słowa należy wybrać właśnie formę ortograficzną. Jeżeli nie jest to istotne, a zależy nam na przykład na określeniu jakimi relacjami składniowymi łączy się słowo *ogród* z innymi słowami, to należy wybrać reprezentację słowa w postaci formy podstawowej. Kiedy

ważne jest rozróżnienie czy *ogród* oznacza miejsce, w którym rosną rośliny, czy może jest to *ogród zoologiczny*, należałoby wybrać *znaczenie* jako sposób reprezentacji słowa.

Scenariusz realizacji metody

Jak już wspomniano, scenariusz jest formalnym opisem konkretnego przebiegu metody, wynikającym z wymagań użytkownika. Przedstawiony na początku tego rozdziału przykładowy przebieg metody, odnoszący się do rozważanego poprzednio przykładu, będzie posiadał następujący zapis:

```
flow_elems = Wsd Sumo Iobber Defender Npsemrel Malt
```

`flow_elems` oznacza listę *kroków do wykonania*. Separatorem między poszczególnymi krokami jest spacja. Kolejność podawanych kroków jest istotna, ponieważ przyjęto założenie, że wejściem do danego kroku jest wyjście z kroku poprzedniego. Dla przykładu wejściem do `Npsemrela` jest wyjście `Defendera`, od którego działanie `Npsemrela` jest zależne. Warto wspomnieć, że każdy wykonany krok dodaje do poprzedniego nowe informacje, nie kasując starych. Oprócz przebiegu metody, scenariusz użytkownika zawiera wszystkie parametry przebiegu metody. Opisują one:

— sposób reprezentacji słowa, ustawiany za pomocą zmiennej `node`; wartości, które mogą zostać przypisane, zostały przedstawione w dodatku B.1.1;

— sposób reprezentacji krawędzi użytych do budowy grafów, zmienna `ebuilders` nie musi posiadać ustawionej wartości, może również zawierać pełny zestaw krawędzi; pełny zestaw krawędzi przedstawiony został w dodatku B.1.2;

— zestaw metod do przebudowy grafów, które opisane zostały w dodatku B.1.3;

— sposób łączenia wiedzy tekstowej ze Słowosiecią; od strony technicznej został on przedstawiony w dodatku B.1.4, od strony teoretycznej w rozdziale 4.2.4;

— możliwość filtrowania (odrzucania) słów podczas budowy grafów; sposoby filtrowania słów przedstawione zostały w dodatku B.1.5.

Jak widzimy, scenariusz realizacji metody określa jak słowo będzie reprezentowane w tekście, jaki zestaw krawędzi będzie dołączany, czy też jakie zasoby zewnętrzne będą dołączone do budowanego grafu. Jest to jeden z istotniejszych punktów wykonania metody. Jedyny, gdzie w interakcję wchodzi użytkownik. Należy również podkreślić, że raz zdefiniowany scenariusz może być wykorzystywany dowolną liczbę razy. Każdorazowe wykorzystanie tego samego scenariusza będzie skutkowało budową grafów o tych samych parametrach budowy. Z drugiej strony, dla dowolnego tekstu można użyć dowolnej liczby różnych scenariuszy, co w efekcie spowoduje budowę różnych grafów dla tego samego tekstu. Innymi słowy, każdy tekst może posiadać wiele reprezentacji semantycznych reprezentowanych za pomocą grafów.

4.2.2. Moduł przetwarzania wstępnego

Przystępując do modelowania tekstu oraz pogłębiania wiedzy zawartej w tekście, konieczne jest wykonanie obróbki tekstu wejściowego i przetransformowanie go do postaci, która będzie mogła być analizowana przez system w dalszych etapach przetwarzania oraz będzie zgodna z wymaganiami użytkownika. Przetwarzanie wstępne obejmuje szereg czynności, które przypisują do *nieoznaczonego tekstu* oznaczonego jako

(blok *Tekst nieoznaczony* na rys. 4.1) anotacje uzyskane na różnych poziomach przetwarzania języka naturalnego. To, jakie anotacje będą przypisywane opisane jest w *scenariuszu realizacji*. Użytkownik określa też poziom szczegółowości informacji, które chce wydobyć z tekstu. Przez poziom szczegółowości informacji, rozumie się bogactwo informacji przypisywane do tekstu, które powiązane jest z poziomami przetwarzania języka naturalnego.

Warto zaznaczyć, że przetwarzanie wstępne wykonywane jest tylko raz dla danego scenariusza. Oznacza to, że określanie wymagań również wykonywane jest tylko raz i nie może być zmieniane podczas uruchomienia metody. Dotyczy to zarówno kroków do wykonania, jak i sposobu reprezentacji słowa w grafie. Wyjściem z modułu *przetwarzania wstępnego* jest zbiór tekstów z przypisanymi anotacjami, który na rys. 4.1) przedstawiony został jako blok *Tekst oznaczony*. Taki zestaw tekstów jest wejściem do modułu *Analizy wiedzy*.

Modelowanie semantyki tekstu zachodzące w module przetwarzania wstępnego wymagało opracowania wielu metod, które zostały zaimplementowane w postaci zestawu użytecznych narzędzi. Dotyczy to wykrywania ról semantycznych i ujednoznaczniania znaczeń leksykalnych.

Wykrywanie ról semantycznych

Opracowanie narzędzia umożliwiającego wykrywanie ról semantycznych wewnątrz frazy nominalnej umożliwia określanie zależności semantycznych między słowami w analizowanym tekście. Wzbogaca to opis semantyczny tekstu o dodatkowe informacje semantyczne. Formalnie zależność taka zapisywana jest jako krawędź skierowana w grafie z przypisaną etykietą zgodną z wykrytą rolą semantyczną. Jest to jedna z następujących ról:

- **Agent** – inicjator działania lub obiekt w specjalnym stanie np. *(człowiek) wykształcony przez Jana*, *wyjący wilk*.
- **Obiekt** – jest to byt podlegający działaniu, zdarzeniu lub zmianie stanu spowodowanej przez innego uczestnika zdarzenia, np. *wykształcenie kogoś*, *(Jan) posiadający majątek*.
- **Instrument** – jest narzędziem, urządzeniem lub środkiem używanym przez kogoś w celu spowodowania czegoś, np. *przeszyty włócznią*, *lina cumownicza*.
- **Materiał** – to byt, który jest używany przez kogoś do produkcji czegoś, materiał ulega zmianie stanu powodując jego zniknięcie i powstanie nowego wytworu: *zrobiony z mosiądzu*, *mosiężna figurka*.
- **Cel** – byt lub sytuacja, do której skierowane jest wydarzenie lub osoba, która odnosi korzyści z wydarzenia: *wręczenie (medali) olimpijczykom*, *sala koncertowa*.
- **Miejsce** jest fizycznym miejscem, w którym dane zdarzenie jest zlokalizowane, miejscem docelowym zdarzenia, ścieżką lub źródłem ruchu, lub po prostu miejscem, w którym znajduje się dana osoba; np. *wręczenie (medali) w auli*, *przedzieranie się przez moczary*.
- **Czas** jest szczególnym momentem lub okresem trwania zdarzenia – lokalizuje sytuację w ciągu zdarzeń lub podaje czas jej trwania, np. *przedzieranie się przez godzinę*, *przedzieranie się w środę*.
- **Wydarzenie temporalne, przestrzenne** – relacja ta wskazuje na przestrzenną

lub czasową część, np. *poniedziałkowy poranek*, *środek zimy*, *koniec drogi*, *stolica kraju*.

— **Atrybut** – jest właściwością osoby lub zdarzenia, takiego jak: kolor, rozmiar, waga, intensywność, czas trwania itd., które mogą być wyrażone za pomocą przymiotnika jakościowego, np. *czerwony samochód*, *głośna muzyka*.
— **Członek rodziny/rodzina** – bezpośrednie lub pośrednie wyrażenie przynależności do rodziny, np. *syn króla*, *moja żona*.
— **Kolejność** podaje pozycję jednostki lub zdarzenia w uporządkowanej sekwencji, np. *druga odpowiedź*, *lata 80*.
— **Ilość** jest ilością czegoś lub licznością danego zbioru: *pięciu panów*, *kieliszek wina*.

Do wykrycia przytoczonego powyżej zbioru ról semantycznych zaproponowane zostało podejście regułowe (Kędzia i Maziarz, 2013). Pojedyncza reguła odpowiedzialna jest za wykrywanie konkretnej roli semantycznej. Reguły zapisane są w języku WCCL, którego opis znajduje się w pracy (Radziszewski i inni, 2011).

Listing 4.1. Przykładowy operator dla Obiektu zapisany w języku WCCL

```
@b:"PROTO-OBIEKT-acc" (
 and(
  equal(class[0], pact),
  or(
   equal(lex(base[0], "acc"), ["1"]),
   not(equal(
   lex(base[0], "frames"), ["1"]))
  ),
  in(class[1], {subst,depr,ger,adj}),
  equal(cas[1],acc)
 )
)
```

Przykładowa reguła służąca wykrywaniu roli *Obiekt* przedstawiona została na listingu 4.1. Jest tu przedstawiony operator o nazwie *PROTO-OBIEKT-acc*, który odwołuje się do dwóch pozycji w tekście, bieżącej oraz kolejnej. Jeżeli wszystkie warunki podane w funkcji *and* będą posiadały wartość *True*, wartość zwrócona przez operator będzie wynosiła *True*. Oznacza to, że w danym miejscu w tekście, pomiędzy wyrazami na pozycji bieżącej i kolejnej po bieżącej, zachodzi rola opisana tym operatorem. Opis pozostałych operatorów języka WCCL znajduje się w pracy (Radziszewski i inni, 2011) oraz dokumentacji technicznej języka WCCL dostępnej pod adresem `http://nlp.pwr.wroc.pl/redmine/projects/joskipi/wiki`.

W Tabeli 4.2.2 przedstawione zostały wyniki (*precyzja* – **P** (wzór 6.3), *kompletność* – **R** (wzór 6.4) oraz *miara-f* – **F**) osiągane z użyciem tej metody dla najczęstszych ról (agent, obiekt, atrybut, ilość). Rozpoznawanie ról zostało ocenione na dwóch zbiorach. Pierwszym z nich był *Korpus Języka Polskiego Politechniki Wrocławskiej* (KPWr) (Broda i inni, 2012) a drugim podkorpus milionowy Narodowego Korpusu Języka Polskiego (NKJP) (Przepiórkowski i inni, 2012).

Wykrywanie ról semantycznych poprzedzone jest procesem wzbogacania wyniku działania istniejącego płytkiego parsera IOBBER (Radziszewski i Pawlaczek, 2013),

Tabela 4.1. *Precyzja* (**P**), *kompletność* (**R**) oraz *miara-f* (**F**) wykrywania ról semantycznych dla Korpusu Politechniki Wrocławskiej **KPWr** oraz Narodowego Korpusu Języka Polskiego **NKJP**

	P		**R**		**F**	
	KPWr	**NKJP**	**KPWr**	**NKJP**	**KPWr**	**NKJP**
Agent	58,3	91,5	30,4	34,3	40,0	50,0
Obiekt	84,9	91,4	39,8	44,0	54,2	49,4
Atrybut	43,8	44,6	58,3	75,8	50,0	56,2
Ilość	83,3	62,4	55,6	43,6	66,7	51,4

polegającym na wykrywaniu zależności składniowych, które dodawane są do struktury opisującej dokument. Można tego dokonać wykorzystując narzędzie `defender`, które dostępne jest pod adresem `https://clarin-pl.eu/dspace/handle/11321/285`. Po uruchomieniu, dokument zostanie wzbogacony o zależności składniowe, do których w dalszym etapie, uruchamiając narzędzie `npsemrel` (dostępne pod adresem `https://clarin-pl.eu/dspace/handle/11321/304`) przypisane zostaną role semantyczne.

Ujednoznacznianie znaczeń leksykalnych

Ujednoznacznianie znaczeń leksykalnych to proces polegający na przypisywaniu do słów w tekście ich znaczeń z wybranego słownika znaczeń. Znaczenie przypisane do słowa, może być wykorzystane w metodzie *PAS* do określenia typu węzła reprezentującego słowo z tekstu. Jest to miejsce, w którym do tekstu przypisywane są znaczenia, co umożliwia wnioskowanie na podstawie znaczeń. Możliwość uogólnienia tekstu pozwala na porównywanie ze sobą różnych tekstów, pomimo tego, że nie posiadają tych samych słów, jednak mogą posiadać te same znaczenia (zjawisko synonimii). Umożliwia to też dalsze wnioskowanie, gdyż jeden tekst może opisywać to samo co w innym tekście, lecz ogólniej czy bardziej szczegółowo.

Aby możliwe było przypisywanie znaczeń do słów w tekście, musi istnieć, która dla dowolnego słowa będzie potrafiła przypisać znaczenie z dostarczonego leksykonu znaczeń. W ramach niniejszej pracy zaadaptowano do języka polskiego metodę zaproponowaną oryginalnie dla języka angielskiego. Metoda ta wykorzystuje jako główny mechanizm przetwarzania algorytm PageRank (Page i inni, 1999) oraz przypisuje do tekstu znaczenia ze Słowosieci (Maziarz i inni, 2013, Piasecki i inni, 2009b) – oryginalna metoda działa na Princeton WordNet.

Wykorzystanie algorytmu PageRank do ujednoznaczniania znaczeń leksykalnych

Do ujednoznaczniania znaczeń leksykalnych w tekście zaadaptowano dla języka polskiego słabo nadzorowaną metodę ujednoznaczniania znaczeń leksykalnych opisaną w (Gutiérrez i inni, 2012) oraz (Agirre i inni, 2009, Agirre i Soroa, 2009, Agirre i inni, 2013). Odwzorowanie zasobu znaczeń, jakim może być Princeton WordNet, czy też dowolna leksykalna sieć semantyczna łącząca ze sobą znaczenia dowolnym zestawem relacji, do struktury grafowej może zostać wykorzystane jako potencjalne źródło wiedzy dla algorytmów nienadzorowanego ujednoznaczniania znaczeń leksykalnych. Zasoby takie znane w literaturze jako *Lexical Knowledge Base* (*LKB*), to zbiór znaczeń oraz relacji pomiędzy nimi (Agirre i inni, 2009, Agirre i Soroa, 2009, Agirre i inni,

2013). *LKB* może być rozumiane jako graf $\mathbf{G} = (V, E)$, gdzie V to zbiór węzłów n_i, $n_i \in V$, reprezentujących znaczenia, E to zbiór krawędzi reprezentujących relacje. Relacje zachodzą między n_i a n_j i reprezentowane są za pomocą nieskierowanej krawędzi $e \in E$. Jak wspomniano w (Agirre i Soroa, 2009) można rozważać relacje różnego typu, na przykład leksykalno-semantyczne. Dodatkowo, relacje mogą posiadać przypisane wagi. Wykorzystując podejście grafowe, stosuje się funkcję rankingowania węzłów w grafie, a następnie wybiera się te węzły grafu, które mają największą wartość funkcji rankingującej.

PageRank (Page i inni, 1999), dalej oznaczany jako *PR*, to metoda rankingowania węzłów odpowiadających tekstom, w której jakość (ranga) tekstu zależna jest od liczby tekstów powołujących się na niego. Zależy ona również od rangi przypisanej do węzła n_j, na który wskazywał węzeł n_i. Im większa jest waga węzła n_i, tym większa jest waga węzła n_j.

Główna właściwość *PR* polega na tym, że jeżeli istnieje w grafie połączenie dla węzłów n_i oraz n_j, to obliczana jest wartość połączenia węzła n_i do n_j. W dalszym procesie przetwarzania oznacza to, że wzrasta wartość funkcji oceniającej dla węzła n_j, na który wskazywał węzeł n_i. Jeżeli n_j posiada przypisaną wagę, jest ona brana pod uwagę w funkcji oceny tego węzła. *PR* może być postrzegany jako wynik losowego procesu przechodzenia przez graf $\mathbf{G}$. Ostateczna wartość Pr rankingu węzła n_i oznacza prawdopodobieństwo trafienia do tego węzła przy losowym przejściu przez graf $\mathbf{G}$. Wartość ta wyznaczana jest uwzględniając wystarczająco długi czas przechodzenia po grafie lub w odpowiednio dużej liczbie iteracji algorytmu.

Niech $\mathbf{G}$ będzie grafem z N węzłami $(n_i, ..., n_N)$ oraz deg_i oznacza stopień wychodzącego węzła n_i. Oryginalna postać *PR* opisana w (Altman i Tennenholtz, 2005) przedstawiona została równaniem 4.1). Warto tutaj zaznaczyć, że w *PR* wartość Pr to wartość rangi strony linkującej (wskazującej na aktualną stronę) do aktualnie przetwarzanej strony.

$$Pr(v_i) = (1 - \alpha) + \alpha \sum_{j \in In(v_i)} \frac{1}{deg_j} Pr(v_j) \tag{4.1}$$

Przyjmując, że M jest macierzą $(N \times N)$ prawdopodobieństwa przejść, w której każdy element M_{ij} definiuje się jako $M_{ij} = \frac{1}{deg_i}$, jeżeli istnieje połączenie z n_i do n_j (w przeciwnym wypadku $M_{ij} = 0$), możemy wyrazić Pr za pomocą wzoru 4.2.

$$Pr = cMPr + (1 - c)\vec{v} \tag{4.2}$$

We wzorze tym Pr oznacza wynik iteracyjnej funkcji rankingowania, $\vec{v}$ jest wektorem o rozmiarze $N \times 1$, którego elementy posiadają wartości $\frac{1}{N}$, c to *współczynnik tłumienia*, którego wartość leży między 0 a 1 i zazwyczaj przyjmuje wartość ze zbioru $\langle 0,85, \ldots, 0,95 \rangle$. Warto tutaj dodać, że w oryginalnej postaci *PR* wartość Pr to wartość rangi strony linkującej (wskazującej na aktualną stronę) do aktualnie przetwarzanej strony. Term $(1-c)\vec{v}$ może być rozumiany jako współczynnik wygładzania, jest on stały we wszystkich iteracjach. W podstawowej wersji PR, wektor $\vec{v}$ przyjmuje wartości $\frac{1}{N}$ (dla każdego elementu), co powoduje, że każdy węzeł ma równe prawdopodobieństwo przeskoczenia do niego. Jednak w pracy (Haveliwala, 2002) pokazano, że wektor $\vec{v}$ może być „personalizowany", przez co doprowadza się do sytuacji, w której pewne przejścia są bardziej prawdopodobne niż inne.

Ocena jakości ujednoznaczniania znaczeń leksykalnych

Przed dalszym wykorzystaniem metody ujednoznaczniania znaczeń leksykalnych w *PAS*, przeprowadzono ocenę jakości jej działania na Korpusie Politechniki Wrocławskiej (KPWr) w wersji 1.1 [1] oraz Narodowym Korpusie Języka Polskiego (NKJP – `http://nkjp.pl/`), który posiada anotacje znaczeniami ze Słowosieci. W tabeli 4.2 znajdują się wyniki oceny jakości działania metody ujednoznaczniania znaczeń leksykalnych (Kędzia i inni, 2015, 2014a) na dwóch zbiorach danych *KPWr* oraz *NKJP*. Jakość działania mierzona była za pomocą precyzji. Kolumny oznaczone jako **Cz.** dotyczą precyzji ujednoznaczniania dla czasowników, kolumny **Rz.** odnoszą się do precyzji ujednoznaczniania rzeczowników, zaś **Śr.** to średnia arytmetyczna precyzji rzeczowników i czasowników.

Tabela 4.2. Precyzja ujednoznaczniania znaczeń leksykalnych na dwóch zbiorach danych – Korpusie Politechniki Wrocławskiej (**KPWr**) oraz Narodowym Korpusie Języka Polskiego (**NKJP**), osobno dla czasowników (**Cz.**) i rzeczowników(**Rz.**) z dodatkową informacją o średniej (**Śr.**)

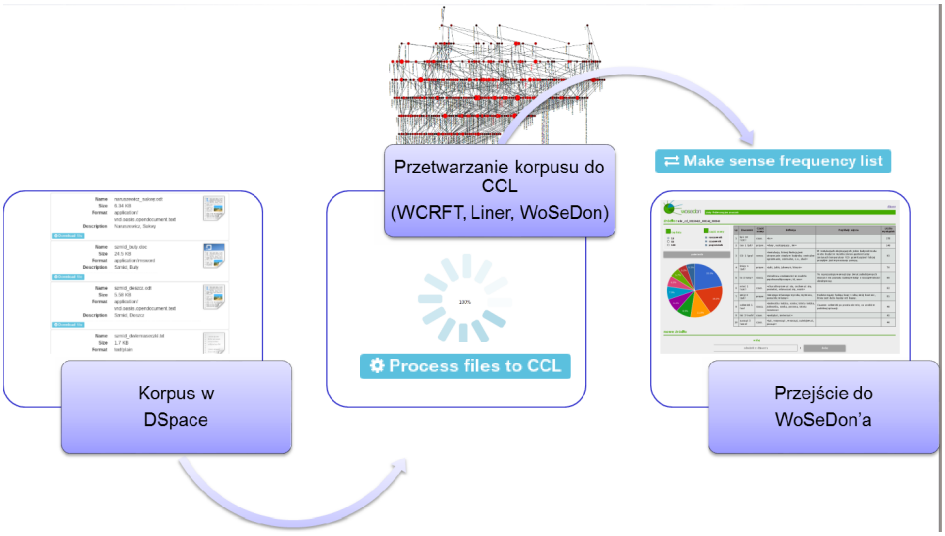
	KPWr			**NKJP**		
	Cz.	**Rz.**	**Śr.**	**Cz**	**Rz.**	**Śr.**
1	32,61	52,22	45,52	49,02	64,02	58,48
2	42,66	47,91	46,12	47,5	61,67	56,16
3	39,76	39,3	39,46	49,28	61,12	56,51

Kolejne wiersze w tabeli 4.2, nazwane jako **1**, **2** oraz **3**, odnoszą się do różnych konfiguracji uruchamianego algorytmu. Wiersz oznaczony jako **1** zawiera wyniki precyzji ujednoznaczniania dla *LKB* zbudowanego z połączenia grafu synsetów z grafem zbudowanym z ontologii SUMO (opisane w rozdziale 4.2.3). Wiersz oznaczony jako **2** zawiera wyniki precyzji algorytmu PageRank (wzór4.1) dla grafu zbudowanego ze Słowosieci, zarówno z relacji synsetów jak i relacji między jednostkami leksykalnymi. W grafie tym, krawędzie odpowiadające relacjom synsetów, posiadają przypisaną arbitralnie wagę $0,7$; krawędzie odpowiadające relacjom między jednostkami leksykalnymi posiadają arbitralnie ustaloną wagę $0,3$ z dodatkową heurystyką ponownego rankingowania wyniku ostatecznego. Polega ono na wyborze 10% najwyżej ocenionych synsetów i wyborze z nich tego znaczenia, które posiada znaczenie o najniższym identyfikatorze. Dla przykładu, otrzymując jako wynik (znaczenia dla słowa *linia*): {*linia.4*, *linia.11*, *linia.9*, *linia.2*}, wybierane jest znaczenie *linia.2*. Wiersz oznaczony jako **3** to wyniki precyzji dla algorytmu oraz grafu zbudowanego analogicznie do podejścia **2**, z tą różnicą, że w ponownym rankingowaniu wyników, wybranych zostało 40% synsetów z rankingu początkowego, a z nich wybrano znaczenie o najmniejszym identyfikatorze znaczenia.

Potrzeba włączenia metody ujednoznaczniania znaczeń leksykalnych do przetwarzania wstępnego w *PAS* spowodowała opracowanie narzędzia WoseDon (dostępne pod adresem `https://clarin-pl.eu/dspace/handle/11321/290`) oraz WebWoSeDon, które dostępne jest pod adresem `http://wosedon.clarin-pl.eu`.

Narzędzie WoSeDon zostało włączone do potoku przetwarzania repozytorium DSPace. Po zalogowaniu się do DSPace, wybraniu dowolnego korpusu, poprzez jedno

[1] Korpus dostępny na stronie `https://clarin-pl.eu/dspace/handle/11321/16`



Rys. 4.3. Proces przetwarzania dokumentów za pomocą DSPace

kliknięcie, możliwe jest automatyczne przypisanie do niego znaczeń leksykalnych ze Słowosieci. Proces ten przedstawiony został na rysunku 4.3. Po wykonaniu procesu ujednoznaczniania, DSPace automatycznie przekierowuje użytkownika do systemu WebWoSeDon, w którym przedstawione są statystyki użycia znaczeń w aktualnie wybranym korpusie tekstów. WebWoSeDon umożliwia również tworzenie statystyk znaczeń leksykalnych ze Słowosieci z korpusów zapisanych w repozytorium DSPace bez logowania się do niego. Po wejściu na stronę WebWoSeDona, wymagane jest jedynie podanie uchwytu zasobu (unikalnego adresu identyfikującego korpus tekstów w DSPace) i uruchomienie analizy.

4.2.3. Moduł wiedzy o świecie

Aby uwzględnić wiedzę pozatekstową wykorzystywany jest moduł *Wiedza o świecie* (rysunek 4.2). Zawiera on procedury wydobywania tej wiedzy, dzięki którym możliwe jest pozyskanie reprezentacji statycznej wiedzy wydobytej nie tylko z analizowanego tekstu, ale także możliwe jest dołączanie reprezentacji wiedzy dotyczącej różnych dziedzin i wydobytej z innych źródeł. W takim przypadku ważne jest, aby do wydobycia wiedzy użyte zostały zasoby jak najbardziej reprezentatywne dla tej dziedziny. Dołączana wiedza może dotyczyć wiedzy ogólnej, ale i w tym przypadku ważne jest, aby zasoby opisywały jak największy wycinek świata rzeczywistego. W pracy, jako źródła statycznej wiedzy zostały wykorzystane Słowosieć (Piasecki i inni, 2016b, Dziob i Piasecki, 2018) oraz ontologia SUMO (Niles i Pease, 2001). Połączenie Słowosieci z SUMO (na rysunku 4.2 oznaczone jako *Rzutowanie*) zostało opracowane przez autora niniejszej pracy i jest szczegółowo przedstawione w tym rozdziale.

Rzutowanie Słowosieci na ontologię SUMO

Z racji ogólności wybranej ontologii, dzięki rzutowaniu Słowosieci na SUMO, możliwe jest przypisanie do słów z analizowanego tekstu bardziej ogólnych pojęć z ontologii. Możliwe jest również dołączenie wiedzy pozatekstowej reprezentowanej w ontologii

SUMO do analizowanego tekstu, dla którego budowana jest reprezentacja semantyczna wiedzy.

Ogólny zarys metody rzutowania. Opracowana metoda rzutowania (Kędzia i Piasecki, 2014) składa się z dwóch etapów. W *etapie pierwszym*, wykorzystując relacje międzyjęzykowe, ustalane jest powiązanie synsetów polskich z angielskimi. Następnie, dla synsetów angielskich, które posiadają relację z synsetem polskim wyszukiwane jest rzutowanie na pojęcie SUMO. Ten krok tworzy potencjalne powiązania synsetu polskiego z pojęciem SUMO. Do określania potencjalnych powiązań oraz ustalania typu relacji, wykorzystywane są zewnętrzne zasoby, którymi są:

1. Relacje międzyjęzykowe ze Słowosieci (relacje między polskimi synsetami, a synsetami występującymi w Princeton WordNet) (Rudnicka i inni, 2012, 2018, 2019),
2. Rzutowanie Princeton WordNet na SUMO (Niles i Pease, 2004), dostępne pod adresem `http://sigmakee.cvs.sourceforge.net/sigmakee/Kbs/`
3. Struktura Słowosieci (Słowosieć),
4. Struktura SUMO (SUMO).

W drugim etapie, dla wydobytych potencjalnych rzutowań, określana jest relacja łącząca polski synset z pojęciem SUMO. Dostępne są trzy typy relacji określonych oraz jedna relacja niedookreślona. Określone relacje, to:

— Ekwiwalencja (*equivalent*) – równoznaczność znaczenia Słowosieci i pojęcia SUMO. Synset odpowiada pojęciu SUMO ze względu na znaczenie synsetu, ze szczególną wagą przykładaną do denotacji synsetu, np. *plant 2* –equivalent– `Plant`.

— Podklasa (*subsumed*) – wskazanie, że znaczenie Słowosieci jest zawężeniem znaczenia pojęcia SUMO, denotacja synsetu zawiera się w denotacji pojęcia SUMO np. {*town 1*} –subsumed– *City*.

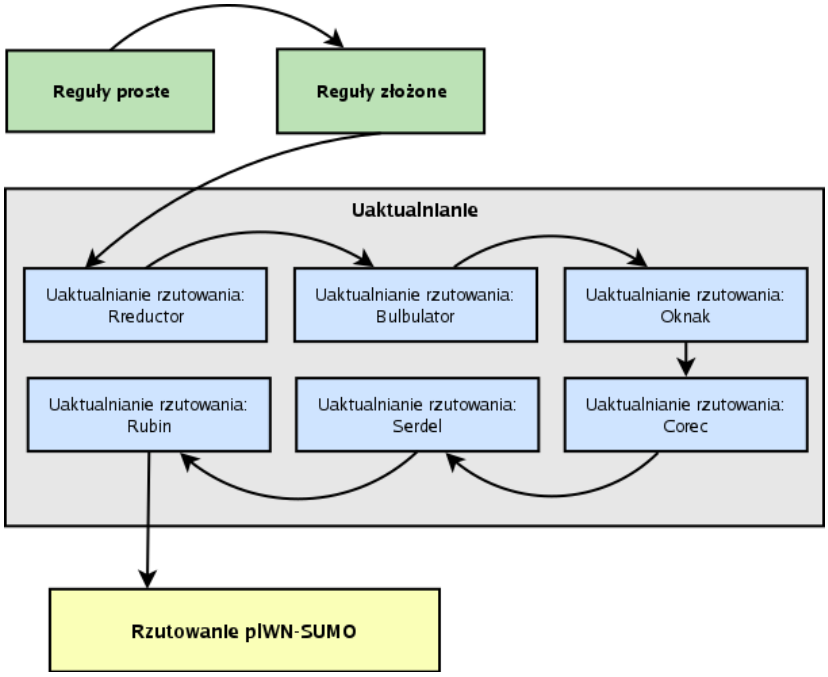
— Instancja (*instance of*) – znaczenie Słowosieci jest instancją (obiektem) klasy/pojęcia SUMO, denotacja synsetu jest instancją pojęcia SUMO, np. {*Aristotle 1}* –instance of– *Man* lub jest elementem kolekcji denotowanej przez pojęcie SUMO, np. {*Eden 2*} –instance of– *Region*

Relacja niedookreślona, oznaczana dalej jako R, wskazuje na niemożliwość podjęcia decyzji odnośnie ustalenia typu relacji pomiędzy potencjalną parą rzutowania, czyli między znaczeniem ze Słowosieci a pojęciem z SUMO. Dotyczy to sytuacji, w których denotacja znaczenia ze Słowosieci jest nadzbiorem denotacji pojęcia z SUMO lub też innych, często błędnych, potencjalnych powiązań. Etap drugi w procesie rzutowania jest podzielony na wiele podetapów opisanych poniżej. Każdy z kolejnych kroków (poza pierwszym), bazuje na wyniku z poprzedniego kroku.

Szczegółowy opis metody rzutowania. Proces rzutowania Słowosieci na ontologię SUMO przedstawiony został na rysunku 4.4. Pierwszym krokiem jest uruchomienie *prostych reguł* rzutujących, których zadaniem jest określenie relacji rzutującej synset Princeton WordNetu na SUMO i przeniesienie jej jako relację rzutującą synset Słowosieci na SUMO. Zapis w pseudokodzie prostych reguł rzutowania opisujących relacje rzutujące przedstawiony jest w algorytmie 1.

Przykład działania prostych reguł dla relacji *subsumed*:

1. {*kark 1 (body)*} – *subsumed* – *BodyPart* → {*kark 1 (body)*} jest i-synonym {*nape 1 (body)*}



Rys. 4.4. Proces rzutowania synsetów Słowosieci na pojęcia SUMO

Algorithm 1 Proste reguły rzutowania

```
if R(PLWN_PWN) = I-synonymy and R(PWN_SUMO) = equivalent then
    R(PLWN_SUMO) = equivalent
end if
if R(PLWN_PWN) = I-synonymy and R(PWN_SUMO) = instance of then
    R(PLWN_SUMO) = instance of
end if
if R(PLWN_PWN) = I-synonymy and R(PWN_SUMO) = subsumed then
    R(PLWN_SUMO) = subsumed
end if
```

2. {*mieczyk 1 (plant)*} – *subsumed* – *FloweringPlant* → {*mieczyk 1 (plant)*} jest i-synonym {*genus Gladiolus 1 (plant)*}

Przykład działania prostych reguł dla relacji *instance of*:

1. {*geometria rzutowa 1 (cogn)*} – *instance of* – *FieldOfStudy* → {*geometria rzutowa 1 (cogn)*} jest i-synonym dla {*projective geometry 1 (cogn)*}
2. {*Ateny 1 (loc)*} – *instance of* – *City* → {*Ateny 1 (cogn)*} jest i-synonym {*Athens 1 (cogn)*}

Przykład działania prostych reguł dla relacji *equivalent*:

1. {*czekolada 1 (food)*} – *equivalent* – *Chocolate* → {*czekolada 1 (food)*} jest i-synonym {*chocolate 2 (food)*}
2. {*wał rozrządu 1 (arte)*} – *equivalent* – *Camshaft* → {*wał rozrządu 1 (arte)*} jest i-synonym {*camshaft 1 (arte)*}

W kolejnym kroku wykonywane są złożone reguły rzutujące, które działają wykorzystując proces *Uaktualniania* rzutowania (rysunek 4.4). Oznacza to wykonanie sekwencji

wielu procedur, których nazwy podane zostały w prostokątach odpowiadających tym procedurom. W ramach sekwencji *uaktualniania* wyjście z poprzedniej procedury jest modyfikowane przez aktualną; przykładowo wyjście z *Bulburatora* (opisanego w dalszej części niniejszego rozdziału) modyfikowane jest przez procedurę *Oknaka*. Krótka charakterystyka charakterystyka procedur ujętych w procesie rzutowania przedstawiona jest poniżej:

- *Rreduktor* ma za zadanie zmniejszenie liczby relacji niedookreślonych R bazując na wykonanym już rzutowaniu.
- *Bulbulator* odnajduje w całym rzutowaniu sytuacje sprzeczne, czyli takie, gdzie pomiędzy konkretną parą $\{synset, pojecie\_sumo\}$ przypisane zostały przynajmniej dwie różne relacje rzutujące i bazując na relacjach międzyjęzykowych, dokonuje korekty rzutowania.
- *Oknak* podobnie jak *Bulbulator*, wyszukuje sprzeczne rzutowania i dokonuje ich poprawy, bazując jednak na innych przesłankach.
- *Corec* czyści wykonane rzutowania. Jeżeli dla znaczenia istnieje więcej niż jedno rzutowanie na różne pojęcia SUMO, to sprawdzana jest hierarchia pojęć SUMO i zostawiane jest tylko to rzutowanie, które wskazuje na najbardziej szczegółowe pojęcie.
- *Serdel* poprawia wykonany do tej pory proces rzutowania przez uwzględnienie znaczeń, które nie posiadają rzutowania na SUMO. Rozważane są relacje: bliskoznaczność, hiperonimia, wartość cechy. Jeśli występuje relacja bliskoznaczności to relacja i pojęcie są kopiowane do aktualnego znaczenia. Jeśli nie występuje relacja bliskoznaczności, to zmieniana jest relacja (PlWN-SUMO) na *subsumed* i kopiowane jest znaczenie.
- *Rubin*, to kolejna procedura, w której czyszczone są istniejące rzutowania. Jeżeli znaczenie ze Słowosieci posiada rzutowanie na przynajmniej dwa pojęcia, a jednym z tych pojęć jest *SubjectiveAssessmentAttribute*, to takie rzutowanie jest usuwane, ze względu na ogólność definicji pojęcia *SubjectiveAssessmentAttribute*. *Rubin* uwzględnia proces rozbudowy Princeton WordNetu (dalej *PWN*) i dodatkowych powiązań pomiędzy Słowosiecią, a dodanymi do *PWN* nowymi znaczeniami. Ten powiększony WordNet został nazwany *enWordNet* (WrUT, 2018).

Warto wspomnieć, że Słowosieć jest ręcznie zrzutowana na *PWN*. Algorytm rzutowania Słowosieci na SUMO wykorzystuje informacje o połączeniu Słowosieci z *PWN*. Nowo dodane znaczenia do *PWN* nie posiadają rzutowania na SUMO, zatem zadaniem procedury *Rubin*, jest przejście po strukturze *PWN*, uwzględniając jedynie wybrane relacje, i znalezienie istniejącego rzutowania synsetu *PWN* na SUMO, na podstawie którego możliwe będzie zrzutowanie znaczenia Słowosieci na SUMO.

Złożone reguły rzutujące uwzględniają wiele aspektów, takich jak:

- typ relacji międzyjęzykowej np. *synonimia międzyjęzykowa*, *hiponimia międzyjezykowa*, *hiperonimia międzyjezykowa*,
- typ relacji rzutowania Princeton WordNet na SUMO np. ekwiwalencja,
- dziedzina synsetu polskiego i/lub angielskiego,
- rodzaj pojęcia, na które odbywa się rzutowanie np. *Currency*, *GroupOfPeople*, *FieldOfStudy*, *Human*,
- informacja o *wielkiej/małej* literze w jednostce synsetu.

Poniżej przedstawione zostały przykładowe złożone reguły rzutujące wraz z efektem ich zastosowania na wybranych przykładach.

Algorithm 2 Przykłady reguł rzutowania, które odwołują się do dziedzin synsetów Słowosieci i WordNetu

```
1: if R(PLWN_PWN) = I-part-of-meronymy and R(PWN_SUMO) = equivalent
   then
2:     if PLWN_SYNSET zaczyna się wielką literą then
3:         if D(PLWN) ∈ {loc} and D(PWN) ∈ {natobj} then
4:             R(PLWN_SUMO) = instance of
5:         end if
6:         if D(PLWN) ∈ {rel} and D(PWN) ∈ {loc} then
7:             R(PLWN_SUMO) = instance of
8:         end if
9:     end if
10: end if
```

Efekty zastosowania złożonych reguł z Algorytmu 2:

1. {*Europa Wschodnia 1 (loc)*} – *instance of* – *Europe* → {*Europa Wschodnia 1 (loc)*} – *I-part-of-meronymy* – {*Europe 1 (natobj)*} – *instance of* – *Europe*
2. {*Europa Zachodnia 1 (loc)*} – *instance of* – *Europe* → {*Europa Zachodnia 1 (loc)*} – *I-part-of-meronymy* – {*Europe 1 (natobj)*} – *instance of* – *Europe*
3. {*Europa Południowo-Wschodnia 1 (loc)*} – *instance of* – *Europe* → {*Europa Południowo-Wschodnia 1 (loc)*} – *I-part-of-meronymy* – {*Europe 1 (natobj)*} – *instance of* – *Europe*
4. {*Amazonia 1 (loc)*} – *instance of* – *SouthAmerica* → {*Amazonia 1 (loc)*} – *I-part-of-meronymy* – {*South America 1 (natobj)*} – *instance of* – *SouthAmerica*
5. {*Hetmańszczyzna 1 (rel)*} – *instance of* – *Ukraine* → {*Hetmańszczyzna 1 (rel)*} – *I-part-of-meronymy* – {*Ukraine 1 (natobj)*} – *instance of* – *Ukraine*

Algorithm 3 Przykłady reguł wykorzystujących dziedziny synsetów Słowosieci i WordNetu oraz pojęcia SUMO

```
1: if R(PLWN_PWN) = I-part-of-holonymy and R(PWN_SUMO) = subsumed
   then
2:     if D(PLWN) ∈ {natphen} and D(PWN) ∈ {st} then
3:         if SUMO_CONCEPT=DiseaseOrSyndrome then
4:             R(PLWN_SUMO) = subsumed
5:         end if
6:     end if
7: end if
```

Efekty zastosowania złożonych reguł z Algorytmu 3:

1. {*dysplazja śmiertelna 1 (natphen)*} – *subsumed* – *DiseaseOrSyndrome* → {*dysplazja śmiertelna 1 (natphen)*} – *I-part-of-holonymy* – {*birth defect 1 (st)*} – *subsumed* – *DiseaseOrSyndrome*
2. {*zespół delecji 13q 1 (natphen)*} – *subsumed* – *DiseaseOrSyndrome* → {*zespół delecji 13q 1 (natphen)*} – *I-part-of-holonymy* – {*birth defect 1 (st)*} – *subsumed* – *DiseaseOrSyndrome*
3. {*fetopatia cukrzycowa 1 (natphen)*} – *subsumed* – *DiseaseOrSyndrome* → {*fetopatia cukrzycowa 1 (natphen)*} – *I-part-of-holonymy* – {*birth defect 1 (st)*} – *subsumed* – *DiseaseOrSyndrome*
4. {*ślepota z Rodrigue 1 (natphen)*} – *subsumed* – *DiseaseOrSyndrome* → {*ślepota z Rodrigue 1 (natphen)*} – *I-part-of-holonymy* – {*birth defect 1 (st)*} – *subsumed* – *DiseaseOrSyndrome*

5. $\{$*pentalogia Cantrella 1 (natphen)*$\}$ – *subsumed* – *DiseaseOrSyndrome* $\rightarrow$ {*pentalogia Cantrella 1 (natphen)*} – *I-part-of-holonymy* – {*birth defect 1 (st)*} – *subsumed* – *DiseaseOrSyndrome*

Może się zdarzyć, że za pomocą relacji międzyjęzykowych oraz relacji rzutujących Princeton WordNet-SUMO, odnaleziona została para $\{Synset_{plwn}, Concept_{sumo}\}$, jednakże relacja zachodząca między nimi nie jest żadną z rozważanych. Powstaje wówczas relacja niedookreślona R, algorytm 4 opisuje sytuację, kiedy typ relacji nie może zostać rozstrzygnięty automatycznie.

Algorithm 4 Wyszukiwanie relacji, której typ nie może być rozstrzygnięty automatycznie

```
1: if R(PLWN_PWN) = I-part-of-meronymy and R(PWN_SUMO) = subsumed
   then
2:     if D(PLWN) ∈ {loc, body, arte, grp, ..., class} then
3:         R(PLWN_SUMO) = manually
4:     end if
5: end if
```

Efekty zastosowania złożonych reguł z Algorytmu 4:

1. $\{$*okno 7 (arte)*$\}$ – *manually* – *Region* $\rightarrow$ {*okno 7 (arte)*} – *I-part-of-meronymy* – {*hotbed 2 (arte)*} – *subsumed* – *Region*
2. $\{$*skrzydło 9 (arte)*$\}$ – *manually* – *Door* $\rightarrow$ {*skrzydło 9 (arte)*} – *I-part-of-meronymy* – {*door 1 (arte)*} – *subsumed* – *Door*
3. $\{$*strzelnica 2 (loc)*$\}$ – *manually* – *Corporation* $\rightarrow$ {*strzelnica 2 (loc)*} – *I-part-of-meronymy* – {*amusement park 1 (loc)*} – *subsumed* – *Corporation*
4. $\{$*biblioteka szkolna 1 (loc)*$\}$ – *manually* – *School* $\rightarrow$ {*biblioteka szkolna 1 (loc)*} – *I-part-of-meronymy* – {*school 2 (arte)*} – *subsumed* – *School*
5. $\{$*oliwka 5 (body)*$\}$ – *manually* – *BodyPart* $\rightarrow$ {*oliwka 5 (body)*} – *I-part-of-meronymy* – {*medulla oblongata 1 (arte)*} – *subsumed* – *BodyPart*
6. $\{$*węzina 1 (body)*$\}$ – *manually* – *ThyroidGland* $\rightarrow$ {*węzina 1 (body)*} – *I-part-of-meronymy* – {*thyroid gland 1 (arte)*} – *subsumed* – *ThyroidGland*

Końcowym krokiem jest próba dookreślenia relacji R – zredukowanie liczby relacji niedookreślonych. Przyjmijmy, że przykładowy wynik po pierwszym kroku rzutowania, w którym otrzymaliśmy R jako relację między znaczeniem Słowosieci a pojęciem SUMO jest następujący:

```
ablucja.1;Removing;subsumed
abnegat.1;Human;R
abnegat.1;Human;subsumed
abnegat.2;Human;subsumed
```

Rozważmy synset zawierający jednostkę leksykalną `abnegat.1`, widzimy tutaj sytuację, że: `abnegat.1` jest w relacji *subsumed* do `Human` oraz `abnegat.1` jest w relacji R do `Human`. Relacja R oznacza, że algorytm nie był w stanie automatycznie określić typu relacji wynikowej, relacja *subsumed* oznacza, że `abnegat.1` jest uszczegółowieniem `Human`. Skoro obie sytuacje dotyczą tej samej jednostki (`abnegat.1`) oraz posiadają relację do tego samego pojęcia SUMO (`Human`) a algorytm potrafił automatycznie rozstrzygnąć typ relacji rzutującej (`abnegat.1` posiada relację *subsumed* do `Human`), można w tym przypadku dookreślić relację R do relacji *subsumed*. Innymi słowy, w pierwszej kolejności wyszukiwana jest niedookreślona relacja R, po czym sprawdzane jest, czy

można ją dookreślić poprzez sprawdzenie czy dla danej jednostki oraz pojęcia SUMO, dla których znaleźliśmy R, istnieje relacja dookreślona. Jeżeli tak, to ustawiany jest typ relacji rzutującej na dookreślony przez znalezione rzutowanie. Dla przytoczonego przykładu, wynik będzie wyglądał następująco:

```
ablucja.1;Removing;subsumed
abnegat.1;Human;subsumed
abnegat.2;Human;subsumed
```

Inny problem pojawia się, gdy daną relację niedookreśloną R można dookreślić przez kilka relacji. W takim wypadku podawane są wszystkie relacje dookreślające (wymienione po przecinku). Załóżmy, że mogliśmy dookreślić R z przytoczonego przykładu również za pomocą relacji *instance of*, zatem wynikiem będzie:

```
ablucja.1;Removing;subsumed
abnegat.1;Human;subsumed, instance_of
abnegat.2;Human;subsumed
```

Ocena rzutowania. Zaproponowana metoda rzutowania została oceniona dwustopniowo. W pierwszej kolejności wykonywana była automatyczna ocena wyników algorytmu rzutowania. Proces ten obejmował rzutowanie wszystkich synsetów ze Słowosieci i polegał na odnalezieniu takich dwójek $\{Synset, Pojecie\}$, które posiadały więcej niż jedno rzutowanie, relacja rzutująca była inna niż R, a pozostałe relacje rzutujące różniły się między sobą. Wykorzystano tutaj fakt, że pomiędzy daną parą $\{Synset, Pojecie\}$ nie może zachodzić więcej niż jedna relacja, ponieważ definicje relacji wykluczają jednoczesne zachodzenie kilku relacji między tą samą parą. Podczas takiej oceny błąd rzutowania wynosił $\approx 0,06\%$, co oznacza, że 0,06% synsetów ze Słowosieci posiadało błędne rzutowanie.

W kolejnym etapie ocena została przeprowadzona ręcznie, przez trzech lingwistów. Do oceny zostało wylosowanych 160 rzutowań znaczeń na pojęcia SUMO. Przyjęto schemat oceny 2+1, co znaczy, że dwoje lingwistów oceniało te same przypadki, a w przypadku niezgodności, trzeci rozstrzygał, który z nich miał rację. Ocena uwzględniała trzy przypadki:

— *poprawne* – rzutowanie dokładnie takie samo,
— *niepoprawne* – rzutowanie niepoprawne,
— *podłączony do hiperonimu* – rzutowanie na hiperonim, nie jednostkę docelową.

Tabela 4.3. Liczba synsetów Słowosieci ***zrzutowanych*** na SUMO z przypisaną relacją określoną oraz liczba rzutowań z relacją niedospecyfikowaną R – kolumna ***Ręcznie***

POS	**Ręcznie**		**Zrzutowane**		**Ręcznie [%]**	
	Maj 2014	**Luty 2016**	**Maj 2014**	**Luty 2016**	**Maj 2014**	**Luty 2016**
Rzeczownik	4 316	5 810	84 607	147 783	4,8	3,8
Czasownik	1	25	17	498	5,5	5,0
Przymiotnik	356	955	3 691	20 564	8,8	4,4

Precyzja rzutowania wynosiła 83,1%. Statystyki rzutowań przedstawione zostały w tabeli 4.3.

— Liczba rzutowań w kwietniu 2014: 87 550
— Liczba rzutowań w maju 2014: 92 810
— Liczba rzutowań w lutym 2016: 175 635

W lutym 2016 roku, liczba zrzutowanych synsetów wynosiła 175 635, co oznaczało 89.2% wszystkich wszystkich synsetów Słowosieci. Tak duży przyrost liczby zrzutowanych synsetów, to efekt rozwoju Słowosieci, a przede wszystkim ręcznego rzutowania Słowosieci na Princeton WordNet.

Narzędzie powstałe w wyniku realizacji zadania dotyczącego rzutowania Słowosieci na ontologię SUMO to `Plumper` (`https://clarin-pl.eu/dspace/handle/11321/279`). Dzięki wykorzystaniu ontologii SUMO, mamy dostęp do pojęć, które modelują świat na bardzo ogólnym poziomie. Umożliwia to interpretację słów na poziomie wyższym niż leksykalny. Narzędzie posiada dwa tryby działania:

1. wykonanie rzutowania wszystkich synsetów Słowosieci,
2. wykonanie rzutowania bazując na synsetach, które znalazły się we wskazanym tekście, wcześniej ujednoznacznionym za pomocą WoSeDona.

Przykładowe efekt działania tego narzędzia znajduje się pod adresem `https://clarin-pl.eu/dspace/handle/11321/307`.

4.2.4. Moduł analizy wiedzy

Moduł *analizy wiedzy* (rysunek 4.1) zgodnie ze *scenariuszem realizacji* metody wydobywa z bazy *oznaczonych tekstów* informacje odpowiednie dla wymagań użytkownika. Są one ze sobą łączone w trakcie iteracyjnego procesu ekstrakcji wiedzy. Analiza wiedzy obejmuje:

1. transformację słów z tekstu do węzłów w grafie,
2. wydobycie odpowiednich informacji z bazy tekstów oznaczonych i dołączenie ich do budowanego grafu,
3. odpytywanie modułu *wiedzy o świecie* o dodatkową wiedzę i dołączanie jej do budowanego grafu,
4. uaktualnianie grafu o nową wiedzę tekstową i pozatekstową,
5. uogólnienie ostatecznej postaci grafu.

Wymienione operacje wydobywania wiedzy i jej łączenia sprowadzają się do wykonywania operacji na grafie reprezentującym semantykę analizowanego tekstu. Na grafach tych wykonywane są trzy operacje:
— budowanie grafu zerowego,
— dodawanie krawędzi do grafu zerowego,
— wycinanie i łączenie grafów, odpowiadające za łączenie wiedzy tekstowej z pozatekstową.

Budowanie grafu zerowego (czyli takiego, który posiada tylko węzły, natomiast nie posiada krawędzi) wykonywane jest zawsze, natomiast pozostałe operacje w razie potrzeby wynikającej z wymagań użytkownika.

Budowanie grafu zerowego

Sposób reprezentacji słowa z analizowanego tekstu określany jest w *scenariuszu realizacji* metody. Z tym sposobem związana jest funkcja transformująca dowolne słowo

w określony węzeł w grafie zerowym Φ[2]. Reprezentacja słowa z tekstu w grafie znajdowana jest przez użycie jednej z następujących funkcji transformujących:

1. *Orth* – transformuje słowo z tekstu do węzła, którego identyfikator odpowiada formie ortograficznej tego słowa,
2. *Base* – słowo z tekstu, reprezentowane jest w grafie za pomocą jego formy podstawowej,
3. *POS* – węzeł w grafie reprezentuje część mowy tego słowa,
4. *PosOrth* – słowo z tekstu posiada reprezentację w postaci wezła opisywanego przez złożenie jego części mowy z formą ortograficzną; transformacja ta łączy ze sobą funkcje *POS* oraz *Orth*,
5. *PosBase* – słowo z tekstu reprezentowane jest przez węzeł opisywany przez złożenie jego części mowy z formą podstawową; transformacja ta łączy ze sobą funkcje: *POS* oraz *Base*,
6. *Synset* – funkcja ta transformuje słowo do węzła w grafie, który jest identyfikowany poprzez znaczenie tego słowa,
7. *Concept* – funkcja ta transformuje słowo z tekstu do węzła w grafie, który jest identyfikowany przez pojęcia ontologii wysokiego poziomu.

Warto wspomnieć, że przyjęta funkcja transformująca może przekształcić różne słowa z tekstu do tego samego węzła. Dla przykładu, jeżeli w tekście występują różne formy ortograficzne słów {*ogród*, *ogrodzie*}, a wybrana została funkcja transformująca *Base*, to słowa te będą odnosiły się do tego samego węzła w grafie: *ogród*.

Rozważmy bardziej skomplikowany przykład:

Kasia piekła ciasto w piecu kaflowym.

Jeżeli jako funkcja transformująca wybrana zostanie funkcja *Base*, to słowa *piekła* oraz *piecu* będą odnosiły się do tego samego węzła w grafie o identyfikatorze *piec*. W przypadku wyboru *PosBase* jako funkcji transformującej, słowa te będą odnosiły się do dwóch różnych węzłów. Będzie to węzeł o identyfikatorze `verb:piec` – czyli czasownik *piec* oraz węzeł o identyfikatorze `subst:piec` – rzeczownik *piec*. Zbiór wszystkich funkcji transformujących słowa do węzłów w grafie oznaczmy jako F_{ntrans}, n oznacza transformację do węzła (ang. *node*):

$$F_{ntrans} = \{Orth, Base, POS, PosBase, Synset, Concept\}$$

Formalizując, każda ze wspomnianych funkcji transformujących f_{nt} należy do zbioru F_{ntrans}, $f_{nt} \in F_{ntrans}$. W oznaczeniu funkcji f_{nt} litera n oznacza, że ta funkcja transformacji odnosi się do węzła, t, że jest to funkcja transformująca. Funkcja f_{nt} przekształca słowo w_i z analizowanego tekstu, do węzła o identyfikatorze k_j w grafie Φ:

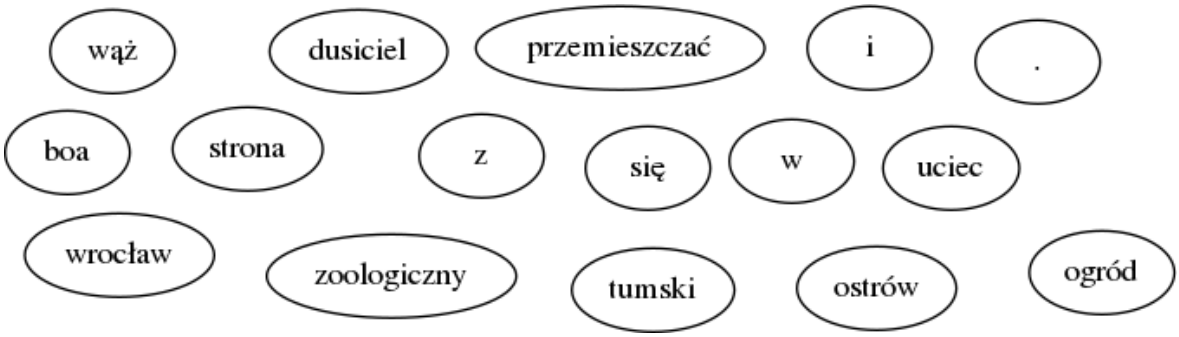
$$f_{nt}(w_i) \rightarrow k_j$$

Oznacza to, że słowo w_i po transformacji trafia do grafu Φ jako węzeł reprezentowany przez unikalny klucz k_j, nadany przez funkcję transformującą f_{nt}. Jeżeli więcej niż jedno słowo zostało przetransformowane do tego samego węzła, to wewnątrz tego

[2] Graf zerowy, oznaczany jako Φ. Nie posiada krawędzi, zawiera jedynie węzły. Stopień każdego węzła równy jest 0.

węzła tworzona jest lista słów z tekstu, które zostały do tego węzła przekonwertowane. Wybór funkcji f_{nt} wpływa na wielkość grafu. Przekształcając tekst za pomocą różnych funkcji f_{nt}, otrzymamy nie tylko różne reprezentacje słów w grafie, ale również wielkość grafów Φ może się znacząco różnić. Jeżeli jako f_{nt} przyjmiemy *Concept*, to odnosząc się do wykorzystanej w pracy ontologii SUMO (Niles i Pease, 2001), liczba węzłów w grafie będzie zazwyczaj mniejsza niż w przypadku $f_{nt} = Orth$. Ma to związek z tym, że liczba pojęć w ontologii SUMO (≈25 000) jest znaczenie mniejsza niż liczba słów w formach ortograficznych języka polskiego (około 6,7 mln form ortograficznych w słowniku Morfeusza[3]).

Na rysunku 4.5 przedstawiony został graf Φ dla zdania *Z Ogrodu Zoologicznego we Wrocławiu uciekł wąż boa dusiciel i przemieszcza się w stronę Ostrowa Tumskiego*. Po zastosowaniu $f_{nt} = Base$ zawiera on węzły odpowiadające słowom z tego zdania. Węzły nie duplikują się, to znaczy słowo *we* jak i *w* posiadają formę bazową *w* i zostały one przetransformowane do tego samego węzła oznaczonego etykietą *w*.



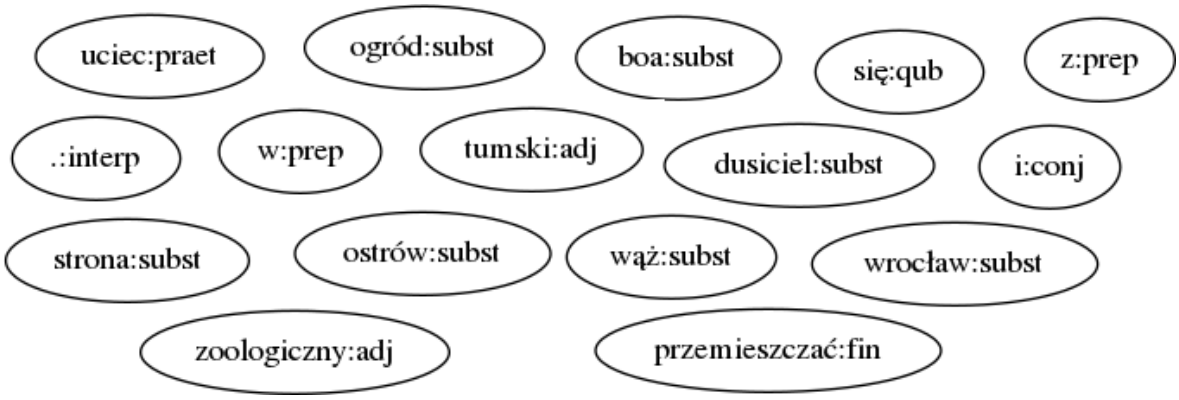
Rys. 4.5. Graf Φ zbudowany ze zdania *Z Ogrodu Zoologicznego we Wrocławiu uciekł wąż boa dusiciel i przemieszcza się w stronę Ostrowa Tumskiego* z $f_{nt} = Base$

Przy zastosowaniu $f_{nt} = PosBase$ graf Φ zbudowany dla tego zdania będzie posiadał tyle samo węzłów, co w przypadku $f_{nt} = Base$, jednak etykiety węzłów będą się różniły. Sytuacja ta została przedstawiona na rysunku 4.6.

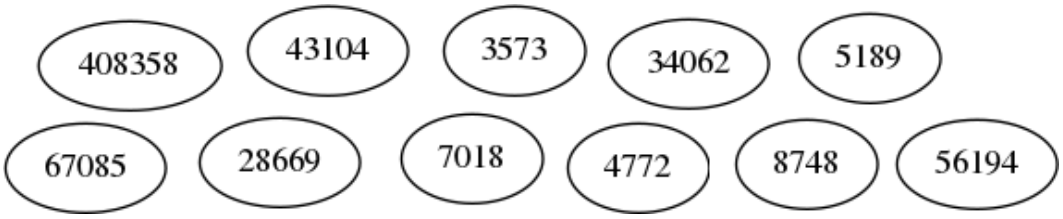
Warto dodać, że funkcja $f_{nt} = PosBase$ wygenerowałaby różne węzły dla słów o tych samych formach bazowych, jednak o różnych częściach mowy. Dla zdania *Kasia piecze ciasto w piecu*, powstałyby dwa węzły zawierające jako część etykiety formę podstawową słowa `piec`, jednak z dwoma różnymi częściami mowy, byłyby to: `(piec:verb)` oraz `(piec:subst)`, przy czym `verb` to oznaczenie czasownika, `subst` oznaczenie rzeczownika.

Jeżeli przyjmiemy $f_{nt} = Synset$, to otrzymamy graf Φ mniejszy od przedstawionych wcześniej. Transformacja zdania *Z Ogrodu Zoologicznego we Wrocławiu uciekł wąż boa dusiciel i przemieszcza się w stronę Ostrowa Tumskiego* do grafu Φ z wykorzystaniem *Synset* jako f_{nt} przedstawiona została na rysunku 4.7. W węzłach grafu umieszczone są identyfikatory synsetów ze Słowosieci. Liczba węzłów jest mniejsza, ponieważ „*w*" oraz „*z*" to przyimki, „*i*" jest spójnikiem, „się" jest kublikiem, a „." to znak interpunkcyjny, a te części mowy nie posiadają znaczeń w Słowosieci, zatem nie jest możliwa ich trans-

[3] Słownik z morfologicznym opisem form wyrazowych – dostępny na stronie `http://sgjp.pl/morfeusz/dopobrania.html`



Rys. 4.6. Graf zerowy Φ zbudowany ze zdania *Z Ogrodu Zoologicznego we Wrocławiu uciekł wąż boa dusiciel i przemieszcza się w stronę Ostrowa Tumskiego* z $f_{nt} = PosBase$



Rys. 4.7. Graf zerowy Φ zbudowany ze zdania *Z Ogrodu Zoologicznego we Wrocławiu uciekł wąż boa dusiciel i przemieszcza się w stronę Ostrowa Tumskiego* z $f_{nt} = Synset$

; w węzłach umieszczone są identyfikatory synsetów.

formacja do węzłów z typem węzła ustawionym jako znaczenie ($Synset$). Wymienione elementy w tym przypadku są pomijane.

Dodawanie krawędzi do grafu Φ

Drugą operacją w procesie budowania grafów, jest uzupełnienie grafu Φ o krawędzie. Zestaw krawędzi, który jest dodawany do grafu wynika ze *scenariusza użycia* i jest zadawany przez użytkownika metody poprzez przypisanie nazw typów krawędzi do zmiennej `ebuilders` w scenariuszu użycia. Każda krawędź e, dodawana do Φ, jest etykietowana zgodnie z nazwą typu relacji zachodzącej między dwoma węzłami reprezentującymi parę słów, pomiędzy którymi zachodzi relacja w tekście. Kierunek zachodzenia relacji w tekście – od którego słowa, do którego – przekładany jest na kierunek krawędzi pomiędzy węzłami w grafie odpowiadającymi tym słowom. Każda krawędź, oznaczona jako e posiada przypisaną etykietę l_e, która jest zgodna z typem krawędzi. Ogólna postać funkcji f_{et} transformującej relację $r(w_i, w_j)$ między słowem w_i a słowem w_j, na krawędź e dodawaną do grafu pomiędzy węzłami z przypisanymi etykietami k_l oraz k_m, odpowiadającymi słowom w_i i w_j, przedstawia się następująco:

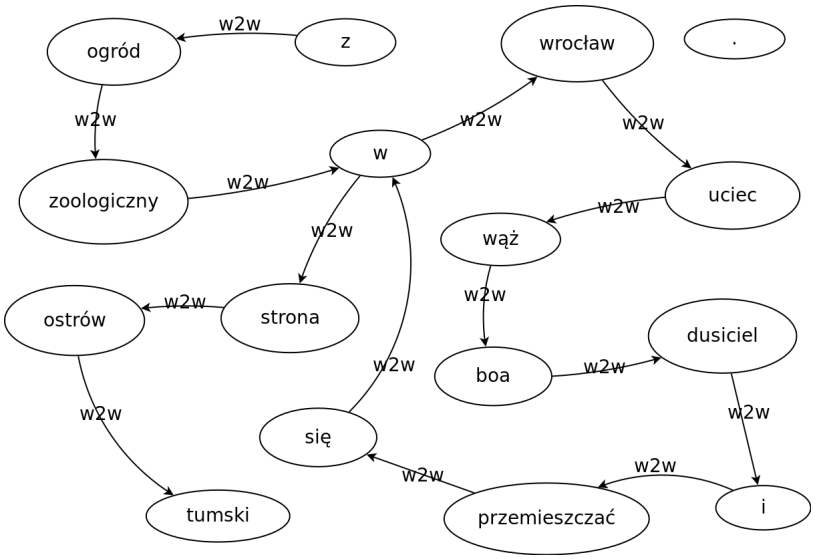
$$f_{et}(r(w_i, w_j)) \rightarrow e(k_l, k_m, l_e)$$

Funkcja transformująca f_{et} tworzy skierowaną krawędź z przypisaną etykietą l_e pomiędzy węzłami z etykietami k_l oraz k_m. W oznaczeniu funkcji f_{et} i l_e, litera e jest

oznaczeniem krawędzi (ang. *edge*). Dostępnych jest 6 funkcji f_{et} transformujących relacje z tekstu do krawędzi w grafie. Są to:

$$f_{et} \in \{dep\_rel, malt\_rel, ne2ne, w2w, h2h, sem\_role\}$$

Będą one szczegółowo opisane w punkcie 5.3.4. Jednak w celu zobrazowania dołączania krawędzi do grafu rozważymy tutaj najprostszą relację zachodzącą między wyrazami, jest to relacja ustalająca kolejność występowania słów w tekście, nazwana $w2w$ – (ang. *word to word*). Łączy ona ze sobą węzły, które reprezentują słowa występujące po sobie w tekście.



Rys. 4.8. Prosty graf zbudowany dla zdania *Z ogrodu zoologicznego we Wrocławiu uciekł wąż Boa Dusiciel i przemieszcza się w stronę Ostrowa Tumskiego.*

Na rysunku 4.8 przedstawiony został prosty graf, w którym węzeł reprezentuje formę podstawową słowa (dla węzłów przyjęto funkcję transformującą $f_{nt} = Base$). Skierowana krawędź określa kolejność występowania słów w tekście (dla krawędzi użyto funkcji transformującej f_{et}=$w2w$). Analiza grafu z rysunku 4.8 pozwala zauważyć przynajmniej dwie sytuacje, które są niepoprawne. Węzeł zaetykietowany *przemieszczać* łączy się z węzłem z etykietą *się*. Jeżeli poddamy analizie pojedyncze słowa, można powiedzieć, że przemieszcza oznacza przemieszczenie jakiejś rzeczy w określonym kierunku, np. *przemieszcza karton w kierunku rogu stołu*, a *się* jest zaimkiem zwrotnym, określającym między innymi[4]:

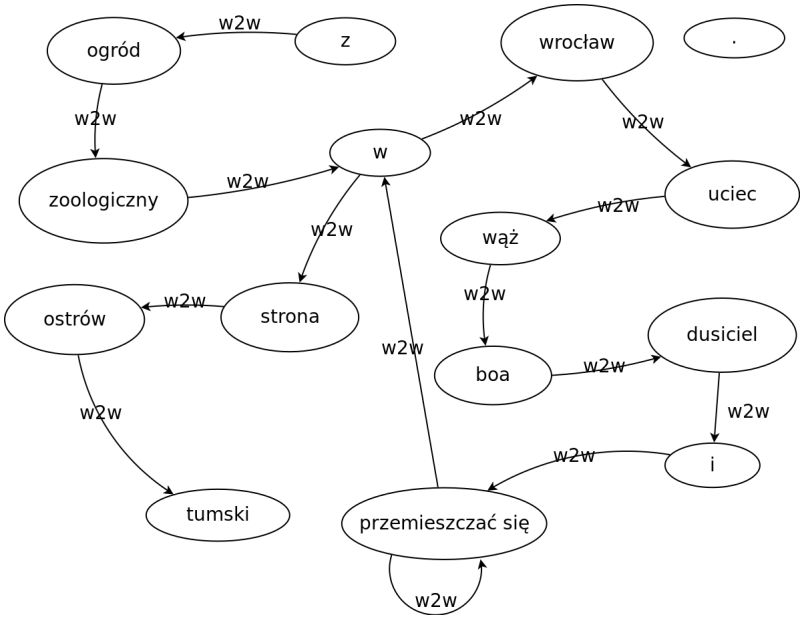
— w stronie zwrotnej – konstrukcję, w której wykonawca lub wykonawcy czynności jednocześnie doznają jej skutków,
— formę bezosobową,
— oznaczenie zbiorowości wykonującej czynność.

Takie podejście wprowadza niejednoznaczność w interpretacji otrzymanego grafu. Aby ją usunąć, kolejnym krokiem jest przebudowa grafu, która polega na analizie tekstu i

[4] Na podstawie artykułu `https://pl.wiktionary.org/wiki/si%C4%99`, ostatni dostęp 18.08.2018

podjęciu decyzji czy *się* powinno być połączone z czasownikiem, czy też występować w grafie osobno. W metodzie *PAS* jest to krok opcjonalny i to użytkownik decyduje czy będzie on wykonywany. Jeżeli użytkownik chciałby, aby *się* było dołączane do czasownika, powinien zaznaczyć to w *scenariuszu realizacji metody*, przypisując do zmiennej `rebuilders` wartość `VerbSieBuilder`.

W przytoczonym przykładzie, wyraz *się* określa w stosunku do węża *przemieszczanie siebie*. Aby znaczenie tego wyrażenia zostało poprawnie odzwierciedlone w grafie, należy w tym wypadku połączyć z sobą oba węzły. Po wykonaniu tego kroku, graf przyjmie postać pokazaną na rysunku 4.9. Węzły odpowiadający słowom *przemieszcza*



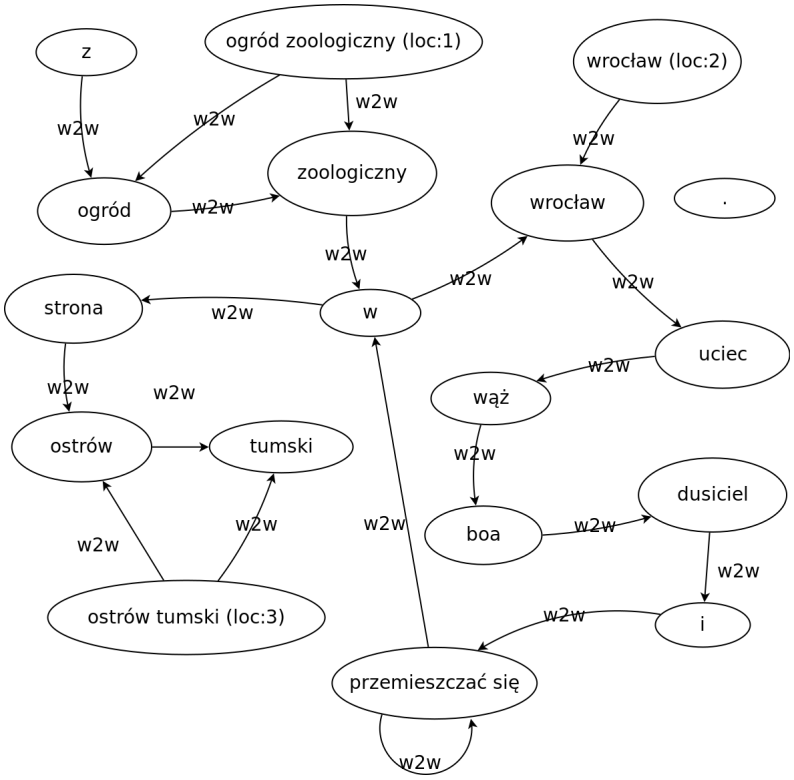
Rys. 4.9. Prosty graf wzbogacony o łączenie „się” z czasownikiem

oraz *się* zostały połączone w jeden *przemieszczać się*, który jednoznacznie wskazuje na przemieszczanie siebie, a oba pojedyncze węzły zostały usunięte. Wszystkie krawędzie wchodzące i wychodzące do węzłów *przemieszczać* jak i *się*, zostały przepięte do nowo dodanego węzła. Zapis s*cenariusza realizacji* dla tego przykładu, znajduje się w dodatku B.2.1.

Uwzględnienie łączenia *się* z czasownikiem daje ich poprawne lematy, jednak nadal pozostają niejednoznaczności związane z opisywaniem złożonych, wielowyrazowych nazw własnych. W przytaczanym przykładzie złożoną nazwą własną jest *Ostrów Tumski* czy też *Ogród Zoologiczny*. Aby rozwiązać ten problem, należy podobnie jak w przypadku łączenia *się* z czasownikiem, połączyć ze sobą węzły odpowiadające wyrazom wielowyrazowej nazwy własnej w jeden węzeł. Wykonanie tego kroku jest opcjonalne. Jeżeli użytkownik metody uzna, że informacja o tym, że węzły dotyczą nazwy własnej jest dla niego istotna, powinien w *scenariuszu realizacji* do zmiennej `rebuilders` przypisać wartość `NEGraphBuilder`.

Należy podkreślić, że dodawanie informacji o nazwie własnej do grafu, różni się jednak od podejścia dotyczącego łączenia *się* z czasownikiem. Różnica polega na tym, że poszczególne węzły odnoszące się do słów wchodzących w skład nazwy własnej oraz

krawędzie, którymi węzły te łączą się z innymi, pozostają w niezmienionej formie w grafie. Do grafu dodawany jest nowy węzeł z identyfikatorem będącym konkatenacją identyfikatorów węzłów odpowiadających słowom składowym wielowyrazowej nazwy własnej. Utworzony węzeł łączy się z tymi węzłami krawędzią, do której przypisana jest relacja określającą klasę nazwy własnej. W przypadku *Ostrowa Tumskiego* oraz *Ogrodu Zoologicznego* jest to relacja *loc*, czyli lokalizacja. W przykładzie przedstawionym na



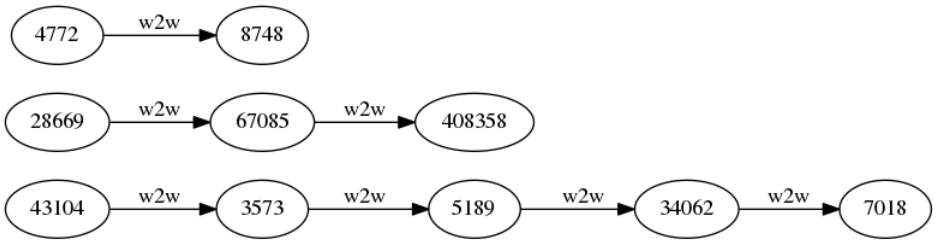
Rys. 4.10. Graf, w którym *się* dołączone jest do czasownika oraz dodano nowy węzeł odpowiadający nazwie własnej

rysunku 4.10 węzły z przypisanymi etykietami *ogród* oraz *zoologiczny* zostały połączone z nowo dodanym węzłem o nazwie *ogród zoologiczny*, węzły z etykietami *ostrów* i *tumski* połączone zostały z węzłem *ostrów tumski*. Etykiety węzłów pisane są z małych liter, ponieważ wykorzystany został typ węzła, który formy podstawowe słów sprowadza do małych liter. W obu wypadkach od nowo dodanych węzłów tworzona jest krawędź skierowana do węzłów wskazujących na słowa składowe wielowyrazowej nazwy własnej, z przypisaną etykietą `loc`. Jeżeli w tekście rozpoznana została jednoelementowa nazwa własna, w przytaczanym wypadku jest to *Wrocław*, to tworzony jest dodatkowy węzeł z etykietowaną krawędzią skierowaną, określającą klasę nazwy własnej. W dodatku B.2.2 przedstawiony został scenariusza użycia, dla omawianego przykładu.

Jak można zauważyć, otrzymywane grafy znacząco się różnią w zależności od użytych funkcji transformujących dla krawędzi f_{et} i węzłów f_{nt} oraz od sposobu traktowania *się* i nazw własnych. Wszystkie te ustawienia zapisane są w *scenariuszu realizacji* metody.

Aby w pełni zobrazować ten fakt, zmieńmy w stosunku do grafu przedstawionego na rysunku 4.8 reprezentację słowa w tekście na *Synset*, czyli funkcja transformująca dla

węzła $f_{nt} = Synset$. Wynikowy graf dla zdania *Z ogrodu zoologicznego we Wrocławiu*



Rys. 4.11. Graf z typem węzła ustawionym na *Synset*, typem krawędzi `w2w`, zbudowany dla zdania „Z ogrodu zoologicznego we Wrocławiu uciekł wąż Boa Dusiciel i przemieszcza się w stronę Ostrowa Tumskiego."

uciekł wąż Boa Dusiciel i przemieszcza się w stronę Ostrowa Tumskiego przedstawiony jest na rysunku 4.11. W stosunku do grafu z rysunku 4.8 posiada on mniej węzłów, nie jest spójny oraz etykietami węzłów są identyfikatory znaczeń zaczerpnięte ze Słowosieci. *Scenariusz realizacji* dla tego przykładu 4.11 zamieszczony został w dodatku B.2.3.

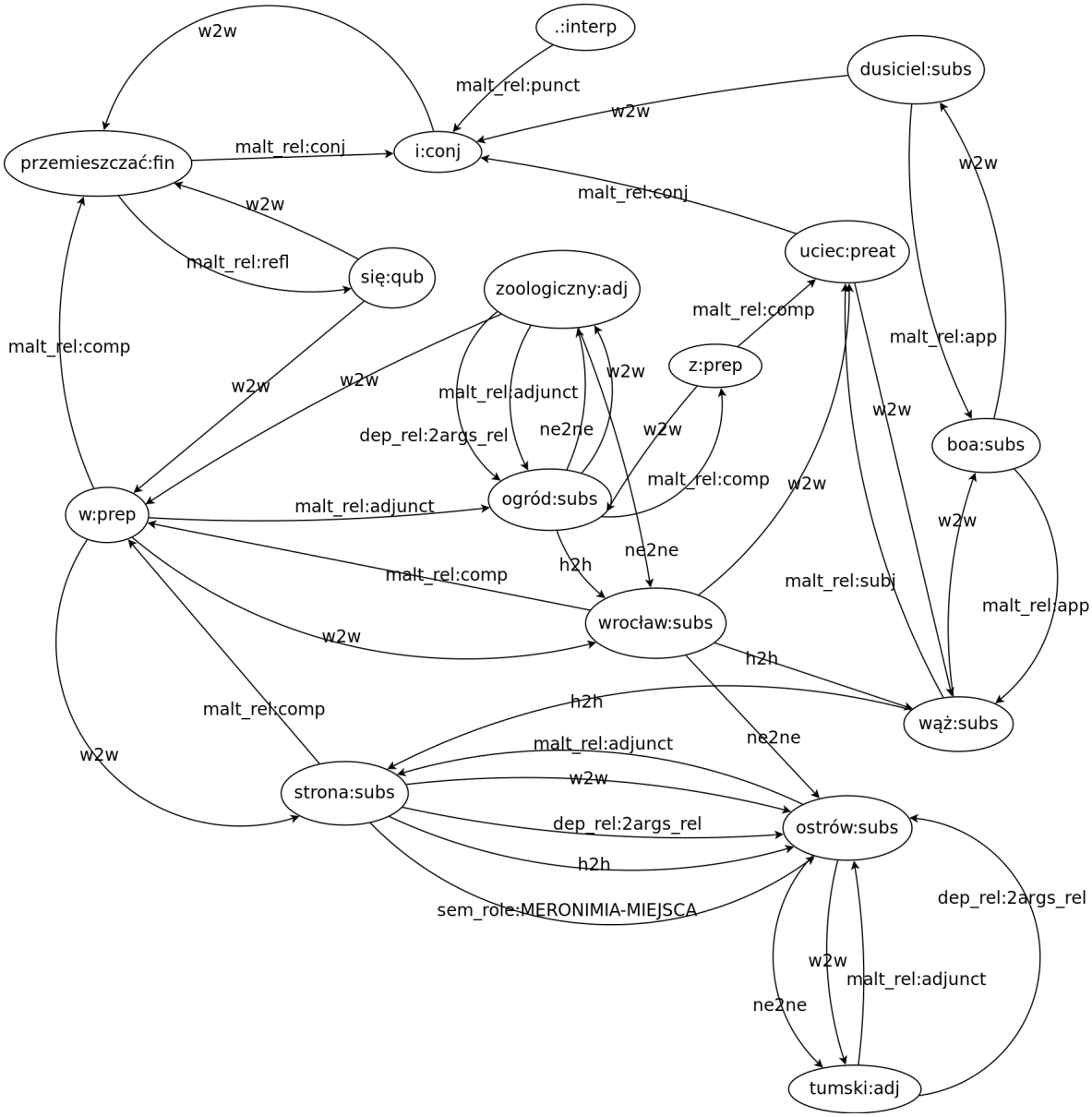
Jeśli w trakcie budowania grafu, dla krawędzi będziemy używać wszystkich funkcji transformujących f_{et}, to otrzymamy graf przedstawiony na rysunku 4.12 z pełnym zestawem typów krawędzi, wynikającym z użycia wszystkich funkcji. Otrzymany graf jest znacznie bardziej rozbudowany w porównaniu do 4.8. Scenariusz realizacji dla tego przykładu znajduje się w dodatku B.2.4.

Wycinanie i łączenie grafów reprezentujących wiedzę tekstową i pozatekstową

Ważnym elementem metody *PAS* jest możliwość dołączania wiedzy pozatekstowej do wiedzy wydobytej z tekstu. W pracy rozważane są dwa *zewnętrzne źródła Wiedzy* – Słowosieć oraz ontologia SUMO. Zanim przejdziemy do istoty problemu, czyli metody wycinania i łączenia grafów reprezentujących zewnętrzne zasoby wiedzy, w pierwszej kolejności krótko scharakteryzujemy sposób dołączania wiedzy ze Słowosieci i z Ontologii SUMO.

Dołączanie wiedzy ze Słowosieci W *Słowosieci* możemy odnaleźć znaczenia słów z tekstu, dzięki wykonaniu procesu *ujednoznaczniania tekstu* (opis w podrozdziale 4.2.2). Pomiędzy znaczeniami słów zachodzą relacje na poziomie synsetów oraz jednostek leksykalnych. Określają one poziom ogólności znaczenia (przy wykorzystaniu *hiponimii/hiperonimii* – czyli relacji podrzędności/nadrzędności między znaczeniami słów, czy *meronimii*, która wskazuje, że jeden byt reprezentowany przez znaczenie słowa jest częścią innego). Są to relacje, które nie występują bezpośrednio w tekście.

Dla słów z tekstu w procesie wydobywania wiedzy pozatekstowej ze Słowosieci wydobywane są znaczenia. Są one wskazywane w grafie zbudowanym ze Słowosieci. Następnie wycinany jest spójny podgraf z grafu reprezentującego Słowosieć (oznaczany dalej jako $\mathbf{G'}_{plwn}$), zawierający wszystkie wskazane węzły (znaczenia, które wystąpiły w tekście). W wyciętym podgrafie znajduje się wiele znaczeń słów oraz relacji między nimi, które nie wystąpiły w tekście, jednak w semantyczny sposób są ze sobą powiązane. Prześledźmy teraz łączenie wiedzy przy użyciu grafów w sposób bardziej szczegółowy.



Rys. 4.12. Graf powstały przy użyciu dla węzła funkcji transformującej $f_{nt} = Base$ a dla krawędzi wszystkich funkcji transformujących (dzięki czemu otrzymywany jest pełny zestaw typów krawędzi). Graf zbudowany jest dla zdania „Z ogrodu zoologicznego we Wrocławiu uciekł wąż Boa Dusiciel i przemieszcza się w stronę Ostrowa Tumskiego.”

Przyjmijmy, że $\mathbf{G}$ to graf zbudowany po przetworzeniu tekstu z funkcją f_{nt} konwertującą słowa z tekstu do węzłów w grafie oraz funkcją f_{et}, która do tego grafu przypisuje krawędzie na podstawie relacji między słowami. W rozważanym obecnie etapie dołączania wiedzy ze Słowosieci w pierwszej kolejności wycinany jest podgraf $\mathbf{G}'_{plwn}$. W tym podgrafie węzły są znaczeniami, a krawędzie odzwierciedlają relacje między nimi. W wyniku połączenia grafu $\mathbf{G}$ z podgrafem $\mathbf{G}'_{plwn}$ powstaje nowy graf, nazwany jako $\mathbf{G}_{merged}$. Zawiera on węzły oraz relacje wydobyte z tekstu oraz węzły i relacje występujące w podgrafie wyciętym ze Słowosieci. Takie połączenie możemy formalnie opisać wzorem 4.3.

$$\mathbf{G}_{merged} = \mathbf{G} \cup \mathbf{G}'_{plwn} \tag{4.3}$$

Jak zwykle, decyzja o tym, czy do grafu z wiedzą tekstową dołączana jest wiedza pozatekstowa ze Słowosieci, należy do użytkownika. W *scenariuszu realizacji* możliwe jest uruchomienie tego procesu, poprzez przypisanie do zmiennej `plwn_merger` wartości `PlWNMerger` lub `PlWNMappingMerger`. Możliwe jest przypisanie tylko jednej wartości, ponieważ wartości te określają sposób wycinania podgrafu $\mathbf{G}'_{plwn}$. Pozostawienie zmiennej `plwn_merger` pustej, skutkuje niedołączeniem do grafu $\mathbf{G}$, grafu $\mathbf{G}'_{plwn}$. Wyjaśnienie różnic między wykorzystaniem `PlWNMerger` a `PlWNMappingMerger` zostało wyjaśnione dalej.

Dołączanie wiedzy z ontologii SUMO Drugim rozważanym w pracy zewnętrznym źródłem wiedzy jest *ontologia SUMO*. SUMO jest ontologią wysokiego poziomu (aczkolwiek posiada wiele dziedzinowych rozszerzeń), zatem wiedza w niej zawarta jest na bardzo wysokim poziomie abstrakcji. Proces wyciągania wiedzy z SUMO jest podobny do wyciągania wiedzy ze Słowosieci z tą różnicą, że najpierw do słów w tekście przypisywane są znaczenia, a następnie do znaczeń słów przypisywane są pojęcia z SUMO. Szczegółowy opis rzutowania Słowosieci na SUMO znajduje się w rozdziale 4.2.3.

Budowa grafu reprezentującego ontologię SUMO przebiega podobnie do budowy grafu dla Słowosieci. Pojęcie z ontologii staje się węzłem, zaś relacje między pojęciami są krawędziami w budowanym grafie. W tak zbudowanym grafie wyszukiwane i oznaczane są te węzły, które posiadają odzwierciedlenie w grafie reprezentującym tekst, a następnie wycinany jest spójny podgraf, zawierający wszystkie te węzły. Podgraf ten nazwany został $\mathbf{G}'_{SUMO}$. W dalszym etapie, graf $\mathbf{G}'_{SUMO}$ dołączany jest do grafu $\mathbf{G}$ reprezentującego rozważany tekst:

$$\mathbf{G}_{merged} = \mathbf{G} \cup \mathbf{G}'_{SUMO}$$

Również i w tym wypadku, użytkownik decyduje o tym, czy do grafu $\mathbf{G}$ dołączany jest graf z wiedzą pozatekstową $\mathbf{G}'_{SUMO}$. Dokonuje tego za pomocą *scenariusza użycia* ustawiając odpowiednio zmienną `sumo_merger` na wartość `SUMOMerger` lub `SUMOMappingMerger`. Użytkownik może przypisać tylko jedną wartość, która wskazuje na sposób wycinania podgrafu $\mathbf{G}'_{SUMO}$. Jeżeli zmienna `sumo_merger` nie będzie posiadała przypisanej wartości, to wiedza pozatekstowa z SUMO nie zostanie dołączona do grafu zbudowanego z wiedzy tekstowej.

Warto zaznaczyć, że do grafu $\mathbf{G}$ można dołączyć jednocześnie wiedzę wydobytą ze Słowosieci i z SUMO. Wówczas wynikowy graf $\mathbf{G}_{merged}$ jest opisany wzorem 4.4. Do grafu tekstowego $\mathbf{G}$ dołączany jest podgraf $\mathbf{G}'_{plwn}$ oraz $\mathbf{G}'_{SUMO}$.

$$\mathbf{G}_{merged} = \mathbf{G} \cup \mathbf{G}'_{plwn} \cup \mathbf{G}'_{SUMO} \tag{4.4}$$

Na rysunku 4.13 przedstawiono graf, w którym węzeł jest wynikiem działania funkcji transformującej $f_{nt} = Base$, zestaw krawędzi utworzony został przy użyciu funkcji transformującej $f_{et} = (\texttt{w2w})$ (relacja kolejności słów w tekście), lecz do tak zbudowanego grafu dołączony został podgraf $\mathbf{G'}_{plwn}$ oraz $\mathbf{G'}_{SUMO}$. Z powodu dużej liczby elementów rysunek 4.13 jest mało czytelny. Jednak daje wyobrażenie o tym jak bardzo wzrosła ilość informacji, które udało się wydobyć z zewnętrznych źródeł wiedzy, w porównaniu do sytuacji przedstawionej na rysunku 4.10, gdzie zawarte są tylko informacje wydobyte z tekstu. Scenariusz realizacji dający w wyniku graf z rysunku 4.13 przedstawiony jest w dodatku B.2.5.

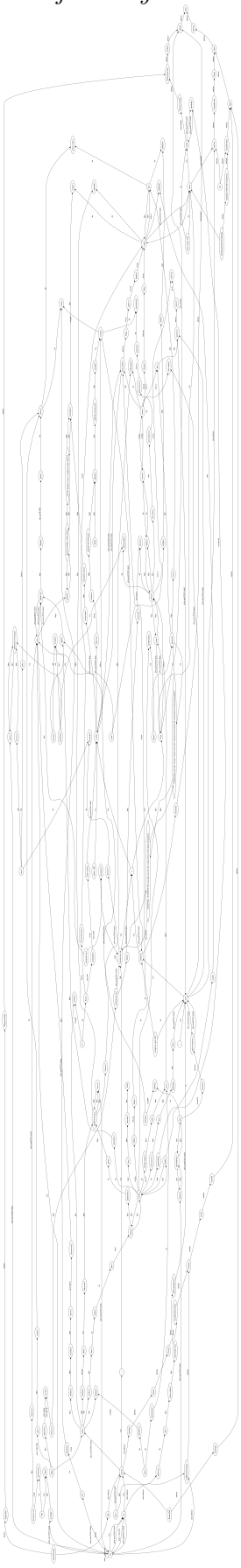
Fragment grafu z rysunku 4.13 przedstawiony został na rysunku 4.14, na którym widoczne jest słowo *wąż* z tekstu oraz relacje tekstowe przypisane do krawędzi na podstawie funkcji transformującej $f_{et} = w2w$ do słowa *boa* oraz $f_{et} = w2w$ od *boa* do *dusiciel*. Widoczne są znaczenia wydobyte ze Słowosieci oraz pojęcia wydobyte z SUMO, które nie wystąpiły w tekście, a są wynikiem łączenia wiedzy pozatekstowej z wiedzą zawartą w tekście, na przykład znaczenia *dusić.7*, *gad.1*, czy też pojęcie *Man*.

Algorytm wycinania i łączenia grafów z zewnętrznych zasobów wiedzy

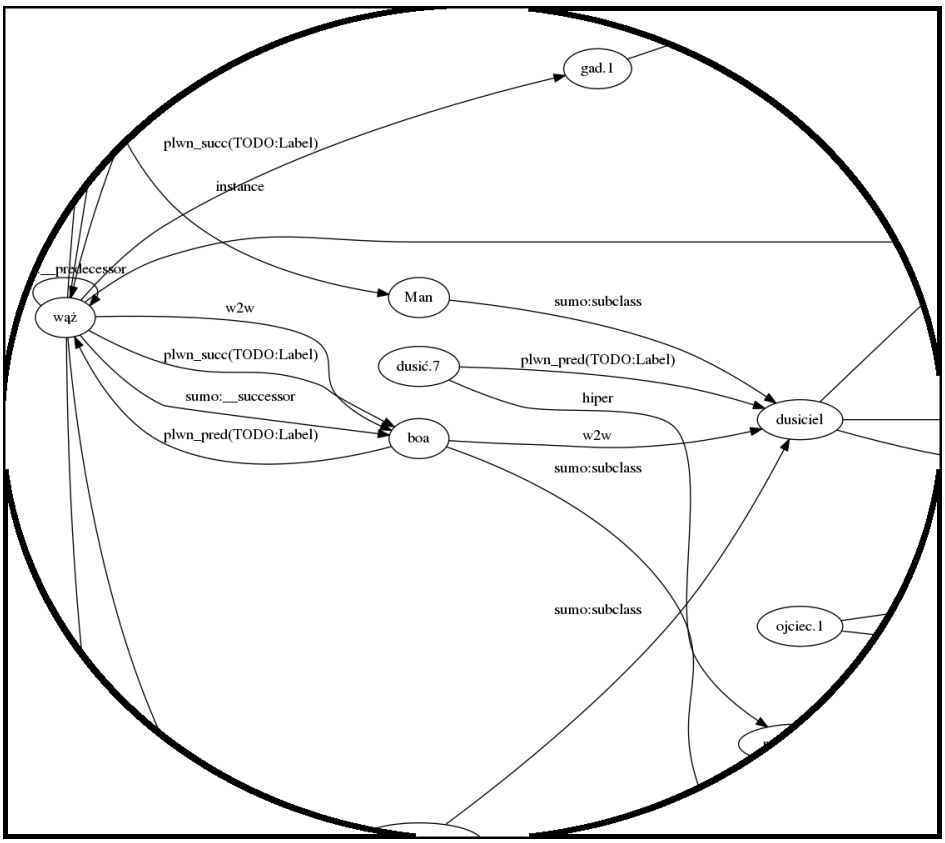
Optymalne wycinanie podgrafu z grafu reprezentującego źródło wiedzy pozatekstowej powinno polegać na znalezieniu i wycięciu *minimalnego spójnego podgrafu* z zadanego grafu. Zadanie to jest problemem NP-trudnym, znanym w literaturze jako *Steiner Tree Problem* (Lin i Xue, 1999). Dlatego, w niniejszej pracy przyjęto uproszczenie, polegające na znajdowaniu spójnego podgrafu, który niekoniecznie jest minimalnym spójnym podgrafem. Oznacza to, że proces dołączania wiedzy pozatekstowej do wiedzy tekstowej, sprowadza się do znalezieniu spójnego podgrafu $\mathbf{G'}_{EKG}$ z grafu oznaczanego dalej $\mathbf{G}_{EKG}$ (EKG - od **E**xternal **K**nowledge **G**raph) reprezentującego wiedzę pozatekstową. Zakładając, że dany jest graf $\mathbf{G}_{EKG}$ będący reprezentacją zewnętrznego źródła wiedzy, algorytm wycinania spójnego podgrafu z wykorzystaniem najkrótszych ścieżek *sp* między węzłami n_i, a n_j, grafu reprezentującego zewnętrzne źródło wiedzy przedstawiony został za pomocą algorytmu 5.
Przebieg algorytmu 5 jest następujący. Przyjmijmy, że dane są grafy $\mathbf{G}$, $\mathbf{G}_{EKG}$ reprezentujące odpowiednio wiedzę tekstową i pozatekstową.

— Dla każdego węzła $n \in \mathbf{G}$ reprezentującego odpowiednie słowo w tekście, znajdź odpowiadający mu węzeł n_{EKG} znajdujący się w grafie $\mathbf{G}_{EKG}$ będącym reprezentacją zewnętrznego źródła wiedzy i zapamiętaj go w zbiorze węzłów V_{EKG}:
  — dla Słowosieci, jest to znaczenie przypisane do słowa w tekście;
  — dla ontologii SUMO, jest to pojęcie, na jakie może być zrzutowane dane słowo z tekstu;
— Dla każdego $n_{EKG} \in V_{EKG}$, znajdź najkrótsze ścieżki $sp(n_{EKG_i}, n_{EKG_j})$, między danym węzłem, a każdym węzłem z V_{EKG}.
— Połącz wszystkie $sp(n_{EKG_i}, n_{EKG_j})$ ze sobą, tworząc spójny podgraf $\mathbf{G'}_{EKG}$, gdzie $\mathbf{G'}_{EKG} = \{\mathbf{G'}_{plwn}, \mathbf{G'}_{SUMO}\}$
— Wykorzystaj $\mathbf{G'}_{EKG}$ w procesie dołączania wiedzy pozatekstowej do wiedzy tekstowej.

Scenariusz realizacji metody, pozwala na dwojaki sposób dołączania wiedzy pozatekstowej do wiedzy tekstowej. Algorytmy dołączania nazywają się `Mapping` oraz `MappingMerger`, zarówno dla Słowosieci (`PlWNMapping` oraz `PlWNMappingMerger`) jak



Rys. 4.13. Graf z typem węzła ustawionym na lemat słowa ($f_{nt} = Base$), łączeniem nazw własnych, dołączaniem *się* do czasownika oraz dołączaniem wiedzy pozatekstowej ze Słowosieci oraz SUMO



Rys. 4.14. Fragment grafu z rysunku 4.13, w którym typ węzła ustawionym został na lemat słowa ($f_{nt} = Base$), wykonane zostało łączenie nazw własnych, dołączanie *się* do czasownika oraz dołączanie wiedzy pozatekstowej ze Słowosieci oraz SUMO do wiedzy tekstowej

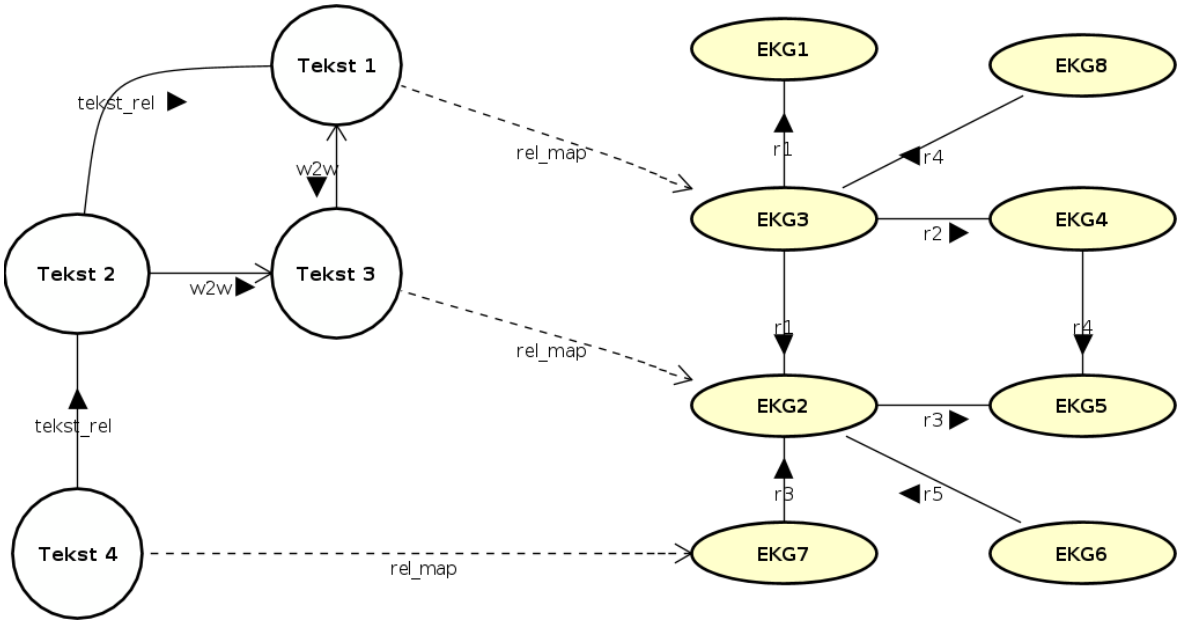
Algorithm 5 Algorytm wycinania spójnego podgrafu $\mathbf{G}'_{EKG} \in \mathbf{G}_{EKG}$

G – graf zbudowany z tekstu
$\mathbf{G}_{EKG}$ – graf reprezentujący wiedzę pozatekstową
$\mathbf{G}_{EKG} = \{\mathbf{G}_{PlWN}, \mathbf{G}_{SUMO}\}$
$V_{EKG} = \{\emptyset\}$
$SP = \{\emptyset\}$
for $n \in \mathbf{G}$ **do**
    `znajdź` $n_{EKG} \in \mathbf{G}_{EKG}$, $n = n_{EKG}$
    $V_{EKG} = V_{EKG} \bigcup \{n_{EKG}\}$
end for
for $n_{EKG_i} \in V_{EKG}$ **do**
    for $n_{EKG_j} \in V_{EKG}$ **do**
        $sp_{i,j} = sp(n_{EKG_i}, n_{EKG_j})$
        $SP = SP \bigcup \{sp_{i,j}\}$
    end for
end for
$\mathbf{G}'_{EKG} = \{\emptyset\}$
for $sp_{i,j} \in SP$ **do**
    $\mathbf{G}'_{EKG} = \mathbf{G}'_{EKG} \bigcup sp_{i,j}$
end for
$\mathbf{G}' = \mathbf{G}'_{EKG} \bigcup \mathbf{G}$

i ontologii SUMO (`SUMOMapping`, `SUMOMappingMerger`). Sposób dołączania podgrafu według algorytmu `Mapping` oraz `MappingMerger` jest taki sam dla obu rozważanych w pracy źródeł wiedzy pozatekstowej.

Na rysunku 4.15 przedstawiony został sposób dołączania wiedzy pozatekstowej, nazwany `Mapping`. Węzły białe, z przypisanymi etykietami $Tekst_i$, to węzły z grafu zbudowanego na podstawie rozważanego tekstu. Relacje między tymi węzłami wynikają z relacji, jakie zachodziły między odpowiadającymi im słowami w tekście. Żółte elipsy, z etykietami EKG_i należą do podgrafu wyciętego z reprezentacji grafowej wiedzy pozatekstowej. Może to być podgraf ze Słowosieci lub SUMO. Między węzłami EKG_i w grafie zachodzą relacje znalezione na podstawie zewnętrznego źródła wiedzy, oznaczone na rysunku jako r_i. Algorytm `Mapping` polega na połączeniu ze sobą węzłów zbudowanych na podstawie tekstu, z odpowiadającymi im węzłami w zewnętrznym źródle wiedzy. Innymi słowy, łączone są ze sobą dwa grafy, których struktura zostaje niezmieniona, a jedynie dodawane są relacje *rel_map* między odpowiadającymi sobie węzłami. Krawędzie modelujące relacje tekstowe, jak i relacje pomiędzy węzłami grafu zbudowanego dla zewnętrznego źródła wiedzy, pozostają bez zmian. Dla sytuacji zobrazowanej na rysunku 4.15 zostały dodane trzy krawędzie o nazwie *rel_map* w wyniku łączenia wiedzy tekstowej z pozatekstową: między węzłami *Tekst 1 – EKG3*, *Tekst 3 – EKG2* oraz *Tekst 4 – EKG7*.

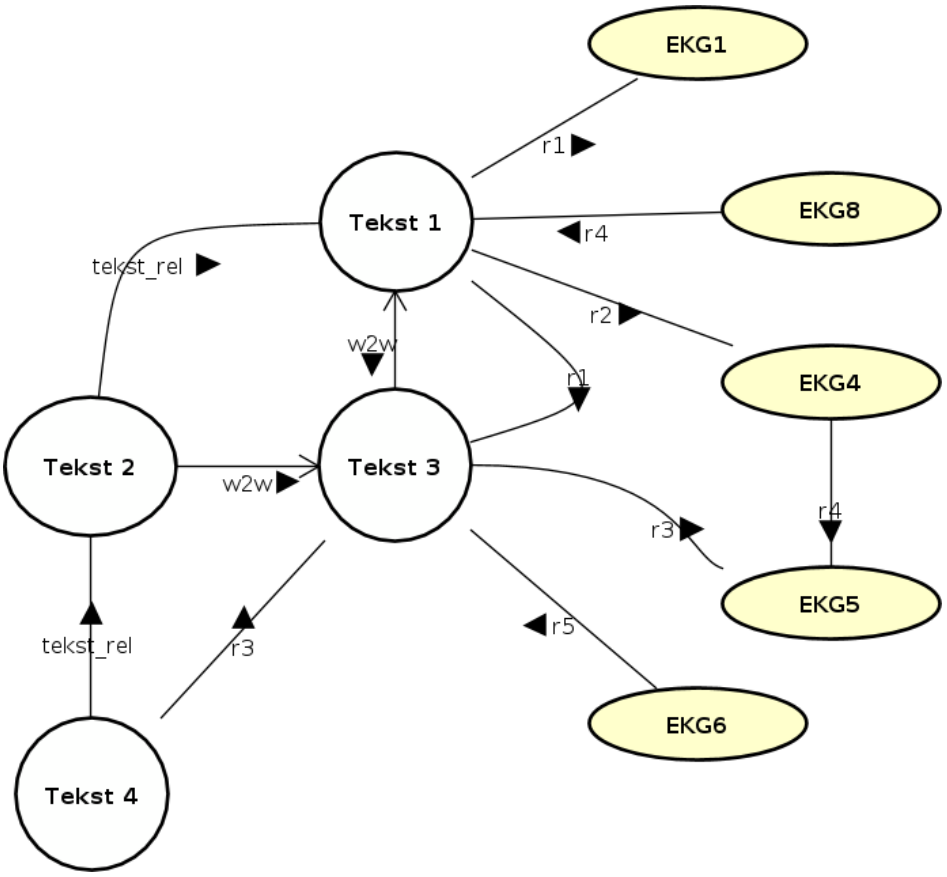
Na rysunku 4.16 przedstawiony został efekt zastosowania algorytmu `MappingMerger` dla grafu zbudowanego z tekstu oraz podgrafu wyciętego z grafu reprezentującego zewnętrzne źródło wiedzy, dla sytuacji początkowej identycznej jak



Rys. 4.15. Łączenie wiedzy pozatekstowej z tekstową za pomocą algorytmu `Mapping`

przed zastosowaniem operacji `Mapping`. Otrzymany wynik jest zupełnie inny. W tym wypadku nie jest dodawana nowa relacja łącząca węzły w grafie reprezentującym tekst z węzłami wyciętego podgrafu. Każdy węzeł w grafie reprezentującym tekst, który posiada relację *rel_map* do węzła w grafie reprezentującym zewnętrzne źródło wiedzy, jest łączony z tym węzłem. Wynikiem takiego łączenia jest usunięcie węzła n_u z $\mathbf{G}'_{EKG}$, z którym połączony był węzeł w grafie reprezentującym tekst oraz przepięcie do niego wszystkich krawędzi, które łączyły węzeł n_u z dowolnym innym z $\mathbf{G}'_{EKG}$. Jako etykieta połączonych węzłów ustawiana jest etykieta węzła pochodzącego z tekstu. Na rysunku 4.16 węzeł *Tekst 1* został połączony z węzłem *EKG3*. Etykieta połączonych węzłów ustawiona została na *Tekst 1*, zaś krawędzie pomiędzy *EKG3*, a *EKG8* i *EKG4*, zostały podpięte do węzła *Tekst 1*. Krawędzie, które łączyły węzeł *EKG2* z węzłami *EKG5*, *EKG6* oraz *EKG7*, zostały przepięte do węzła *Tekst 3*. Tym samym, między węzłem *Tekst 1*, a *Tekst 3* dodana została nowa krawędź *r1*, która pierwotnie łączyła ze sobą *EKG3* z *EKG2*. W wyniku połączenia węzłów reprezentujących słowa z tekstu z węzłami występującymi w zewnętrznym źródle wiedzy, węzły *EKG2*, *EKG3* oraz *EKG7* nie występują w połączonym grafie. Krawędzie łączące te węzły z pozostałymi węzłami z $\mathbf{G}'_{EKG}$ zostały przepięte kolejno do węzłów *Tekst 3*, *Tekst 1* oraz *Tekst 4*.

Na koniec warto przypomnieć, że obligatoryjne jest wykonanie jedynie kroku pierwszego, czyli budowa grafów Φ. Grafy Φ zbudowane na podstawie tekstu mogą być od razu wykorzystane w procesie dalszego przetwarzania. Jednak wykonanie kroku drugiego oraz trzeciego, pozwala w dużo lepszy sposób zamodelować semantykę tekstu.



Rys. 4.16. Łączenie wiedzy pozatekstowej z tekstową za pomocą algorytmu `MappingMerger`

Rozdział 5

Zadanie rozpoznawania relacji semantycznych

Niniejszy rozdział opisuje rozpoznawanie relacji semantycznych między fragmentami tekstów dla języka polskiego, które zdefiniowano jako zadanie klasyfikacji. Przedstawiono wykorzystane w pracy metody klasyfikacji oraz zestaw cech służący do budowy wektorów wejściowych dla zastosowanych klasyfikatorów. Opisano również zbiór danych uczących.

5.1. Zdefiniowanie problemu rozpoznawania relacji semantycznych

Rozpoznawanie relacji semantycznych jest realizowane w niniejszej pracy jako zadanie klasyfikacji. Rozpoznawane są wszystkie relacje zdefiniowane w modelu *CST* (rozdział 3.1.1). Problem rozpoznawania relacji między fragmentami tekstów sprowadza się więc do rozpoznawania jednej z 16 klas oznaczających zachodzenie relacji z modelu *CST* oraz jednej klasy, nazwanej *Brak_relacji*, która oznacza, że pomiędzy daną parą zdań nie rozpoznano żadnej relacji z modelu CST. Zatem, ostatecznie rozpoznawanych jest 17 klas, $\mathcal{C}_1, ..., \mathcal{C}_{17}$.

Na wejście klasyfikatora podawane są wektory cech $\vec{v}$ opisujące fragmenty tekstów, pomiędzy którymi wykrywana jest relacja, $\vec{v} \in \mathcal{V}$, $\mathcal{V} = \{\vec{v}_1, ..., \vec{v}_N\}$, gdzie $\mathcal{V}$ oznacza zbiór danych uczących, a N to liczba przykładów uczących. Wektor cech $\vec{v}$ definiowany jest jako zbiór cech f_i, $\vec{v} = [f_1, ..., f_L]$, gdzie L to liczba cech f_i w wektorze $\vec{v}$. W opisywanych dalej badaniach L, czyli wymiar wektora cech zależny jest od badanego zestawu cech.

W pracy przyjęto kodowanie wyjść za pomocą podejścia *jeden z n*, co oznacza, że klasyfikator na wyjściu zawsze podaje jedną klasę (tę, która wskazywana jest przez wyjście bliskie jeden), pozostałe wyjścia powinny być bliskie zeru. Używając opisu probabilistycznego, rozpoznanie relacji semantycznej dla danego wektora $\vec{v}$ reprezentującego porównywane teksty możemy określić jako znalezienie klasy $\mathcal{C}_k$, której prawdopodobieństwo p wystąpienia jest najwyższe, pod warunkiem, że na wejście został podany wektor wejściowy $\vec{v}$ i uczono klasyfikator na zbiorze uczącym $\mathcal{V} = \{\vec{v}_1, ..., \vec{v}_N\}$.

Formalnie, zgodnie z (Murphy, 2013), klasyfikację możemy zapisać wzorem 5.1.

$$\mathcal{C}_k = arg \max_{i=1}^{17} p\Big(y_i = \mathcal{C}|\vec{v}, \mathcal{V}\Big) \tag{5.1}$$

Takie podejście wymaga jednak sumowania wszystkich wyjść do wartości 1. Jeśli ten warunek nie jest spełniony wyjścia z klasyfikatora przestają mieć interpretację probabilistyczną, ale klasa dalej wyznaczana jest przez wyjście z maksymalną wartością. Jakość wyuczonego klasyfikatora jest sprawdzana na zbiorze testowym $\mathcal{V}_{test}$, którego elementy nie należą do zbioru uczącego $\mathcal{V}$

Oczywistym jest, że jakość klasyfikacji (rozpoznawania relacji semantycznych) zależy od jakości cech, zawartych we wzorcu wejściowym reprezentowanym przez wektor $\vec{v}$, a dokładniej od zdolności dyskryminowania klas przy użyciu tego wektora. Do tej pory najczęściej wykorzystywano cechy opisane w (Maziero i inni, 2014), które w dalszej części pracy nazwano cechami bazowymi, tworzącymi *Zestaw cech 1.* Dzięki zastosowaniu metody PAS, która wydobywa wiedzę w procesie pogłębianej analizy semantycznej tekstu, reprezentowanej za pomocą grafu, możliwe stało się zaproponowanie nowych cech. W tym wypadku cechy określają podobieństwo między grafami zbudowanymi dla fragmentów tekstów. Wykorzystując podobieństwo grafów zaproponowano dwa zbiory cech, nazwane `cechy grafowe(1)` tworzące *Zestaw cech 2* oraz `cechy grafowe(2)`, tworzące *Zestaw cech 4.*

5.2. Opracowanie zbioru danych do klasyfikacji

Wyuczenie klasyfikatora rozpoznawania relacji semantycznych między fragmentami tekstów, wymaga posiadania zaanotowanych danych, które będą służyły do uczenia i testowania. Klasyfikator uczony jest metodą nadzorowaną, a więc każdy element tego zbioru to para – wektor cech i odpowiadająca mu etykieta relacji. Został zbudowany na podstawie zbioru tekstów w języku polskim (par zdań z przypisaną informacją o zachodzącej relacji między nimi). Dla każdej pary zdań tworzony był wektor cech (różny w zależności od badanego zestawu cech) z przypisaną etykietą relacji.

Zbiór tekstów, dla których określone zostały relacje semantyczne powstał w ramach projektu CLARIN-PL na przełomie lat 2015-2016. Jest pierwszym publicznie dostępnym, a zarazem do tej pory jedynym takim zestawem danych. Dane te zawierają pary zdań z dokumentów dostępnych na otwartej licencji. Pochodzą z podzbioru serwisu WikiNews [1]. Proces anotowania danych był dwuetapowy i zostanie szczegółowo przedstawiony w kolejnych podrozdziałach.

5.2.1. Pierwszy etap znakowania danych

Każdy z dokumentów $\mathcal{D}_i$ dostępnych w wybranym podzdbiorze WikiNews'ów (dalej zwany zbiór $WNews$) został w pierwszej kolejności automatycznie podzielony na zdania. Osoby anotujące, miały za zadanie odnaleźć w tym zbiorze pary zdań, między którymi zachodzi relacja semantyczna oraz odpowiednio oznaczyć. Ze względu na bardzo dużą liczbę zdań (164 697 zdań w 11 949 dokumentach) ręczny proces wyszukiwania

[1] `https://pl.wikinews.org/`

potencjalnych par, pomiędzy którymi może zajść jedna z relacji modelu CST, został częściowo zautomatyzowany.

W tym celu każdy z dokumentów został przekształcony w kolekcję podstawowych form słów ze zdań z tego dokumentu (ang. Bag of Words – $BOW(\mathcal{D}_i)$). Następnie każda z kolekcji słów przekształcona została do wektora reprezentującego słowa w kontekście całej kolekcji $WNews$. W kolejnym kroku pomiędzy tymi wektorami obliczana była miara kosinusowa. Dla danego dokumentu $\mathcal{D}_{i0}$ wyszukiwane były dwa dokumenty najbardziej podobne pod względem wartości miary kosinusowej. Dzięki temu, osoby ręcznie anotujące zbiór danych otrzymywały jedynie paczki składające się z 3 dokumentów, w których do dokumentu $\mathcal{D}_{i0}$ wskazane były dwa najbardziej podobne: $\mathcal{D}_{i1}$ oraz $\mathcal{D}_{i2}$. Zadaniem osoby anotującej było odszukanie dla każdego zdania $\mathcal{S}_i$ z $\mathcal{D}_{i0}$, zdań w dokumentach $\mathcal{D}_{i1}$ oraz $\mathcal{D}_{i2}$, pomiędzy którymi zachodzi jedna z relacji z modelu CST. Analogicznie do wyszukiwania zdań podobnych między dokumentami, wyszukiwanie zdań podobnych do siebie odbywało się w obrębie dokumentu $\mathcal{D}_{i0}$.

W procesie anotacji uczestniczyło 5 osób anotujących. Każda z osób była specjalnie wyszkolona do tego celu. Jedna z tych osób była koordynatorem całego procesu znakowania i to ona podejmowała końcowe decyzje w przypadku niezgodności między anotatorami. Poprawność procesu anotacji sprawdzona została na próbce 580 zdań wylosowanych z zaanotowanego zbioru. Do oceny wykorzystana została miara $SA(+)$ (ang. *Positive Specific Agreement*) (Hripcsak i Rothschild, 2005). Miara $SA(+)$ pozwala na zbadanie poprawności anotacji pomiędzy dwoma anotatorami, dlatego ocena poprawności znakowania sprawdzana była na zasadzie każdym z każdym. W zależności od relacji, osiągnięto zgodność między anotatorami od 33% do 100% miary $SA(+)$, obliczanej zgodnie ze wzorem 5.2.

$$SA(+) = \frac{2TP}{2TP + FP + FN} \tag{5.2}$$

We wzorze tym TP oznacza zgodność obu anotatorów co do decyzji pozytywnej (zachodzi relacja między dwoma fragmentami tekstów). FP oznacza liczbę takich sytuacji, w których tylko pierwszy z anotatorów wskazał, że między parą zdań zachodzi relacja. FN oznacza liczbę takich sytuacji, w których pierwszy z anotatorów nie wskazał relacji między daną parą zdań, zaś drugi z anotatorów wskazał taką relację. Jeżeli anotatorzy zgodnie będą podejmować decyzje, wtedy wartość miary $SA(+)$ wynosi 1. Jeżeli zaś decyzje będą zupełnie rozbieżne, wartość będzie wynosiła 0. W Tabeli 5.1 przedstawiona została zgodność anotacji z podziałem na poszczególne relacje. Kolumna **SA**$(+)$ zawiera wartość miary $SA(+)$ dla danej relacji, kolumna **Liczba relacji** zawiera informację o liczności ocenianej relacji. Miara $SA(+)$ pozwala na ocenę jakości znakowania pomiędzy dwoma anotatorami, dlatego wykonanie porównania każdy z każdym, wymaga dalszej szczegółowej analizy uzyskanych wartości miary $SA(+)$, która powinna być wykonana przez koordynatora procesu znakowania.

Część danych oceniona została niezależnie przez cztery osoby oraz koordynatora. Kiedy dla pary zdań $\mathcal{S}_1, \mathcal{S}_2$, wystąpiła niezgodność co do typu relacji zachodzącej między nimi, decyzję co do końcowej relacji podejmował koordynator. Jeżeli pomiędzy zadaną parą zdań nie zachodziła żadna relacja, informacja taka nie była anotowana. Bez udziału koordynatora ocenionych zostało 461 instancji relacji przez cztery osoby. Dana instancja relacji pomiędzy $\mathcal{S}_1$ a $\mathcal{S}_2$ uznawana była za poprawną, tylko wtedy, kiedy każda z czterech osób dokonała tego samego wyboru co do relacji zachodzącej między

Tabela 5.1. Przykładowe wartości jakości znakowania relacji między fragmentami tekstów mierzonej za pomocą miary $SA(+)$, między dwoma anotatorami

Relacja	**SA**$(+)$	**Liczba relacji**
Dalsze informacje	0,84	68
Krzyżowanie się	0,88	267
Mowa zależna	1	4
Opis	0,6	64
Parafraza	0,5	6
Spełnienie	0,88	24
Sprzeczność	0,33	3
Streszczenie	0,65	17
Tło historyczne	0,8	55
Tożsamość	1	6
Uszczegółowienie	0,88	24
Zawieranie	0,4	29
Zmiana poglądu	1	2
Źródło	0,64	11

tą parą. Liczności poszczególnych relacji, po pierwszym etapie znakowania przedstawione zostały w tabeli 5.2. Jak można zauważyć niektóre z relacji są bardzo częste (jak na przykład *Krzyżowanie się*), inne zaś bardzo rzadkie (np. *Zmiana poglądu*). Przewaga liczności jednych relacji nad innymi stwarza pewne problemy w automatycznym rozpoznawaniu relacji semantycznych. Niezrównoważenie danych, powoduje, że relacje mało liczne, mogą być słabo (lub mogą nie być wcale) rozpoznawane w stosunku do tych wielolicznych. Z tego też względu proces znakowania danych posiadał swoją kontynuację w postaci *drugiego etapu znakowania*, w którym, podobnie jak w *etapie pierwszym*, znakowane były zdania ze zbioru *WNews*.

5.2.2. Drugi etap znakowania danych

Ręczny proces anotacji (nawet obrębie automatycznie dobranych paczek dokumentów) był bardzo czasochłonny. Dlatego zdecydowano, aby w dalszym procesie znakowania wesprzeć osoby znakujące, mechanizmem wyszukującym w paczkach dokumentów konkretne pary zdań, między którymi może zachodzić jakaś relacja semantyczna. Praca anotatora ograniczała się jedynie do stwierdzenia, czy między daną parą zdań zachodzi relacja semantyczna i jeżeli tak, to anotator wskazywał jaka relacja zachodzi. Jeżeli między daną parą zdań nie zachodziła żadna z relacji, taka informacja również była odnotowywana.

Drugi etap znakowania relacji między zdaniami bazował na podobieństwie badanych zdań do siebie. Wykorzystano paczki dokumentów, które nie były użyte w pierwszym etapie znakowania. Każda paczka zawierała dokument $\mathcal{D}_{i0}$, do którego w *etapie 1* wyszukiwane były dwa najbardziej podobne dokumenty: $\mathcal{D}_{i1}$ oraz $\mathcal{D}_{i2}$. Z nich wyszukane zostały konkretne pary zdań, które przedstawiane były osobie znakującej. Jej zadaniem było podjęcie decyzji czy pomiędzy przedstawioną parą zdań zachodzi relacja semantyczna, czy też nie. Jeżeli zachodzi, miała ona wskazać typ relacji. W przypadku, gdy

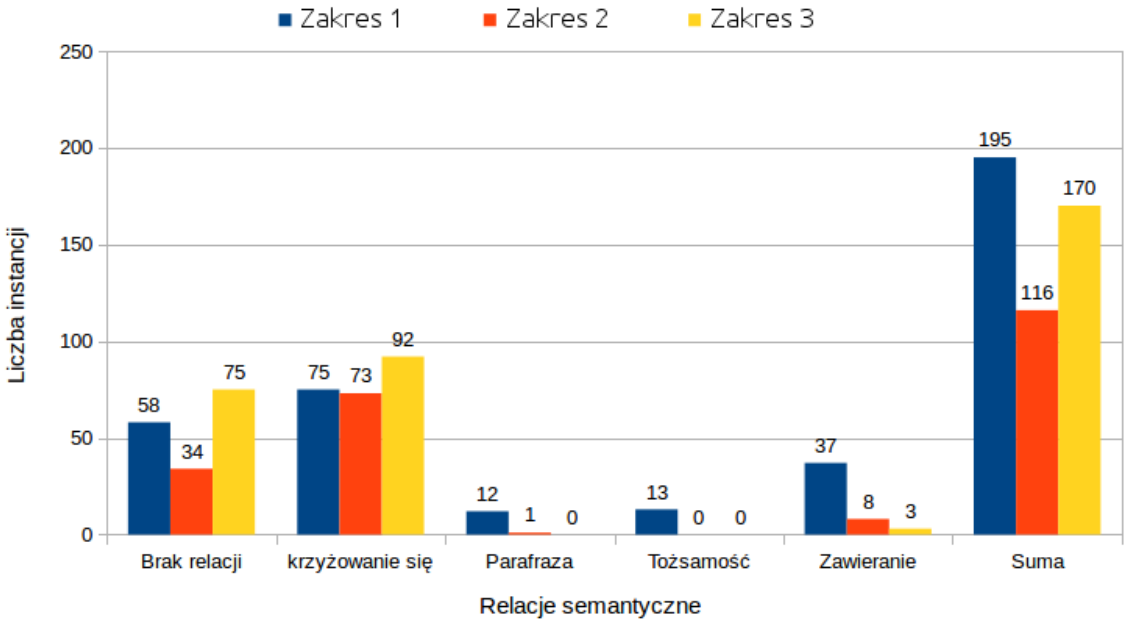
Tabela 5.2. Liczności poszczególnych relacji, po *pierwszym etapie* znakowania

Relacja	**Liczność**
Mowa zależna	26
Streszczanie	54
Spełnienie	160
Uszczegółowienie	119
Krzyżowanie się	1299
Zawieranie	144
Zmiana poglądu	6
Tożsamość	47
Opis	372
Źródło	46
Cytowanie	3
Sprzeczność	17
Parafraza	34
Modalność	1
Tło historyczne	311
Dalsze informacje	306

relacja nie zachodziła, taka informacja również była anotowana. Uproszczeniem w stosunku do podejścia z *pierwszego etapu* znakowania, było przedstawienie już konkretnych par zdań z paczki (pary te były dobierane zgodnie z zasadą: dokument $\mathcal{D}_{i0} \rightarrow \mathcal{D}_{i1}$, $\mathcal{D}_{i0} \rightarrow \mathcal{D}_{i2}$ oraz wewnątrz $\mathcal{D}_{i0}$). Do wyszukiwania potencjalnych zdań, do których przypisywano relację semantyczną wykorzystano podobieństwo grafów. Dla każdego ze zdań utworzono graf, który reprezentował zależności składniowe oraz kolejność występowania słów w tym zdaniu. Węzeł grafu reprezentował formę podstawową słowa połączoną z jego częścią mowy. Jako podobieństwo, wykorzystana została odległość edycyjna grafów (ang. *Graph edit distance*) – s_{GED} (wzór 3.5). Dla każdego zdania określone zostało podobieństwo do każdego innego z rozpatrywanej paczki dokumentów. Dla pary zdań $(\mathcal{S}_1, \mathcal{S}_2)$ określona została wartość podobieństwa grafów zbudowanych dla $\mathcal{S}_1$ oraz $\mathcal{S}_2$. Pary zdań, uszeregowane zostały malejąco względem wartości s_{GED} oraz ograniczone zostały tylko do tych, które wykazywały podobieństwo wyższe od eksperymentalnie przyjętej wartości równej $0,29$.

Ze względu na dużą liczbę par zdań podzielono je na trzy zbiory uwzględniając trzy przedziały wartości podobieństwa. Granice tych przedziałów dobrane zostały eksperymentalnie na podstawie obserwacji rozkładu wartości podobieństwa w całym zbiorze par zdań, z uwzględnieniem liczności relacji w poszczególnych przedziałach podobieństwa.

W pierwszej kolejności anotowano te pary, które wykazywały wysoką wartość podobieństwa *zakres1* (podobieństwo w przedziale $\langle 0,64; 1\rangle$). Liczba par w tym przedziale podobieństwa wynosiła 195. Następnie anotowano te pary zdań, między którymi wartość podobieństwa była „średnia" *zakres2* – z przedziału $\langle 0,48; 0,497\rangle$. Takich par było 105. W ostatnim kroku, anotacji poddano takie pary zdań, które wykazywały stosunkowo niską wartość podobieństwa *zakres3* – z przedziału $\langle 0,2903; 0,2909\rangle$. Tych było 169. Po zakończonym procesie anotacji otrzymano 481 nowych instancji relacji.



Rys. 5.1. Zestawienie liczności relacji semantycznych między fragmentami tekstów z drugiego etapu znakowania z podziałem odnoszącym się do trzech wyróżnionych zakresów podobieństwa

W tabeli 5.3 przedstawiony został procentowy udział relacji w każdym z przyjętych przedziałów podobieństwa. Dla podobieństwa z *Zakresu 1* najliczniejszymi relacjami

Tabela 5.3. Procentowe zestawienie relacji w poszczególnych przedziałach podobieństwa dla anotacji relacji semantycznych między fragmentami tekstów

Relacja	**Zakres 1** [%]	**Zakres 2** [%]	**Zakres 3** [%]	**Sumarycznie** [%]
Brak relacji	29,74%	29,31%	44,12%	34,72%
Krzyżowanie się	38,46%	62,93%	54,12%	49,9%
Parafraza	6,15%	0,86%	0%	2,7%
Tożsamość	6,67%	0%	0%	2,7%
Zawieranie	18,97	6,9%	1,76%	9,98%

okazały się *Krzyżowanie się* oraz *Brak relacji*. *Zakres 2* również wykazuje podobną tendencję, jednak istnieje tutaj znacząca przewaga liczby relacji *Krzyżowania się* nad pozostałymi relacjami. Warto również zauważyć, że w *zakresie 2* podobieństwa nie wystąpiła ani jedna instancja relacji *Tożsamości* – czyli sytuacji, gdy podobieństwo dwóch identycznych tekstów musi być równe 1. Z kolei dla *Zakresu 3* zarówno relacja *Tożsamości*, jak i *Parafrazy* nie wystąpiły ani razu. Relacja *Krzyżowanie* stanowi ponad połowę wszystkich przykładów poddanych analizie, a większość z pozostałych *44% całości* jest *Brakiem relacji*. Sumarycznie dla wszystkich trzech przedziałów, najczęściej pojawiały się relacje *Krzyżowanie się* oraz *Brak relacji*. Relacja *Zawieranie* stanowiła ~10% wszystkich przykładów, a relacje *Tożsamość* oraz *Parafraza* były małym odsetkiem całego zbioru.

Na rysunku 5.1 przedstawiono liczności zaanotowanych relacji, dla wszystkich zakresów podobieństwa. Można zauważyć, że *Brak relacji* często występuje przy wysokiej oraz niskiej wartości podobieństwa. Może to świadczyć o tym, że zdania podobne do siebie, różnią się jednym, znaczącym słowem, które zmienia całkowicie semantykę zdania np.: *Dzisiaj pada deszcz — Dzisiaj nie pada deszcz*, słowo *nie*, całkowicie zmienia obraz sytuacji. Przy niskim poziomie podobieństwa, zdania różnią się od siebie na tyle, że mogą dotyczyć innej sytuacji np. *Wszystkie domy posiadają okna i drzwi — Karol wszedł przez okno.*

Po *drugim etapie* znakowania, liczności instancji poszczególnych relacji wzrosły. Są one przedstawione w tabeli 5.4. W przyjętych arbitralnie przedziałach podobieństwa znalazły się pary zdań, pomiędzy którymi zachodziła jedna z relacji CST, a rozkład tych relacji zgodny był z rozkładem relacji przypisywanych ręcznie. Zbiór danych powstały w ramach *drugiego* etapu znakowania połączony został ze zbiorem z *pierwszego* etapu i wykorzystany w badaniach eksperymentalnych.

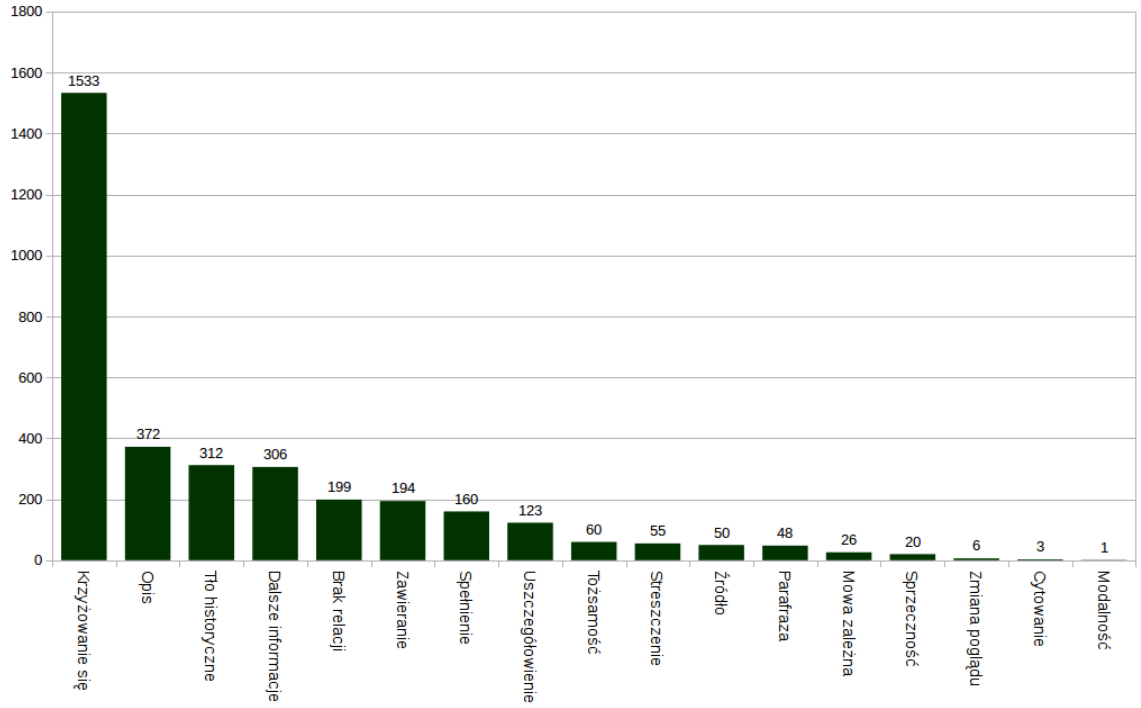
Tabela 5.4. Liczności poszczególnych relacji, po *drugim etapie* znakowania

Relacja	**Liczność**
Mowa zależna	26
Streszczanie	**55**
Spełnienie	160
Uszczegółowienie	**123**
Krzyżowanie się	**1533**
Zawieranie	**194**
Zmiana poglądu	6
Tożsamość	**60**
Opis	372
Źródło	**50**
Cytowanie	3
Sprzeczność	**20**
Parafraza	**48**
Modalność	1
Tło historyczne	**312**
Dalsze informacje	306

5.3. Zestawy cech użyte w procesie klasyfikacji

Dane opisane w poprzednim podrozdziale wykorzystane zostały do utworzenia zbiorów treningowo-testowych do klasyfikacji relacji między fragmentami tekstów. W badaniach uwzględniono pięć różnych zestawów cech (*Zestaw 1* - *Zestaw 5*), które posłużyły do zbudowania pięciu różnych wektorów cech dla klasyfikatora:

1. *Zestaw 1* – wektory wejściowe zbudowane przy wykorzystaniu cech literaturowych,
2. *Zestaw 2* – wektory wejściowe, w których wykorzystano pierwszy zestaw cech grafowych, nazwany `cechy grafowe (1)`,



Rys. 5.2. Liczności poszczególnych relacji CST dla zbioru łącznego powstałego po *pierwszym* i *drugim* etapie oznaczania

3. *Zestaw 3* – wektory wejściowe budowane na bazie połączenia *cech literaturowych* z `cechami grafowymi (1)`,
4. *Zestaw 4* – wektory wejściowe bazujace na drugim zestawie cech grafowych nazwanych `cechy grafowe (2)`,
5. *Zestaw 5* – wektory cech powstałe z połączenia *cech literaturowych* z `cechami grafowymi (2)`.

5.3.1. Zestaw 1 – cechy literaturowe

W tym przypadku do utworzenia wektorów cech posłużyły cechy opisane w (Maziero i inni, 2014). Dotyczą one przede wszystkim postaci gramatycznej zdań wchodzących w skład rozpoznawanej relacji. Wyróżniono 9 cech, które przedstawiono poniżej posługując się następującym schematem opisu: Nazwa cechy *(nazwa angielska)* – krótki opis cechy.

— Wspólne lematy (ang. *common lemmas*) – liczba wspólnych podstawowych form słów dla pary fragmentów tekstów;
— Wspólne nazwy własne (ang. *proper names*) – liczba wspólnych nazw własnych w obu fragmentach tekstu. Cecha wymaga, aby analizowane fragmenty tekstów posiadały przypisaną informację o wykrytych nazwach własnych;
— Najdłuższy wspólny podciąg (ang. *longest common substring*) – długość najdłuższego wspólnego podciągu (leżącego koło siebie), który zawiera się w obu fragmentach tekstów;

— Najdłuższa wspólna podsekwencja (ang. *longest common subsequence*) – długość najdłuższej wspólnej podsekwencji, która zawiera się w obu fragmentach;
— Podobieństwo kosinusowe (ang. *cosine similarity*) – cecha określa podobieństwo wektorów, budowanych na bazie modelu worka słów. Wszystkie słowa, z których tworzony jest worek słów, sprowadzane są do form podstawowych. Wykorzystując worek słów dla pierwszego i drugiego fragmentu tekstu tworzone są wektory, które reprezentują liczność słów a następnie obliczany jest kosinus kąta między tymi wektorami, zgodnie z równaniem 2.4;
— Dłuższe zdanie (ang. *longer sentence*) – cecha przyjmuje wartość 1, jeżeli fragment pierwszy tekstu jest dłuższy od drugiego, wartość -1 jeżeli fragment pierwszy jest krótszy od drugiego oraz wartość 0, gdy oba fragmenty są tej samej długości;
— Część wspólna zbiorów znaczeń (ang. *synsets intersection*) – część wspólna zbiorów zawierających formy bazowe słów występujących w każdym z fragmentów. Słowa można zastąpić formami synonimicznymi (synsetami) ze Słowosieci. Miara jest znormalizowana długością krótszego z fragmentów;
— Liczności części mowy (ang. *POS amounts*) – cecha określa liczbę wystąpień poszczególnych części mowy w każdym z fragmentów. Tworzy z nich wektory, pomiędzy którymi obliczana jest miara kosinusowa;
— podmiot-orzeczenie-dopełnienie (ang. *SVO*) – podobieństwo Jaccarda (wzór 3.14) wyznaczane na podstawie zbiorów trójek (podmiot, orzeczenie, dopełnienie).

Pełny wektor cech przyjmuje postać przedstawioną w dodatku A.1. Przedstawione powyżej cechy wykorzystano jako bazowy zestaw cech w procesie klasyfikacji. Względem tego zestawu, porównywana była jakość klasyfikacji przy zastosowaniu wektorów cech budowanych dzięki użyciu metody PAS.

5.3.2. Zestaw 2 – cechy grafowe (1)

Wektory cech w tym wypadku budowane są przy wykorzystaniu reprezentacji grafowej uzyskanej za pomocą metody PAS. Określają one podobieństwo między dwoma grafami, zbudowanymi na podstawie fragmentów tekstów, pomiędzy którymi badane jest zachodzenie danej relacji. Zestaw krawędzi E w budowanych grafach jest wynikiem zastosowania wszystkich dostępnych funkcji transformujących f_{et} dostępnych w metodzie PAS. Przez zastosowanie różnych funkcji transformujących węzłów f_{nt} zastosowanych do danego zdania, otrzymuje się 16 różnych reprezentacji danego zdania – 16 grafów (możliwe jest dołączanie wiedzy pozatekstowej ze Słowosieci i/lub SUMO), które są dalej wykorzystane do budowy zestawu cech. Grafy te oznaczono następująco:

1. *Lemma lower* – graf budowany jest na bazie podstawowej formy słów (sprowadzonej do małych liter) występujących w tekście.
2. *Lemma lower + sumo* – graf budowany jak *Lemma lower*, jednak z dołączonym do niego podgrafem wyciętym z SUMO. Wycięcie podgrafu SUMO poprzedzone jest odniesieniem słowa do pojęcia za pomocą rzutowania Słowosieć $\rightarrow$ SUMO. Do słów przypisane są ich znaczenia, następnie dla znaczeń przypisanych do słów wybierane są pojęcia zrzutowane na to znaczenie, które są przypisywane do słowa. Z SUMO wycinany jest spójny podgraf zawierający wszystkie pojęcia zrzutowane na tekst.

3. *Lemma lower + plwn* – graf budowany jak *Lemma lower* z dołączonym podgrafem wyciętym ze Słowosieci zawierającym wszystkie znaczenia, przypisane do słów w tekście.
4. *Lemma lower + plwn + sumo* – graf budowany jak *Lemma lower*, jednak posiada dołączone podgrafy wycięte ze Słowosieci oraz z SUMO.
5. *Lemma pos lower* – graf budowany jest podstawie formy podstawowej słowa (sprowadzonej do małych liter) połączonej z informacją o części mowy tego słowa.
6. *Lemma pos lower + sumo* – graf budowany jak w przypadku *Lemma pos lower* z dołączonym podgrafem wyciętym z SUMO (analogicznie do typu grafu *Lemma lower + sumo*).
7. *Lemma pos lower + plwn* – graf budowany jak w przypadku *Lemma pos lower* z dołączonym podgrafem wyciętym ze Słowosieci (analogicznie do typu grafu *Lemma lower + plwn*).
8. *Lemma pos lower + plwn + sumo* – graf budowany jak *Lemma pos lower*, posiada jednak dołączone podgrafy wycięte z SUMO oraz ze Słowosieci.
9. *Concept* – graf zbudowany z pojęć SUMO, słowa z tekstu rzutowane są na pojęcia SUMO, po czym zestaw krawędzi E między słowami, przenoszony jest na odpowiadające im pojęcia ontologii SUMO.
10. *Concept + sumo* – podobnie jak w *Concept*, jednak do tak zbudowanego grafu, dołączany jest spójny podgraf SUMO zawierający wszystkie pojęcia z tekstu.
11. *Concept + plwn* – podobnie jak w *Concept*, jednak do tak zbudowanego grafu, dołączany jest spójny podgraf Słowosieci. Wycięcie podgrafu Słowosieci możliwe jest dzięki rzutowaniu synsetów Słowosieci na pojęcia SUMO.
12. *Concept + plwn + sumo* – graf budowany jest podobnie jak w przypadku *Concept*, dołączane są podgrafy SUMO oraz Słowosieci.
13. *Synset* – graf zbudowany ze znaczeń Słowosieci, słowa z tekstu posiadają przypisane znaczenia, po czym zestaw krawędzi E między słowami przenoszony jest na odpowiadające im znaczenia Słowosieci.
14. *Synset + sumo* – budowanie grafu zgodne z *Synset*, jednak dołączony jest podgraf SUMO, budowany dla pojęć przypisanych do znaczeń.
15. *Synset + plwn* – graf zbudowany w sposób zgodny z *Synset* wraz z przypisanym podgrafem wyciętym ze Słowosieci, zawierającym wszystkie znaczenia z analizowanego tesktu.
16. *Synset + plwn + sumo* – graf zbudowany jest jak *Synset* z dołączonymi podgrafami wyciętymi z SUMO oraz ze Słowosieci.

Podsumowując, każde zdanie ze wszystkich par zdań, dla których zachodzi relacja, posiadało odwzorowanie za pomocą 16 grafów różniących się typem węzła (różne funkcje f_{nt} dla każdego grafu) z dołączaną (lub nie) wiedzą pozatekstową, z pełnym zestawem krawędzi (dla każdego grafu zastosowano wszystkie rodzaje funkcji f_{et}).

Dla każdej z 16 par grafów reprezentujących dwa zdania, pomiędzy którymi zachodzi dana relacja semantyczna, określono podobieństwo za pomocą miar podobieństwa grafów, szczegółowo opisanych w rozdziale 3.2.1. Miary wykorzystane do pomiaru podobieństwa, to: d_{MCS}, d_{WGU}, d_{UGU} , d_{MMCS}, d_{MMCSN}, s_{GED}, $s_{JACCARD}$ oraz s_{CTXBow}. Wektor wejściowy w *Zestawie 2* zawiera 128 cech (16 różnych reprezentacji grafowych x 8 miar podobieństwa). Zbiór tych cech został nazwany `cechy grafowe (1)`. Wzorce uczące to para $\langle wektorcech, C_k \rangle$, gdzie C_k to etykieta klasy relacji semantycznej. Zbiór

uczący dla klasyfikatora zawiera zarówno pozytywne jak i negatywne przykłady uczące. Liczba pozytywnych przykładów uczących, równa jest liczbie ręcznie zaanotowanych relacji korpusie uczącym. Przykładowe cechy z wektora cech zostały umieszczone w dodatku A.2.

5.3.3. Zestaw 3 – połączenie `cech grafowych (1)` z *cechami literaturowymi*

Wektor cech w *zestawie 3* tworzony jest z połączonych cech literaturowych *Zestaw 1* oraz `cech grafowych (1)`. Motywacją do utworzenia takiego wektora cech była chęć sprawdzenia czy uzupełnienie 128 cech z *Zestawu 2* o 9 cech z *Zestawu 1* poprawi jakość klasyfikacji, i jeżeli tak, to czy ta poprawa jest istotna statystycznie.

5.3.4. Zestaw 4 – cechy grafowe (2)

Zestaw `cechy grafowe (2)` jest rozszerzeniem zestawu cech opisanego w rozdziale 5.3.2. Zestaw ten opiera się na modelowaniu pary zdań, pomiędzy którymi zachodzi relacja, za pomocą 16 różnych reprezentacji grafowych z dołączaniem (lub nie) wiedzy pozatekstowej. W przeciwieństwie do `cechy grafowe (1)`, w każdym grafie zastosowano inny zestaw funkcji f_{nt} transformacji słów wchodzących w relacje do węzłów.

W metodzie *PAS* zdefiniowano 6 funkcji transformujących f_{et} (mogą one utworzyć 6 typów krawędzi w grafie):

- `dep_rel` – modeluje relacje zależności składniowych między słowami, wykorzystuje pogłębiony parser opisany w rozdziale 4.2.2;
- `h2h` – odwzorowuje występowanie po sobie głów składniowych oznaczonych za pomocą parsera IOBBER; w grafie dodawana jest krawędź między tymi węzłami, które dotyczą głów składniowych fraz w analizowanych zdaniach; kolejność występowania głów odzwierciedlana jest za pomocą skierowanej relacji;
- `malt_rel` – wykorzystuje relacje zależności składniowych nadawane przez MaltParser (Wróblewska, 2014) i przenosi je na krawędzie; każda relacja przypisana do tekstu między dwoma słowami, staje się krawędzią w grafie z nazwą zgodną z nazwą relacji nadaną przez MaltParser[2] między węzłami, które reprezentują słowa w zdaniu, dla których ta relacja zachodzi; kierunek krawędzi zgodny jest z kierunkiem relacji przypisanej do tekstu;
- `ne2ne` – relacja modeluje kolejność występowania po sobie nazw własnych; dodaje krawędź pomiędzy węzłami dotyczącymi nazw własnych w kierunku zgodnym z kolejnością występowania tych nazw w tekście;
- `sem_role` – wykorzystując płytki parser semantyczny, opisany w rozdziale 4.2.2 dodawane są krawędzie pomiędzy tymi węzłami, które reprezentują słowa z tekstu, między którymi wystąpiła rola semantyczna; nazwa krawędzi oraz jej kierunek jest zgodny z nazwą oraz kierunkiem roli semantycznej;
- `w2w` – dodaje krawędzie pomiędzy węzłami, które reprezentują słowa z tekstu występujące po sobie; jest to najprostsza relacja, lecz dzięki niej możliwe jest odtworzenie szyku zdania;

[2] Wykaz relacji nadawanych przez MaltParser znajduje się na stronie `http://zil.ipipan.waw.pl/PDB/DepRelTypes`

Dla każdej wybranej funkcji f_{nt} transformującej słowo do węzła, generowane są wszystkie możliwe kombinacje zestawów krawędzi, czyli dla 6 typów krawędzi, możliwe jest wygenerowanie 63 różnych zestawów krawędzi, jest to suma wszystkich kombinacji bez powtórzeń ($C_n^k = \frac{n!}{k!(n-k)!}$) wartości $1-6$ z 6-ścio elementowego zbioru: $\sum_{n=6,k=1}^{k=6} \frac{n!}{k!(n-k)!}$. Każda możliwa kombinacja krawędzi dodawana jest do każdego typu węzła i w kolejnych grafach łączona z wiedzą pozatekstową ze Słowosieci, ontologii SUMO bądź obu.

Przykładowe kombinacje krawędzi, nazwane również zestawami krawędzi, to:
- `w2w` – w grafie znajdują się tylko krawędzie oznaczające kolejność występowania słów po sobie;
- `malt_rel-sem_role-ne2ne` – graf zawiera krawędzie odnoszące się do relacji składniowych nadawanych przez *MaltParser* (`malt`), krawędzie odzwierciedlające role semantyczne z tekstu (`sem_role`) oraz krawędzie oznaczające kolejność występowania po sobie nazw własnych (`ne2ne`);
- `dep_rel-h2h-malt_rel-sem_role-ne2ne` – graf składa się z krawędzi odnoszących się do zależności składniowych (`dep_rel`), kolejności występowania po sobie głów składniowych (`h2h`), relacji składniowych *Maltowych* (`malt`), ról semantycznych wykrytych w tekście (`sem_role`) oraz kolejności występowania po sobie nazw własnych (`ne2ne`).

Każdy ze wspomnianych 63 zestawów krawędzi w połączeniu z 16 różnymi typami węzłów opisanych w rozdziale 5.3.2 daje możliwość reprezentacji tego samego zdania za pomocą 1008 różnych grafów. Każda para zdań wchodząca w relację, transformowana jest do dwóch grafów o takim samym zestawie krawędzi oraz typie węzłów. Dla każdego zdania z badanej pary zdań otrzymujemy 1008 różnych reprezentacji grafowych. Wektor cech dla pary zdań z relacji obliczany jest na podstawie obliczania podobieństwa między odpowiadającymi sobie reprezentacjami grafowymi przy wykorzystaniu każdej z dostępnych 8 miar podobieństwa. Wynikiem tego procesu jest wektor cech o wymiarze 8064 ($8 \cdot 1008 = 8064$). Taki wektor jest wektorem wejściowym dla klasyfikatora w procesie ucznia i rozpoznawania relacji semantycznych między dwoma zdaniami. Liczba wektorów uczących równa jest liczbie ręcznie zaanotowanych przykładów w korpusie.

Wybrany fragment wektora uczącego z przykładowymi cechami z zestawu `grafowych cech (2)` przedstawiony został w dodatku A.3. Pojedyncza cecha wektora, określa wartość podobieństwa obliczonego między parą grafów zbudowanych dla zdań, między którymi zachodzi relacja.

5.3.5. Zestaw 5 – połączenie `cech grafowych (2)` z *cechami literaturowymi*

Wektory cech tworzone w *zestawie 5* uwzględniają cechy literaturowe (szczegółowo opisane w rozdziale 5.3.1) połączone z `cechami grafowymi (2)` (podrozdział 5.3.4). Utworzenie takiego wektora cech miało na celu sprawdzeniem czy rozszerzenie 8064 cech z *Zestawu 4* o cechy z *Zestawu 1* (9 cech), poprawia jakość klasyfikacji, i jeżeli tak, to czy ta poprawa jest istotna statystycznie.

5.3.6. Założenia odnośnie budowanych grafów

Podczas budowania grafów, do obliczania cech w *Zestawie 2*, *Zestawie 3 Zestawie 4* oraz *Zestawie 5* przyjęto, że:

— węzły jednej wieloelementowej nazwy własnej (np. *Zakład Ubezpieczeń Społecznych*) łączone były do jednego węzła, zamiast rozbijania na niezależne węzły (tzn. {Zakład, Ubezpieczenie, Społeczny});

— wyraz *się* dołączany był do odpowiedniego czasownika, zamiast wydzielania osobnego węzła w grafie (np. *Robić się* (np. na bóstwo) zamiast {Robić, się});

— dla znaków interpunkcyjnych nie były tworzone węzły (na przykład dla zdania: *Mówi to ten, który się nie zna!* utworzone zostały węzły: {Mówi, to, ten, który, się, nie, zna});

— jako metoda łączenia wiedzy tekstowej z pozatekstową, zarówno do grafu wycinanego ze Słowosieci jak i SUMO, wykorzystany został algorytm `MappinhMerger`.

5.4. Wybór klasyfikatorów

Po opracowaniu cech, które będą tworzyć wektor cech dla klasyfikatora, rozpoznanie relacji semantycznych sprowadza się do użycia (a wcześniej wyboru) jednej z wielu dostępnych metod klasyfikacji. Przygotowane wektory cech podawane są na wejście klasyfikatora jako wektory wejściowe. Po wykonaniu przeglądu literaturowego i dostępnych implementacji, w eksperymentach użyto następujących czterech klasyfikatorów:

Naiwny klasyfikator Bayesa (Rish, 2001), dalej oznaczany*NB*. Jest to prosty klasyfikator probabilistyczny, oparty na twierdzeniu Bayesa. Zakłada wzajemną niezależność zmiennych wejściowych. *NB* zwraca dla każdej klasy prawdopodobieństwo jej zajścia przy zadanym wektorze wejściowym. Jego domyślne ustawienia parametrów w środowisku *Weka* pokazane są na rysunku C.1 w dodatku C.1. We wszystkich przeprowadzonych eksperymentach, *NB* posiadał takie same ustawienia.

Maszyna wektorów nośnych (Steinwart i Christmann, 2008), oznaczana dalej *SVM*. U podstaw tej metody leży koncepcja przestrzeni decyzyjnej, której granice dzielą wzorce z dwóch różnych klas (klasyfikacja binarna). Klasyfikacja wieloklasowa dla klasyfikatora *SVM* odbyła się z zastosowaniem podejścia *one-vs-all*. Oznacza to, że dla każdej rozpoznawanej relacji budowany był klasyfikator binarny, który jako pozytywne przykłady uczące wykorzystywał wektory zbudowane dla tej relacji, zaś jako przykłady negatywne, wektory zbudowane dla pozostałych relacji. Dla 17 relacji zbudowanych zostało 17 klasyfikatorów binarnych rozpoznających pojedyncze relacje. Ostateczna klasyfikacja traktowana jest jako odpowiedź całej rodziny klasyfikatorów SVM. Rozpoznanie jaka relacja zachodzi między dwoma wektorami reprezentującymi zdania odbywa się przez wybór najwyższej wartości funkcji decyzyjnej spośród wszystkich zwracanych wartości wyjściowych SVM. Domyślne wartości parametrów przedstawione są na rysunku C.2 w dodatku C.2.

Drzewo decyzyjne *J48*, uczone algorytmem C4.5 (Quinlan, 1993). Węzły w drzewie decyzyjnym odpowiadają atrybutom (cechom z wektora cech), liście zaś są klasami. Budowanie drzewa decyzyjnego metodą *C4.5* polega na wybieraniu najbardziej informatywnych węzłów (atrybutów) i dołączanie ich do drzewa jako pierwszych. Kolejno dołączane są węzły z coraz niższą wartością „informacyjności". Po utworzeniu

modelu można łatwo prześledzić, które wartości cech miały największy wpływ na klasyfikację. Na podstawie modelu można również łatwo wygenerować reguły decyzyjne opisujące sposób klasyfikacji. Domyślne ustawienia parametrów dla klasyfikatora w systemie *Weka*, przedstawione zostały w dodatku C.3 na rysunku C.3.

Drzewo modeli logistycznych (Landwehr i inni, 2005, Sumner i inni, 2005), oznaczane dalej *LMT*. Jest ono połączeniem drzewa decyzyjnego z regresją logistyczną. Liście w drzewie zawierają fragmentaryczne modele regresji liniowych. Jak sugerują autorzy tego modelu drzewo uczone algorytmem C4.5 może się przeuczać pewnych wzorców, dlatego wydaje się korzystne użycie modelu logistycznego, zamiast procedury wzrostu drzewa. Ustawienia domyślne parametrów zaproponowane w systemie *Weka*, przedstawiono na rysunku C.4 w dodatku C.4.

Wymienione klasyfikatory, poza *LMT*, były już używane w problemie rozpoznawania relacji semantycznych między fragmentami tekstów, dlatego podjęto decyzję o ich wykorzystaniu w badaniach eksperymentalnych wykonanych w niniejszej pracy. Klasyfikator *LMT* został dodany do zbioru używanych klasyfikatorów, ponieważ w przypadku n-arnej klasyfikacji posiada duże możliwości dostosowania zestawu cech do konkretnej klasy. *LMT*, podobnie jak *J48*, daje możliwość interpretacji podejmowanych decyzji dla poszczególnych klas, co jest niewątpliwą zaletą i umożliwia osobie przeprowadzającej badania podgląd kroków, jakie wykonywał klasyfikator przy podejmowaniu decyzji.

Aby proces przeprowadzonych badań były wiarygodny i porównywane wyniki nie zależały od zmieniających się ustawień klasyfikatora, we wszystkich eksperymentach przyjęto takie same wartości parametrów wszystkich klasyfikatorów, niezależnie od liczby cech w wektorze wejściowym. Są to ustawienia domyślne dostępne w systemie *Weka*.

5.5. Środowisko testowe

Opisane w rozdziale 6 badania, dotyczące budowy grafów oraz określania podobieństwa między nimi, wykonane zostały przy użyciu autorskiego środowisku testowego, które dla dowolnego zdania/fragmentu tekstu umożliwia budowę grafu o dowolnym typie węzłów i krawędzi. Środowisko to umożliwia obliczenie podobieństwa między dwoma zbudowanymi wcześniej grafami i jest dostępne na darmowej licencji w repozytorium *dSpace* pod adresem `https://clarin-pl.eu/dspace/handle/11321/280`. Proces uczenia i testowania klasyfikatorów, zaimplementowano w języku *Java* z wykorzystaniem klas narzędzia *Weka*.

Rozdział 6

Badania

Niniejszy rozdział zawiera opis oraz wyniki eksperymentów mających na celu ocenę skuteczności rozpoznawania relacji semantycznych między fragmentami tekstów przy użyciu metody pogłębianej analizy semantycznej *PAS*. Przypomnijmy, że ideą opracowanej metody było zbliżenie jej działania do parsera głębokiego, dlatego oceniając otrzymane rezultaty najlepiej byłoby odnieść się do efektów otrzymywanych przy użyciu parsera głębokiego. Takie porównanie byłoby jednak niemiarodajne, ponieważ opracowana metoda umożliwia przypisanie reprezentacji informacji semantycznej dla dowolnego zdania w języku polskim, natomiast istniejący parser ENIAM (Jaworski i Kozakoszczak, 2016) potrafi analizować stosunkowo niewielki podzbiór zdań języka polskiego. Dlatego ocena jakości rozpoznawania relacji semantycznych z wykorzystaniem metody *PAS*, została wykonana w stosunku do podejścia bazującego na cechach zaproponowanych w literaturze, opisanych w rozdziale 5.3.1.

6.1. Procedura badawcza

Procedura badawcza obejmowała wykonanie następujących kroków:

1. Utworzenie zbiorów uczących dla badanych zestawów cech. W pierwszej kolejności dla par zdań budowano grafy oraz dołączono informacje wydobyte spoza tekstu. Wykorzystując zbudowane grafy obliczano wartości miar podobieństwa tych grafów, które to wartości są uwzględniane w wektorze cech.
2. Przeprowadzenie eksperymentów, w których obliczane były miary jakości klasyfikacji poszczególnych klasyfikatorów dla każdego zestawu cech.
3. Ocena skuteczności klasyfikacji.
4. Analiza statystyczna uzyskanych wyników.

W eksperymentach wykorzystano cztery różne klasyfikatory. Ich lista i motywacje stojące za ich wyborem opisane zostały w poprzednim rozdziale. Jakość klasyfikacji mierzona była przy wykorzystaniu *miary-f*. Ogólna postać miary przedstawiona jest za pomocą wzoru 6.1.

$$F_\beta = (1 + \beta^2) \cdot \frac{precyzja \cdot czulosc}{\beta^2 \cdot precyzja + czulosc} \tag{6.1}$$

Tabela 6.1. Macierz pomyłek w binarnej klasyfikacji

	$\widetilde{\mathcal{C}_1}$	$\widetilde{\mathcal{C}_2}$
$\mathcal{C}_1$	TP	FP
$\mathcal{C}_2$	FN	TN

Parametr β we wzorze 6.1, umożliwia określenie wagi nadanej dla precyzji w stosunku do czułości. Najczęściej β ustawiana jest na wartość równą 1. Otrzymujemy wówczas miarę *F1*, przedstawioną wzorem 6.2.

$$F1 = 2 \cdot \frac{precyzja \cdot czulosc}{precyzja + czulosc} \tag{6.2}$$

Będziemy ją dalej oznaczać *miara-f* . Poszczególne składniki wzoru miary *miara-f*, to:

$$precyzja = \frac{TP}{TP + FP} \tag{6.3}$$

$$czulosc = \frac{TP}{TP + FN} \tag{6.4}$$

Czułość wskazuje na ile klasyfikator jest w stanie rozpoznać obiekty z danej klasy. *Precyzja*, to dokładność klasyfikacji w obrębie rozpoznanej klasy. Oba te parametry zdefiniowane są przy użyciu macierzy pomyłek (ang. *Confusion matrix*). W tabeli 6.1 przedstawiona została macierz pomyłek dla klasyfikacji binarnej do dwóch klas $\mathcal{C}_1$ oraz $\mathcal{C}_2$. Przyjmijmy, że klasa $\mathcal{C}_1$ będzie klasą *pozytywną*, a klasa $\mathcal{C}_2$ *negatywną*, W tabeli oryginalne klasy oznaczono jako $\mathcal{C}_1$, $\mathcal{C}_2$, zaś jako $\widetilde{\mathcal{C}_1}$ i $\widetilde{\mathcal{C}_2}$ oznaczono klasy predykowane. Tabela zawiera następujące miary:

— **TP** (ang. *true positive*) liczba obiektów klasy $\mathcal{C}_1$ poprawnie zaklasyfikowanych jako $\widetilde{\mathcal{C}_1}$;
— **TN** (ang. *true negative*), czyli liczba obiektów klasy $\mathcal{C}_2$ poprawnie zaklasyfikowanych jako klasa $\widetilde{\mathcal{C}_2}$;
— **FP** (ang. *false positive*), to liczba obiektów klasy $\mathcal{C}_1$ zaklasyfikowanych jako $\widetilde{\mathcal{C}_2}$;
— **FN** (ang. *false negative*), to liczba obiektów klasy $\mathcal{C}_2$, które zaklasyfikowane zostały jako klasa $\widetilde{\mathcal{C}_1}$;

W przypadku klasyfikacji n-arnej, macierz pomyłek przyjmuje wielkość $N \times N$, gdzie N oznacza liczbę klas a na przekątnej wpisane są wartości **TP**.

Jako metoda walidacji jakości klasyfikacji, wybrana została dziesięciokrotna walidacja krzyżowa (ang. *10-fold Cross-Validation*). Wyniki z poszczególnych foldów (części zbioru walidacyjnego) dziesięciokrotnej walidacji krzyżowej, wykorzystano także w procesie obliczania istotności statystycznej poprawy jakości uzyskanych wyników.

Dla miary s_{GED} przedstawionej w rozdziale 3.2.2, określającej podobieństwo dwóch grafów za pomocą sumy kosztów operacji atomowych przekształcających $\mathbf{G}_1$ w $\mathbf{G}_2$, przyjęto arbitralnie następujące wartości kosztu c operacji:

1. Koszt operacji wstawiania c_i węzła n_i lub krawędzi e_j wynosi 2.0.
2. Koszt operacji usunięcia c_d dla n_i lub e_j wynosi 2.5.
3. Koszt zmiany c_s dla węzła n_i lub e_j wynosi 1.0.

Do określenia istotności statystycznej uzyskanych wyników wykorzystano dwa testy: t-Studenta oraz U Manna-Whitneya. Test t-studenta służy do porównywania dwóch grup obserwacji. W eksperymentach za jego pomocą chcemy zbadać czy wartość średnia zmiennej (w naszym wypadku miary jakości klasyfikacji), w dwóch grupach jest równa. Grupy wyznaczane są przez wyniki klasyfikacji otrzymywane przy różnych zestawach cech (różnej budowie wektorów uczących). Ponieważ jest to test parametryczny, więc aby można go było zastosować, muszą być spełnione warunki o normalności rozkładu zmiennej, o równości wariancji oraz o podobnej liczebności grup. Jeśli te warunki nie były spełnione, do analizy użyto nieparametrycznego testu U Manna-Whitneya opisanego w pracy (Neuhäuser, 2011, Nachar, 2008).

Test t-Studenta wymaga zdefiniowania hipotezy zerowej H_0 oraz hipotezy alternatywnej H_1:

1. H_0 – nie ma istotnych statystycznie różnic między średnią z badanej próby zawierającej wyniki osiągnięte przy wykorzystaniu zaproponowanego zestawu cech *Zestawu I*, a wynikami osiągniętymi przy wykorzystaniu cech z *Zestawu J*; $I, J \in \{1, \ldots, 5\}$ i $I \neq J$.
2. H_1 – między średnią z badanej próby zawierającej wyniki osiągnięte przy wykorzystaniu zaproponowanego zestawu cech *Zestawu I*, a wynikami osiągniętymi przy wykorzystaniu cech z *Zestawu J*; $I, J \in \{1, \ldots, 5\}$ i $I \neq J$ występuje istotna statystycznie różnica.

Jeśli wynik testu będzie wyższy od wartości krytycznej dla przyjętego poziomu istotności, to istnieją podstawy do odrzucenia H_0 oraz nie ma podstaw do odrzucenia H_1. Odrzucenie H_0 a brak podstaw do odrzucenia H_1 oznacza, że wynik w rozumieniu testu t-Studenta, jest istotnie różny. Do określenia wartości krytycznej testu t-Studenta należy ustalić wymagany poziom istotności α. Na potrzeby niniejszej pracy wartość α ustawiono na $0,05$, co oznacza, że istnieje mniej niż 5% szans na to, że uzyskane wyniki są przypadkowe. Inaczej mówiąc, poziom ufności otrzymanych wyników wynosi $1 - \alpha$, czyli $1 - 0,05 = 0,95$.

Oprócz wartości α wymagane jest ustalenie stopni swobody, czyli liczby niezależnych wyników obserwacji, pomniejszonej o liczbę związków łączących wyniki ze sobą. Dla testu t-Studenta dla dwóch prób niezależnych o takich samych licznościach n, liczba stopni swobody k ustalana jest jako $N - 2$, gdzie $N = 2 \cdot n$. Porównując ze sobą wyniki dwóch niezależnych klasyfikatorów z oceną za pomocą dziesięciokrotnej walidacji krzyżowej, liczba obserwacji wynosi $2 \cdot 10$, zatem liczba stopni swobody to $k = 2 \cdot 10 - 2 = 18$. W naszych badaniach mamy do czynienia z małą próbą ($n < 30$). W teście obliczana jest wartość statystki t, dla dwóch prób niezależnych, którą definiuje wzór 6.5.

$$t = \frac{\bar{X}_1 - \bar{X}_2}{S_{x_1 - x_2}} \tag{6.5}$$

gdzie $\bar{X}_1$ jest średnią wartością z pierwszej (badanej) grupy, $\bar{X}_2$ to średnia wartość z drugiej (kontrolnej) grupy. We wzorze tym $S_{x_1 - x_2}$ obliczane jest zgodnie ze wzorem 6.6.

$$S_{x_1 - x_2} = \sqrt{\frac{(n_1 - 1) \cdot s_1^2 + (n_2 - 1) \cdot s_2^2}{n_1 + n_2 - 2} \cdot (\frac{1}{n_1} + \frac{1}{n_2})} \tag{6.6}$$

We wzorze tym s_1^2 to wariancja dla badanej próby, s_2^2 jest wariancją w próbie kontrolnej, n_1 to liczność próby badanej, zaś n_2 jest licznością próby kontrolnej. Po przyjęciu wartości istotności statystycznej α oraz liczby stopni k należy ustalić wartość krytyczną testu t-Studenta. Wartość tę można odczytać z tablic testu t-Studenta. Dla podanych parametrów wynosi ona $2,1009$. Wykonując test t-Studenta, wartość t otrzymana z testu porównywana jest z wartością krytyczną. Jeżeli wartość t jest większa od wartości krytycznej, wtedy można powiedzieć, że istnieją podstawy do twierdzenia, iż otrzymany wynik jest istotnie statystycznie lepszy.

Jednym z założeń testu t-Studenta jest **normalność rozkładu** porównywanych ze sobą prób. Jako metoda sprawdzania normalności rozkładu badanej próby wykorzystany został test Shapiro-Wilka (Shapiro i Wilk, 1965), dalej oznaczany *S-W*. Podobnie jak test t-Studenta, wymaga on zdefiniowania hipotezy zerowej oraz alternatywnej, które zdefiniowane są następująco:

1. H_0 – Rozkład badanej zmiennej jest rozkładem normalnym,
2. H_1 – Rozkład badanej zmiennej nie jest rozkładem normalnym.

Test *S-W* wymaga również określenia poziomu istotności α, który został ustalony na $\alpha = 0,05$. Wartość krytyczna testu *S-W* odczytana została z tablic rozkładu *Shapiro-Wilka*. Do sprawdzenia normalności rozkładu uzyskanych wyników wykorzystane zostały wyniki z poszczególnych foldów dziesięciokrotnej walidacji krzyżowej, zatem mamy 10 takich wyników. Z tablic rozkładu *Shapiro-Wilka*, dla 10 obserwacji oraz $\alpha = 0,05$ odczytujemy wartość równą $0,842$. Jest to wartość krytyczna testu, której przekroczenie oznacza brak podstaw do odrzucenia postawionej H_0. Wartość testu *S-W* mniejsza lub równa wartości krytycznej dla $\alpha = 0,05$ oznacza, że istnieją podstawy do odrzucenia H_0.

Kolejnym założeniem, które powinno być spełnione aby wykonać test t-Studenta, jest **równość wariancji** porównywanych ze sobą prób. Do określenia równości wariancji porównywanych wyników (*precyzji*, *kompletności*, *miary-f*) wykorzystany został test Levene'a (Levene, 1960). Test Levene'a wymaga zdefiniowania hipotezy zerowej H_0 oraz hipotezy alternatywnej H_1:

1. H_0 – Wariancje badanych zmiennych są równe,
2. H_1 – Wariancje badanych zmiennych nie są równe.

Przy testowaniu hipotezy o równości wariancji należy przyjąć poziom istotności statystycznej, który tak jak w pozostałych testach, przyjmuje wartość $\alpha = 0,05$. Uzyskanie statystyki testowej o wartości p mniejszej, bądź równej przyjętemu α, oznacza odrzucenie H_0 a przyjęcie H_1. Uzyskanie statystyki testowej p większej od przyjętego α, to brak podstaw do odrzucenia H_0, co należy interpretować, że porównywane ze sobą obserwacje mają spełniony warunek równości wariancji.

Jak już wspomniano, jeżeli nie zostaną spełnione założenia testu t-Studenta dotyczące normalności rozkładów oraz równości wariancji porównywanych grup, nie jest możliwe jego wykorzystanie. W tych przypadkach należy użyć nieparametrycznego odpowiednika dla testu t-Studenta dla dwóch prób niezależnych. W elsperymentach

wykorzystano **test U Manna-Whitneya**, w którym obliczana statystyka obliczana jest za pomocą równania 6.7.

$$U = R_{min(k)} - \frac{n_k(n_k + 1)}{2} \tag{6.7}$$

We wzorze tym $R_{min(k)}$ to mniejsza suma rang z obu prób, n_k to liczebność grupy z mniejszą sumą rang. W badanych próbach występują rangi wiązane, zatem należy uwzględnić poprawkę dla rang wiązanych na test U Manna-Whitneya poprzez obliczenie statystyki testowej Z, przedstawionej w równaniu 6.8.

$$Z = \frac{U - \frac{n_1 n_2}{2}}{\sqrt{\frac{n_1 n_2(n_1 + n_2 + 1)}{12} - \frac{n_1 n_2 \sum\limits_{i=1}(w_i^3 - w_i)}{12(n_1 + n_2)(n_1 + n_2 - 1)}}} \tag{6.8}$$

gdzie U to wynik testu U Manna-Whitneya, n_1 to liczność pierwszej grupy, n_2 jest licznością grupy drugiej, zaś w jest liczbą obserwacji posiadających taką samą rangę. Podobnie jak test t-Studenta wymaga on zdefiniowania hipotezy zerowej H_0 oraz hipotezy alternatywnej H_1:

1. H_0 – nie ma istotnych statystycznie różnic między średnią rangą z badanej próby zawierającej wyniki osiągnięte przy wykorzystaniu zaproponowanego zestawu cech *Zestawu I*, a wynikami osiągniętymi przy wykorzystaniu cech z *Zestawu J*; $I, J \in \{1, \ldots, 5\}$ i $I \neq J$.
2. H_1 – między średnią rangą z badanej próby zawierającą wyniki osiągnięte przy wykorzystaniu zaproponowanego zestawu cech *Zestawu I*, a wynikami osiągniętymi przy wykorzystaniu cech z *Zestawu J*; $I, J \in \{1, \ldots, 5\}$ i $I \neq J$ występuje istotna statystycznie różnica.

Obliczoną wartość statystyki testowej należy porównać z wartością krytyczną testu U Manna-Whitneya odczytaną z tablic statystycznych. Dla $n_1 = 10$ oraz $n_2 = 10$ wartość krytyczna wynosi 23. Jeżeli otrzymana wartość statystyki U będzie mniejsza od 23 istnieją podstawy do odrzucenia hipotezy zerowej H_0 z przyjętym poziomem istotności statystycznej $\alpha = 0,05$ oraz przyjęcia hipotezy alternatywnej H_1.

Wartości testów statystycznych t-Studenta oraz U Manna-Whitneya, wykorzystanych do zbadania istotności różnic w otrzymanych wynikach, zostały użyte w procesie metaanalizy wyników podczas określania **wielkości efektu**. W tym wypadku wielkość efektu należy rozumieć jako wielkość różnic w porównywanych próbach. Dla testu t-Studenta wykorzystana została statystyka testowa r zaproponowana w pracy (Rosenthal, 1994), która obliczana jest zgodnie z równaniem 6.9. W równaniu, r to wynik statystyki testowej, t jest wartością testu t-Studenta, n_1 to liczność grupy badanej, n_2 jest licznością grupy kontrolnej.

$$r = \frac{t}{\sqrt{t^2 + n_1 + n_c - 2}} \tag{6.9}$$

Wykorzystując dziesięciokrotną walidację krzyżową, porównywane grupy wyników uwzględniających miary *precyzji*, *kompletności* oraz *miary-f* posiadają po 10 obserwacji. Zatem n_1 oraz n_2 są równe 10. Podczas analizy przyjęta została kategoryzacja

wielkości efektu według pracy (Cohen, 1988). W pracy tej zaproponowano kategoryzację wartości bezwzględnych r, jako: $r \in \langle 0, \ldots, 0,3)$ oznacza *małą* wielkość efektu, $r \in \langle 0,3, \ldots, 0,5)$ to *średnia* wielkość efektu, zaś $r \in \langle 0,5, \ldots, 1\rangle$ oznacza *dużą* wielkość efektu. Dla wyników otrzymanych za pomocą testu U Manna-Whitneya, wykorzystana została statystyka testowa z opisana wzorem 6.10.

$$z = \frac{Z}{\sqrt{N}} \tag{6.10}$$

W statystyce tej Z to wartość statystyki testowej Z opisanej wzorem 6.8, N to liczba obserwacji w obu grupach, czyli 20. Interpretacja wyniku statystyki z, według pracy (Cohen, 1988), jest taka sama jak w przypadku statystyki r, czyli: $z \in \langle 0, \ldots, 0,3)$ oznacza *małą* wielkość efektu, $z \in \langle 0,3, \ldots, 0,5)$ to *średnia* wielkość efektu, zaś $z \in \langle 0,5, \ldots, 1\rangle$ oznacza *dużą* wielkość efektu.

6.2. Przeprowadzone eksperymenty

Przeprowadzone eksperymenty miały na celu sprawdzenie wpływu wykorzystanych zestawów cech na jakość klasyfikacji relacji semantycznych pomiędzy dwoma zdaniami, mierzoną za pomocą *precyzji*, *kompletności* oraz *miary-f*. Za tak ustawionym celem kryło się pytanie, czy zestaw cech, który wprowadzono dzięki opracowaniu metody *PAS* poprawia jakość rozpoznawania relacji semantycznych między dwoma zdaniami w stosunku do cech literaturowych. We wszystkich przeprowadzonych eksperymentach rozpoznawanie przeprowadzono dla wszystkich 4 klasyfikatorów (*SVM*, *J48*, *Naiwny klasyfikator Bayesa* oraz *Drzewo modeli logistycznych*). ich parametry miały ustawione wartości domyślne i nie ulegały zmianie. W kolejnych eksperymentach rozpoznawanie relacji odbywało się z użyciem innych wektorów cech, które były budowane na podstawie zestawów cech uczących opisanych jako *Zestaw 1* do *Zestaw 5*. W celu określenia, czy zaproponowane nowe cechy poprawiają jakość klasyfikacji dokonano analizy porównawczej wyników.

Badano również czy wybór klasyfikatora mógł mieć wpływ na jakość klasyfikacji dla różnych wektorów cech. W tym celu porównano ze sobą wyniki wszystkich klasyfikatorów dla każdego z dostępnych zestawów cech. Sprawdzono czy uzyskane wyniki istotnie różnią się między sobą. Na podstawie tych porównań wybrany został klasyfikator, który z wykorzystaniem opracowanych zestawów cech wykazuje najlepszą jakość rozpoznawania relacji semantycznych między dwoma zdaniami.

W dalszej części tego rozdziału znajdują się wyniki przeprowadzonych eksperymentów z komentarzem do nich. Przedstawione zostały wyniki dla wszystkich klas, zarówno mało licznych (o liczności niższej od 50) jak i tych, których liczba wektorów uczących była wyższa/równa 50. Przy tych założeniach średnia ważona podawana w wynikach uwzględnia również mało liczne klasy. Ważenie odbyło się za pomocą liczności klasy. Każdy z przedstawionych eksperymentów posiada opis w postaci schematu zawierającego następujące elementy: cel badania, otrzymane wyniki, test sprawdzający normalność rozkładu uzyskanych wyników, test sprawdzający równość wariancji porównywanych grup oraz statystyczną analizę wyników.

6.2.1. Eksperyment 1. Określenie poziomu odniesienia – zbadanie jakości klasyfikacji relacji semantycznych dla dwóch zdań przy wektorze cech zbudowanym na podstawie *Zestawu 1*

Cel badania

Badanie miało na celu określenie poziomu bazowego jakości klasyfikacji relacji semantycznych między dwoma zdaniami, w kontekście którego porównywane będą wyniki kolejnych eksperymentów dla proponowanych nowych zestawów cech. *Zestaw 1* używany do zbudowania wektora cech, zawiera cechy do rozpoznawania relacji semantycznych proponowane w literaturze. Opisany jest szczegółowo w rozdziale 5.3.1.

Otrzymane wyniki

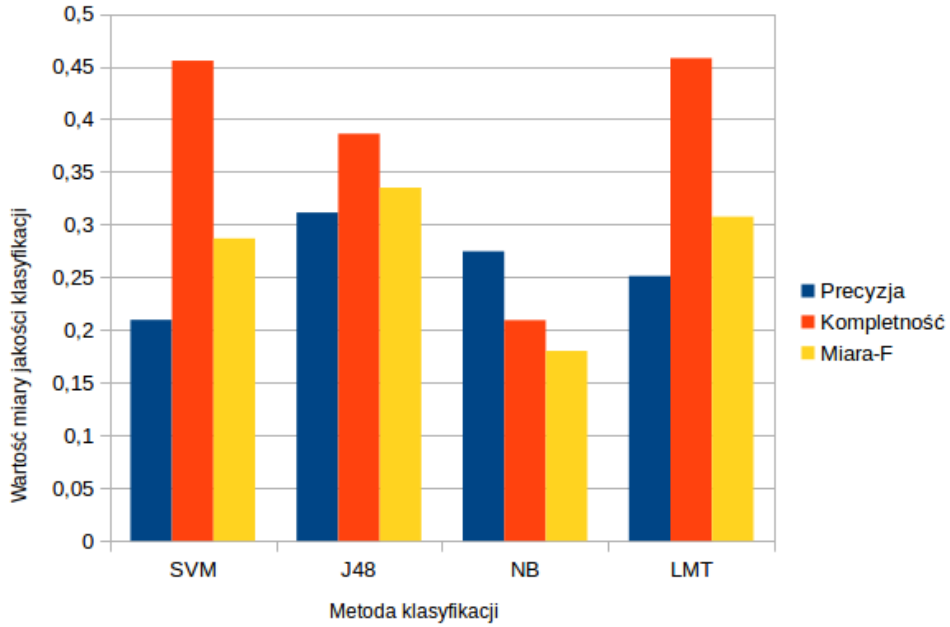
Tabela 6.2 pokazuje szczegółowe wyniki jakości klasyfikacji dla klasyfikatorów *NB*, *J48*, *SVM* oraz *LMT*. Kolejne wiersze zawierają wyniki dla poszczególnych relacji opisanych w ostatniej kolumnie. Ostatni wiersz zawiera średnią ważoną. Wyniki wyrażone są za pomocą precyzji **P**, kompletności **R** oraz miary-f **F** liczonej osobno dla każdego klasyfikatora.

Tabela 6.2. Wyniki jakości klasyfikacji mierzonej za pomocą *precyzji* (**P**), *kompletności* (**R**) oraz *miary-f* (**F**) w eksperymencie służącym do określenia poziomu odniesienia (wektor cech opisany jako *Zestaw 1*), dla czterech klasyfikatorów: **NB**, **J48**, **SVM** oraz **LMT**

NB			**J48**			**SVM**			**LMT**			
P	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**	**Klasa**
0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	Cytowanie
0,080	0,014	0,022	0,029	0,018	0,022	0,000	0,000	0,000	0,000	0,000	0,000	Dalsze informacje
0,478	0,186	0,266	0,460	0,682	0,549	0,456	0,994	0,626	0,456	0,994	0,626	Krzyżowanie się
0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	Modalność
0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	Mowa zależna
0,173	0,790	0,284	0,309	0,241	0,270	0,000	0,000	0,000	0,000	0,000	0,000	Opis
0,096	0,680	0,168	0,228	0,215	0,219	0,000	0,000	0,000	0,000	0,000	0,000	Parafraza
0,097	0,026	0,038	0,062	0,032	0,042	0,000	0,000	0,000	0,000	0,000	0,000	Spełnienie
0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	Sprzeczność
0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	Streszczenie
0,101	0,120	0,106	0,222	0,170	0,191	0,000	0,000	0,000	0,000	0,000	0,000	Tło historyczne
0,387	0,850	0,520	0,403	0,592	0,464	0,439	0,850	0,565	0,439	0,850	0,565	Tożsamość
0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	Uszczegółowienie
0,219	0,077	0,112	0,253	0,176	0,202	0,000	0,000	0,000	0,000	0,000	0,000	Zawieranie
0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	Zmiana poglądu
0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	Źródło
0,033	0,006	0,011	0,463	0,313	0,359	0,000	0,000	0,000	0,000	0,000	0,000	Brak relacji
0,275	**0,209**	**0,180**	**0,311**	**0,386**	**0,335**	**0,210**	**0,455**	**0,287**	**0,251**	**0,458**	**0,308**	**Średnia ważona**

Jak można zauważyć, przy zastosowanym zestawie cech *Zestaw 1*, nie było możliwe rozpoznanie relacji *Cytowania*, *Modalności*, *Mowy zależnej*, *Sprzeczności*, *Streszczenia*, *Uszczegóławiania*, *Zmiany poglądu* oraz *Źródła*. Żaden z klasyfikatorów w kontekście rozważanych miar (*precyzji*, *kompletności* jak i *miary-f*) nie wykazał żadnych zdolności do rozpoznania obiektów tej klasy. Relacja *Parafrazy* rozpoznana została tylko przez *NB* oraz *J48*. Klasyfikatory *SVM* oraz *LMT* nie były w stanie rozpoznać relacji *Parafrazy*, co może mieć związek z małą licznością przykładów uczących, których dla tej relacji jest 48 oraz ubogim zestawem cech. Podobną sytuację można zaobserwować dla relacji *Spełnienia*, *Tła historycznego* oraz *Zawierania*, jednak w tym przypadku liczba

przykładów uczących wynosi odpowiednio 160, 312 i 194. Stąd można wysnuć wniosek, że cechy z *Zestawu 1* mogą być zbyt mało dyskryminujące dla *SVM* oraz *LMT*.



Rys. 6.1. Wyniki jakości klasyfikacji dla wszystkich klasyfikatorów przy użyciu wektora cech opisanego przez *Zestaw 1*, wyrażone jako średnie ważone *precyzji*, *kompletności* oraz *miary-f*

Najwyższą wartość średniej ważonej *precyzji* oraz *miary-f* osiągnął klasyfikator *J48*, najwyższą jakość wyrażoną za pomocą średniej ważonej kompletności – *LMT*. Na rysunku 6.1 przedstawione zostały średnie ważone *precyzji*, *kompletności* oraz *miary-f* dla wszystkich klasyfikatorów. Można zauważyć, że *LMT* oraz *SVM* wykazują podobną tendencję w jakości rozpoznawania relacji z wykorzystaniem pierwszego zestawu cech. Dla *J48* *precyzja* i *kompletność* są do siebie bardziej zbliżone niż w pozostałych przypadkach. *NB* wykazuje najmniejszą kompletność rozpoznawanych relacji, jednak jego precyzja jest wyższa niż *LMT* i *SVM*.

Testowanie hipotezy o normalności rozkładu uzyskanych wyników z wykorzystaniem testu Shapiro-Wilka

W tabeli 6.3 przedstawione zostały wyniki testu Shapiro-Wilka, mającego na celu sprawdzanie normalności rozkładu uzyskanych wyników. Uzyskane wartości testu Shapiro-Wilka, przedstawione w tabeli w wierszach nazwanych W, dla każdego klasyfikatora są wyższe od wartości krytycznej, równej $0,842$, odczytanej z tablic rozkładu $S\text{-}W$. Zatem nie ma podstaw do odrzucenia hipotezy H_0 mówiącej o istnieniu rozkładu normalnego badanej próby.

Tabela 6.3. Wyniki obliczonych statystyk Shapiro-Wilka (W) dla wszystkich klasyfikatorów, osobno dla *precyzji* **P**, *kompletności* – **R** oraz *miary-f* – **F**, dla wyników uzyskanych z wykorzystaniem cech z *Zestawu 1*

SVM	**P**	**R**	**F**	**J48**	**P**	**R**	**F**
W	0.90321	0.9254	0.91195	W	0.94863	0.95744	0.90767
NB	**P**	**R**	**F**	**LMT**	**P**	**R**	**F**
W	0.94352	0.96067	0.92587	W	0.92229	0.95768	0.89654

6.2.2. Eksperyment 2. Rozpoznawanie relacji semantycznych między dwoma zdaniami z wykorzystaniem wektora cech zbudowanego z użyciem cech z *Zestawu 2*

Cel badania

Badanie miało na celu określenie jakości klasyfikacji z użyciem wektora cech bazującego na zestawie cech z *Zestawu 2* (rozdział 5.3.2), porównanie otrzymanych wyników z wynikami bazowymi oraz określenie istotności otrzymanych wyników.

Otrzymane wyniki

W tabeli 6.4 przedstawione zostały szczegółowe wyniki klasyfikacji z wykorzystaniem `cech grafowych (1)` ujętych w *Zestawie 2*. W tabeli tej przedstawione zostały wyniki dla każdego wykorzystanego klasyfikatora: naiwnego klasyfikatora Bayesa – **NB**, drzewa decyzyjnego uczonego algorytmem C.45 – **J48**, maszyny wektorów nośnych – **SVM** oraz drzewa modeli logistycznych – **LMT**. W kolejnych wierszach tabeli znajdują się wyniki dla poszczególnych relacji semantycznych, zaś ostatni wiersz zawiera

Tabela 6.4. Wyniki jakości klasyfikacji relacji semantycznych między dwoma zdaniami, mierzonej za pomocą *precyzji* (**P**), *kompletności* (**R**) oraz *miary-f* (**F**) z wykorzystaniem cech z *Zestawu 2*, dla czterech klasyfikatorów: **NB**, **J48**, **SVM** oraz **LMT**

NB			**J48**			**SVM**			**LMT**			
P	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**	**Klasa**
0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,100	0,100	0,100	Cytowanie
0,300	0,010	0,020	0,242	0,197	0,216	0,764	0,965	0,852	0,691	0,810	0,743	Dalsze informacje
0,950	0,127	0,224	0,894	0,944	0,918	0,947	1,000	0,973	0,966	0,994	0,980	Krzyżowanie się
0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	Modalność
0,036	0,200	0,059	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	Mowa zależna
0,192	0,715	0,302	0,346	0,420	0,376	0,527	0,715	0,605	0,555	0,745	0,633	Opis
0,089	0,575	0,153	0,139	0,130	0,123	0,217	0,085	0,117	0,367	0,140	0,198	Parafraza
0,290	0,065	0,100	0,095	0,103	0,097	0,442	0,052	0,091	0,521	0,192	0,277	Spełnienie
0,071	0,250	0,109	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	Sprzeczność
0,028	0,157	0,047	0,000	0,000	0,000	0,000	0,000	0,000	0,183	0,077	0,097	Streszczenie
0,234	0,239	0,230	0,281	0,315	0,293	0,519	0,721	0,602	0,594	0,714	0,646	Tło historyczne
0,325	0,875	0,466	0,491	0,675	0,556	0,785	0,767	0,758	0,892	0,850	0,849	Tożsamość
0,000	0,000	0,000	0,042	0,035	0,035	0,960	0,148	0,247	0,767	0,311	0,436	Uszczegółowienie
0,366	0,061	0,102	0,277	0,239	0,254	0,521	0,596	0,553	0,488	0,458	0,465	Zawieranie
0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	Zmiana poglądu
0,000	0,000	0,000	0,073	0,060	0,065	0,400	0,080	0,133	0,783	0,415	0,521	Źródło
0,665	0,192	0,296	0,489	0,339	0,389	0,771	0,677	0,710	0,726	0,703	0,706	Brak relacji
0,568	**0,199**	**0,191**	**0,311**	**0,561**	**0,547**	**0,738**	**0,764**	**0,724**	**0,756**	**0,765**	**0,746**	**Średnia ważona**

średnią ważoną *precyzji* (**P**), *kompletności* (**R**) oraz *miary-f* (**F**) dla wszystkich wykorzystanych metod klasyfikacji. Pierwsze spostrzeżenie jakie możemy poczynić to fakt, że poza kompletnością **R** dla naiwnego klasyfikatora Bayesa (**NB**), wszystkie uzyskane wyniki dla średnich ważonych są wyższe niż w przypadku wyników z *Eksperymentu 1.* Warto jednak zwrócić uwagę, że dwie relacje, to jest*Modalność* oraz *Zmiana poglądu* nadal nie są rozpoznawane przez żaden z klasyfikatorów. Przyczyny tego stanu można upatrywać w małej liczbie przykładów uczących, odpowiednio 1 oraz 6. W przypadku *Mowy zależnej* z 26 przykładami uczącymi, *Sprzeczności* z 20 przykładami uczącymi oraz *Streszczania* z 55 przykładami uczącymi oprócz klasyfikatora probabilistycznego *NB*, żadna z pozostałych metod klasyfikacji nie była w stanie rozpoznać tych relacji. Relacje *Uszczegółowienia* oraz *Źródła* nie zostały rozpoznane przez *NB*, zaś pozostałe klasyfikatory, pomimo niskiej *kompletności*, rozpoznały je. *Precyzja* wykrywania relacji *Uszczegółowienia*, zarówno dla *LMT* oraz *SVM* jest na wysokim poziomie. W przypadku *Źródła*, którego liczność przykładów uczących jest równa 50, *precyzja* rozpoznawania tej relacji dla *LMT* jest na wysokim poziomie $0,783$. To, co warto jeszcze zauważyć, to przypadkowość rozpoznania relacji *Cytowanie* przez *LMT*; *precyzja*, *kompletność* oraz *miara-f* wynoszą $0,100$, co oznacza, że średnio przy jednym podziale zbioru testowego podczas dziesięciokrotnej walidacji krzyżowej, udało się rozpoznać relację *Cytowania.*

Testowanie hipotezy o normalności rozkładu uzyskanych wyników z wykorzystaniem testu Shapiro-Wilka

W tabeli 6.5 przedstawione zostały wyniki testu Shapiro-Wilka, sprawdzającego normalność rozkładu uzyskanych wyników w kontekście wszystkich używanych do oceny miar: *precyzji* (**P**), *kompletności* (**R**) oraz *miary-f* (**F**). Wyniki testu Shapiro-Wilka

Tabela 6.5. Wyniki obliczonych statystyk Shapiro–Wilka (W) dla wszystkich klasyfikatorów, osobno dla *precyzji* **P**, *kompletności* – **R** oraz *miary-f* – **F**, dla wyników uzyskanych z wykorzystaniem cech z *Zestawu 2*

SVM	**P**	**R**	**F**	**J48**	**P**	**R**	**F**
W	0.93416	0.97729	0.96276	W	0.95733	0.92418	0.94464
NB	**P**	**R**	**F**	**LMT**	**P**	**R**	**F**
W	0.96488	0.97683	0.91458	W	0.92998	0.9319	0.95312

przedstawiono w wierszach oznaczonych jako W. Dla każdego klasyfikatora są one wyższe od wartości krytycznej, równej $0,842$, odczytanej z tablic rozkładu S-W. Zatem nie ma podstaw do odrzucenia H_0 mówiącej o istnieniu rozkładu normalnego badanej próby.

Testowanie założenia o równości wariancji w porównywanych grupach z wykorzystaniem testu Levene'a

Jak już wspominano zastosowanie testu t-Studenta, oprócz normalności rozkładu wyników porównywanych grup, wymaga sprawdzenia równości wariancji w porównywanych grupach. W kolumnach tabeli 6.6 przedstawione zostały wyniki testu Levene'a dla wykorzystanych klasyfikatorów (**NB**, **J48**, **SVM**, **LMT**) osobno dla wszystkich rozważanych miar: *precyzji* – **P**, *kompletności* – **R** oraz *miary-f* – **F**. W tym teście porównywane

są ze sobą średnie ważone wyników klasyfikacji wyrażonych przez **P**, **R**, **F** (ważenie odbywa się za pomocą liczby relacji) z wykorzystaniem wektorów cech zbudowanych na podstawie *Zestawu 1* oraz `cech grafowych (1)` z *Zestawu 2*. Pogrubione zostały te

Tabela 6.6. Wyniki statystyki testowej p testu Levene'a dla *precyzji* (**P**), *kompletności* (**R**) oraz *miary-f* (**F**), dla każdego z klasyfikatorów (**NB**, **J48**, **SVM** oraz **LMT**) obliczone dla wyników klasyfikacji z wykorzystaniem cech z *Zestawu 1* oraz *Zestawu 2*

	NB	**J48**	**SVM**	**LMT**
P	**0,0096**	0,9088	**0,0001**	0,6651
R	0,324	0,9067	**0,0083**	**0,0255**
F	0,8938	0,4363	**0,0019**	0,1067

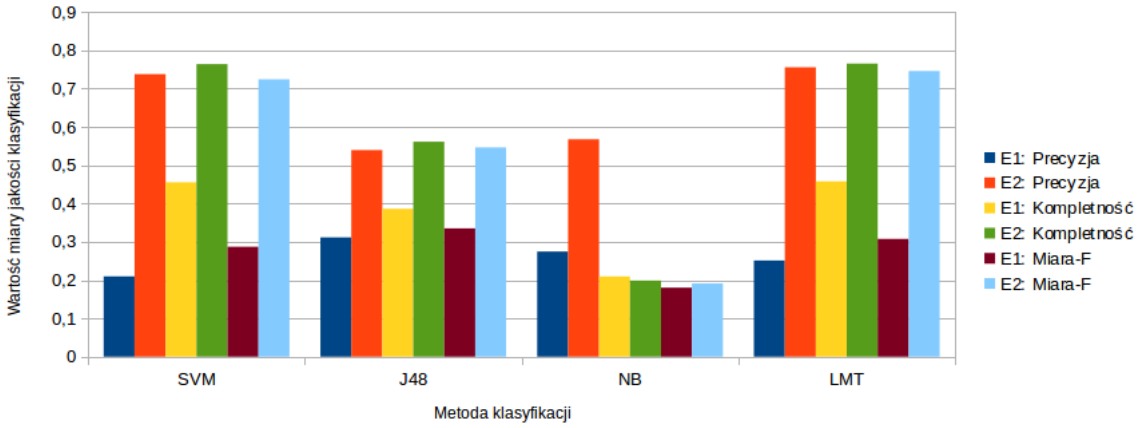
wyniki testu, których wartość obliczonej statystyki testowej p, nie przekracza zadanego poziomu istotności statystycznej wynoszącego $0,05$. Jest pięć przypadków, dla których nie ma spełnionego warunku o równości wariancji porównywanych ze sobą grup, są to: *precyzja* dla naiwnego klasyfikatora bayesowskiego, *precyzja*, *kompletność* i $miara-f$ dla SVM oraz *kompletność* dla LMT. Dla tych przypadków należy odrzucić hipotezę H_0 mówiącą o jednorodności wariancji w porównywanych grupach i przyjąć hipotezę alternatywną H_1, mówiącą o braku równości wariancji porównywanych grup. Dla tych pięciu przypadków nie jest możliwe wykonanie testu t-Studenta do określenia istotności statystycznej uzyskanych wyników i dla nich został wykonany nieparametryczny test U Manna-Whitneya.

Analiza uzyskanych wyników

Na rysunku 6.2 porównane zostały ze sobą wyniki klasyfikacji z eksperymentu pierwszego (oznaczonego jako *E1*) z wynikami eksperymentu drugiego (*E2*) dla średnich ważonych *precyzji*, *kompletności* oraz *miary-F*, dla każdego klasyfikatora. Warto zauważyć tendencję do dużo wyższej *precyzji*, *kompletności* oraz *miary-f* w eksperymencie wykorzystującym cechy z *Zestawu 2* dla klasyfikatorów *SVM*, *J48* oraz *LMT*. Klasyfikator *NB* nie wykazuje tej tendencji, *kompletność* uzyskana w eksperymencie drugim jest niższa od *kompletności* uzyskanej w eksperymencie pierwszym.
Tabele 6.7 oraz 6.9 oraz 6.10 zawierają wyniki testu t-Studenta dla wszystkich uwzględnionych miar jakości klasyfikacji (*precyzji*, *kompletności* oraz *miary-f*) dla każdego z klasyfikatorów, dla wszystkich wypadków, w których wyniki spełniają założenia do wykonania testu, z podaniem informacji czy osiągnięty wynik dla danego klasyfikatora jest istotnie statystycznie lepszy. Tabela 6.8 zawiera wyniki testu U Manna-Whitneya. Wpisane w komórkę wartości *nie dotyczy* oznaczają, że dla danego przypadku nie został wykonany test U Manna-Whitneya, a został wykonany test t-Studenta, ponieważ dla tych przypadków były spełnione założenia testu t-Studenta.

Tabela 6.7 zawiera wyniki testu t-Studenta dla jakości mierzonej za pomocą *precyzji*. Ten test mógł być wykonany jedynie dla klasyfikatorów *J48* oraz *LMT*, dla których wartość testu jest powyżej wartości krytycznej $t = 2,1009$. Wyniki oceny istotności statystycznej uzyskanych wyników *precyzji* dla *SVM* oraz *NB* znajdują się w tabeli 6.8.

Dla klasyfikatorów *SVM* oraz *NB* wartości statystyki testowej U testu U Manna-Whitneya są poniżej wartości krytycznej wynoszącej 23. Zatem istnieją pod-



Rys. 6.2. Zestawienie wyników jakości klasyfikacji dla `cech grafowych (1)` z *Zestawu 2*, z eksperymentu *E2* oraz *cech bazowych* opisanych *Zestawem 1* uzyskanych w eksperymencie *E1*, wyrażonych za pomocą średnich ważonych miar: *precyzji*, *kompletności* i *miary-f*

Tabela 6.7. Wartości testu t-Studenta dla *precyzji*, obliczone dla wyników uzyskanych w eksperymencie pierwszym (wektor cech z *Zestawu 1*) i drugim (wektor cech z *Zestawu 2*)

	Zestaw 1		**Zestaw 2**		**Wynik testu**	
Klasyfikator	**Średnia precyzja**	**Wariancja**	**Średnia precyzja**	**Wariancja**	**Wartość testu**	**Czy wynik jest istotnie lepszy**
J48	0,311	0,001	0,540	0,001	10,093	TAK
LMT	0,251	0,001	0,756	0,001	28,637	TAK

stawy do odrzucenia hipotezy zerowej H_0 mówiącej o braku różnic uzyskanych wyników, zarówno dla testu t-Studenta jak i U Manna-Whitneya. Można zatem powiedzieć, że wykorzystanie `cech grafowych (1)` istotnie poprawia jakość klasyfikacji mierzoną za pomocą *precyzji*.

Tabela 6.8. Wartości statystyki testowej U obliczonej dla testu U Manna-Whitneya dla klasyfikatorów **SVM**, **NB** oraz **LMT**, w których porównywana jest jakość klasyfikacji mierzona za pomocą *precyzji* (**P**), *kompletności* (**R**) i *miary-f* (**F**) z zaznaczonymi za pomocą pogrubienia wartościami nieprzekraczającymi poziomu istotności statystycznej $\alpha = 0,05$

	P	**R**	**F**
SVM	**0,00**	**0,00**	**0,00**
NB	**0,00**	*nie dotyczy*	*nie dotyczy*
LMT	*nie dotyczy*	**0,00**	*nie dotyczy*

W tabeli 6.9 przedstawione zostały wyniki testu t-Studenta dla *kompletności* obliczonej dla klasyfikatorów *J48* oraz *NB*. Wyniki klasyfikatorów *SVM* oraz *LMT* nie spełniają wymagań do wykonania testu t-Studenta, dlatego dla nich wykonany został test U Manna-Whitneya, którego wyniki zostały przedstawione w tabeli 6.8. Dla klasyfikatorów *J48*, *SVM* oraz *LMT*, wykorzystanie `cech grafowych (1)` poprawiło jakość

Tabela 6.9. Wartości testu t-Studenta dla *kompletności*, obliczone pomiędzy wynikami uzyskanymi w eksperymencie pierwszym i drugim, wykorzystującymi cechy z *Zestawu 1* oraz *Zestawu 2*

	Zestaw 1		**Zestaw 2**		**Wynik testu**	
Klasyfikator	**Średnia kompletność**	**Wariancja**	**Średnia kompletność**	**Wariancja**	**Wartość testu**	**Czy wynik jest istotnie lepszy**
J48	0,386	0,001	0,561	0,001	8,350	TAK
NB	0,209	0,001	0,199	0,001	0,619	**NIE**

mierzoną za pomocą *kompletności*. Uzyskana poprawa jakości klasyfikacji w każdym wypadku była istotna statystycznie. Wynik testu t-Studenta dla *J48* był wyższy od poziomu krytycznego $t = 2,1009$, zaś wynik statystyki U testu U Manna-Whitneya obliczony dla *SVM* oraz *LMT* był niższy od wartości krytycznej 23. W obu przypadkach istnieją podstawy do odrzucenia hipotezy H_0, a przyjęcia hipotezy H_1. Kompletność rozpoznawanych relacji dla klasyfikatora *NB* z wykorzystaniem cech z *Zestawu 2* jest niższa od kompletności uzyskanej z wykorzystaniem cech z *Zestawu 1*. Wynik testu t-Studenta jest negatywny, uzyskana wartość testu (0,619) jest niższa od wartości krytycznej testu $t = 2,1009$, co oznacza, że nie ma podstaw do stwierdzenia, że uzyskany wynik jest istotnie statystycznie inny, w tym przypadku gorszy.

W przypadku porównania jakości klasyfikacji mierzonej za pomocą *miary-f*, wyniki klasyfikatora *SVM* osiągnięte z użyciem cech z *Zestawu 1* oraz `cech grafowych (1)`, nie spełniają warunku testu t-Studenta odnośnie równości wariancji w porównywanych grupach, dlatego do określenia istotności uzyskanych wyników wykorzystany został test u Manna-Whitneya. W pozostałych przypadkach, dla klasyfikatorów *J48*, *NB* oraz *LMT*, użyto testu t-Studenta. Jego wyniki przedstawiono w W tabeli 6.10 Wartości te-

Tabela 6.10. Wartości testu t-Studenta dla *miary-f*, obliczone pomiędzy wynikami uzyskanymi w eksperymencie pierwszym i drugim, wykorzystującymi cechy z *Zestawu 1* oraz *Zestawu 2*

	Zestaw 1		**Zestaw 2**		**Wynik testu**	
Klasyfikator	**Średnia miara-f**	**Wariancja**	**Średnia miara-f**	**Wariancja**	**Wartość testu**	**Czy wynik jest istotnie lepszy**
J48	0,335	0,001	0,547	0,001	12,067	TAK
NB	0,180	0,001	0,191	0,001	0,733	**NIE**
LMT	0,308	0,001	0,746	0,001	39,810	TAK

stów dla *J48* oraz *LMT* są wyższe od założonej wartości krytycznej testu $t = 2,1009$, zatem istnieją podstawy do odrzucenia hipotezy H_0, a przyjęcia hipotezy H_1, co oznacza, że uzyskane wyniki z wektorem cech z *Zestawu 2*, są istotnie statystycznie lepsze od wyników uzyskanych za pomocą cech z *Zestawu 1*. Wartość testu studenta dla *NB* jest niższa od wartości krytycznej testu $t = 2,1009$, zatem nie ma podstaw do odrzucenia H_0, co oznacza, że uzyskane wyniki za pomocą cech z *Zestawu 2* nie są istotnie statystycznie lepsze od wyników uzyskanych z wykorzystaniem cech z *Zestawu 1*. Obliczona wartość statystyki U w przeprowadzonym teście U Manna-Whitneya dla *miary-f* dla klasyfikatora *SVM*, jest niższa od wartości krytycznej 23 dla poziomu istotności

statystycznej $\alpha = 0,05$. Skutkuje to odrzuceniem hipotezy zerowej H_0, a przyjęciem H_1, co oznacza istotną różnicę w otrzymanych wynikach.

Meta-analiza uzyskanych wyników – określanie wielkości efektu

W kolejnych wierszach tabeli 6.11 przedstawione zostały wyniki obliczonych statystyk r oraz z dla wykorzystanych klasyfikatorów (*SVM*, *J48*, *NB*, *LMT)*. W kolumnach **Wrt.** , na niebiesko oznaczone zostały wyniki statystyki z, na czarno wyniki statystyki r. Kolumna **Wlk.** zawiera interpretację wartości statystyk r oraz z, zgodnie z propozycją Cohena (Cohen, 1988). Kolumna **P** zawiera wyniki wielkości efektu dla *precyzji*, kolumna **R** dla *kompletności*, a kolumna **F** dla *miary-f*.

Tabela 6.11. Wartości statystyk (kolumny **Wrt.**), dla statystyki r (czarny kolor czcionki) i z (niebieski kolor czcionki) oraz interpretacja w postaci wielkości różnicy (kolumny **Wlk.**) między wynikami uzyskanymi w *Eksperymencie 1* i *Eksperymencie 2* dla *precyzji* (**P**), *kompletności* (**R**) oraz *miary-f* (**F**), dla klasyfikatorów *SVM*, *J48*, *NB* oraz *LMT*

	P		**R**		**F**	
	Wrt.	**Wlk.**	**Wrt.**	**Wlk.**	**Wrt.**	**Wlk.**
SVM	0,837	Duża	0,837	Duża	0,837	Duża
J48	0,922	Duża	0,892	Duża	0,943	Duża
NB	0,837	Duża	0,144	Mała	0,170	Mała
LMT	0,989	Duża	0,837	Duża	0,994	Duża

Wielkości efektu dla *precyzji* (kolumna **P**) we wszystkich klasyfikatorach są *duże*, co oznacza, że różnica między wynikami uzyskanymi w *Eksperymencie 1* (wektor cech tworzony na bazie *Zestawu 1*) oraz *Eksperymencie 2* (wektor cech tworzony na bazie *Zestawu 2*) jest *duża*. Wielkość efektu koreluje z wartościami testów istotności statystycznych.

Dla *kompletności* (kolumna **R**), dla *SVM*, *J48* oraz *LMT* wielkość efektu jest również *duża*. Dla klasyfikatora *NB*, obliczoną wartość statystyki r należy zinterpretować jako *małą* różnicę między otrzymanymi wynikami. Wielkość efektu, w przypadku *kompletności* również koreluje z testami istotności statystycznej. W testach istotności statystycznej dla *SVM*, *J48* oraz *LMT* wyniki mierzone za pomocą *kompletności*, otrzymane za pomocą cech z *Zestawu 1* różnią się istotnie od wyników otrzymanych za pomocą cech z *Zestawu 2*. W przypadku *NB*, nie ma istotnej różnicy w otrzymanych wynikach.

Wynik analizy efektu dla jakości mierzonej za pomocą *miary-f* (kolumna **F**) jest analogiczny jak w przypadku *kompletności*. Dla klasyfikatorów *SVM*, *J48* oraz *LMT* jest *duża* różnica pomiędzy wynikami osiągniętymi z użyciem wektorów cech z *Zestawu 1* oraz cech z *Zestawu 2*. Dla *NB*, obliczoną wartość statystyki r należy zinterpretować jako *małą* różnicę między otrzymanymi wynikami. Wielkości efektu dla *miary-f* korelują z testami istotności statystycznej. W testach istotności statystycznej dla *SVM*, *J48* oraz *LMT* wyniki *miary-f* otrzymane za pomocą cech z *Zestawu 1* różnią się istotnie od wyników otrzymanych za pomocą cech z *Zestawu 2*. W przypadku *NB*, nie ma istotnej różnicy w otrzymanych wynikach.

6.2.3. Eksperyment 3. Badanie jakości rozpoznawanie relacji semantycznych między dwoma zdaniami z wykorzystaniem wektora cech zbudowanego na podstawie *Zestawu 3*

Cel badania

Badanie miało na celu zbadanie jakości klasyfikacji z wykorzystaniem wektorów cech budowanych na podstawie *Zestawu 3* (rozdział 5.3.3) i porównanie otrzymanych wyników z wynikami otrzymanymi w *Eksperymencie 2*.

Otrzymane wyniki

W tabeli 6.12 przedstawione zostały szczegółowe wyniki klasyfikacji relacji semantycznej między dwoma zdaniami z wykorzystaniem wektora cech z *Zestawu 3*, będącego połączeniem *cech literaturowych* opisanych *Zestawem 1* z `cechami grafowymi (1)` opisanymi *Zestawem 2*. W tabeli tej przedstawione zostały wyniki dla każdego wykorzystanego klasyfikatora: naiwnego klasyfikatora Bayesa – **NB**, drzewa decyzyjnego uczonego algorytmem C.45 – **J48**, maszyny wektorów nośnych – **SVM** oraz drzewa modeli logistycznych – **LMT**. W kolejnych wierszach tabeli znajdują się wyniki dla po-

Tabela 6.12. Wyniki jakości klasyfikacji relacji semantycznych między dwoma zdaniami, mierzone za pomocą *precyzji* (**P**), *kompletności* (**R**) oraz *miary-f* (**F**) z wykorzystaniem cech z *Zestawu 3*, dla czterech klasyfikatorów: **NB**, **J48**, **SVM** oraz **LMT**

NB			**J48**			**SVM**			**LMT**			
P	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**	**Klasa**
0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,100	0,100	0,100	Cytowanie
0,400	0,017	0,033	0,146	0,151	0,147	0,788	0,965	0,866	0,735	0,793	0,760	Dalsze informacje
0,955	0,100	0,180	0,886	0,932	0,908	0,947	1,000	0,973	0,970	0,992	0,981	Krzyżowanie się
0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	Modalność
0,025	0,200	0,045	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	Mowa zależna
0,187	0,728	0,298	0,375	0,409	0,389	0,554	0,712	0,621	0,571	0,751	0,646	Opis
0,084	0,595	0,147	0,235	0,200	0,210	0,050	0,020	0,029	0,276	0,185	0,210	Parafraza
0,257	0,070	0,110	0,127	0,122	0,124	0,683	0,104	0,177	0,522	0,229	0,310	Spełnienie
0,094	0,300	0,140	0,050	0,050	0,050	0,000	0,000	0,000	0,000	0,000	0,000	Sprzeczność
0,025	0,160	0,044	0,050	0,040	0,044	0,000	0,000	0,000	0,375	0,100	0,151	Streszczenie
0,212	0,253	0,228	0,297	0,277	0,281	0,537	0,717	0,612	0,587	0,720	0,645	Tło historyczne
0,416	0,930	0,564	0,525	0,695	0,586	0,882	0,835	0,845	0,942	0,810	0,864	Tożsamość
0,000	0,000	0,000	0,025	0,017	0,020	0,922	0,252	0,392	0,655	0,328	0,430	Uszczegółowienie
0,475	0,093	0,154	0,363	0,340	0,349	0,499	0,649	0,563	0,500	0,524	0,505	Zawieranie
0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	Zmiana poglądu
0,100	0,020	0,033	0,000	0,000	0,000	0,500	0,120	0,190	0,758	0,375	0,489	Źródło
0,498	0,095	0,158	0,481	0,396	0,430	0,801	0,740	0,767	0,774	0,722	0,742	Brak relacji
0,575	**0,191**	**0,171**	**0,311**	**0,561**	**0,546**	**0,756**	**0,777**	**0,742**	**0,764**	**0,765**	**0,756**	**Średnia ważona**

szczególnych relacji semantycznych, zaś ostatni wiersz zawiera średnią ważoną (ważoną za pomocą liczności relacji) dla wszystkich rozważanych miar *precyzji* (**P**), *kompletności* (**R**) oraz *miary-f* (**F**) dla wszystkich metod klasyfikacji. Można zauważyć, że zdolność do rozpoznawania poszczególnych relacji (lub jej brak) są identyczne jak w *Eksperymencie 2*, w którym w wektorze cech nie były uwzględniane cechy literaturowe.

Testowanie hipotezy o normalności rozkładu uzyskanych wyników z wykorzystaniem testu Shapiro-Wilka

W tabeli 6.13 przedstawione zostały wyniki testu Shapiro-Wilka, sprawdzającego normalność rozkładu uzyskanych wyników wyrażonych za pomocą miar: *precyzji* (**P**),

kompletności (**R**) oraz *miary-f* (**F**). Przedstawione wartości testu Shapiro-Wilka, w wierszach nazwanych *W*, dla każdego klasyfikatora, poza *kompletnością* (kolumna **R**) dla *J48*, są wyższe od wartości krytycznej, równej 0, 842, odczytanej z tablic rozkładu *S-W*. Jedynie dla *kompletności* klasyfikatora *J48*, istnieją podstawy do odrzucenia H_0 i

Tabela 6.13. Wyniki obliczonych statystyk Shapiro–Wilka (*W*) dla wszystkich klasyfikatorów, osobno dla *precyzji* **P**, *kompletności* – **R** oraz *miary-f* – **F**, dla wyników uzyskanych z użyciem wektora cech zbudowanego na podstawie *Zestawu 3*

SVM	**P**	**R**	**F**	**J48**	**P**	**R**	**F**
W	0.91725	0.94272	0.97857	*W*	0.95136	**0.81705**	0.91152
NB	**P**	**R**	**F**	**LMT**	**P**	**R**	**F**
W	0.97502	0.91849	0.88797	*W*	0.97256	0.9024	0.98378

przyjęcia hipotezy H_1, Oznacza to, że dla *kompletności* klasyfikatora *J48* nie ma możliwości zastosowania testu t-Studenta do określenia istotności statystycznej uzyskanych wyników oraz wymusza to zastosowanie nieparametrycznego testu U Manna-Whitneya. Dla pozostałych przypadków nie ma podstaw do odrzucenia H_0 dotyczącej rozkładu normalnego badanej próby, co oznacza spełnienie założenia testu t-Studenta odnośnie wymaganego rozkładu normalnego porównywanych obserwacji.

Testowanie założenia o równości wariancji w porównywanych grupach z wykorzystaniem testu Levene'a

W kolumnach tabeli 6.14 przedstawione zostały wyniki testu Levene'a dla wykorzystanych klasyfikatorów (**NB**, **J48**, **SVM**, **LMT**) osobno dla *precyzji* – **P**, *kompletności* – **R** oraz *miary-f* – **F**. Porównywane są ze sobą średnie ważone (ważenie za pomocą liczby relacji) dla **P**, **R**, **F** obliczone na podstawie wyników klasyfikacji relacji semantycznych między parą zdań z wykorzystaniem wektorów cech zbudowanych na podstawie cech z *Zestawu 2* oraz wektora cech budowanego na bazie *Zestawu 3*. Wszystkie wartości statystyki testowej *p* są wyższe od przyjętego poziomu istotności statystycznej 0, 05, zatem nie ma podstaw do odrzucenia hipotezy zerowej H_0 mówiącej o równości wariancji w porównywanych grupach.

Analiza uzyskanych wyników

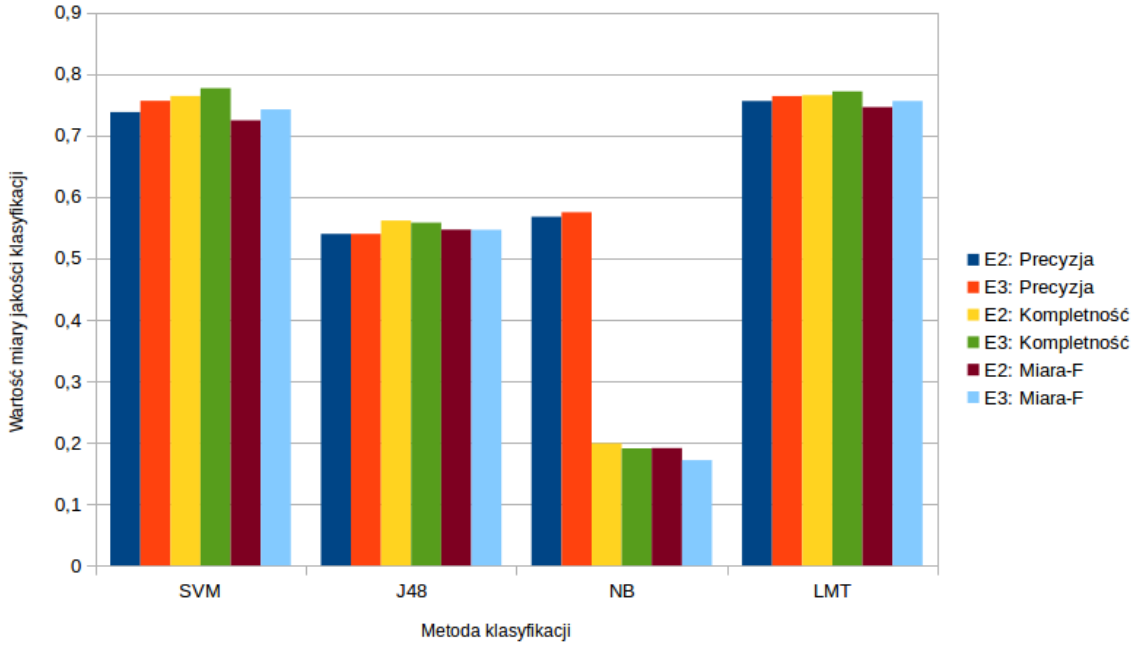
Tabele 6.15, 6.16 oraz 6.17 zawierają wyniki testu t-Studenta dla miar: *precyzji*, *kompletności* oraz *miary-f* dla każdego z klasyfikatorów, którego wyniki spełniają założenia tego testu, z informacją, czy osiągnięty wynik dla danego klasyfikatora jest istotnie statystycznie lepszy. Jedynie dla *kompletności* dla drzewa decyzyjnego *J48*

Tabela 6.14. Wyniki statystyki testowej *p* testu Levene'a dla *precyzji* (**P**), *kompletności* (**R**) oraz *miary-f* (**F**), dla każdego z klasyfikatorów (**NB**, **J48**, **SVM** oraz **LMT**) obliczone dla wyników klasyfikacji z wykorzystaniem cech z *Zestawu 2* oraz *Zestawu 3*

	NB	**J48**	**SVM**	**LMT**
P	0,6445	0,637	0,6764	0,8413
R	0,8312	0,171	0,4427	0,1651
F	0,9046	0,3124	0,3629	0,484

założenie o normalności rozkładu nie zostało spełnione. W tym przypadku został wykonany nieparametryczny test U Manna-Whitneya.

Na rysunku 6.3 porównano wyniki klasyfikacji z eksperymentu drugiego (oznaczonego na rysunku jako *E2*) oraz eksperymentu trzeciego – *E3* dla średnich ważonych *precyzji*, *kompletności* oraz *miary-f*, dla wszystkich wykorzystanych klasyfikatorów (*SVM*, *J48*, *NB* oraz *LMT*). Warto zauważyć fakt, że *kompletność* oraz *miara-f* dla



Rys. 6.3. Zestawienie wyników jakości klasyfikacji dla wektora cech utworzonego na podstawie *Zestawu 2*, uzytego w eksperymencie *E2* oraz wyników dla wektora cech utworzonego na bazie *Zestawu 3* oznaczonych przez *E3*, wyrażonych za pomocą średnich ważonych *precyzji*, *kompletności* i *miary-f*

klasyfikatora *NB*, w eksperymencie *E3* jest niższa niż w *E2*. W tym wypadku, zamiast mówić o istotnej poprawie uzyskanych wyników, należy mówić o istotnym (lub nie) pogorszeniu wyników.

Założenia testu t-Studenta dla jakości mierzonej za pomocą *precyzji* spełniają wyniki wszystkich klasyfikatorów, dlatego znalazły się one w tabeli 6.15. Można zauważyć, że liczbowe wyniki uzyskane przy użyciu wektora cech z *Zestawu 3* są wyższe niż te otrzymane z użyciem wektora cech z *Zestawu 2*. Jednak wartości testu t-Studenta we wszystkich przypadkach są niższe od krytycznej wartości $t = 2,1009$. Zatem nie ma podstaw, aby odrzucić hipotezę zerową H_0 i przyjąć hipotezę alternatywną H_1. Taki wynik możemy zinterpretować jako brak istotnych różnic w uzyskanych wynikach z użyciem obu wektorów cech.

W tabeli 6.16 przedstawione zostały wyniki testu t-Studenta dla jakości klasyfikacji relacji semantycznych między dwoma zdaniami, mierzonej za pomocą *kompletności* dla klasyfikatorów *SVM*, *NB* oraz *LMT*. We wszystkich przypadkach wartość testu była niższa od odczytanej z tablic wartości krytycznej testu $t = 2,1009$. Oznacza

Tabela 6.15. Wartości testu t-Studenta dla *precyzji*, obliczone pomiędzy wynikami uzyskanymi w eksperymencie drugim i trzecim, wykorzystującymi odpowiednio wektor cech z *Zestawu 2* oraz wektor cech z *Zestawu 3*

	Zestaw 2		**Zestaw 3**		**Wynik testu**	
Klasyfikator	**Średnia precyzja**	**Wariancja**	**Średnia precyzja**	**Wariancja**	**Wartość testu**	**Czy wynik jest istotnie lepszy**
SVM	0,738	0,001	0,756	0,001	0,795	**NIE**
J48	0,540	0,001	0,540	0,001	0,000	**NIE**
NB	0,568	0,004	0,575	0,003	0,138	**NIE**
LMT	0,756	0,001	0,764	0,001	0,403	**NIE**

to, że nie ma podstaw do odrzucenia hipotezy zerowej H_0 mówiącej o braku różnic między porównywanymi obserwacjami. Wyniki dla *kompletności* dla *J48* nie spełniają

Tabela 6.16. Wartości testu t-Studenta dla *kompletności*, obliczone pomiędzy wynikami uzyskanymi w eksperymencie drugim i trzecim, z wektorami cechy zbudowanymi odpowiednio na podstawie *Zestawu 2* lub *Zestawu 3*

	Zestaw 2		**Zestaw 3**		**Wynik testu**	
Klasyfikator	**Średnia kompletność**	**Wariancja**	**Średnia kompletność**	**Wariancja**	**Wartość testu**	**Czy wynik jest istotnie lepszy**
SVM	0,764	0,001	0,777	0,001	1,203	**NIE**
NB	0,199	0,001	0,191	0,001	0,624	**NIE**
LMT	0,765	0,001	0,772	0,001	0,406	**NIE**

założenia odnośnie normalności rozkładu.W tym wypadku przeprowadzony został nieparametryczny test U Manna-Whitneya, którego wartość wyniosła 54. Dla przyjętego poziomu istotności $\alpha = 0,05$ otrzymany wynik jest wyższy od poziomu krytycznego wynoszącego 23, co oznacza, że nie ma podstaw do odrzucenia hipotezy zerowej H_0 mówiącej o braku różnic między porównywanymi obserwacjami. Podsumowując, we wszystkich przypadkach przeprowadzonych testów t-Studenta oraz U Manna-Whitneya nie ma statystycznie istotnych różnic między precyzją osiągniętą z wykorzystaniem cech z *Zestawu 2*, a precyzją osiągniętą za pomocą cech z *Zestawu 3*.

Założenia testu t-Studenta dla jakości mierzonej za pomocą *miary-f* spełniają wyniki wszystkich klasyfikatorów. Zostały one ujęte w tabeli 6.17. W każdym wypadku

Tabela 6.17. Wartości testu t-Studenta dla *miary-f*, obliczone pomiędzy wynikami uzyskanymi w eksperymencie drugim i trzecim, wykorzystującymi wektory cech z *Zestawu 2* oraz *Zestawu 3*

	Zestaw 2		**Zestaw 3**		**Wynik testu**	
Klasyfikator	**Średnia miara-f**	**Wariancja**	**Średnia miara-f**	**Wariancja**	**Wartość testu**	**Czy wynik jest istotnie lepszy**
SVM	0,724	0,001	0,742	0,001	1,502	**NIE**
J48	0,547	0,001	0,546	0,001	0,021	**NIE**
NB	0,191	0,001	0,171	0,001	1,236	**NIE**
LMT	0,746	0,001	0,756	0,001	0,634	**NIE**

wartość przeprowadzonego testu była poniżej wartości krytycznej $t = 2,1009$. Zatem

nie ma podstaw do odrzucenia H_0 mówiącej o braku różnic między porównywanymi wynikami. Średnie ważone wartości *miary-f* otrzymane w eksperymencie 3 są niższe od wyników z eksperymentu 2. Jednak biorąc pod uwagę wyniki przeprowadzonych testów statystycznych, nie są one istotnie niższe.

Podsumowując, w żadnym przypadku nie było istotnej statystycznie różnicy między wynikami mierzonymi za pomocą *precyzji*, *kompletności* oraz *miary-f* z Eksperymentu 2 w porównaniu z wynikami osiągniętymi za pomocą cech z *Zestawu 3*.

Meta-analiza uzyskanych wyników – określanie wielkości efektu

W kolejnych wierszach tabeli 6.18 przedstawione zostały wyniki obliczonych statystyk r oraz z dla miar obliczonych na podstawie wyników uzyskiwanych przez klasyfikatory: (*SVM*, *J48*, *NB*, *LMT)*. W kolumnach **Wrt.** oznaczającej wartość statystyki – na niebiesko oznaczone zostały wyniki statystyki z, na czarno wyniki statystyki r. Kolumna **Wlk.** zawiera interpretację wartości statystyk r oraz z zgodnie z propozycją Cohena (Cohen, 1988). Kolumna **P** zawiera wyniki obliczania wielkości efektu dla *precyzji*, kolumna **R** dla *kompletności*, a kolumna **F** dla *miary-f*.

Tabela 6.18. Wartości statystyk (kolumny **Wrt.**), dla statystyki r (czarny kolor czcionki) i z (niebieski kolor czcionki) oraz interpretacja w postaci wielkości różnicy (kolumny **Wlk.**) między wynikami uzyskanymi w *Eksperymencie 2* i *Eksperymencie 3* dla *precyzji* (**P**), *kompletności* (**R**) oraz *miary-f* (**F**), dla klasyfikatorów *SVM*, *J48*, *NB* oraz *LMT*

	P		**R**		**F**	
	Wrt.	**Wlk.**	**Wrt.**	**Wlk.**	**Wrt.**	**Wlk.**
SVM	0,184	Mała	0,273	Mała	0,334	Średnia
J48	0,000	Mała	0,093	Mała	0,005	Mała
NB	0,032	Mała	0,146	Mała	0,280	Mała
LMT	0,095	Mała	0,095	Mała	0,148	Mała

Wielkości efektu dla *precyzji* (kolumna **P**), *kompletności* (kolumna **R**) dla wszystkich klasyfikatorów są *małe*. Innymi słowy, różnica między wynikami uzyskanymi w *Eksperymencie 2* oraz *Eksperymencie 3* jest *mała*. Ten wynik możemy podsumować jako brak statystycznie istotnych różnic pomiędzy wynikami dla miary *precyzji* otrzymanymi za pomocą wektora cech z *Zestawu 1* oraz wynikami uzyskanymi dla wektora cech z *Zestawu 2*.

Podobne stwierdzenie możemy sformułować w wypadku *kompletności* (kolumna **F**). Dla klasyfikatorów *J48*, *NB* oraz *LMT* jest *mała* różnica pomiędzy wynikami mierzonymi za pomocą *miary-f*, osiągniętymi za pomocą wektora cech z *Zestawu 2* oraz wektora cech z *Zestawu 3*. Jedynie wartość testu r dla *miary-f* dla klasyfikatora *SVM* mieści się w przedziale liczbowym, który interpretowany jest jako średnia wielkość efektu. Co oznacza, że wyniki mierzone za pomocą *miary-f* osiągnięte za pomocą wektora cech z *Zestawu 2* są średnio różne od wyników osiągniętych za pomocą wektora cech z *Zestawu 3*. W tym wypadku test istotności statystycznej nie wykazał istotnej różnicy w osiągniętych wynikach.

6.2.4. Eksperyment 4. Rozpoznawanie relacji semantycznych między dwoma zdaniami z wykorzystaniem cech z *Zestawu 4*

Cel badania

Badanie miało na celu określenie jakości klasyfikacji z użyciem wektora wejściowego z *Zestawu 4*, opisanego w rozdziale 5.3.4 oraz porównanie wyników osiągniętych z jego wykorzystaniem z wynikami osiągniętymi w *Eksperymencie 2*.

Otrzymane wyniki

W tabeli 6.19 przedstawione zostały szczegółowe wyniki klasyfikacji z wykorzystaniem `cech grafowych (2)` opisanych jako *Zestaw 4*. W tabeli tej przedstawione zostały wyniki dla każdego wykorzystanego klasyfikatora: naiwnego klasyfikatora Bayesa – **NB**, drzewa decyzyjnego uczonego algorytmem C.45 – **J48**, maszyny wektorów nośnych – **SVM** oraz drzewa modeli logistycznych – **LMT**.

Tabela 6.19. Wyniki jakości klasyfikacji relacji semantycznych między dwoma zdaniami, mierzonej za pomocą *precyzji* (**P**), *kompletności* (**R**) oraz *miary-f* (**F**) z wykorzystaniem cech z *Zestawu 4*, dla czterech klasyfikatorów: **NB**, **J48**, **SVM** oraz **LMT**

NB			**J48**			**SVM**			**LMT**			
P	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**	**Klasa**
0,000	0,000	0,000	0,000	0,000	0,000	0,100	0,100	0,100	0,000	0,000	0,000	Cytowanie
0,300	0,011	0,021	0,284	0,149	0,192	0,877	0,965	0,918	0,682	0,923	0,783	Dalsze informacje
0,994	0,419	0,589	0,889	0,947	0,917	0,971	0,988	0,980	0,977	0,990	0,983	Krzyżowanie się
0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	Modalność
0,054	0,450	0,096	0,000	0,000	0,000	0,183	0,250	0,207	0,325	0,350	0,307	Mowa zależna
0,567	0,808	0,664	0,345	0,599	0,435	0,782	0,879	0,827	0,862	0,890	0,874	Opis
0,262	0,605	0,359	0,236	0,155	0,164	0,580	0,450	0,494	0,808	0,530	0,605	Parafraza
0,475	0,090	0,148	0,111	0,078	0,090	0,763	0,590	0,661	0,815	0,640	0,714	Spełnienie
0,071	0,700	0,128	0,000	0,000	0,000	0,500	0,300	0,367	0,150	0,150	0,150	Sprzeczność
0,055	0,747	0,103	0,000	0,000	0,000	0,503	0,477	0,468	0,723	0,597	0,613	Streszczenie
0,425	0,681	0,519	0,257	0,305	0,277	0,743	0,788	0,764	0,763	0,826	0,791	Tło historyczne
0,402	0,917	0,551	0,520	0,767	0,606	0,885	0,892	0,883	0,975	0,867	0,907	Tożsamość
0,100	0,009	0,017	0,162	0,044	0,060	0,756	0,545	0,599	0,739	0,401	0,506	Uszczegółowienie
0,851	0,370	0,503	0,300	0,198	0,236	0,748	0,707	0,724	0,812	0,712	0,755	Zawieranie
0,005	0,100	0,010	0,000	0,000	0,000	0,000	0,000	0,000	0,300	0,300	0,300	Zmiana poglądu
0,600	0,200	0,294	0,400	0,105	0,164	0,773	0,505	0,597	0,777	0,550	0,627	Źródło
0,971	0,461	0,620	0,403	0,385	0,384	0,925	0,899	0,910	0,902	0,893	0,895	Brak relacji
0,713	**0,431**	**0,473**	**0,548**	**0,580**	**0,551**	**0,862**	**0,866**	**0,858**	**0,869**	**0,868**	**0,860**	**Średnia ważona**

Analizując wyniki przedstawione w tabeli 6.19 warto zwrócić uwagę, że większość relacji jest rozpoznawana przez *LMT*. Jedynym wyjątkiem jest relacja *Modalności*, dla której istnieje tylko jeden przykład uczący. Warto też zauważyć, że dla klasyfikatora *SVM* relacja *Cytowania* została jednokrotnie rozpoznana (wartość $0,1$ dla P, R oraz F). Najniższą wartość *precyzji* rozpoznawanych relacji osiągnęło drzewo decyzyjne *J48*, najniższą wartość *kompletności* oraz *miary-f* – naiwny klasyfikator bayesowski. Najwyższe wartości, jeżeli chodzi o *precyzję*, *kompletność* oraz *miarę-f* osiągnął klasyfikator *LMT*, jednak różnica względem *SVM* jest niewielka. W dalszej części niniejszego rozdziału przedstawione zostały badania odnośnie istotności różnicy w otrzymanych wynikach.

Testowanie hipotezy o normalności rozkładu uzyskanych wyników z wykorzystaniem testu Shapiro-Wilka

W tabeli 6.20 przedstawione zostały wyniki testu Shapiro-Wilka, sprawdzającego normalność rozkładu uzyskanych wyników wyrażonych za pomocą *precyzji* (**P**), *kompletności* (**R**) oraz *miary-f* (**F**). Otrzymane wartości testu Shapiro-Wilka oznaczone

Tabela 6.20. Wyniki obliczonych statystyk Shapiro–Wilka (W) dla wszystkich klasyfikatorów, osobno dla *precyzji* **P**, *kompletności* – **R** oraz *miary-f* – **F**, dla wyników uzyskanych z wykorzystaniem cech z *Zestawu 4*

SVM	**P**	**R**	**F**	**J48**	**P**	**R**	**F**
W	0.94264	0.97649	0.94561	W	0.94404	0.93562	0.96662
NB	**P**	**R**	**F**	**LMT**	**P**	**R**	**F**
W	0.96662	0.90907	0.96348	W	0.96405	0.97515	0.97028

za pomocą W, dla każdego klasyfikatora są wyższe od wartości krytycznej $W = 0,842$. Oznacza to, że nie ma podstaw do odrzucenia H_0 mówiącej o istnieniu rozkładu normalnego badanej próby. Możemy więc przyjąć, że spełnione jest założenie testu t-Studenta odnośnie wymagania dotyczącego rozkładu normalnego porównywanych obserwacji.

Testowanie założenia o równości wariancji w porównywanych grupach z wykorzystaniem testu Levene'a

W kolumnach tabeli 6.21 przedstawione zostały wyniki testu Levene'a dla wykorzystanych klasyfikatorów (**NB**, **J48**, **SVM**, **LMT**) osobno dla *precyzji* – **P**, *kompletności* – **R** oraz *miary-f* – **F**. Porównywane są ze sobą średnie ważone (ważenie za pomocą liczby relacji) **P**, **R**, **F** wyników klasyfikacji z wykorzystaniem cech z *Zestawu 2* oraz cech z *Zestawu 4*. Wszystkie wartości obliczonej statystyki testowej p przekraczają

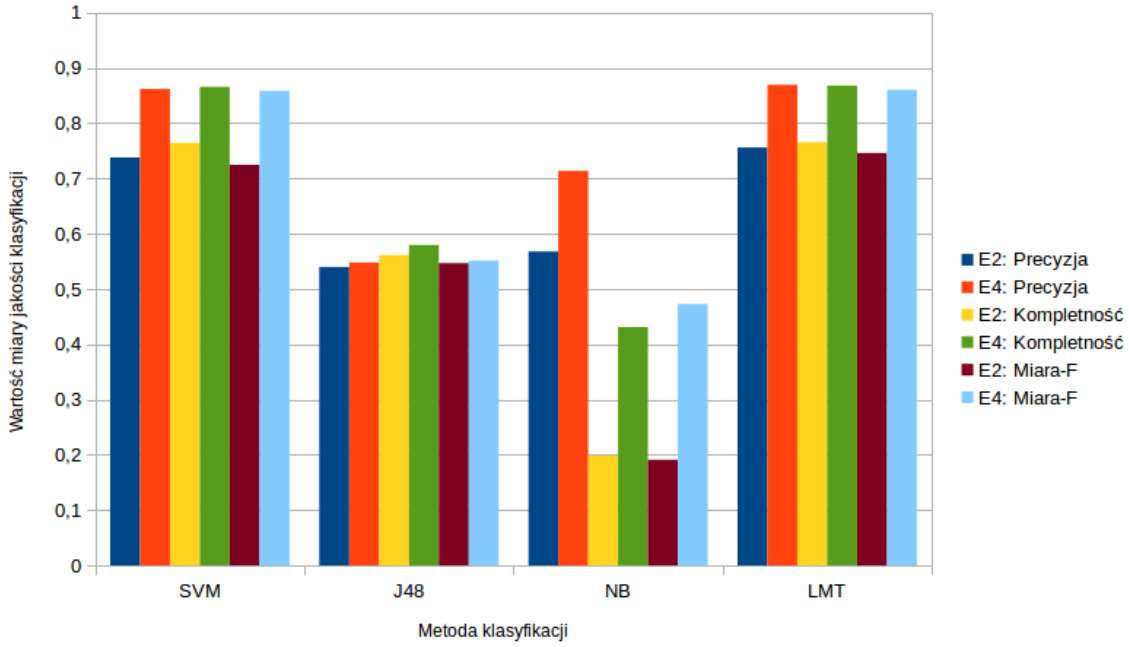
Tabela 6.21. Wyniki statystyki testowej p testu Levene'a dla *precyzji* (**P**), *kompletności* (**R**) oraz *miary-f* (**F**), dla każdego z klasyfikatorów (**NB**, **J48**, **SVM** oraz **LMT**) obliczone dla wyników klasyfikacji z wykorzystaniem cech z *Zestawu 2* oraz *Zestawu 4*

	NB	**J48**	**SVM**	**LMT**
P	0,8117	0,3515	0,9882	0,6713
R	0,1335	0,6395	0,3103	0,8959
F	0,7196	0,9057	0,3489	0,5964

zadany poziom istotności statystycznej $\alpha = 0,05$. Co oznacza, że nie ma podstaw do odrzucenia hipotezy H_0, która mówi o jednorodności wariancji wyników w porównywanych grupach.

Analiza uzyskanych wyników

Na rysunku 6.3 porównane zostały ze sobą wyniki klasyfikacji z eksperymentu drugiego (nazwanego jako *E2*) oraz eksperymentu czwartego – *E4* dla średnich ważonych *precyzji*, *kompletności* oraz *miary-f*, dla wszystkich wykorzystanych klasyfikatorów (*SVM*, *J48*, *NB* oraz *LMT*). Dla każdego klasyfikatora oraz wszystkich miar jakości P, R oraz F wyniki osiągnięte za pomocą cech z *Zestawu 4* są wyższe od



Rys. 6.4. Zestawienie wyników jakości klasyfikacji dla cech z *Zestawu 2*, oznaczonych jako *E2* oraz cech z *Zestawu 4* oznaczonych jako *E4*, wyrażonych za pomocą średnich ważonych *precyzji*, *kompletności* i *miary-f*

wyników osiągniętych z wykorzystaniem cech z *Zestawu 2*. Warto jednak zauważyć, że wykorzystanie cech z *Zestawu 4* do trenowania klasyfikatora *J48*, nie przyniosło tak zauważalnej poprawy jak w przypadku pozostałych klasyfikatorów. Może to świadczyć o zbyt dużej liczbie cech w wektorze wejściowym, z którą klasyfikator *J48* nie potrafił sobie poradzić.
Przeprowadzone testy sprawdzające normalność rozkładu oraz równość porównywanych grup wykazały, że założenia testu t-Studenta, do oceny istotności uzyskanych wyników są spełnione. W tabelach 6.22, 6.23 oraz 6.24 przedstawione zostały wyniki testu t-Studenta dla *precyzji*, *kompletności* oraz *miary-f* dla każdego z klasyfikatorów.

Tabela 6.22 zawiera wyniki testu t-Studenta dla *precyzji*, obliczone dla każdego z wykorzystywanych klasyfikatorów. W każdym przypadku można zaobserwować poprawę wyników (kolumny **Średnia precyzja**, w **Zestaw 2** oraz **Zestaw 4**). Jednak wartość statystyki dla testu t-Studenta dla klasyfikatora *J48*, wyniosła $0,269$ jest wartością niższą od poziomu krytycznego testu $t = 2,1009$. W pozostałych wypadkach wartości testu t-Studenta są wyższe od poziomu krytycznego. Co oznacza, że jedynie dla *J48* nie ma podstaw do odrzucenia hipotezy zerowej H_0 mówiącej o braku różnic między badanymi próbami, czyli otrzymane wyniki nie są istotne statystycznie. W pozostałych przypadkach wartość testu t-Studenta była wyższa od poziomu $t = 2,1009$, zatem istnieją podstawy do odrzucenia H_0, a przyjęcia hipotezy alternatywnej H_1. Taki efekt pozwala stwierdzić, że jakość rozpoznawania relacji semantycznych z wykorzystaniem cech z *Zestawu 4*, mierzona za pomocą *precyzji*, dla *NB*, *SVM* oraz *LMT* jest istotnie statystycznie wyższa wszystkich badanych klasyfikatorów, oprócz *J48*.

Tabela 6.22. Wartości testu t-Studenta dla *precyzji*, obliczone pomiędzy wynikami uzyskanymi w eksperymencie drugim i czwartym, wykorzystującymi cechy z *Zestawu 2* oraz *Zestawu 4*

	Zestaw 2		**Zestaw 4**		**Wynik testu**	
Klasyfikator	**Średnia precyzja**	**Wariancja**	**Średnia precyzja**	**Wariancja**	**Wartość testu**	**Czy wynik jest istotnie lepszy**
SVM	0,738	0,001	0,862	0,001	4,730	TAK
J48	0,540	0,001	0,548	0,001	0,296	**NIE**
NB	0,568	0,004	0,713	0,004	2,438	TAK
LMT	0,756	0,001	0,869	0,001	4,769	TAK

Tabela 6.23 zawiera wartości testu t-Studenta dla *kompletności*, obliczone dla każdego z wykorzystywanych klasyfikatorów. W każdym przypadku można zaobserwować poprawę wyników (kolumny **Średnia kompletność**, w **Zestaw 2** oraz **Zestaw 4**). Podobnie jak w przypadku *precyzji*, tak i dla *kompletności*, wynik testu t-Studenta dla

Tabela 6.23. Wartości statystyki testu t-Studenta dla *kompletności*, obliczone pomiędzy wynikami uzyskanymi w eksperymencie drugim i trzecim, wykorzystującymi wektory cech zbudowane na podstawie z *Zestawu 2* oraz *Zestawu 4*

	Zestaw 2		**Zestaw 4**		**Wynik testu**	
Klasyfikator	**Średnia kompletność**	**Wariancja**	**Średnia kompletność**	**Wariancja**	**Wartość testu**	**Czy wynik jest istotnie lepszy**
SVM	0,764	0,001	0,866	0,001	4,172	TAK
J48	0,561	0,001	0,580	0,001	0,924	**NIE**
NB	0,199	0,001	0,431	0,001	10,606	TAK
LMT	0,765	0,001	0,868	0,001	5,000	TAK

J48 o wartości $0,924$ jest niższy od poziomu krytycznego testu $t = 2,1009$. W pozostałych przypadkach wartości testu t-Studenta są wyższe od poziomu krytycznego. W przypadku *J48* nie ma podstaw do odrzucenia hipotezy zerowej H_0 mówiącej o braku różnic między badanymi próbami, czyli otrzymane wyniki nie są istotne statystycznie. Dla pozostałych klasyfikatorów, wartość testu t-Studenta była wyższa od poziomu $t = 2,1009$, zatem istnieją podstawy do odrzucenia H_0, a przyjęcia hipotezy alternatywnej H_1. Możemy zatem wnioskować, że jakość rozpoznawania relacji semantycznych z wykorzystaniem cech z *Zestawu 4*, mierzona za pomocą *kompletności*, dla *NB*, *SVM* oraz *LMT* jest istotnie statystycznie wyższa.

Tabela 6.24 zawiera wartości statystyki testu t-Studenta dla *miary-f*, przeprowadzonego dla każdego z wykorzystywanych klasyfikatorów. W każdym przypadku można zaobserwować poprawę wyników (kolumny **Średnia miara-f**, w **Zestaw 2** oraz **Zestaw 4**). Analogicznie do *precyzji* oraz *kompletności*, dla *miary-f* wynik testu t-Studenta dla *J48* o wartości $0,231$ jest niższy od poziomu krytycznego testu $t = 2,1009$. Dla *J48* nie ma podstaw do odrzucenia hipotezy zerowej H_0 mówiącej o braku różnic między badanymi próbami. Oznacza to, że otrzymane wyniki nie są istotne statystycznie. W pozostałych przypadkach wartości testu t-Studenta są wyższe od poziomu krytycznego $t = 2,1009$. Zatem dla *NB*, *SVM* oraz *LMT* istnieją podstawy do odrzucenia H_0, a przyjęcia hipotezy alternatywnej H_1. Oznacza to, że jakość rozpoznawania relacji

Tabela 6.24. Wartości testu t-Studenta dla *miary-f*, obliczonymi pomiędzy wynikami uzyskanymi w eksperymencie drugim i trzecim, wykorzystującymi cechy z *Zestawu 2* oraz *Zestawu 4*

	Zestaw 2		**Zestaw 4**		**Wynik testu**	
Klasyfikator	**Średnia miara-f**	**Wariancja**	**Średnia miara-f**	**Wariancja**	**Wartość testu**	**Czy wynik jest istotnie lepszy**
SVM	0,724	0,001	0,858	0,001	4,847	TAK
J48	0,547	0,001	0,551	0,001	0,231	**NIE**
NB	0,191	0,001	0,473	0,001	2,438	TAK
LMT	0,746	0,001	0,860	0,001	5,128	TAK

semantycznych z wykorzystaniem cech z *Zestawu 4*, mierzona za pomocą *miary-f* dla tych klasyfikatorów jest istotnie statystycznie wyższa.

Meta-analiza uzyskanych wyników – określanie wielkości efektu

W kolejnych wierszach tabeli 6.25 przedstawiono wyniki obliczonych statystyk r dla wykorzystanych klasyfikatorów (*SVM*, *J48*, *NB*, *LMT)*. Kolumna **Wrt.** oznacza wartość statystyki r. Kolumna **Wlk.** zawiera interpretację wartości statystyki r zgodnie z propozycją Cohena (Cohen, 1988). Kolumna **P** zawiera wyniki obliczania wielkości efektu dla *precyzji*, kolumna **R** dla *kompletności*, a kolumna **F** dla *miary-f*.

Tabela 6.25. Wartości statystyki r (kolumny **Wrt.**) oraz interpretacja w postaci wielkości różnicy (kolumny **Wlk.**) między wynikami uzyskanymi w *Eksperymencie 2* i *Eksperymencie 4* dla *precyzji* (**P**), *kompletności* (**R**) oraz *miary-f* (**F**), dla klasyfikatorów *SVM*, *J48*, *NB* oraz *LMT*

	P		**R**		**F**	
	Wrt.	**Wlk.**	**Wrt.**	**Wlk.**	**Wrt.**	**Wlk.**
SVM	0,744	Duża	0,701	Duża	0,752	Duża
J48	0,070	Mała	0,213	Mała	0,054	Mała
NB	0,498	Średnia	0,928	Duża	0,974	Duża
LMT	0,747	Duża	0,762	Duża	0,770	Duża

Między wynikami mierzonymi za pomocą *precyzji* (kolumna **P**) dla klasyfikatorów *SVM* oraz *LMT* występuje duża różnica. Wartość statystyki r dla *precyzji* klasyfikatora *J48* jest interpretowana jako mała różnica dla wyników klasyfikacji mierzonych *precyzją*, z użyciem wektorów cech z *Zestawu 2* oraz *Zestawu 4*. Dla klasyfikatora *NB* wartość statystyki r należy zinterpretować jako średnią różnicę między wynikami. Zauważmy jednak, że wartość ta (równa $0,498$) jest bardzo blisko wartości granicznej $0,5$, która w interpretacji Cohena wskazuje na dużą wielkość efektu. Warto dodać, że wartość testu t-Studenta dla *precyzji* klasyfikatora *NB* wyniosła $2,43$, co jest wartością niewiele wyższą od wartości krytycznej testu $t = 2,1009$.

W przypadku *kompletności* (kolumna **R**), dla *SVM*, *J48* oraz *LMT* jest *duża* różnica między wynikami. Dla *J48* ta różnica jest *mała*. Wielkości efektu korelują z wynikami istotności statystycznej, przedstawionymi w tabeli 6.23. W wynikach klasyfikacji mierzonych za pomocą *kompletności*, dla klasyfikatorów *SVM*, *NB* oraz *LMT* osiągniętych przy użyciu wektorów cech budowanych na podstawie *Zestawu 2* oraz *Zestawu 4* występują istotne statystycznie różnice. Dla klasyfikatora *J48* nie ma istotnej różnicy w

osiągniętych wynikach. Dla *kompletności* klasyfikatora *J48*, wielkość efektu koreluje z wartością testu t-Studenta.

Interpretacja wielkości różnicy między wynikami mierzonymi za pomocą *miary-f* (kolumna **F**) jest podobna do interpretacji *kompletności*. Dla *SVM*, *J48* oraz *LMT* jest *duża* różnica między wynikami. Dla *J48* różnica ta jest *mała*. Wielkości efektu korelują z wynikami istotności statystycznej, przedstawionymi w tabeli 6.24. Między wynikami mierzonymi za pomocą *miary-f*, dla klasyfikatorów *SVM*, *NB* oraz *LMT* osiągniętymi za pomocą cech z *Zestawu 2* oraz *Zestawu 4* są istotne statystycznie różnice. Takie stwierdzenie nie jest prawdziwe dla klasyfikatora *J48*.

6.2.5. Eksperyment 5. Rozpoznawanie relacji semantycznych między dwoma zdaniami z wykorzystaniem wektorów cech z *Zestawu 5*

Cel badania

Badanie miało na celu określenie jakości klasyfikacji z wykorzystaniem wektora cech bazującego na *Zestawie 5* (rozdział 5.3.5) oraz porównanie otrzymanych wyników z wynikami osiągniętymi w *Eksperymencie 4*.

Otrzymane wyniki

W tabeli 6.26 przedstawione zostały wyniki klasyfikacji z wykorzystaniem wektora cech zbudowanego na podstawie cech z *Zestawu 5*. Tabela ta przedstawia wyniki dla każdego wykorzystanego klasyfikatora ujmując wszystkie rozważane miary. W kolejnych wierszach znajdują się wyniki dla relacji semantycznych, zaś ostatni wiersz zawiera średnią ważoną *precyzji*, *kompletności* oraz *miary-f* dla wykorzystanych metod klasyfikacji.

Tabela 6.26. Wyniki jakości klasyfikacji relacji semantycznych między dwoma zdaniami, mierzonej za pomocą *precyzji* (**P**), *kompletności* (**R**) oraz *miary-f* (**F**) z wykorzystaniem cech z *Zestawu 5*, dla czterech klasyfikatorów: **NB**, **J48**, **SVM** oraz **LMT**

NB			**J48**			**SVM**			**LMT**			
P	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**	**Klasa**
0,000	0,000	0,000	0,000	0,000	0,000	0,100	0,100	0,100	0,100	0,100	0,100	Cytowanie
0,200	0,011	0,020	0,202	0,175	0,185	0,877	0,965	0,919	0,679	0,895	0,770	Dalsze informacje
0,992	0,401	0,567	0,881	0,928	0,903	0,976	0,992	0,984	0,976	0,995	0,985	Krzyżowanie się
0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	Modalność
0,057	0,600	0,102	0,000	0,000	0,000	0,217	0,300	0,247	0,517	0,500	0,457	Mowa zależna
0,556	0,799	0,654	0,373	0,509	0,423	0,797	0,879	0,834	0,863	0,893	0,876	Opis
0,250	0,640	0,353	0,265	0,200	0,205	0,665	0,495	0,531	0,832	0,645	0,692	Parafraza
0,441	0,116	0,180	0,116	0,088	0,099	0,797	0,623	0,695	0,831	0,622	0,705	Spełnienie
0,066	0,700	0,121	0,000	0,000	0,000	0,400	0,250	0,300	0,233	0,250	0,240	Sprzeczność
0,055	0,747	0,102	0,067	0,040	0,047	0,440	0,413	0,409	0,622	0,507	0,551	Streszczenie
0,403	0,673	0,502	0,307	0,313	0,305	0,762	0,830	0,792	0,756	0,817	0,783	Tło historyczne
0,400	0,950	0,559	0,532	0,615	0,551	0,873	0,950	0,900	0,975	0,875	0,918	Tożsamość
0,100	0,008	0,015	0,031	0,027	0,028	0,692	0,574	0,617	0,672	0,417	0,504	Uszczegółowienie
0,872	0,368	0,499	0,337	0,301	0,310	0,768	0,689	0,724	0,809	0,743	0,771	Zawieranie
0,008	0,100	0,015	0,000	0,000	0,000	0,100	0,100	0,100	0,233	0,300	0,250	Zmiana poglądu
0,400	0,090	0,147	0,000	0,000	0,000	0,682	0,470	0,530	0,763	0,390	0,488	Źródło
0,956	0,402	0,551	0,482	0,427	0,447	0,948	0,917	0,930	0,924	0,887	0,902	Brak relacji
0,697	**0,420**	**0,457**	**0,542**	**0,570**	**0,551**	**0,862**	**0,873**	**0,866**	**0,868**	**0,868**	**0,860**	**Średnia ważona**

Oczywistym jest, że relacja *Modalność*, ze względu na pojedyncze wystąpienie w zbiorze uczącym, nie została rozpoznana przez żaden z klasyfikatorów. Klasyfikatory

NB oraz *J48*, wykazują taką samą tendencję do rozpoznawania oraz nierozpoznawania relacji semantycznych między dwoma zdaniami jak w przypadku wektorów cech opisanych *Zestawem 4*. Klasyfikator *LMT* dla relacji *Cytowanie* jednokrotnie poprawnie rozpoznał tę relację, natomiast *SVM* dla relacji *Zmiana poglądu* w jednym na 10 podziałów dobrze rozpoznał wszystkie instancje tej relacji.

Testowanie hipotezy o normalności rozkładu uzyskanych wyników z wykorzystaniem testu Shapiro-Wilka

W tabeli 6.27 przedstawione zostały wyniki testu Shapiro-Wilka, sprawdzające normalność rozkładu uzyskanych wyników. Przedstawione wartości testu Shapiro-Wilka, w wierszach oznaczonych jako *W*, dla każdego klasyfikatora, poza *precyzją* (kolumna **P**) dla *NB*, są wyższe od wartości krytycznej, równej $0,842$, odczytanej z tablic rozkładu *S-W*. Dla *precyzji* klasyfikatora *NB*, istnieją podstawy do odrzucenia H_0 i przyjęcia

Tabela 6.27. Wyniki obliczonych statystyk Shapiro–Wilka (W) dla wszystkich klasyfikatorów, osobno dla *precyzji* **P**, *kompletności* – **R** oraz *miary-f* – **F**, dla wyników uzyskanych z wykorzystaniem cech z *Zestawu 5*

SVM	**P**	**R**	**F**	**J48**	**P**	**R**	**F**
W	0.99077	0.9739	0.96698	*W*	0.95772	0.86707	0.90856
NB	**P**	**R**	**F**	**LMT**	**P**	**R**	**F**
W	**0.82227**	0.90232	0.90391	*W*	0.91589	0.89391	0.91262

hipotezy H_1, co oznacza niespełnienie założenia o istnieniu rozkładu normalnego dla wyników mierzonych za pomocą *precyzji* dla *NB*. W konsekwencji wymusza to zastosowanie nieparametrycznego testu U Manna-Whitneya. Dla pozostałych przypadków nie ma podstaw do odrzucenia H_0 dotyczącej rozkładu normalnego badanej próby.

Testowanie założenia o równości wariancji w porównywanych grupach wyników z wykorzystaniem testu Levene'a

W kolumnach tabeli 6.28 przedstawione zostały wyniki testu Levene'a dla wykorzystanych klasyfikatorów (**NB**, **J48**, **SVM**, **LMT**) osobno dla *precyzji* – **P**, *kompletności* – **R** oraz *miary-f* – **F**. Porównywane są ze sobą średnie ważone (ważenie za pomocą liczby

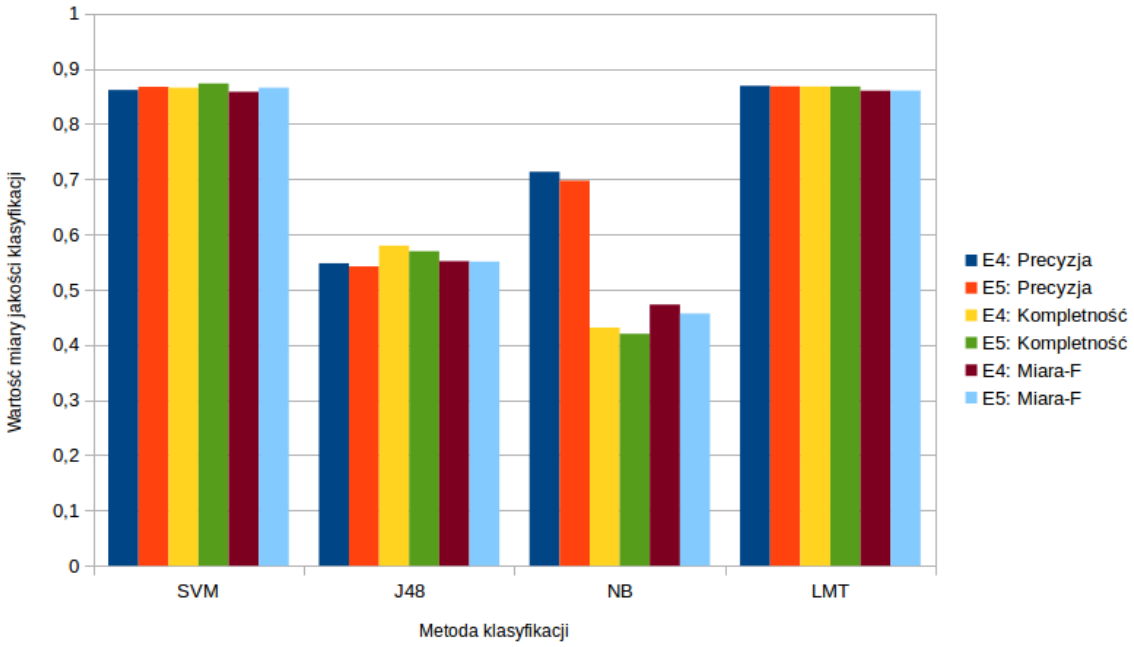
Tabela 6.28. Wyniki statystyki testowej p testu Levene'a dla *precyzji* (**P**), *kompletności* (**R**) oraz *miary-f* (**F**), dla każdego z klasyfikatorów (**NB**, **J48**, **SVM** oraz **LMT**) obliczone dla wyników klasyfikacji z wykorzystaniem wektorów cech z *Zestawu 4* oraz *Zestawu 5*

	NB	**J48**	**SVM**	**LMT**
P	0,5044	0,2056	0,68	0,7849
R	0,3948	0,8882	0,2537	0,8368
F	0,0776	0,8159	0,3454	0,8138

relacji) **P**, **R**, **F** klasyfikacji otrzymane przy użyciu cech z *Zestawu 4* oraz *Zestawu 5*. Wszystkie wartości statystyki testowej p są wyższe od przyjętego poziomu istotności statystycznej równej $0,05$, zatem nie ma podstaw do odrzucenia hipotezy zerowej H_0 mówiącej o jednorodności wariancji w porównywanych grupach.

Analiza uzyskanych wyników

Na rysunku 6.5 porównane zostały ze sobą wyniki klasyfikacji z eksperymentu czwartego (nazwanego jako *E4*) oraz eksperymentu piątego – *E5* dla średnich ważonych (za pomocą liczności relacji w zbiorze uczącym) *precyzji*, *kompletności* oraz *miary-f*, dla wszystkich wykorzystanych klasyfikatorów (*SVM*, *J48*, *NB* oraz *LMT*). Warto podkreślić, że wyniki osiągnięte za pomocą cech z *Zestawu 5*, dla *NB*, *J48* oraz



Rys. 6.5. Zestawienie wyników jakości klasyfikacji relacji semantycznych z użyciem wektorów cech z *Zestawu 4*, oznaczonych na rysunku jako *E4* oraz cech z *Zestawu 5* oznaczonych jako *E5*, wyrażonych za pomocą średnich ważonych *precyzji*, *kompletności* i *miary-f*

LMT są niższe w porównaniu z wynikami osiągniętymi za pomocą cech z *Zestawu 4*. Przeprowadzone testy sprawdzające normalność rozkładu oraz równość wariancji porównywanych grup wykazały, że założenia testu t-Studenta, dotyczące oceny istotności statystycznej uzyskanych wyników są spełnione we wszystkich przypadkach, poza jednym – wyniki *precyzji* dla klasyfikatora *NB* nie posiadają rozkładu normalnego. Dla tego przypadku przeprowadzony został nieparametryczny test U Manna-Whitneya, dla pozostałych – w tabelach 6.29, 6.30 oraz 6.31 przedstawione zostały wyniki testu t-Studenta dla *precyzji*, *kompletności* oraz *miary-f*.

W tabeli 6.29 przedstawione zostały wyniki testu t-Studenta dla *precyzji* dla *J48*, *SVM* oraz *LMT*. Warto zwrócić uwagę, że średnia ważona precyzji (kolumna **Średnia precyzja**), dla *J48* oraz *LMT* jest niższa dla *Zestawu 5*. Wartość testu t-Studenta dla tych przypadków jest niższa od poziomu krytycznego testu $t = 2,1009$, zatem nie ma podstaw do odrzucenia hipotezy zerowej H_0 mówiącej o braku różnic w porównywanych ze sobą próbach. Uzyskane wyniki dla *J48* oraz *LMT* należy interpretować jako nieistotnie niższe. W przypadku *SVM* uzyskany wynik *precyzji* przy użyciu cech z *Zestawu 5* jest wyższy od *precyzji* uzyskanej za pomocą cech z *Zestawu 4*. Wartość

Tabela 6.29. Wartości testu t-Studenta dla *precyzji*, obliczone dla wyników klasyfikacji uzyskanych w eksperymencie czwartym, wykorzystującym wektory cech z *Zestawu 4* oraz w eksperymencie piątym z wektorem cech z *Zestawu 5*

	Zestaw 4		**Zestaw 5**		**Wynik testu**	
Klasyfikator	**Średnia precyzja**	**Wariancja**	**Średnia precyzja**	**Wariancja**	**Wartość testu**	**Czy wynik jest istotnie lepszy**
SVM	0,862	0,001	0,867	0,001	0,216	**NIE**
J48	0,548	0,001	0,542	0,001	0,212	**NIE**
LMT	0,869	0,001	0,868	0,001	0,050	**NIE**

testu t-Studenta dla tego przypadku jest jednak niższa od wartości krytycznej testu $t = 2,1009$. Oznacza to, że w przypadku poprawy *precyzji* dla klasyfikatora *SVM*, nie można mówić o istotnej poprawie, ponieważ nie ma podstaw do odrzucenia hipotezy H_0, mówiącej o braku różnic między badanymi grupami. Dla klasyfikatora *NB* porównanie *precyzji* osiągniętej przy użyciu wektora cech z *Zestawu 4* oraz wektora cech z *Zestawu 5* odbyło się za pomocą nieparametrycznego testu U Manna-Whitneya. Obliczona wartość statystyki U testu U Manna Whitneya wyniosła $55,5$, co jest wartością wyższą od wartości krytycznej wynoszącej 23. Zatem, w przypadku *precyzji* uzyskanej za pomocą klasyfikatora *NB* nie ma podstaw do odrzucenia hipotezy zerowej H_0 mówiącej o braku różnic w porównywanych grupach, a przyjęcia hipotezy H_1, która mówi, że istnieją różnice w porównywanych grupach.

Wyniki mierzone za pomocą *kompletności*, dla każdego z klasyfikatorów spełniają założenia testu t-Studenta. Dlatego do oceny istotności uzyskanych wyników został wykorzystany ten test. Tabela 6.30 przedstawia wyniki testu t-Studenta odnośnie *kompletności* dla klasyfikatorów *SVM*, *J48*, *NB* oraz *LMT*. Klasyfikatory *J48* oraz

Tabela 6.30. Wartości testu t-Studenta dla *kompletności*, obliczonej pomiędzy wynikami uzyskanymi w eksperymencie czwartym wykorzystującym cechy z *Zestawu 4* i piątym wykorzystującym wektory cech *Zestawu 5*

	Zestaw 4		**Zestaw 5**		**Wynik testu**	
Klasyfikator	**Średnia kompletność**	**Wariancja**	**Średnia kompletność**	**Wariancja**	**Wartość testu**	**Czy wynik jest istotnie lepszy**
SVM	0,866	0,001	0,867	0,001	0,216	**NIE**
J48	0,580	0,001	0,570	0,001	0,506	**NIE**
NB	0,431	0,001	0,420	0,001	0,490	**NIE**
LMT	0,868	0,001	0,868	0,001	0,009	**NIE**

NB osiągnęły wynik *precyzji* wyższy przy wykorzystaniu wektora cech z *Zestawu 4*. Wartości testu t-Studenta dla tych klasyfikatorów są jednak niższe od poziomu krytycznego testu $t = 2,1009$. Oznacza to, że nie ma podstaw do odrzucenia hipotezy zerowej mówiącej o braku różnic w badanych grupach. Innymi słowy, wyniki mierzone za pomocą *precyzji* osiągnięte z wykorzystaniem wektora cech z *Zestawu 5*, pomimo tego że niższe, to nie są istotnie statystycznie gorsze. Wyniki *precyzji* dla *LMT* są takie same dla klasyfikacji z wektorami cech dla obu zestawów cech, zatem nie ma między nimi różnic. Klasyfikator *SVM* osiągnął wyższy wynik z wykorzystaniem wektora cech z *Zestawu 5*. Wartość testu t-Studenta dla tego przypadku jest niższa od wartości krytycznej $t = 2,1009$, co podobnie jak w pozostałych przypadkach, uniemożliwia

odrzucenie hipotezy zerowej H_0. Podsumowując, wyniki jakości klasyfikacji, mierzonej za pomocą *precyzji*, z wykorzystaniem wektora cech z *Zestawu 5* nie różnią się w istotny sposób od wyników osiągniętych za pomocą cech z *Zestawu 4*.

Podobnie jak w przypadku *precyzji*, wyniki klasyfikacji mierzone za pomocą *miary-f*, dla każdego wykorzystanego klasyfikatora spełniają założenia testu t-Studenta. W tabeli 6.31 przedstawione zostały wyniki testu t-Studenta, które obliczone zostały na podstawie *miary-f* z wyników klasyfikacji z wektorem cech z *Zestawu 4* oraz z wektorem cech z *Zestawu 5*. Średnia ważona wartość *miary-f* dla klasyfikatora *SVM* jest wyższa w

Tabela 6.31. Wartości testu t-Studenta dla *miary-f*, obliczone dla wyników uzyskanych w eksperymencie czwartym (wektor cech tworzony na podstawie z *Zestawu 4*) i piątym, wykorzystującymi wektor cechy z *Zestawu 5*

	Zestaw 4		**Zestaw 5**		**Wynik testu**	
Klasyfikator	**Średnia miara-f**	**Wariancja**	**Średnia miara-f**	**Wariancja**	**Wartość testu**	**Czy wynik jest istotnie lepszy**
SVM	0,858	0,001	0,866	0,001	0,264	**NIE**
J48	0,551	0,001	0,551	0,001	0,036	**NIE**
NB	0,473	0,001	0,457	0,001	0,942	**NIE**
LMT	0,860	0,001	0,860	0,001	0,009	**NIE**

przypadku wykorzystania wektora cech z *Zestawu 5*. Jednak wartość testu t-Studenta $(0,264)$ jest niższa od poziomu krytycznego testu $t = 2,1009$. Oznacza to, że nie ma podstaw do odrzucenia hipotezy zerowej H_0, czyli różnica w otrzymanych wynikach nie jest statystycznie istotna. Klasyfikatory *J48* oraz *LMT* osiągnęły średnią ważoną *miary-f* taką samą w eksperymencie 4 oraz 5, zatem z definicji nie ma między nimi różnic. Dla naiwnego klasyfikatora bayesowskiego wartość *miary-f* otrzymana przy użyciu cech z *Zestawu 5* jest niższa w porównaniu z wynikami osiągniętymi z wykorzystaniem cech z *Zestawu 4*. Wartość testu t-Studenta wyniosła $0,942$, co jest wartością niższą od poziomu krytycznego testu $t = 2,1009$. Taki wynik pozwala stwierdzić, że nie ma podstaw do odrzucenia hipotezy zerowej H_0 mówiącej o braku różnic w porównywanych grupach.

Meta-analiza uzyskanych wyników – określanie wielkości efektu

W kolejnych wierszach tabeli 6.32 przedstawiono wyniki obliczonych statystyk r (czarny kolor czcionki) oraz z (niebieski kolor czcionki) dla wykorzystanych klasyfikatorów (*SVM*, *J48*, *NB*, *LMT)*. Kolumna **Wrt.** oznacza wartość statystyki r lub z. Kolumna **Wlk.** zawiera interpretację wartości statystyk zgodnie z propozycją Cohena (Cohen, 1988). Kolumna **P** zawiera wyniki obliczania wielkości efektu dla *precyzji*, kolumna **R** dla *kompletności*, a kolumna **F** dla *miary-f*.

Wielkości efektu obliczone dla *precyzji* (kolumna **P**) dla wszystkich klasyfikatorów są *małe*, co oznacza, że różnica między wynikami mierzonymi za pomocą *precyzji*, uzyskanymi w *Eksperymencie 4* oraz *Eksperymencie 5* jest *mała*.

W przypadku *kompletności* (kolumna **R**), dla klasyfikatorów *SVM*, *J48*, *NB* oraz *LMT* występuje *mała* różnica między wynikami osiągniętymi za pomocą wektora cech z *Zestawu 4* i wektora cech z *Zestawu 5*.

Tabela 6.32. Wartości statystyk (kolumny **Wrt.**), dla statystyki r (czarny kolor czcionki) i z (niebieski kolor czcionki) oraz interpretacja w postaci wielkości różnicy (kolumny **Wlk.**) między wynikami uzyskanymi w *Eksperymencie 4* i *Eksperymencie 5* dla *precyzji* (**P**), *kompletności* (**R**) oraz *miary-f* (**F**), dla klasyfikatorów *SBM*, *J48*, *NB* oraz *LMT*

	P		**R**		**F**	
	Wrt.	**Wlk.**	**Wrt.**	**Wlk.**	**Wrt.**	**Wlk.**
SVM	0,051	Mała	0,073	Mała	0,062	Mała
J48	0,050	Mała	0,118	Mała	0,009	Mała
NB	0,140	Mała	0,115	Mała	0,217	Mała
LMT	0,012	Mała	0,002	Mała	0,002	Mała

Dla wyników mierzonych za pomocą *miary-f* (kolumna **F**), dla wszystkich klasyfikatorów występuje *mała* różnica między wynikami osiągniętymi za pomocą wektora cech z *Zestawu 4* i wektora cech z *Zestawu 5*.

6.2.6. Podsumowanie eksperymentów E1 – E5

Tabela 6.33 w sposób syntetyczny pokazuje istotności oraz wielkość efektu dla klasyfikacji przy użyciu wektorów cech budowanych na podstawie zestawów *Zestaw 1*, *Zestaw 2*, *Zestaw 3*, *Zestaw 4* oraz *Zestaw 5*. W wierszach nazwanych **Istotność**, na przecięciu wierszy reprezentujących kolejne klasyfikatory: *NB*, *J48*, *SVM* oraz *LMT* i kolumn dotyczących *precyzji* (**P**), *kompletności* (**R**) oraz *miary-f* (kolumny **F**) znajduje się informacja, czy pomiędzy porównywanymi wynikami osiągniętymi w eksperymentach (**E1**–**E5**) jest istotna różnica w uzyskanych wynikach, zaś w wierszach nazwanych **Efekt**, znajduje się informacja o wielkości efektu (wielkości różnicy uzyskanych wyników); **D** oznacza *dużą* wielkość efektu, **Ś** średnią wielkość, zaś **M** małą wielkość.

Dla przypomnienia, w eksperymencie **E2** wykorzystane zostały wektory budowane na podstawie `cech grafowych (1)`, w **E4** `cech grafowych (2)`. **E3** oraz **E5** to eksperymenty, w których wykorzystane zostały wektory budowane odpowiednio na bazie `cech grafowych (1)` i `cech grafowych (2)` z dołączonymi cechami bazowymi (literaturowymi). Warto zauważyć, że dla klasyfikatora *SVM* oraz *LMT* uwzględnienie w

Tabela 6.33. Podsumowanie wyników istotności (wiersz nazwany **Istotność**) oraz wielkości efektu (wiersz nazwany **Efekt**) w uzyskanych wynikach dla *precyzji* (kolumny oznaczone **P**), *kompletności* (**R**) oraz *miary-f* (**F**) w przeprowadzonych eksperymentach **E1**—**E5** dla klasyfikatorów **NB**, **J48**, **SVM** oraz **LMT** z zaznaczonymi przypadkami, kiedy wielkość efektu różni się od istotności statystycznej

		E1:E2			**E2:E3**			**E2:E4**			**E4:E5**		
Klasyfikator	**Miara**	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**
NB	**Istotność**	Tak	Nie	Nie	Nie	Nie	Nie	**Tak**	Tak	Tak	Nie	Nie	Nie
	Efekt	D	M	M	M	M	M	**Ś**	D	D	M	M	M
J48	**Istotność**	Tak	Tak	Tak	Nie	Nie	Nie	Nie	Nie	Nie	Nie	Nie	Nie
	Efekt	D	D	D	M	M	M	M	M	M	M	M	M
SVM	**Istotność**	Tak	Tak	Tak	Nie	Nie	**Nie**	Tak	Tak	Tak	Nie	Nie	Nie
		D	D	D	M	M	**Ś**	D	D	D	M	M	M
LMT	**Istotność**	Tak	Tak	Tak	Nie	Nie	Nie	Tak	Tak	Tak	Nie	Nie	Nie
	Efekt	D	D	D	M	M	M	D	D	D	M	M	M

wektorze wejściowym klasyfikatora cech grafowych z *Zestawu 2* lub *Zestawu 4* istotnie poprawia otrzymane wyniki i, co ważne, wielkość różnicy otrzymanych wyników jest duża. Sytuacje te dotyczą porównań **E1:E2** oraz **E2:E4**. W każdym wypadku dla **P**, **R** oraz **F** była istotna poprawa jakości działania klasyfikatorów oraz wielkość różnicy była duża. Kolumny **E2:E3** i **E4:E5** obrazują porównanie wyników klasyfikacji dla wektorów wejściowych budowanych na bazie cech grafowych oraz wektorów cech bazujących na połączonym zestawie cech grafowych i literaturowych. Dla *SVM* oraz *LMT* nie ma istotnych różnic w otrzymanych wynikach. Poza wielkością efektu *miary-f* dla *SVM* w **E2:E3**, która w tym przypadku jest *średnia*, wielkość różnicy otrzymanych wyników jest *mała*.

Interesująco przedstawia się sytuacja dla klasyfikatora *J48*. Istotna poprawa osiąganych wyników oraz duża siła efektu, jest tylko w przypadku wykorzystania cech z *Zestawu 2*. Wykorzystanie cech z zestawów *Zestaw 3*–*Zestaw 5* nie poprawia w istotny sposób wyników uzyskanych za pomocą wektorów cech z *Zestawu 2*, przy tym warto zauważyć, że wartość wielkości efektu jest mała. Sytuacja ta prawdopodobnie jest

związana ze zwiększoną liczbą cech w wektorze wejściowym oraz poziomem przycięcia drzewa, który zawsze pozostaje taki sam.

Nieco odmiennie zachowuje się naiwny klasyfikator Bayesa. Uwzględnienie `cech grafowych (1)` istotnie poprawia jedynie *precyzję*, dla której wielkość różnicy jest *duża*. Dla *kompletności* oraz *miary-f* nie można zaobserwować istotnej różnicy w otrzymanych wynikach oraz wielkość efektu jest *mała*. Uwzględnienie `cech grafowych (2)` w istotny sposób poprawia jakość klasyfikacji relacji semantycznych między dwoma zdaniami, mierzoną przez (**P**, **R** oraz **F**) w porównaniu do wyników klasyfikacji z wektorami cech tworzonymi z `cech grafowych (1)` z *Zestawu 2* (kolumna **E2:E4**). Wielkość efektu również jest *duża*, poza przypadkiem *precyzji*, dla którego jest *średnia*. Jednak jak wspomniano w rozdziale 6.2.4, wartość statystyki r wyniosła $0,498$, co jest bardzo bliską wartością do wartości granicznej $0,5$ oznaczającej dużą wielkość efektu.

Podsumowując, nie można jednoznacznie powiedzieć, że wykorzystanie cech grafowych dla wszystkich klasyfikatorów poprawia jakość klasyfikacji mierzoną za pomocą *precyzji*, *kompletności* oraz *miary-f*. Można jednak powiedzieć, że w większości przypadków jakość znacząco rośnie. Uwzględniając porównanie wyników osiągniętych w eksperymentach **E1:E2**, na 12 przypadków (4 klasyfikatory · 3 oceny – P, R, F) w 10 była istotna poprawa. Tylko naiwny klasyfikator Bayesa nie wykazywał istotnej różnicy w otrzymanych wynikach dla R oraz F.

To, co jest warte podkreślenia, dotyczy dołączania cech literaturowych z *Zestawu 1* do cech z *Zestawu 2* i porównanie jakości wykrywania relacji semantycznych do klasyfikacji z wektorami cech z *Zestawu 2* (kolumna **E2:E3**). W żadnym wypadku rozszerzenie zbioru cech nie poprawiło w istotny sposób wyników otrzymanych w ramach *Eksperymentu 2*. Wykorzystanie do tworzenia wektora wejściowego cech grafowych z *Zestawu 4* znacząco poprawia jakość klasyfikacji w porównaniu do klasyfikacji z wektorem cech grafowych z *Zestawu 2* (w 9 na 12 przypadków). Jedynie klasyfikator *J48* nie wykazywał istotnej różnicy w otrzymanych wynikach. Dołączenie cech literaturowych do cech grafowych z *Zestawu 4* w istotny sposób poprawia *precyzję* rozpoznawania relacji semantycznych jedynie dla naiwnego klasyfikatora bayesowskiego.

Wartości wielkości efektu, korelują w 46 na 48 przypadków z oceną istotności statystycznej. Zatem można powiedzieć, że w 46 przypadkach jest pełna korelacja między wielkością efektu, a istotnością statystyczną. W dwóch przypadkach jest ona częściowa.

6.3. Badanie wrażliwości klasyfikatorów na wektory wejściowe o różnych zestawach cech

Analiza opisana w niniejszym rozdziale miała dać odpowiedź na pytanie czy jakość klasyfikacji istotnie różni się dla któregoś z wykorzystanych w tym badaniu klasyfikatorów, w zależności użytych zestawów cech do tworzenia wektorów wejściowych. Aby sprawdzić, czy dany klasyfikator, przy danym zestawie cech wykazywał wyższą jakość rozpoznawanych relacji semantycznych od pozostałych, należy porównać ze sobą uzyskane wyniki klasyfikacji poszczególnych klasyfikatorów dla tego samego zestawu cech tworzących wektor wejściowy dla klasyfikatora i sprawdzić czy różnice są istotne statystycznie. Podobnie jak w eksperymentach porównujących ze sobą zestawy cech, wykorzystano testy t-Studenta oraz U Manna-Whitneya. w celu znalezienia najlep-

szego klasyfikatora dla danego zestawu cech dokonano porównania wyników każdego klasyfikatora z każdym innym. W przeprowadzonych eksperymentach wykorzystano: *SVM*, *J48*, *NB* oraz *LMT*, co daje 6 porównań każdego z każdym, bez powtórzeń. Ustalając poziom α dla istotności statystycznej należy pamiętać o propagacji błędów przy wielu porównaniach. Dokonując 6 porównań, poprzez propagację błędów, zadany błąd istotności statystycznej $\alpha = 0,05$, wzrósłby znacząco. Przy założeniu $\alpha = 0,05$ dla sprawdzenia danej pary, błąd dla wytypowania najlepszego klasyfikatora, byłby równy $6 \cdot 0,05 = 0,3$. Aby otrzymać odpowiedź na pytanie, który z klasyfikatorów, dla danego zestawu cech posiada najwyższą jakość, z maksymalnym prawdopodobieństwem podania błędnej odpowiedzi $\alpha = 0,05$, należy zastosować poprawkę do wielokrotnych porównań, na przykład *poprawkę Bonferroniego* (Dunn, 1961).

Aby całe badanie posiadało poziom istotności $\alpha = 0,05$, w każdym wykonanym porównaniu, poziom istotności powinien być ustawiony na $\alpha = \frac{0,05}{6}$ i dla tak ustalonego poziomu istotności statystycznej należy ustawić wartość krytyczną testu t-Studenta oraz dla porównań niespełniających założenia testu t-Studenta – testu U Manna-Whitneya. W przeprowadzonych porównaniach, poziom istotności statystycznej został ustalony na $\alpha = 0.005$, co dla sześciu porównań daje odpowiedź z błędem mniejszym niż $0,05$, a dokładniej z maksymalnym błędem równym $0,03$ $(6 \cdot 0,005 = 0,03)$. Po zastosowaniu poprawki Bonferroniego, wartość krytyczna testu t-Studenta odczytana z tablic posiada wartość $t = 3,1966$. Zaś wartość krytyczna testu U Manna-Whitneya, odczytana z tablic, wynosi 13. Hipoteza zerowa H_0 oraz alternatywna H_1 dla obu testów zostały zdefiniowane następująco:

— H_0: Pomiędzy porównywanymi wynikami nie ma statystycznie istotnej różnicy,
— H_1: Pomiędzy porównywanymi wynikami jest istotna statystycznie różnica.

Osiągnięcie wartości testu t-Studenta wyższej od $t = 3,1966$ oraz w przypadku U Manna-Whitneya niższej od 13, oznacza, że istnieją podstawy do odrzucenia H_0, przy założonym błędzie $\alpha = 0,005$. Co należy interpretować, że porównywane wyniki dla pary klasyfikatorów istotnie się różnią.

W dalszej części rozdziału przedstawione zostały wyniki porównań dla *precyzji*, *kompletności* oraz *miary-f*.

6.3.1. Porównanie precyzji

W tabeli 6.34 znajdują się wyniki dokonanych porównań jakości klasyfikatorów, mierzonej za pomocą *precyzji*, wykorzystanych w badaniach. Pierwsza kolumna zawiera informację o numerze eksperymentu *E1 – E5*. Wiersz oznaczony jako **E1**, to porównanie ze sobą jakości klasyfikacji dla średniej ważonej precyzji w eksperymencie pierwszym. Kolejne kolumny, to zestawienie każdego klasyfikatora z każdym innym. Przykładowo, kolumna nazwana jako *SVM:LMT*, oznacza, że w komórkach na przecięciu tej kolumny z kolejnymi wierszami oznaczającymi eksperymenty, znajduje się odpowiedź na pytanie czy między *SVM*, a *LMT* w danym eksperymencie jest istotna różnica w wynikach. W komórkach wpisane są wartości *Tak* lub *Nie*. Wartości te należy interpretować następująco: *Tak* – jest istotna różnica w wynikach, *Nie* – nie ma istotnej różnicy w osiągniętych wynikach pomiędzy daną parą klasyfikatorów. Warto zauważyć, że niezależnie od przyjętego zbioru danych, między *SVM*, a *J48* oraz *SVM*, a *NB* są istotne różnice w otrzymanych wynikach.

Tabela 6.34. Istotność różnicy wyników (*Tak* – jest istotna różnica, *Nie* – nie ma istotnej różnicy) mierzonych za pomocą *precyzji*, uzyskanych pomiędzy wykorzystanymi w badaniach klasyfikatorami (**SVM**, **LMT**, **NB**, **J48**) w poszczególnych eksperymentach: **E1**, **E2**, **E3**, **E4** oraz **E5**

	SVM:J48	**SVM:LMT**	**SVM:NB**	**J48:NB**	**J48:LMT**	**NB:LMT**
E1	Tak	Tak	Tak	Nie	Nie	Nie
E2	Tak	Nie	Tak	Nie	Tak	Tak
E3	Tak	Nie	Tak	Nie	Tak	Tak
E4	Tak	Nie	Tak	Tak	Tak	Nie
E5	Tak	Nie	Tak	Tak	Tak	Tak

6.3.2. Porównanie kompletności

W tabeli 6.35 znajdują się wyniki dokonanych porównań jakości klasyfikatorów, mierzonej za pomocą *kompletności*. Pierwsza kolumna zawiera informację o numerze eksperymentu, a kolejne kolumny, to zestawienie danego klasyfikatora z każdym innym. W tym przypadku pomiędzy *SVM*, a *LMT* w żadnym wykorzystanym zbiorze uczą-

Tabela 6.35. Istotność różnicy wyników (*Tak* – jest istotna różnica, *Nie* – nie ma istotnej różnicy) mierzonych za pomocą *kompletności*, uzyskanych pomiędzy wykorzystanymi w badaniach klasyfikatorami (**SVM**, **LMT**, **NB**, **J48**) w poszczególnych eksperymentach: **E1**, **E2**, **E3**, **E4** oraz **E5**

	SVM:J48	**SVM:LMT**	**SVM:NB**	**J48:NB**	**J48:LMT**	**NB:LMT**
E1	Tak	Nie	Tak	Tak	Tak	Tak
E2	Tak	Nie	Tak	Tak	Tak	Tak
E3	Tak	Nie	Tak	Tak	Tak	Tak
E4	Tak	Nie	Tak	Tak	Tak	Tak
E5	Tak	Nie	Tak	Tak	Tak	Tak

cym nie było istotnej różnicy w wynikach. Można na tej podstawie wysnuć wniosek, że wyniki obu klasyfikatorów, zgodnie z przyjętym błędem $\alpha = 0,005$, są na tym samym poziomie jakości. Pomiędzy pozostałymi parami (*SVM:J48*, *SVM:NB*. *J48:LMT*, *NB:LMT*), wynik zawsze był istotnie inny, w zależności od interpretacji: lepszy dla jednego klasyfikatora, gorszy dla drugiego. Biorąc pod uwagę osiągane wyniki *SVM* i *LMT* wykazują pewne podobieństwo, które może wynikać z podobnego przetwarzania wzorców. SVM został użyty tutaj jako rodzina klasyfikatorów binarnych a LMT jest drzewem decyzyjnym, które w liściach ma zaimplementowane modele logistyczne. Jednak trzeba pamiętać, że istnieje między nimi wiele różnic, takich jak różne funkcje straty, czas zbiegania do rozwiązania czy odporność na wartości odstające. Klasyfikator *J48*, w badaniach zawsze posiada taki sam próg przycięcia drzewa, co może powodować, że wraz ze wzrostem liczby cech traci zdolność uogólniania i dostosowuje się za bar-

dzo do danych uczących. Naiwny klasyfikator bayesowski określa prawdopodobieństwo zajścia danej klasy przy zadanym wektorze cech. Co w powiązaniu z coraz większą liczbą cech, może powodować coraz dokładniejsze rozpoznawanie klas, przy stracie na kompletności.

6.3.3. Porównanie miary-f

Tabela 6.36 zawiera porównanie jakości klasyfikatorów mierzonej za pomocą *miary-f* w każdym przeprowadzonym eksperymencie. Pierwsza kolumna zawiera informację o numerze eksperymentu, a kolejne kolumny, to zestawienie każdego klasyfikatora z każdym innym, na przykład kolumna **SVM:NB** oznacza porównanie klasyfikatora **SVM** z **NB**.

Tabela 6.36. Istotność różnicy wyników (*Tak* – jest istotna różnica, *Nie* – nie ma istotnej różnicy) mierzonych za pomocą *miary-f*, uzyskanych pomiędzy wykorzystanymi w badaniach klasyfikatorami (**SVM**, **LMT**, **NB**, **J48**) w poszczególnych eksperymentach: **E1**, **E2**, **E3**, **E4** oraz **E5**

	SVM:J48	**SVM:LMT**	**SVM:NB**	**J48:NB**	**J48:LMT**	**NB:LMT**
E1	Tak	Tak	Tak	Tak	Nie	Tak
E2	Tak	Tak	Tak	Tak	Tak	Tak
E3	Tak	Nie	Tak	Tak	Tak	Tak
E4	Tak	Nie	Tak	Tak	Tak	Tak
E5	Tak	Nie	Tak	Tak	Tak	Tak

Porównując wyniki w kontekście *miary-f* dla pary klasyfikatorów **SVM:J48** możemy stwierdzić, że każdy zastosowany zestaw cech grafowych wnosił istotne różnice w uzyskanych wynikach. Precyzja *J48* przy zastosowaniu cech z *Zestawu 1*, wiersz oznaczony jako *E1* w tabeli 6.36 jest istotnie statystycznie wyższa od *SVM*. Jednak zastosowanie pierwszego zestawu cech grafowych, linia oznaczona jako *E2*, powoduje istotnie lepszą jakość klasyfikacji *SVM* w stosunku do *J48*. Sytuacja ta powtarza się w *E3*, *E4* oraz *E5*. Zatem można powiedzieć, że zmiana cech w wektorze cech, tak by uwzględniane były cechy grafowe, spowodowała poprawę jakości działania *SVM* w stosunku do *J48*. Porównując klasyfikatory **SVM:LMT**, zaobserwowana została zależność, że dla zestawów *E1* oraz *E2* jakość mierzona za pomocą *miary-f* jest wyższa (istotnie lepsza) dla *LMT*. Jednak przy zastosowaniu zbiorów *E3* oraz *E4* *miara-f* dla *LMT* jest wyższa, ale różnica w stosunku do *SVM* nie jest statystycznie istotna. Zaś dla zbioru *E5* wynik *miary-f* jest wyższy dla *SVM*, lecz różnica nie jest istotna statystycznie. Porównując **SVM:NB**, dla każdego zbioru danych w eksperymentach *E1-E5*, *SVM* osiągnął wyższą wartość *miary-f* i w każdym przypadku ta różnica jest statystycznie istotna. Podobna sytuacja ma miejsce dla pary klasyfikatorów **J48:NB**, w tym przypadku *J48* osiąga wyższe wartości *miary-f* i w każdym przypadku, poprawa jest statystycznie istotna. Klasyfikatory **J48:LMT** wykazują nieco odmienną tendencję w jakości klasyfikacji, *J48* osiąga wyższą wartość *miary-f* w stosunku do *LMT*, jednak wartość ta nie jest statystycznie wyższa. Dla przypomnienia, różnica *J48* w stosunku

do *SVM* była istotna statystycznie na korzyść *J48*. Zestawy cech w eksperymentach *E2-E5*, to już istotna statystycznie przewaga *LMT* nad *J48*. Ostatnia para klasyfikatorów, to **NB:LMT**. W tym wypadku jakość mierzona *miarą-f* uzyskana przez *LMT* zawsze była wyższa od *NB*, co więcej różnica w wynikach na korzyść *LMT* w każdym przypadku była statystycznie istotna.

6.4. Wnioski z przeprowadzonych eksperymentów

Wykorzystanie cech zbudowanych na podstawie wartości miar podobieństwa pomiędzy grafami utworzonymi dla fragmentów tekstów, możliwe dzięki użyciu metody pogłębianej analizy semantycznej, znacząco poprawia jakość wykrywanych relacji. Zastosowanie `cech grafowych (1)` z *Zestawu 2* istotnie poprawia jakość klasyfikacji dla większości wykorzystanych klasyfikatorów w stosunku do cech literaturowych. Zaś wykorzystanie `cech grafowych (2)` z *Zestawu 4*, istotnie poprawia jakość klasyfikacji w odniesieniu do `cech grafowych (1)`.

Dołączenie cech literaturowych do cech grafowych (zarówno `grafowych (1)` oraz `grafowych (2)`) w rezultacie daje mniejszy rozrzut pomiędzy *precyzją* i *kompletnością*, a co za tym idzie również *miarą-f*. Chociaż nie poprawia istotnie wyników, jednak ma pozytywny wpływ na stosunek *precyzji* wykrywania do *kompletności*. Po wykonaniu szczegółowej analizy istotności statystycznej poprawy jakości klasyfikacji we wszystkich eksperymentach, dla poszczególnych relacji okazało się, że dla niektórych relacji, dołączenie cech literaturowych do cech grafowych istotnie poprawia wyniki, jednak w przypadku średniej ważonej nie ma istotnej poprawy. Można więc wysnuć wniosek, że jeżeli chcielibyśmy równie precyzyjnie, co dokładnie wykrywać relacje, warto rozważyć połączenie obu zestawów cech.

Porównanie ze sobą jakości klasyfikacji poszczególnych klasyfikatorów na tych samych zestawach cech pokazało, że między *Logistic Model Tree*, a *Supported Vector Machine* nie ma statystycznie istotnych różnic. Zaletą *LMT* w stosunku do *SVM* jest automatyczne dokonywanie selekcji cech. Każda z funkcji logistycznych, opisujących poszczególne klasy, budowana jest na podstawie podzbioru cech dostarczonych do wektora wejściowego dla klasyfikatora. Jednak porównanie *SVM* z *J48* w eksperymencie *E1*, z korzyścią dla *J48*, wykazało, że uzyskane wyniki istotnie statystycznie się różnią. Zaś porównanie *LMT* z *J48* w tym eksperymencie wykazało, że różnica uzyskanych wyników między *J48*, a *LMT*, z przewagą *J48* nie jest istotna statystycznie. Dlatego łącząc oba te fakty ze sobą można uznać, że jednak *LMT* okazał się w tym zagadnieniu lepszy, czego bezpośrednie porównanie nie pokazało.

Rozdział 7

Podsumowanie pracy

Niniejszy rozdział jest podsumowaniem całej pracy, zawiera ocenę osiągnięcia celu i zadań badawczych. W dalszej jego części przedstawiono kierunki rozwoju metody, w tym propozycje rozszerzenia badań nad metodą *PAS*. Przedstawiono także projekty badawcze, w ramach których powstawała niniejsza praca.

7.1. Realizacja celu rozprawy

Głównym celem pracy było *opracowanie metody pogłębianej analizy semantycznej, którą można byłoby wykorzystać w zadaniu wykrywania relacji semantycznych między fragmentami tekstów w języku polskim.* Tak postawiony cel udało się osiągnąć dzięki realizacji opisanych niżej zadań badawczych.

Opracowano płytki parser semantyczny, wykrywający role semantyczne w obrębie frazy nominalnej. Jest to system regułowy będący rozszerzeniem istniejącego *chunkera* o nazwie IOBBER. Działanie chunkera IOBBER w obrębie frazy rzeczownikowej oraz przymiotnikowej zostało wzbogacone o wykrywanie zależności składniowych zachodzących między słowami w tych frazach. Do wykrytych zależności składniowych przypisywane są nazwy ról, które pełnią słowa w danej zależności składniowej. Tak powstały mechanizm nazywany jest w literaturze *płytkim parserem semantycznym.* Proces budowy płytkiego prasera semantycznego oraz ocena skuteczności jego działania przedstawiona została w rozdziale 4.2.2.

Dostosowano do języka polskiego metodę ujednoznacznia znaczeń słów oryginalnie opracowaną dla języka angielskiego. Metoda wykorzystuje nienadzorowane podejście oparte o błądzenie po grafie. W pierwszym kroku, Słowosieć, która jest największym na świecie relacyjnym słownikiem semantycznym, została przekształcona do grafu. W grafie tym, węzły reprezentują synsety ze Słowosieci, zaś krawędzie to relacje między znaczeniami. Na tak zbudowanym grafie, w iteracyjny sposób wykonywany jest algorytm nadawania rang węzłom tego grafu – *PageRank*. Sposób adaptacji oraz jakość zaadaptowanego systemu, zostały przedstawione w rozdziale 4.2.2.

Opracowano metodę oraz narzędzia do rzutowania Słowosieci na Ontologię SUMO. SUMO posiada wiele rozszerzeń na ontologie dziedzinowe, np. (Vanderhulst, 2005), które również zostały wykorzystane. Dodatkowym argumentem za wyborem SUMO

jako ontologii wysokiego poziomu jest rzutowanie angielskiego Princeton WordNet na ontologię SUMO oraz ręczne rzutowanie Słowosieci za pomocą relacji międzyjęzykowych na Princeton WordNet, co było pomocne w opracowaniu algorytmu rzutowania Słowosieci na SUMO. Opracowane rzutowanie sprowadza się do wieloetapowego użycia hierarchicznie ułożonych reguł, które podzielone zostały na reguły proste oraz złożone. Reguły wykorzystują jako przesłanki wiele informacji o znaczeniach oraz pojęciach. Cały proces rzutowania, wynik oraz jakość osiągniętego automatycznego rzutowania Słowosieci na SUMO zostały przedstawione w rozdziale 4.2.3.

Opracowano metodę wycinania spójnych podgrafów. Wycinanie spójnych, minimalnych podgrafów jest problemem NP-trudnym. Dlatego w pracy przyjęto uproszczenie. Metoda w sposób deterministyczny wycina podgrafy, jednak nie zakłada się, że wycięty spójny podgraf musi być grafem minimalnym. Brak wymagań względem grafu minimalnego powoduje, że opracowana metoda, oparta o wyszukiwanie najkrótszych ścieżek posiada pesymistyczną złożoność obliczeniową $\frac{1}{2}O(N^2) \cdot O(V^2)$, gdzie $O(V^2)$ to pesymistyczna złożoność wyszukiwania najkrótszej ścieżki w grafie z wykorzystaniem algorytmu Dijkstry, między każdym a każdym węzłem z $(\frac{1}{2}O(N^2))$ węzłów występujących w tekście. Ten problem został omówiony w rozdziale 4.2.4.

Opracowano metodę dołączania informacji pozatekstowej do informacji wydobytej z tekstu. Informacje pozatekstowe, to takie, które nie są wyrażone w całości bezpośrednio w tekście, jedynie fragmenty. Opracowane w ramach rozprawy dwa algorytmy dołączania wiedzy pozatekstowej do tekstowej przedstawione zostały w rozdziale 4.2.4.

Wykonanie przedstawionych wyżej zadań złożyło się na opracowanie metody pogłębianej analizy semantycznej *PAS*. Metoda podczas swojego działania, wydobywa z tekstu coraz to dokładniejsze informacje, które dopisywane są do istniejącej wiedzy reprezentowanej za pomocą grafu.

Jedną z głównych zalet metody *PAS*, jest możliwość zapisu informacji niekompletnej do grafu. Inną jest możliwość uzupełniania dotychczasowej wiedzy wyekstrahowanej z tekstu o informacje wydobyte z zewnętrznego źródła wiedzy, w tym przypadku Słowosieci oraz ontologii SUMO. Do zalet można również zaliczyć elastyczność metody. Przebieg metody jest definiowany w *scenariuszu realizacji* i może być dopasowywany do konkretnych potrzeb użytkownika (użytkownik określa poziom szczegółowości informacji, które wydobywane są z tekstu).

Zadanie wykrywania relacji semantycznych między fragmentami tekstów, ujęte zostało jako problem klasyfikacji. W badaniach eksperymentalnych użyto czterech klasyfikatorów uczonych oraz testowanych na tych samych zbiorach uczących i testowych: *naiwny klasyfikator Bayesa*, *Supported Vector Machine*, *Logistic Model Tree* oraz *drzewo J48*. Sprawdzona została istotność różnic osiąganych przez poszczególne klasyfikatory wyników klasyfikacji (dla miar: *precyzja*, *kompletność* oraz *miara-f*) z cechami zbudowanymi dzięki użyciu metody *PAS*. Otrzymane wyniki zostały porównane z wynikami osiągniętymi przy wykorzystaniu cech bazowych, czyli tych, które najczęściej wykorzystywane są w literaturze. Wnioskiem z przeprowadzonych badań jest poprawa jakości klasyfikacji relacji semantycznych przy wykorzystaniu tych zestawów cech, do budowy których wykorzystana została metoda *PAS*.

Jako oryginalny wkład pracy należy również wymienić nowo powstałą miarę do określania podobieństwa dwóch grafów. Bada ona podobieństwo dwóch grafów na zasadzie badania podobieństwa sąsiedztwa węzłów o tych samych identyfikatorach, wy-

stępujących w porównywanych grafach. Została ona użyta jako jedna z cech w wektorze cech dla klasyfikatora. Inne miary podobieństwa porównywanych ze sobą grafów, wchodzące w skład wektora cech, opisane zostały w rozdziale 3.2.2.

7.2. Kierunki dalszych badań

Opracowana metoda posiada kilka ograniczeń, które mogą być rozwiązane w ramach dalszych prac. Podstawowym ograniczeniem jest sposób wycinania spójnych podgrafów z zewnętrznych źródeł wiedzy. W ramach pracy zaproponowane zostało deterministyczne podejście, które jest przybliżeniem *minimalnego spójnego podgrafu*. Jednak w wielu przypadkach, wiedza, która jest wydobywana z zewnętrznych źródeł wiedzy jest nadmiarowa. Dlatego pierwszym rozszerzeniem, które poprawiłoby jakość samej metody, jest zastosowanie algorytmu do wycinania *minimalnych spójnych podgrafów*.

Zaproponowana metoda *PAS* nie uwzględnia problemu kompozycyjności/niekompozycyjności wyrażeń. *Wyrażenie kompozycyjne* to takie, którego znaczenie jest funkcją znaczeń elementów tego wyrażenia, przykładowo *białe wino*. Wyrażenie *niekompozycyjne*, to takie, którego znaczenie nie jest deterministyczną funkcją znaczeń składowych wyrażenia. Przykładem wyrażenia niekompozycyjnego jest *czerwona kartka* (w meczach piłki nożnej), która oprócz tego, że jako symbol jest czerwoną kartką, przede wszystkim niesie ze sobą znaczenie *kary*. Sytuacje te nie są modelowane za pomocą metody *PAS*, dlatego rozszerzeniem metody, mogłoby być uwzględnienie kompozycyjności/niekompozycyjności wyrażeń wielowyrazowych wraz z odpowiednim ich modelowaniem.

Kolejne ograniczenie, to miary podobieństwa grafowego. Wykorzystane w pracy miary nie stanowią pełnego zestawu miar porównujących ze sobą grafy. Dlatego interesujące wydaje się być rozszerzenie aktualnego zestawu miar, o inne występujące w literaturze, i sprawdzenie jak to wpływa na jakość klasyfikacji relacji między fragmentami tekstów.

Podczas dokonanego przeglądu literaturowego, zauważono brak miary, która porównywałaby ze sobą grafy na poziomie oznaczeń oraz kierunków krawędzi. Kolejnym rozszerzeniem, mogłaby być propozycja własnej miary, dostosowanej do problemu porównywania ze sobą grafów semantycznych.

W założeniach projektowych opracowana metoda nie została ograniczona do przetwarzania tekstów w języku polskim, a zadanie klasyfikacji relacji między fragmentami tekstów, było jedynie przykładem zadania, do rozwiązania którego zastosowano metodę *PAS*. Dlatego celowym wydaje się sprawdzenie jakości działania metody do wykrywania relacji semantycznych między fragmentami tekstów w innych językach.

Metoda *PAS* nie jest dedykowana jedynie do problemu wykrywania relacji między fragmentami tekstów. Jej przeznaczenie jest bardzo ogólne – może być zastosowana wszędzie tam, gdzie wymagane jest porównanie między sobą dwóch fragmentów tekstów lub też wymagane jest wykorzystanie ustrukturalizowanej informacji wyekstrahowanej z tekstu. Zatem kolejnym kierunkiem dalszych badań może być jej wykorzystanie do innych zadań inżynierii języka naturalnego, na przykład w problemie odpowiadania na pytania. Problem ten polega na tym, że użytkownik systemu, zadaje pytanie w języku naturalnym, system zaś wyszukuje w repozytorium tekstów taki fragment, który jest najlepszą odpowiedzią na jego pytanie. Wykorzystując metodę *PAS*, zarówno akapity tekstów, które są odpowiedziami na pytanie, jak i samo pytanie, mogłyby być przed-

stawione za pomocą grafów. Wykorzystując miary podobieństwa grafowego, możliwe byłoby porównanie ze sobą grafu zbudowanego dla zapytania ze wszystkimi grafami reprezentującymi akapity tekstu i wybranie grafu reprezentującego akapit, który byłby najbardziej podobny do grafu zbudowanego z pytania.

Podczas przeprowadzonych eksperymentów, celowo nie została wykonana selekcja cech. Wynikało to z dążenia do sprawdzenia jakości działania klasyfikatorów z tak dużymi wektorami cech, które reprezentują podobieństwo między grafami zbudowanymi dla konkretnej instancji relacji. Oprócz wykorzystanego klasyfikatora *LMT*, który sam dokonuje selekcji cech, żaden z pozostałych klasyfikatorów nie posiada wbudowanego mechanizmu selekcji. Wydaje się interesującym wykonanie selekcji cech na całym zbiorze i zbadanie czy ma to wpływ na wyniki klasyfikacji relacji pomiędzy fragmentami tekstów. Przedstawiono wiele kierunków rozwoju metody i przeprowadzenia dalszych badań z wykorzystaniem metody *PAS*. Najważniejszymi wydają się być modelowanie wyrażeń niekompozycyjnych w grafie oraz rozszerzenie zestawu miar i/lub dodanie własnej miary do określania podobieństwa grafów.

7.3. Realizacja metody PAS w kontekście realizowanych projektów badawczych

Niniejsza praca powstawała podczas realizacji następujących projektów naukowo-badawczych:

— **SyNaT**, zadanie badawcze o nazwie *Utworzenie uniwersalnej, otwartej, repozytoryjnej platformy hostingowej i komunikacyjnej dla sieciowych zasobów wiedzy dla nauki, edukacji i otwartego społeczeństwa wiedzy.* Zadanie badawcze było częścią Programu Strategicznego Narodowego Centrum Badań i Rozwoju noszącego nazwę: „Interdyscyplinarny system interaktywnej informacji naukowej i naukowo technicznej". Projekt był finansowany przez Narodowe Centrum Badań i Rozwoju, Nr Umowy SP/I/1/77065/10. Podczas realizacji tego zadania badawczego powstał *płytki parser semantyczny*, który jest elementem zaproponowanej metody *PAS*.

— **CLARIN-PL**: *„Polska część infrastruktury naukowej CLARIN ERIC: Wspólne zasoby językowe i infrastruktura technologiczna."* oraz *Opracowanie metodologii prac nad rozwojem infrastruktury technologii językowych CLARIN oraz upowszechnienie wytworzonej w ramach CLARIN infrastruktury badawczej.*
Common Language Resources & Technology Infrastructure, jest to ogólnoeuropejska infrastruktura naukowa umożliwiająca badaczom z dziedziny nauk humanistycznych i społecznych pracę z dużymi zbiorami tekstów. Podczas realizacji zadań badawczych w projekcie CLARIN-PL zaadaptowana została do języka polskiego metoda do *ujednoznaczniania znaczeń leksykalnych* słów, opracowany został algorytm *rzutowania Słowosieci na ontologię SUMO* oraz powstał szkielet metody *PAS*, z pierwszym jej wykorzystaniem do *zadania wykrywania relacji semantycznych między fragmentami tekstów.*

Słownik pojęć i oznaczeń

Słownik pojęć używanych w pracy

W rozdziale przedstawione zostały wyrażenia wykorzystywane w pracy, które mogą powodować problemy ze zrozumieniem treści.

1. **chunk** – fragment tekstu, który posiada jakąś interpretację strukturalną lub semantyczną. *Chunkiem* może być fraza z określonym typem, na przykład *fraza rzeczownikowa*, czyli taka, której element główny jest *rzeczownikiem*.
2. **chunker** – narzędzie dzielące tekst na *chunki*.
3. **hiponim** – synset o znaczeniu węższym od innego, z którym wchodzi w relację *hiponimii*. Przykładowo *pies* jest hiponimem słowa *ssak*.
4. **hiponimia** – najważniejsza relacja synsetów, stanowi $\approx$ 66% relacji synsetów w Słowosieci. Jest to relacja zawężania znaczenia, np. *pies* jest w relacji hiponimii do *ssaka*.
5. **jednostka leksykalna** – to znaczenie słowa, na przykład `zamek.1(msc)` wskazuje na zamek, który jest budynkiem, `zamek.5(wytw)` to pojęcie z informatyki. Jednostka leksykalna jest reprezentowana przez lemat (*zamek*), numer znaczenia (kolejno *1* oraz *5*), część mowy (tutaj nie jest widoczna) oraz dziedzina (*msc* – miejsce, *wytw* – wytówr). Jednostka leksykalna jest elementem *synsetu*.
6. **korpus** – na podstawie (Przepiórkowski i inni, 2012) jest to komputerowy zbiór autentycznych tekstów językowych, mówionych i pisanych, reprezentują różne odmiany, style i typy tekstów.
7. **meronim** – znaczenie, którego reprezentacja semantycznie jest częścią innego znaczenia, np. *palec* jest *meronimem* znaczenia *ręka*.
8. **meronimia** – relacja „część-całość" zachodząca między dwoma synsetami, przykładowo *palec* jest *częścią ręki*. Mówi się wtedy, że znaczenie reprezentujące *palec* jest *meronimem* (**patrz:** *meronim*) znaczenia *ręka*..
9. **Słowosieć** – sieć relacji semantycznych i leksykalnych dla języka polskiego. Znaczenia jednostek leksykalnych opisywane są poprzez odpowiednie ich umieszczenie w sieci powiązań z innymi jednostkami.

10. **synset** (ang. *Set of synonyms*) – zbiór jednostek leksykalnych należących do tej samej części mowy, wchodzących w takie same relacje semantyczne.
11. **token** – najmniejszy, niepodzielny fragment tekstu, który posiada przypisane informacje morfologiczne, takie jak część mowy, czas, osoba, rodzaj.
12. **tokenizacja** – proces podziału tekstu na tokeny (**patrz:** *token*). Po wykonaniu *tokenizacji*, podczas dalszego przetwarzania część tokenów może odpowiadać wyrazom.

Przyjęte oznaczenia

W niniejszym rozdziale przedstawione zostały skróty oraz oznaczenia wraz z ich wyjaśnieniem, wykorzystywane w pracy.

1. $\mathcal{C}$ – korpus/kolekcja dokumentów.
2. $\mathcal{D}$ – dokument w kolekcji dokumentów, $\mathcal{D} \in \mathcal{C}$.
3. $\mathcal{T}$ – przetwarzany tekst, tekst może być częścią lub całością dokumentu, $\mathcal{T} \in \mathcal{D}$.
4. $\mathcal{S}$ – zdanie, to element tekstu, $\mathcal{S} \in \mathcal{T}$.
5. w – słowo znajdujące się w zdaniu, $w \in \mathcal{S}$.
6. $\vec{v}$ – oznacza wektor cech, wykorzystany podczas procesu uczenia i testowania klasyfikatorów.
7. $\mathbf{G}$ – to graf zbudowany z tekstu.
8. Φ – oznacza graf zerowy.
9. n – oznaczenie węzła w grafie.
10. V – zbiór węzłów z grafu $\mathbf{G}$.
11. L_V – oznaczenie zbioru etykiet węzłów z V.
12. e – krawędź w grafie.
13. E – zbiór krawędzi z grafu $\mathbf{G}$
14. L_E – zbiór etykiet krawędzi z E.
15. f_{nt} – oznaczenie funkcji transformującej słowo w węzeł w grafie.
16. f_{et} – funkcja transformująca relację zachodzącą między dwoma słowami, do krawędzi w grafie.
17. $\mathbf{G'}_{plwn}$ – to graf zbudowany ze Słowosieci.
18. $\mathbf{G'}_{SUMO}$ – graf zbudowany z ontologii SUMO.
19. $\mathbf{R}$ – relacja niedospecyfikowana w procesie rzutowania Słowosieci na ontologię SUMO.
20. $\mathbf{M}$ – oznaczenie macierzy M, reprezentującej dokument $\mathcal{D}$.

Dodatek A

Przykładowe fragmenty wektorów cech wykorzystanych w badaniach

A.1. Cechy z literatury

Pełen wektor cech, wykorzystany w eksperymencie pierwszym mającym na celu określenie poziomu odniesienia do dalszych badań. Eksperyment bazowy przedstawiony został w rozdziale 6.2.1.

```
(
    posamount, intersectsynsets, longersentence, grapheditdist,
    cosine, subseq, substr, commonlemmas, propernames, relation_name
)
```

Gdzie jako ostatnia wartość, nazwana `relation_name`, podawana jest nazwa relacji opisywana za pomocą tego wektora. Łącznie rozpoznawanych jest 17 klas, z czego 16 to relacje z modelu CST oraz jedna dodatkowa, określająca brak zachodzenia relacji między daną parą zdań:
$relation\_name \in \{$

$$Cytowanie, Dalsze\_informacje, Krzyzowanie\_sie, Modalnosc, Mowa\_zalezna, Opis, Parafraza, Spelnienie, Sprzecznosc, Streszczenie, Tlo\_historyczne, Tozsamosc, Uszczegolowienie, Zawieranie, Zmiana\_pogladu, Zrodlo, Brak\_relacji \tag{A.1}$$

$\}$

A.2. Zestaw 2 – `cechy grafowe (1)`

Fragment zestawu cech grafowych, nazwanego `cechami grafowymi (1)`, wykorzystanego w eksperymencie drugim, przedstawionym w rozdziale 6.2.2.

```
(
    (...),
    lemma_lower_plwn_sumo:DMMCSSimilarityMeasure,
```

```
    concept_plwn:SameNodeSimilarityMeasure,
    concept:ContextualBOWSimilarityMeasur,
    lemma_pos_lower_plwn:GraphEditDistanceMeasure,
    (...)
)
```

Przedstawione powyżej przykładowe cechy z pełnego wektora 128 cech, należy interpretować jako podobieństwo dwóch grafów (zdań) wchodzących w konkretną relację:

- `lemma_lower_plwn_sumo:DMMCSSimilarityMeasure` – wartość podobieństwa mierzona za pomocą miary `DMMCSSimilarityMeasure` dla grafów, w których typ węzła ustawiony został na `lemma_lower` (czyli forma bazowa słowa, sprowadzona do małych liter) z dołączoną wiedzą pozatekstową ze Słowosieci (`_plwn`) oraz ontologii SUMO (`_sumo`);
- `concept_plwn:SameNodeSimilarityMeasure` – podobieństwo mierzone za pomocą miary `SameNodeSimilarityMeasure` pomiędzy grafami z ustawionym typem węzła na `concept` (słowo z tekstu reprezentowane jest jako pojęcie w grafie), z dołączaną wiedzą pozatekstową ze Słowosieci (`_plwn`);
- `concept:ContextualBOWSimilarityMeasur` – typ węzła w grafach ustawiony został na `concept` – pojęcie ontologii SUMO, podobieństwo grafów mierzone za pomocą miary `ContextualBOWSimilarityMeasur`;
- `lemma_pos_lower_plwn:GraphEditDistanceMeasure` – podobieństwo między grafami liczone jest za pomocą miary `GraphEditDistanceMeasure`, grafy posiadają typ węzła ustawiony na `lemma_pos_lower` – czyli forma bazowa słowa sprowadzona do małych liter w połączeniu z częścią mowy tego słowa z dołączoną wiedzą pozatekstową ze Słowosieci – postfiks `_plwn`;

A.3. Zestaw 4 – `cechy grafowe (2)`

Fragment zestawu `cech grafowych (2)`, wykorzystanego w czwartym eksperymencie, opisanym w rozdziale 6.2.4.

```
(
    (...),
    synset_dep_rel:DMMCSNSimilarityMeasure,
    concept_h2h-malt_rel-w2w:DUGUSimilarityMeasure,
    lemma_lower_sumo_dep_rel-malt_rel-sem_role-ne2ne:ContextualBOWSim,
    (...)
)
```

Poszczególne składowe tego wektora reprezentują wartości podobieństwa między parą grafów zbudowanych dla konkretnej relacji:

- `synset_dep_rel:DMMCSNSimilarityMeasure` – oznacza, że para grafów porównana została miarą `DMMCSNSimilarityMeasure`, grafy posiadały typ węzła `synset`, a zestaw krawędzi, to zależności składniowe (`dep_rel`);
- `concept_h2h-malt_rel-w2w:DUGUSimilarityMeasure` – podobieństwo grafów, które jako typ węzła ustawione mają pojęcie SUMO, krawędzie odzwierciedlają kolejność występowania głów składniowych po sobie (`h2h`), relacje nadawane przez

parser *Malt* (`malt_rel`) oraz kolejność występowania słów po sobie – `w2w`, zmierzone zostało za pomocą miary `DUGUSimilarityMeasure`;

— `lemma_lower_sumo_dep_rel-malt_rel-sem_role-ne2ne:ContextualBOWSim` – wartość tej cechy oznacza podobieństwo grafów zmierzone za pomocą miary `ContextualBOWSimilarityMeasure`, grafy posiadają typ węzła ustawiony jako forma podstawowa słowa sprowadzona do małych liter (`lemma_lower`) z dołączoną wiedzą pozatekstową z ontologii SUMO, krawędzie zaś odzwierciedlają zależności składniowe z tekstu (`dep_rel`), relacje *Maltowe* (`malt_rel`), role semantyczne (`sem_role`) oraz kolejność występowania po sobie nazw własnych (`ne2ne2`);

Dodatek B

Scenariusz użytkownika w metodzie PAS

B.1. Opis zmiennych ze scenariusza użycia

B.1.1. Sposób reprezentacji słowa w grafie

```
node = {
    LemmaNode LemmaNodeLower LemmaNodeUpper LemmaPosNode
    LemmaPosNodeLower LemmaPosNodeUpper SynsetNode ConceptNode}
```

Zmienna ta musi posiadać ustawioną dokładnie jedną wartość z dostępnych. Oznacza to, że każdy węzeł w budowanym grafie, definiowanym za pomocą danego scenariusza, będzie posiadał taki sam typ węzła, np. *LemmaNode* – czyli węzeł w każdym grafie będzie reprezentowany przez formę podstawową słowa.

B.1.2. Reprezentacja relacji tekstowej w grafie

```
ebuilders = {
    NESemRelEdgeBuilder NEToNEEdgeBuilder MaltEdgeBuilder
    SemRelEdgeBuilder MlNpSemRelEdgeBuilder DepRelEdgeBuider
    HeadToHeadEdgeBuilder WordToWordEdgeBuilder}
```

Jeżeli zmienna `ebuilders` nie będzie posiadała przypisanej żadnej wartości, wtedy zbudowane zostaną grafy zerowe, czyli grafy bez krawędzi, posiadające same węzły. Wszystkie budowane grafy w oparciu o dany scenariusz będą posiadały taki sam zestaw krawędzi.

B.1.3. Metody przebudowy grafu

Nie jest wymagane podanie żadnej metody przebudowującej grafy, jak również możliwe jest podanie wszystkich:

```
rebuilders = {
```

```
    NEGraphBuilder VerbSieBuilder PlwnSpikeGraphBuilder
    SumoSpikeGraphBuilder MergeBaseGraphToOne FilterByNodeEdgeCount}
```

Podobnie jak w przypadku typu węzła oraz zestawu krawędzi, wybrane metody przebudowy grafów zostaną zastosowane do wszystkich zbudowanych grafów w oparciu o dany scenariusz. Przebudowa grafów może być pożądana przez użytkownika w sytuacjach, kiedy przykładowo nazwy własne wieloelementowe powinny być traktowane jako jeden, a nie kilka węzłów – zamiast trzech węzłów dla *Zakład Ubezpieczeń Społecznych*, kolejno *Zakład*, *Ubezpieczenie*, *Społeczny*, wykorzystując *NEGraphBuilder* powstanie jeden węzeł, nazwany jako *Zakład Ubezpieczeń Społecznych*. W tym miejscu warto dodać. iż wykorzystanie `MergeBaseGraphToOne` spowoduje połączenie wszystkich grafów zbudowanych dla zdań z analizowanego tekstu do jednego, zawierającego węzły i krawędzie ze wszystkich.

B.1.4. Dołączanie wiedzy pozatekstowej

Dołączanie wiedzy ze Słowosieci.

```
plwn_merger = {
    PlWNMerger PlWNMappingMerger}
```

Jeżeli użytkownik zakłada możliwość połączenia wiedzy tekstowej z wiedzą pozatekstową, w tym przypadku z wiedzą zawartą w Słowosieci, musi wybrać sposób łączenia. Jeżeli nie zostanie ustalony sposób łączenia wiedzy ze Słowosieci z wiedzą tekstową, wiedza pozatekstowa ze Słowosieci nie zostanie dołączona – zbudowane grafy będą posiadały jedynie informacje zawarte w tekście. W przypadku chęci dołączenia wiedzy ze Słowosieci, należy określić jeden sposób łączenia.
Sposób łączenia wiedzy tekstowej z **ontologią SUMO**.

```
sumo_merger = {
    SUMOMerger SUMOMappingMerger}
```

Podobnie jak w przypadku łączenia wiedzy tekstowej z wiedzą zawartą w Słowosieci, użytkownik posiada możliwość określenia sposobu łączenia wiedzy tekstowej z wiedzą zawartą w ontologii SUMO. Jeżeli żaden sposób łączenia nie zostanie wybrany, wiedza z ontologii SUMO nie zostanie dołączona do grafu zbudowanego w oparciu o tekst.
Sposób wycinania podgrafu spójnego dla zasobów wiedzy zewnętrznej (Słowosieć, SUMO):

```
reducers = {
    SimpleReducer, SimpleTraversalCostReducer,
    KMeansClusteringReducer}
```

Jest to wybór algorytmu wycinającego podgraf ze źródeł zewnętrznych. W przypadku użycia dołączania wiedzy pozatekstowej wymagane jest ustawienie tej wartości na jedną z dostępnych. Ze względu na złożoność metod wycinających podgraf, proponowane jest ustawienie tej wartości na `SimpleReducer`.

B.1.5. Filtrowanie słów

```
filters = {
```

```
POSInterpFilter POSFilter}
```

Użytkownik metody posiada możliwość filtrowania słów, z których budowany jest graf. Jeżeli graf ma być budowany ze wszystkich słów łącznie ze znakami interpunkcyjnymi, zmienna ta powinna zostać pusta, bez ustawionej żadnej wartości. Jeżeli znaki inteprunkcyjne nie powinny znaleźć się w grafie, wtedy należy ustawić zmienna `filters` na wartość `POSInterpFilter`.

B.2. Przykłady zapisów scenariuszy użycia

B.2.1. Przykład 1

Przykład prezentowany na rysunku 4.9 może być zdefiniowany za pomocą scenariusza użycia jako:

```
node = LemmaNodeLower
ebuilders = WordToWordEdgeBuilder
rebuilders = VerbSieBuilder
plwn_merger =
sumo_merger =
filters =
```

B.2.2. Przykład 2

W celu otrzymania grafu zgodnego z przedstawionym na rysunku 4.10 użytkownik musi w scenariuszu użycia ustawić zmienne w następujący sposób:

```
node = LemmaNodeLower
ebuilders = WordToWordEdgeBuilder
rebuilders = VerbSieBuilder NEGraphBuilder
plwn_merger =
sumo_merger =
filters =
```

B.2.3. Przykład 3

Aby otrzymać graf zgodny z przedstawionym na rysunku 4.11, należy ustawić zmienne w następujący sposób:

```
node = SynsetNode
ebuilders = WordToWordEdgeBuilder
rebuilders =
plwn_merger =
sumo_merger =
filters =
```

B.2.4. Przykład 4

Aby otrzymać graf zgodny z rysunkiem 4.12, za pomocą formalizmu scenariusza użycia, należy ustawić:

```
node = LemmaNodeLower
ebuilders = WordToWordEdgeBuilder HeadToHeadEdgeBuilder \
            DepRelEdgeBuider SemRelEdgeBuilder \
            MaltEdgeBuilder NEToNEEdgeBuilder
rebuilders =
plwn_merger =
sumo_merger =
filters =
```

B.2.5. Przykład 5

Aby dołączyć do wiedzy tekstowe wiedzę pozatekstową i otrzymać wynik podobny do prezentowanego na rysunku 4.10, w scenariuszu użytkownika należy ustawić:

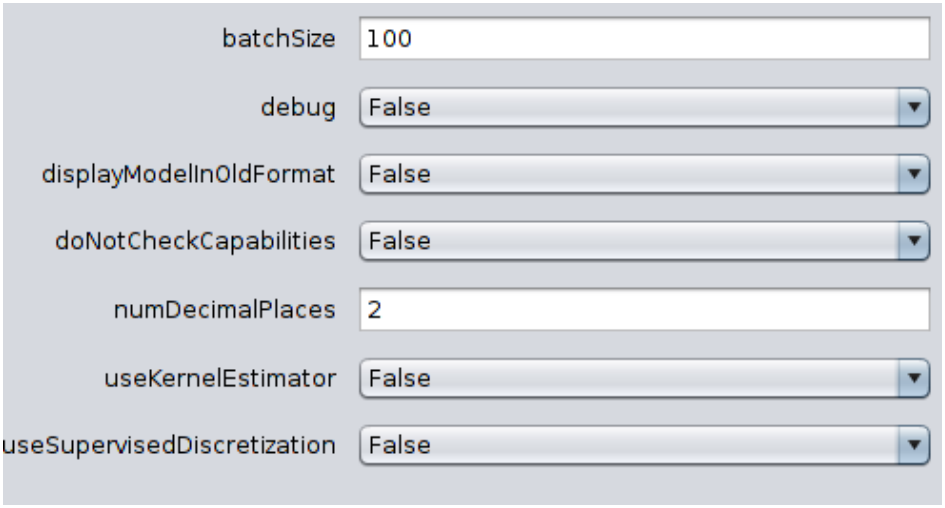
```
node = LemmaNodeLower
ebuilders = WordToWordEdgeBuilder
rebuilders = NEGraphBuilder VerbSieBuilder
plwn_merger = PlWNMappingMerger
sumo_merger = SUMOMappingMerger
filters =
```

Dodatek C

Parametry klasyfikatorów wykorzystanych w badaniach

C.1. Parametry Naiwnego Klasyfikatora Bayesowskiego

Na rysunku C.1 przedstawione zostały parametry *naiwnego klasyfikatora Bayesa*, wykorzystane podczas przeprowadzonych eksperymentów. Są to domyślne ustawienia

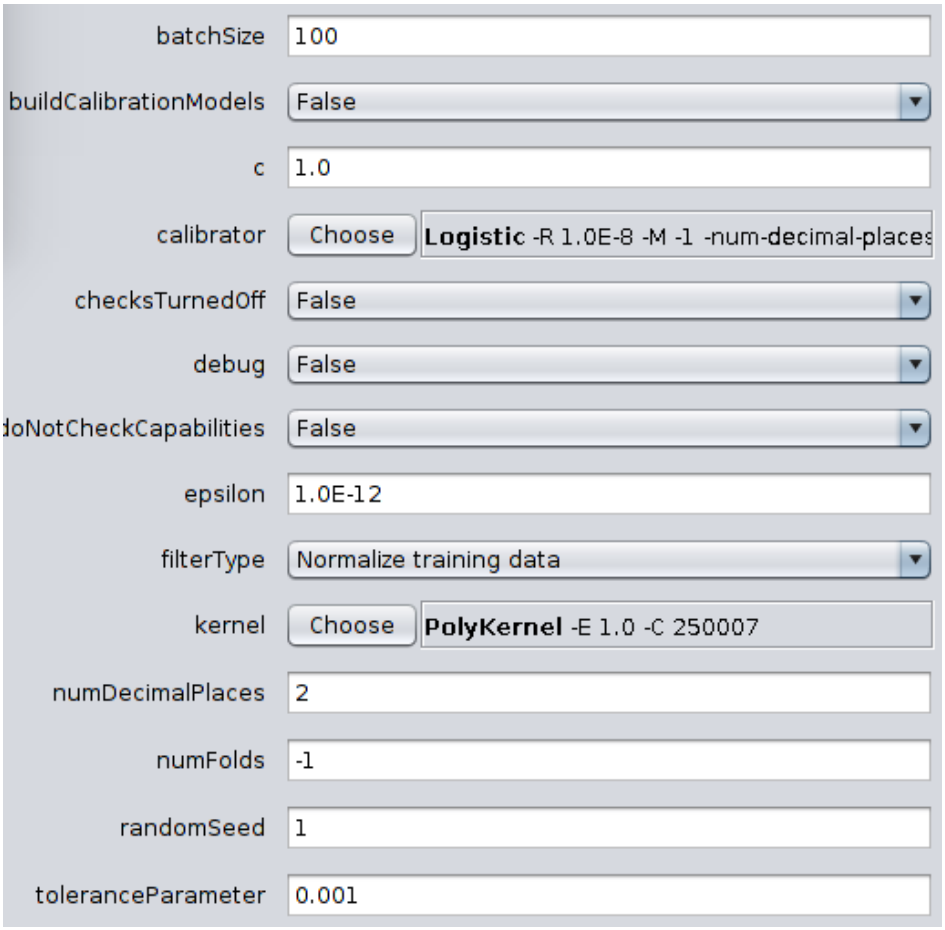


Rys. C.1. Ustawienia klasyfikatora *NB* w środowisku Weka

w Wece.

C.2. Parametry klasyfikatora SVM

Rysunek C.2 przedstawia domyślne ustawienia klasyfikatora *SVM* w systemie Weka. C.2, Ustawienia te zostały wykorzystane we wszystkich przeprowadzonych eksperymentach.



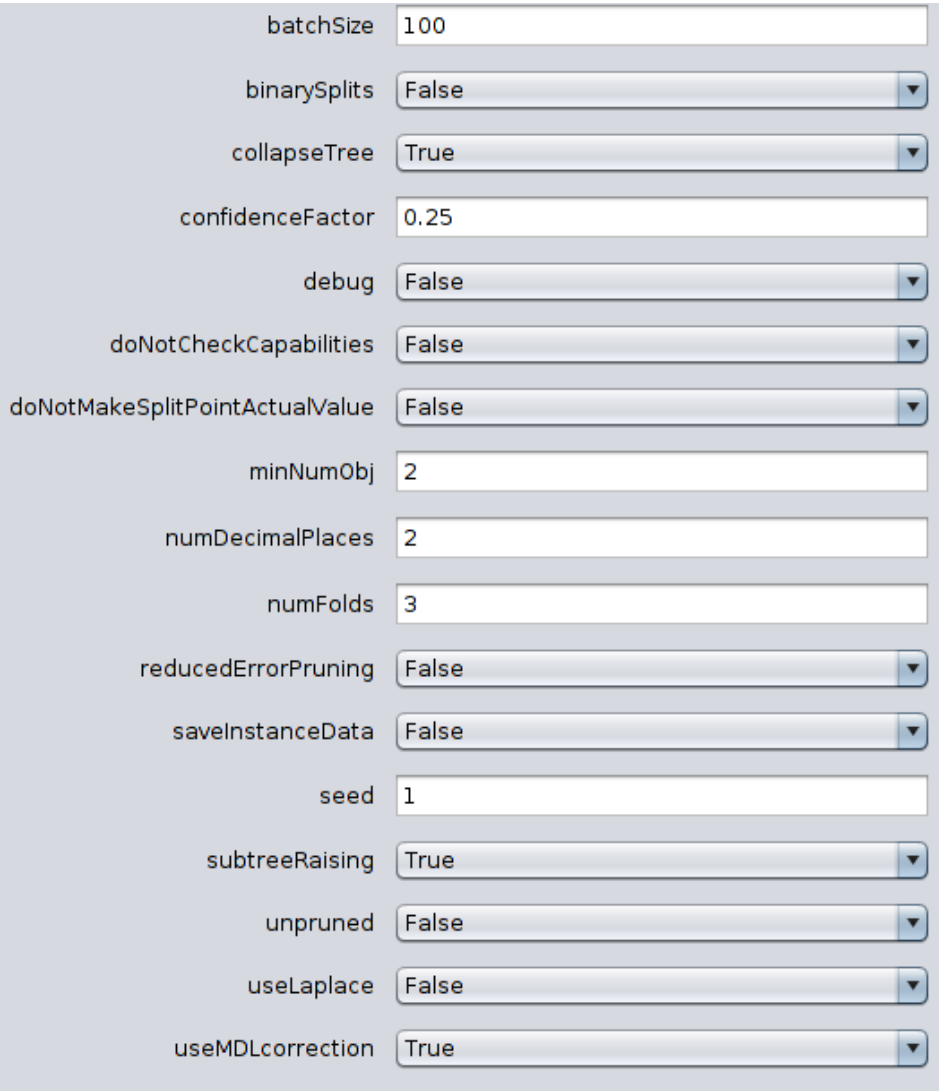
Rys. C.2. Ustawienia klasyfikatora *SVM* w środowisku Weka

C.3. Parametry klasyfikatora J48

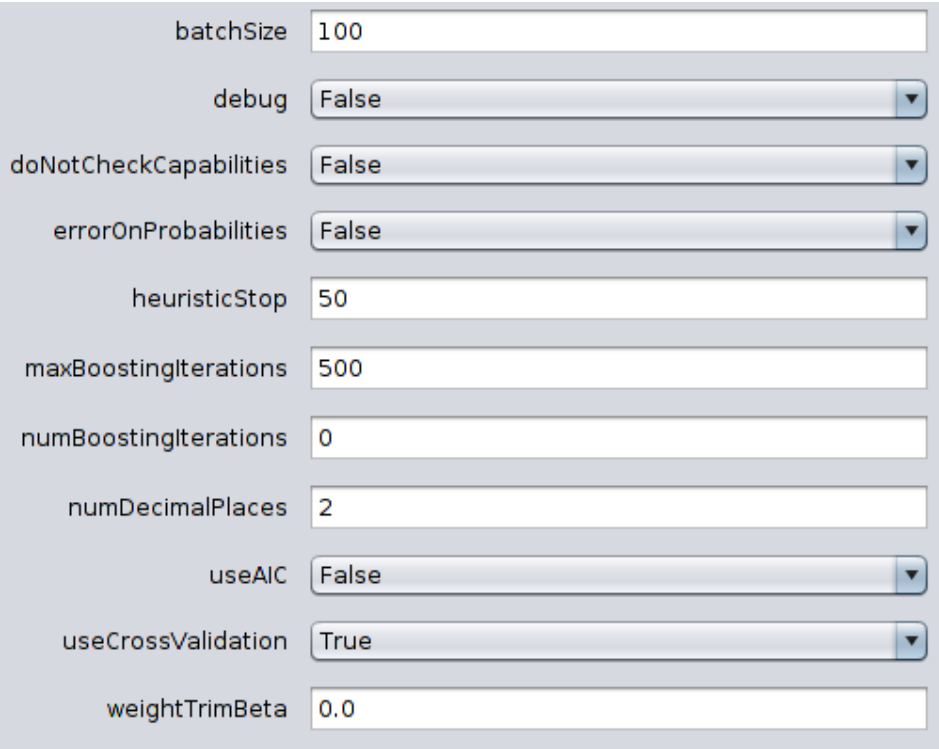
Na rysunku C.2 przedstawione zostały ustawienia klasyfikatora *J48* w systemie Weka. Ustawienia te są domyślnymi w Wece. W każdym przeprowadzonym eksperymencie są takie same.

C.4. Parametry klasyfikatora LMT

Na rysunku C.4 przedstawione zostały domyślnie ustawienia klasyfikatora *LMT* w systemie Weka. W każdym przeprowadzonym eksperymencie ustawienia pozostawały takie same.



Rys. C.3. Ustawienia klasyfikatora *J48* w środowisku Weka



Rys. C.4. Ustawienia klasyfikatora *LMT* w środowisku Weka

Bibliografia

(2017). Wiedza. W: *Encyklopedia PWN*. Państwowe Wydawnictwo Naukowe.

Abdulsahib, A. i Kamaruddin, S. (2015). Graph based text representation for document clustering. *Journal of Theoretical and Applied Information Technology*, 76.

Acedański, S. (2010). A morphosyntactic brill tagger for inflectional languages. W: Loftsson, H., Rögnvaldsson, E., i Helgadóttir, S., red., *Advances in Natural Language Processing*, volume 6233 of *Lecture Notes in Computer Science*, strony 3–14. Springer.

Acedański, S. i Gołuchowski, K. (2009). A Morphosyntactic Rule-Based Brill Tagger for Polish. W: *Recent Advances in Intelligent Information Systems*, strony 67–76, Kraków, Poland. Academic Publishing House EXIT.

Agirre, E., De Lacalle, O. L., i Soroa, A. (2009). Knowledge-based wsd on specific domains: Performing better than generic supervised wsd. W: *Proceedings of the 21st International Jont Conference on Artifical Intelligence*, IJCAI'09, strony 1501–1506, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Agirre, E., De Lacalle, O. L., i Soroa, A. (2013). Random walks for knowledge-based word sense disambiguation. MIT Press.

Agirre, E. i Soroa, A. (2009). Personalizing pagerank for word sense disambiguation. W: *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '09, strony 33–41, Stroudsburg, PA, USA. Association for Computational Linguistics.

Aleixo, P. i Pardo, T. A. S. (2008). Finding related sentences in multiple documents for multidocument discourse parsing of brazilian portuguese texts. W: *Companion Proceedings of the XIV Brazilian Symposium on Multimedia and the Web*, WebMedia '08, strony 298–303, New York, NY, USA. ACM.

Allan, J. (1996). Automatic hypertext link typing. W: *Proceedings of the the Seventh ACM Conference on Hypertext*, HYPERTEXT '96, strony 42–52, New York, NY, USA. ACM.

Altman, A. i Tennenholtz, M. (2005). Ranking systems: The pagerank axioms. W: *Proceedings of the 6th ACM Conference on Electronic Commerce*, EC '05, strony 1–8, New York, NY, USA. ACM.

Antoniou, G. i van Harmelen, F. (2004). *Web Ontology Language: OWL*, strony 67–92. Springer Berlin Heidelberg, Berlin, Heidelberg.

Barker, K., Agashe, B., Chaw, S. Y., Fan, J., Friedland, N. S., Glass, M. R., Hobbs, J. R., Hovy, E. H., Israel, D. J., Kim, D. S., Mulkar-Mehta, R., Patwardhan, S., Porter, B. W., Tecuci, D., i Yeh, P. Z. (2007). Learning by reading: A prototype system, performance baseline and lessons learned. W: *Proceedings of the Twenty-Second AAAI Conference on Artificial Intelligence, July 22-26, 2007, Vancouver, British Columbia, Canada*, strony 280–286.

Bas, D., Broda, B., i Piasecki, M. (2008). Towards word sense disambiguation of polish. W: *IMCSIT*.

Bobrow, D. G., Kaplan, R. M., Kay, M., Norman, D. A., Thompson, H., i Winograd, T. (1986). Gus: A frame driven dialog system. W: Grosz, B. J., Sparck Jones, K., i Webber, B. L., red., *Natural Language Processing*, strony 595–604. Kaufmann, Los Altos, CA.

Brill, E. (1992). A simple rule-based part of speech tagger. W: *Proceedings of the Third Conference on Applied Natural Language Processing*, ANLC '92, strony 152–155, Stroudsburg, PA, USA. Association for Computational Linguistics.

Broda, B. i Kędzia, P. (2011). *Text, Speech and Dialogue: 14th International Conference, TSD 2011, Pilsen, Czech Republic, September 1-5, 2011. Proceedings*, rozdział Finding the Optimal Number of Clusters for Word Sense Disambiguation, strony 388–394. Springer Berlin Heidelberg, Berlin, Heidelberg.

Broda, B., Marcińczuk, M., Maziarz, M., Radziszewski, A., i Wardyński, A. (2012). KPWr: Towards a free corpus of Polish. W: Calzolari, N., Choukri, K., Declerck, T., Doğan, M. U., Maegaard, B., Mariani, J., Odijk, J., i Piperidis, S., red., *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

Broda, B. i Piasecki, M. (2011). Evaluating lexicographer controlled semi-automatic word sense disambiguation method in a large scale experiment. *Control and Cybernetics*, Vol. 40, no 2:419–436.

Bunke, H. (1997). On a relation between graph edit distance and maximum common subgraph. *Pattern Recogn. Lett.*, 18(9):689–694, Elsevier Science Inc.

Bunke, H. i Shearer, K. (1998). A graph distance metric based on the maximal common subgraph. *Pattern Recogn. Lett.*, 19(3-4):255–259, Elsevier Science Inc.

Candan, K. S., Liu, H., i Suvarna, R. (2001). Resource description framework: Metadata and its applications. *SIGKDD Explor. Newsl.*, 3(1):6–19, ACM.

Champin, P.-A. i Solnon, C. (2003). *Measuring the Similarity of Labeled Graphs*, strony 80–95. Springer Berlin Heidelberg, Berlin, Heidelberg.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. Lawrence Erlbaum Associates.

Collins, A. M. i Quillian, M. R. (1995). Computation & intelligence. rozdział Retrieval Time from Semantic Memory, strony 191–201. American Association for Artificial Intelligence, Menlo Park, CA, USA.

Dehak, N., Dehak, R., Glass, J., Reynolds, D., i Kenny, P. (2010). Cosine Similarity Scoring without Score Normalization Techniques. W: *Odyssey The Speaker and Language Recognition*, Brno, Czech Republic.

Deliyanni, A. i Kowalski, R. A. (1979). Logic and semantic networks. *Commun. ACM*, 22(3):184–192, ACM.

Dunn, O. J. (1961). Multiple comparisons among means. *Journal of the American*

Statistical Association, 56(293):52–64, Taylor Francis.

Dziob, A. i Piasecki, M. (2018). Implementation of the Verb Model in plWordNet 4.0. W: Bond, F., Fellbaum, C., i Vossen, P., red., *Proceedings of the 9th Global Wordnet Conference, Singapore, 8-12 January 2018*. Global WordNet Association.

Fernández, M.-L. i Valiente, G. (2001). A graph distance metric combining maximum common subgraph and minimum common supergraph. *Pattern Recogn. Lett.*, 22(6-7):753–758, Elsevier Science Inc.

Gower, J. C. (1982). Euclidean distance geometry. *The Mathematical Scientist*, 7:1–14.

Gutiérrez, Y., Vázquez, S., i Montoyo, A. (2012). A graph-based approach to wsd using relevant semantic trees and n-cliques model. W: Gelbukh, A. F., red., *CICLing (1)*, volume 7181 of *Lecture Notes in Computer Science*, strony 225–237. Springer.

Gärdenfors, P. (2004). *Conceptual spaces: The geometry of thought*. The MIT Press.

Haveliwala, T. H. (2002). Topic-sensitive pagerank. W: *Proceedings of the 11th International Conference on World Wide Web*, WWW '02, strony 517–526, New York, NY, USA. ACM.

Hobbs, J. R. (2008). Deep lexical semantics. W: *Text, Speech and Dialogue, 11th International Conference, TSD 2008, Brno, Czech Republic, September 8-12, 2008. Proceedings*, strona 13.

Hobbs, J. R., Stickel, M. E., Appelt, D. E., i Martin, P. (1993). Interpretation as abduction. *Artif. Intell.*, 63(1-2):69–142, Elsevier Science Publishers Ltd.

Hripcsak, G. i Rothschild, A. S. (2005). Agreement, the f-measure, and reliability in information retrieval. *J. of Am. Medical Informatics Association*, 12(3):296–298.

Jaccard, P. (1912). The distribution of the flora in the alpine zone. *New Phytologist*, 11(2):37–50.

Janz, A. (2016). System do rozpoznawania relacji semantycznych pomiędzy fragmentami tekstów. praca magisterska, Politechnika Wrocławska.

Jassem, K. i Pawluczuk, L. (2015). Automatic summarization of polish news articles by sentence selection. W: *2015 Federated Conference on Computer Science and Information Systems, FedCSIS 2015, Lódz, Poland, September 13-16, 2015*, strony 337–341.

Jaworski, W. i Kozakoszczak, J. (2016). Eniam: Categorial syntactic-semantic parser for polish. W: *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations*, strony 243–247. The COLING 2016 Organizing Committee.

Jiang, J. J. i Conrath, D. W. (1997). Semantic similarity based on corpus statistics and lexical taxonomy. W: *Proc of 10th International Conference on Research in Computational Linguistics, ROCLING'97*.

Kashyap, V., Bussler, C., i Moran, M. (2008). *The Semantic Web: Semantics for Data and Services on the Web*. Springer Publishing Company, Incorporated, wydanie 1.

Kobyliński, Ł., Wasiluk, M., i Wojdyga, G. (2018). Improving part-of-speech tagging by meta-learning. W: *International Conference on Text, Speech, and Dialogue*, strony 144–152. Springer.

Krasnowska-Kieraś, K. (2017). Morphosyntactic disambiguation for polish.

Kumar, Y., Salim, N., Hamza, A., i Abuobieda, A. (2012). *Automatic identification of cross-document structural relationships*, strony 26–29.

Kędzia, P. i Maziarz, M. (2013). Recognizing semantic relations within polish noun

phrase: A rule-based approach. W: Angelova, G., Bontcheva, K., i Mitkov, R., red., *Recent Advances in Natural Language Processing 2013*, strony 342–349, Hissar, Bulgaria. RANLP 2011 Organising Committee / ACL.

Kędzia, P. i Piasecki, M. (2014). Ruled-based, interlingual motivated mapping of plwordnet onto SUMO ontology. W: Calzolari, N., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., i Piperidis, S., red., *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014), Reykjavik, Iceland, May 26-31, 2014.*, strony 4351–4358.

Kędzia, P., Piasecki, M., Kocoń, J., i Indyka-Piasecka, A. (2014a). Distributionally extended network-based word sense disambiguation in semantic clustering of polish texts. *IERI Procedia*, 10(Complete):38–44.

Kędzia, P., Piasecki, M., Kocoń, J., i Indyka-Piasecka, A. (2014b). Distributionally extended network-based word sense disambiguation in semantic clustering of polish texts. W: *International Conference on Future Information Engineering (FIE 2014), Beijing, China 7-8 July, 2014*, strony 38–44.

Kędzia, P., Piasecki, M., i Orlińska, M. J. (2015). Word sense disambiguation based on large scale polish clarin heterogeneous lexical resources. *Cognitive Studies*, 14(To appear).

Landwehr, N., Hall, M., i Frank, E. (2005). Logistic model trees. 95(1-2):161–205.

Lehmann, F. (1992). Semantic networks. *Computers & Mathematics with Applications*, 23(2):1 – 50.

Levene, H. (1960). *Robust Tests for Equality of Variances*, strony 278–292. Stanford University Press, Stanford, Calif.

Levi, G. (1973). A note on the derivation of maximal common subgraphs of two directed or undirected graphs. *CALCOLO*, 9(4):341.

Lin, G.-H. i Xue, G. (1999). Steiner tree problem with minimum number of steiner points and bounded edge-length. *Information Processing Letters*, 69(2):53 – 57.

Mann, W. C. i Thompson, S. A. (1987). Rhetorical structure theory: A framework for the analysis of texts. Technical Report ISI/RS-87-185, Information Sciences Institute.

Mann, W. C. i Thompson, S. A. (1988). Rhetorical structure theory: Toward a functional theory of text organization. *Text*, 8(3):243–281.

Manning, C. D., Raghavan, P., i Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA.

Marcinczuk, M. (2015). Automatic construction of complex features in conditional random fields for named entities recognition. W: *RANLP*.

Marciniak, M. i Mykowiecka, A. (2014). Terminology extraction from medical texts in Polish. *Journal of Biomedical Semantics*, 5, BioMed Central.

Marcińczuk, M., Kocoń, J., i Janicki, M. (2013). Liner2 – a customizable framework for proper names recognition for Polish. W: Bembenik, R., Skonieczny, L., Rybinski, H., Kryszkiewicz, M., i Niezgodka, M., red., *Intelligent Tools for Building a Scientific Information Platform*, strony 231–253.

Maziarz, M., Piasecki, M., i Rudnicka, E. (2014). Słowosieć – polski wordnet. Proces tworzenia tezauru. *Polonica*, XXXIV:79–97.

Maziarz, M., Piasecki, M., Rudnicka, E., Szpakowicz, S., i Kędzia, P. (2016). PlWordNet 3.0 – a Comprehensive Lexical-Semantic Resource. W: Calzolari, N., Matsumoto, Y., i Prasad, R., red., *COLING 2016, 26th International Conference on Computational*

Linguistics, Proceedings of the Conference: Technical Papers, December 11-16, 2016, Osaka, Japan, strony 2259–2268. ACL, ACL.

Maziarz, M., Piasecki, M., i Szpakowicz, S. (2012). Approaching plWordNet 2.0. W: *Proceedings of the 6th Global Wordnet Conference*, Matsue, Japan.

Maziarz, M., Piasecki, M., i Szpakowicz, S. (2013). The chicken-and-egg problem in wordnet design: synonymy, synsets and constitutive relations. *Language Resources and Evaluation*, 47(3):769–796.

Maziero, E. G., del Rosario Castro Jorge, M. L., i Pardo, T. A. S. (2010). Identifying multidocument relations. W: Sharp, B. i Zock, M., red., *NLPCS*, strony 60–69. SciTePress.

Maziero, E. G., Jorge, M. L. D. R. C., i Pardo, T. A. S. (2014). Revisiting cross-document structure theory for multi-document discourse parsing. *Inf. Process. Manage.*, 50(2):297–314, Pergamon Press, Inc.

McKeown, K. i Radev, D. R. (1995). Generating summaries of multiple news articles. W: *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '95, strony 74–82, New York, NY, USA. ACM.

Miłkowski, M. i Lipski, J. (2009). Using srx standard for sentence segmentation in languagetool. *Human Language Technologies as a Challenge for Computer Science and Linguistics*, strony 556–560.

Minsky, M. (1981). A framework for representing knowledge. W: Haugeland, J., red., *Mind Design: Philosophy, Psychology, Artificial Intelligence*, strony 95–128. MIT Press, Cambridge, MA.

Murphy, K. P. (2013). *Machine learning : a probabilistic perspective*. MIT Press, wydanie 1.

Nachar, N. (2008). The mann-whitney u: A test for assessing whether two independent samples come from the same distribution. *Tutorials in Quantitative Methods for Psychology*, 4.

Neuhäuser, M. (2011). *Wilcoxon–Mann–Whitney Test*, strony 1656–1658. Springer Berlin Heidelberg, Berlin, Heidelberg.

Niles, I. i Pease, A. (2001). Towards a standard upper ontology. W: *Proceedings of the International Conference on Formal Ontology in Information Systems - Volume 2001*, FOIS '01, strony 2–9, New York, NY, USA. ACM.

Niles, I. i Pease, A. (2004). Linking lexicons and ontologies: Mapping wordnet to the suggested upper merged ontology.

Page, L., Brin, S., Motwani, R., i Winograd, T. (1999). The pagerank citation ranking: Bringing order to the web.

Pazienza, M. T., Pennacchiotti, M., i Zanzotto, F. M. (2005). Terminology extraction: An analysis of linguistic and statistical approaches. W: Sirmakessis, S., red., *Knowledge Mining*, strony 255–279, Berlin, Heidelberg. Springer Berlin Heidelberg.

Pease, A. (2011). *Ontology: A Practical Guide*. Articulate Software Press, Angwin, CA.

Piasecki, M. (2014). User-driven language technology infrastructure – the case of clarin-pl. W: *Proceedings of the Ninth Language Technologies Conference*, Ljubljana, Slovenia.

Piasecki, M., Kędzia, P., i Orlińska, M. (2016a). plWordNet in word sense disam-

biguation task. W: Mititelu, V. B., Forăscu, C., Fellbaum, C., i Vossen, P., red., *Proceedings of the 8th Global Wordnet Conference, Bucharest, 27-30 January 2016*, strony 280–289. Global Wordnet Association.

Piasecki, M., Szpakowicz, S., i Broda, B. (2009a). *A Wordnet from the Ground Up*. Oficyna Wydawnicza Politechniki Wroclawskiej, Wrocław.

Piasecki, M., Szpakowicz, S., i Broda, B. (2009b). *A wordnet from the ground up*. Oficyna Wydawnicza Politechniki Wroclawskiej.

Piasecki, M., Szpakowicz, S., Maziarz, M., i Rudnicka, E. (2016b). PlWordNet 3.0 – Almost There. W: Mititelu, V. B., Forăscu, C., Fellbaum, C., i Vossen, P., red., *Proceedings of the 8th Global Wordnet Conference, Bucharest, 27-30 January 2016*, strony 290–299. Global Wordnet Association.

Przepiórkowski, A. (2004). *Korpus IPI PAN. Wersja wstępna*. Instytut Podstaw Informatyki, Polska Akademia Nauk, Warszawa.

Przepiórkowski, A., Bańko, M., Górski, R., i Lewandowska-Tomaszczyk, B. (2012). *Narodowy Korpus Języka Polskiego*. Wydawnictwo Naukowe PWN, Warszawa.

Przepiórkowski, A., Hajnicz, E., Patejuk, A., i Woliński, M. Extended phraseological information in a valence dictionary for NLP applications. strony 83–91.

Przepiórkowski, A., Hajnicz, E., Patejuk, A., Woliński, M., Skwarski, F., i Świdziński, M. Walenty: Towards a comprehensive valence dictionary of Polish. strony 2785–2792.

Przepiórkowski, A., Skwarski, F., Hajnicz, E., Patejuk, A., Świdziński, M., i Woliński, M. (2014). Modelowanie własności składniowych czasowników w nowym słowniku walencyjnym języka polskiego. *Polonica*, XXXIII:159–178.

Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Radev, D., Otterbacher, J., i Zhang, Z. (2003). CSTBank: Cross-document Structure Theory Bank. http://tangra.si.umich.edu/clair/CSTBank.

Radev, D. R. (2000). A common theory of information fusion from multiple text sources step one: Cross-document structure. W: *Proceedings of the 1st SIGdial Workshop on Discourse and Dialogue - Volume 10*, SIGDIAL '00, strony 74–83, Stroudsburg, PA, USA. Association for Computational Linguistics.

Radev, D. R., Otterbacher, J., i Zhang, Z. (2004). Cst bank: A corpus for the study of cross-document structural relationships. W: *LREC*. European Language Resources Association.

Radziszewski, A. (2013). A tiered CRF tagger for Polish. W: Bembenik, R., Skonieczny, Ł., Rybiński, H., Kryszkiewicz, M., i Niezgódka, M., red., *Intelligent Tools for Building a Scientific Information Platform: Advanced Architectures and Solutions*, strona to appear. Springer Verlag.

Radziszewski, A., Maziarz, M., i Wieczorek, J. (2012). Shallow syntactic annotation in the Corpus of Wrocław University of Technolog. *Cognitive Studies*, 12.

Radziszewski, A. i Pawlaczek, A. (2012). Large-scale experiments with NP chunking of Polish. W: Sojka P., Horák A., K. I. P. K., red., *Proceedings of Text, Speech and Dialogue 2012*, volume 7499, strony 143–149, Brno, Czech Republic. Springer.

Radziszewski, A. i Pawlaczek, A. (2013). *Language Processing and Intelligent Information Systems: 20th International Conference, IIS 2013, Warsaw, Poland, June 17-18, 2013. Proceedings*, rozdział Incorporating Head Recognition into a CRF Chunker,

strony 22–27. Springer Berlin Heidelberg, Berlin, Heidelberg.

Radziszewski, A. i Śniatowski, T. (2011). Maca — a configurable tool to integrate Polish morphological data. W: *Proceedings of the Second International Workshop on Free/Open-Source Rule-Based Machine Translation*.

Radziszewski, A., Wardyński, A., i Śniatowski, T. (2011). Wccl: A morpho-syntactic feature toolkit. W: Habernal, I. i Matoušek, V., red., *Text, Speech and Dialogue*, strony 434–441, Berlin, Heidelberg. Springer Berlin Heidelberg.

Rish, I. (2001). An empirical study of the naive bayes classifier. W: *IJCAI 2001 workshop on empirical methods in artificial intelligence*, volume 3, strony 41–46. IBM New York.

Rosenthal, R. (1994). *Parametric measures of effect size. The handbook of research synthesis.* Number 6. Russell Sage Foundation.

Rudnicka, E., Bond, F., Grabowski, , Piasecki, M., i Piotrowski, T. (2018). Lexical Perspective on Wordnet to Wordnet Mapping. W: Bond, F., Fellbaum, C., i Vossen, P., red., *Proceedings of the 9th Global Wordnet Conference, Singapore, 8-12 January 2018*. Global WordNet Association.

Rudnicka, E., Bond, F., Grabowski, , Piotrowski, T., i Piasecki, M. (2019). Sense Equivalence in plWordNet to Princeton WordNet Mapping.

Rudnicka, E., Maziarz, M., Piasecki, M., i Szpakowicz, S. (2012). A Strategy of Mapping Polish WordNet onto Princeton WordNet. W: Kay, M. i Boitet, C., red., *Proceedings of COLING 2012: Posters*, strony 1039–1048, Mumbai, India. The COLING 2012 Organizing Committee. ACL Anthology.

Rybak, P. i Wróblewska, A. (2018). Semi-supervised neural system for tagging, parsing and lematization. W: *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, strony 45–54, Brussels, Belgium. Association for Computational Linguistics.

Rychlikowski, P., Zapotoczny, M., i Chorowski, J. (2017). Character-based neural pos tagger. strony 382–385.

Sanfeliu, A. i Fu, K. (1983). A distance measure between attributed relational graphs for pattern recognition. *IEEE Trans. Systems, Man, and Cybernetics*, 13(3):353–362.

Schapire, R. E. i Singer, Y. (2000). Boostexter: A boosting-based system for text categorization. *Machine Learning*, 39(2):135–168.

Schenker, A., Bunke, H., Last, M., i Kandel, A. (2005). *Graph-Theoretic Techniques for Web Content Mining*. World Scientific.

Shapiro, S. S. i Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, 52(3/4):591–611.

Smywiński-Pohl, A. (2013). Knowledge-based named entity recognition in polish. W: Ganzha, M., Maciaszek, L., i Paprzycki, M., red., *Proceedings of the 2013 Federated Conference on Computer Science and Information Systems (FedCSIS)*, strony 145–151. Polskie Towarzystwo Informatyczne, Warszawa.

Steinwart, I. i Christmann, A. (2008). *Support Vector Machines*. Springer Publishing Company, Incorporated, wydanie 1st.

Sumner, M., Frank, E., i Hall, M. (2005). Speeding up logistic model tree induction. W: *9th European Conference on Principles and Practice of Knowledge Discovery in Databases*, strony 675–683. Springer.

Trigg, R. (1983). A network-based approach to text handling for the online scientific

community. *Technical report TR-1346*.

Trigg, R. H. i Weiser, M. (1986). Textnet: A network-based approach to text handling. *ACM Trans. Inf. Syst.*, 4(1):1–23, ACM.

Vanderhulst, G. (2005). Mapping real domain ontologies to sumo: a case study.

Wagner, R. A. i Fischer, M. J. (1974). The string-to-string correction problem. *J. ACM*, 21(1):168–173, ACM.

Walas, M. i Jassem, K. (2010). Named entity recognition in a polish question answering system. *Intelligent Information Systems*, strony 181–192, Citeseer.

Wallis, W. D., Shoubridge, P., Kraetz, M., i Ray, D. (2001). Graph distances using graph union. *Pattern Recogn. Lett.*, 22(6-7):701–704, Elsevier Science Inc.

Waszczuk, J., Głowińska, K., Savary, A., Przepiórkowski, A., i Lenart, M. (2013). Annotation tools for syntax and named entities in the National Corpus of Polish. *International Journal of Data Mining, Modelling and Management*, 5(2):103–122.

Widdows, D. (2004). *Geometry and Meaning*. Center for the Study of Language and Information/SRI.

Woliński, M. (2006). Morfeusz—a practical tool for the morphological analysis of polish. W: *Intelligent information processing and web mining*, strony 511–520. Springer Berlin Heidelberg.

Wróbel, K. (2017). Krnnt: Polish recurrent neural network tagger. *Vetulani [12]*.

WrUT (2018). enWordNet 1.0. CLARIN-PL digital repository.

Wróblewska, A. (2014). *Polish Dependency Parser Trained on an Automatically Induced Dependency Bank*. Ph.D. dissertation, Institute of Computer Science, Polish Academy of Sciences, Warsaw.

Yogan, J. K. i Naomie, S. (2013). Feed-forward network model for multi-document relation classification. *International Journal of Innovative Computing*, 1(1):1–4, UTM.

Zahri, N. A. H. B. i Fukumoto, F. (2011). *Multi-document Summarization Using Link Analysis Based on Rhetorical Relations between Sentences*, strony 328–338. Springer Berlin Heidelberg, Berlin, Heidelberg.

Zhang, Z., Otterbacher, J., i Radev, D. (2003). Learning cross-document structural relationships using boosting. W: *Proceedings of the Twelfth International Conference on Information and Knowledge Management*, CIKM '03, strony 124–130, New York, NY, USA. ACM.

Zhang, Z. i Radev, D. (2005). Combining labeled and unlabeled data for learning cross-document structural relationships. W: *Proceedings of the First International Joint Conference on Natural Language Processing*, IJCNLP'04, strony 32–41, Berlin, Heidelberg. Springer-Verlag.