Politechnika Warszawska

WYDZIAŁ ELEKTRONIKI
I TECHNIK INFORMACYJNYCH

Praca dyplomowa magisterska

na kierunku Informatyka
w specjalności Systemy informacyjno-decyzyjne

Odwzorowanie pojęć Słowosieci w taksonomię Linked Open Data

Hubert Święciński
Numer albumu 259119

promotor
dr inż. Mariusz Kamola

Warszawa 2020

Odwzorowanie pojęć Słowosieci w taksonomię Linked Open Data

Streszczenie

W ramach pracy podjęta została próba powiązania wyrażeń w zadanym tekście z odpowiadającymi i możliwie najlepiej dopasowanymi węzłami Link Open Data. Krokiem początkowym do realizacji powiązania była Słowosieć – największy polski wordnet, natomiast oczekiwanym efektem było powiązanie wyrażenia z adekwatnym adresem URL stron w DBpedii, która jest największą interdyscyplinarną bazą wiedzy wchodząca w skład LOD. Mapowanie uwzględnia kontekst znaczeniowy wyrażenia oraz w możliwie najmniejszy sposób korzysta z narzędzi wykorzystujących sieci neuronowe. Obranie takiego podejścia ma na celu umożliwienie wykorzystania otrzymanych wyników w weryfikacji rezultatów prac innych narzędzi wykorzystujących sieci neuronowe. W pracy zawarto także opis metodologii przeprowadzania mapowań oraz analizę otrzymanych wyników.

Słowa kluczowe: Słowosieć, infrastruktura Clarin, Link Open Data, DBpedia

Mapping the concepts of Słowosieć into the Linked Open Data taxonomy

Abstract

As part of the work, an attempt was made to link expressions in a given text to the corresponding and best-matched Link Open Data nodes. The initial step for the implementation of the linking was Słowosieć – the largest Polish wordnet, while the expected effect was the linking of the expression with an adequate URL of pages in DBpedia, which is the largest interdisciplinary knowledge base included in LOD. Mapping takes into account the semantic context of the expression and uses the tools using neural networks as little as possible. This approach is aimed at enabling the use of obtained results in verifying the results of the work of other tools using neural networks. The work also includes a description of the mapping methodology and analysis of the obtained results.

Key words: Słowosieć, Clarin infrastructure, Link Open Data, DBpedia

załącznik do zarządzenia nr 28/2016 r.
Rektora PW

Politechnika Warszawska

„załącznik nr 3 do zarządzenia nr 24/2016 Rektora PW

.....
miejscowość i data

.....
imię i nazwisko studenta

.....
numer albumu

.....
kierunek studiów

OŚWIADCZENIE

Świadomy/-a odpowiedzialności karnej za składanie fałszywych zeznań oświadczam, że niniejsza praca dyplomowa została napisana przeze mnie samodzielnie, pod opieką kierującego pracą dyplomową.

Jednocześnie oświadczam, że:

- niniejsza praca dyplomowa nie narusza praw autorskich w rozumieniu ustawy z dnia 4 lutego 1994 roku o prawie autorskim i prawach pokrewnych (Dz.U. z 2006 r. Nr 90, poz. 631 z późn. zm.) oraz dóbr osobistych chronionych prawem cywilnym,
- niniejsza praca dyplomowa nie zawiera danych i informacji, które uzyskałem/-am w sposób niedozwolony,
- niniejsza praca dyplomowa nie była wcześniej podstawą żadnej innej urzędowej procedury związanej z nadawaniem dyplomów lub tytułów zawodowych,
- wszystkie informacje umieszczone w niniejszej pracy, uzyskane ze źródeł pisanych i elektronicznych, zostały udokumentowane w wykazie literatury odpowiednimi odnośnikami,
- znam regulacje prawne Politechniki Warszawskiej w sprawie zarządzania prawami autorskimi i prawami pokrewnymi, prawami własności przemysłowej oraz zasadami komercjalizacji.

Oświadczam, że treść pracy dyplomowej w wersji drukowanej, treść pracy dyplomowej zawartej na nośniku elektronicznym (płycie kompaktowej) oraz treść pracy dyplomowej w module APD systemu USOS są identyczne.

.....
czytelny podpis studenta"

Spis treści

1 Wstęp

1.1 Rys historyczny

Już w 2007 roku Tim Berners-Lee, główny twórca światowej rozległej sieci komputerowej (ang. World Wide Web, w skrócie WWW) [21], mówiąc o jej fundamentach [2], stwierdził, że tak ogromny sukces sieci jako otwartego środowiska informacyjnego zawdzięcza ona głównie trzem czynnikom:

- nieograniczonym możliwościom łączenia dowolnych dokumentów w sieci,
- otwartym i udostępnionym standardom będącymi podstawą do dalszego rozwoju aplikacji,
- separacji warstw sieci, umożliwiającej niezależny rozwój każdej z warstw.

Aktualnie istniejący World Wide Web to sieć dokumentów – statycznych lub generowanych przez aplikacje, które w znacznej mierze powstały w celu odbioru ich treści przez człowieka, przez co nie są one przystosowane do tego by były w łatwy sposób przetwarzalne przez inne aplikacje w sposób automatyczny [16]. Trudność ta wynika głównie z faktu, iż w hipertekstowym języku znaczników (ang. HyperText Markup Language, w skrócie HTML), za pomocą którego opisywane są strony internetowe, zarówno informacje zawarte w dokumencie, jak i jego struktura stanowią nierozerwalną całość. Zmiany tego podejścia na przestrzeni ostatnich lat zaowocowały wykształceniem się różnych generacji sieci WWW: Web 1.0, Web 2.0 oraz Web 3.0. [13], z których to Web 3.0. rozumie World Wide Web jako strukturę zbudowaną z połączonych dokumentów wzbogaconą o metadane, które ułatwiają jej automatyczną eksplorację.

Koncepcja danych powiązanych wysunięta przez Tima Berners'a-Lee znana jako Linked Data (w skrócie LD) polega właśnie na wykorzystaniu sieci WWW wraz z jej technologiami do budowania sformalizowanych połączeń między danymi pochodzącymi z różnych zbiorów i co za tym idzie reprezentującymi różne dziedziny wiedzy. W ten sposób możliwe jest prezentowanie informacji jako swoistej bazy wiedzy, w której z każdym pojęciem związane są inne pokrewne lub powiązane pojęcia. Same metody reprezentacji, łączenia i współdzielenia danych w sieci wykorzystują istniejące standardy i narzędzia sieciowe [16] takie jak RDF (ang. Resource Description Framework) czy OWL (ang. Web Ontology Language).

Jeżeli dane powiązane są udostępnione w sposób wolny, do dowolnego wykorzystania i na licencjach umożliwiających nieograniczony dostęp, wówczas możemy mówić o nich jako o danych otwartych, czyli Link Open Data (w skrócie LOD) [14].

Systemem łączącym się z Linked Data jest leksykalna baza danych, zwana także wordnetem, która w sposób łatwodostępny dla automatycznych aplikacji przechowuje dane na temat słów oraz ich znaczeń. Pierwszym wordnetem był wordnet języka angielskiego Princeton WordNet, nad którym prace rozpoczęły się już w 1985 roku na uniwersytecie w Princeton pod kierownictwem George Armitage Millera [12]. Szybko idea wordnetów rozprzestrzeniła się na inne języki. W przypadku języka polskiego największym wordnetem jest Słowosieć – część infrastruktury CLARIN rozwijana pod wodzą Politechniki Wrocławskiej.

O ile Princeton WordNet posiada powiązania z węzłami LOD, o tyle w przypadku Słowosieci takie bezpośrednie powiązania dla aktualnej wersji nie istnieją. W celu umożliwienia wykorzystania całego kontekstu informacyjnego zawartego w Linked Open Data potrzebne było wytworzenie takich połączeń.

Aby takie mapowanie było możliwe, wymagany jest zbiór danych, baza wiedzy, która jest zarazem częścią Link Open Data oraz w zadowalający sposób pokrywa zakres pojęć zbioru danych, który chcemy przemapować, czyli w prezentowanym przykładzie pojęć Słowosieci. Jednym z największych zbiorów danych wchodzących w skład Link Open data jest DBpedia – projekt mający na celu odwzorowanie największej internetowej encyklopedii, czyli Wikipedii, do postaci danych powiązanych i to właśnie ona została wybrana jako punkt docelowy mapowania.

Innym zbiorem danych spełniającym wymienione warunki jest Wikidata. To tworzona przez społeczność baza wiedzy opierająca się o Wikipedię oraz centralna platforma do zarządzania Wikipedią i większością jej siostrzanych projektów [23]. Ponieważ projekt ten jest publiczny i został uruchomiony już pod koniec 2012 roku, to strona zgromadziła już dane o ponad 15 milionach podmiotów, w tym ponad 34 milionach jednostek oraz ponad 80 milionach etykiet i opisów w ponad 350 językach. Jest to praca ponad 40 tysięcy zarejestrowanych użytkowników, którzy ją nieustannie rozwijają [4].

W pracy przybliżone zostały cel i organizacja infrastruktury CLARIN z naciskiem na Słowosieć. Opisano strukturę i powiązania w niej występujące oraz uwzględniono sposób jej budowania. W kolejnych rozdziałach zaprezentowano pokrótce historię Danych Powiązanych i to, z czego wyewoluowały. Ponadto scharakteryzowano Link Open Data z zaakcentowaniem roli DBpedii oraz Wikidaty. Po wstępie teoretycznym opisano proces realizacji pracy: metodologię odwzorowywania pojęć, architekturę aplikacji realizującej mapowanie oraz sposoby testowania rezultatów. Kolejno przedstawiono otrzymane wyniki wraz z ich analizą i interpretacją. Praca skwitowana została podsumowaniem wraz ze wskazaniem potencjalnych kierunków jej dalszego rozwoju.

1.2 Kontekst badawczy

W zakresie mapowania pojęć Słowosieci w inne międzynarodowe i ogólnodostępne zasoby w ostatnim czasie były przeprowadzane liczne próby. Większość z nich dotyczyła mapowania polskiego wordnetu w Princeton WordNet (PWN). Przykładem takich działań może być praca podjęta przez badaczy z Politechniki Wrocławskiej: Pawła Kędzia, Macieja Piaseckiego, Ewę Rudnicką oraz Konrada Przybycienia [9]. W zaproponowanym przez nich algorytmie dane wymagały ręcznego preprocesing, który w przypadku procesu opisywanego w niniejszej pracy nie był konieczny.

Innym podejściem mającym także na celu zmapowanie Słowosieci w PWN była próba podjęta ponownie przez badaczy z PWr [17]. Tym razem mapowanie było przeprowadzane na podstawie analizy otoczenia węzłów Słowosieci i angielskiego wordnetu. W ramach pracy porównywane były cechy charakterystyczne samych węzłów oraz węzłów sąsiadujących.

W jeszcze innym podejściu do rozszerzenia Princeton WordNet o informacje ze Słowosieci, ponownie zaproponowanym przez Politechnikę Wrocławską [18], naukowcy posłużyli się podziałem pojęć na grupy oraz ich ręcznym przyporządkowaniem do odpowiadających pojęć z PWN.

Wszystkie z powyższych prac dotyczą odwzorowania w Princeton WordNet, a nie bezpośrednio w Link Open Data jak to ma miejsce w niniejszej pracy. Ponadto część przedstawionych podejść opiera się o ręczne metody mapowania, które są niewykorzystywane w przedstawianym procesie.

2 Cel pracy

Celem pracy było opracowanie metodologii mapowania pojęć z zadanego tekstu w taksonomię Link Open Data.

Na potrzeby realizacji celu należało stworzyć aplikację, która dla zadanego tekstu mapowałaby pojęcia w nim występujące na pojęcia w Link Open Data. Mapowanie to miało odbywać się z wykorzystaniem zasobów i narzędzi infrastruktury CLARIN-PL oraz z uwzględnieniem kontekstu znaczeniowego słów towarzyszących danemu pojęciu. Specyfika problemu oraz późniejsze potencjalne zastosowania aplikacji w weryfikacji działań narzędzi wykorzystujących sieci neuronowe wymagały, aby dane wynikowe charakteryzowały się wysokim poziomem ufności, a sama aplikacja nie korzystała bezpośrednio z sieci neuronowych.

Ponadto w ramach pracy należało także wypracować sposób testowania oraz oceny wynikowego mapowania, a następnie przeprowadzić same testy, które ocenią jakość mapownia oraz wskażą czy obrana ścieżka łączenia pojęć ma potencjał do dalszych prac w tym kierunku. Granulacja testów oraz ich rezultaty powinny dawać możliwość wskazania najbardziej zawodnych i niepewnych elementów procesu mapowania, oraz powinna dawać możliwość sprecyzowania płaszczyzn, na których niedomagające narzędzia powinny być udoskonalone.

3 Słowosieć

3.1 Opis infrastruktury CLARIN

CLARIN (ang. Common Language Resources and Technology Infrastructure – powszechne zasoby językowe i infrastruktura technologiczna) to ogólnoeuropejski projekt, którego celem jest udostępnianie trwałych i bezpiecznych usług zapewniających łatwy dostęp do narzędzi służących do przetwarzania języka [22]. Jest to część Europejskiej Mapy Drogowej Infrastruktury Naukowej (ESFRI – European Roadmap for Research Infrastructures, European Strategy Forum on Research Infrastructures) która jest strategicznym instrumentem służącym rozwojowi integracji naukowej Europy i wzmocnieniu jej zasięgu międzynarodowego [24].

Organem zarządzającym i koordynującym CLARIN jest ERIC (ang. European Research Infrastructure Consortium – Konsorcjum na rzecz Europejskiej Infrastruktury Badawczej). ERIC to międzynarodowy podmiot prawny ustanowiony przez Komisję Europejską w 2009 r. Członkami CLARIN ERIC są rządy państw i organizacje międzyrządowe, których celem jest objęcie infrastrukturą nie tylko wszystkich krajów Unii Europejskiej i krajów stowarzyszonych, ale także krajów niepowiązanych z UE [25][7].

Sama infrastruktura CLARIN składa się z sieci centrów, w których zależnie od typu centra prowadzone są prace nad innymi zagadnieniami systemu. Rozróżnia się cztery podstawowe typy centrów:

- typ A - odpowiedzialne za rozwój technologiczny systemów. W nich tworzone są usługi i funkcjonalności sieci,
- typ B - których zadaniem jest dostarczanie zasobów i narzędzi do przetwarzania języka naturalnego,
- typ C - zajmujące się opisem zasobów, czyli uzupełnianiem metadanych,
- typ K - wspierające bezpośrednio użytkowników infrastruktury, zapreniające im wsparcie merytoryczne.

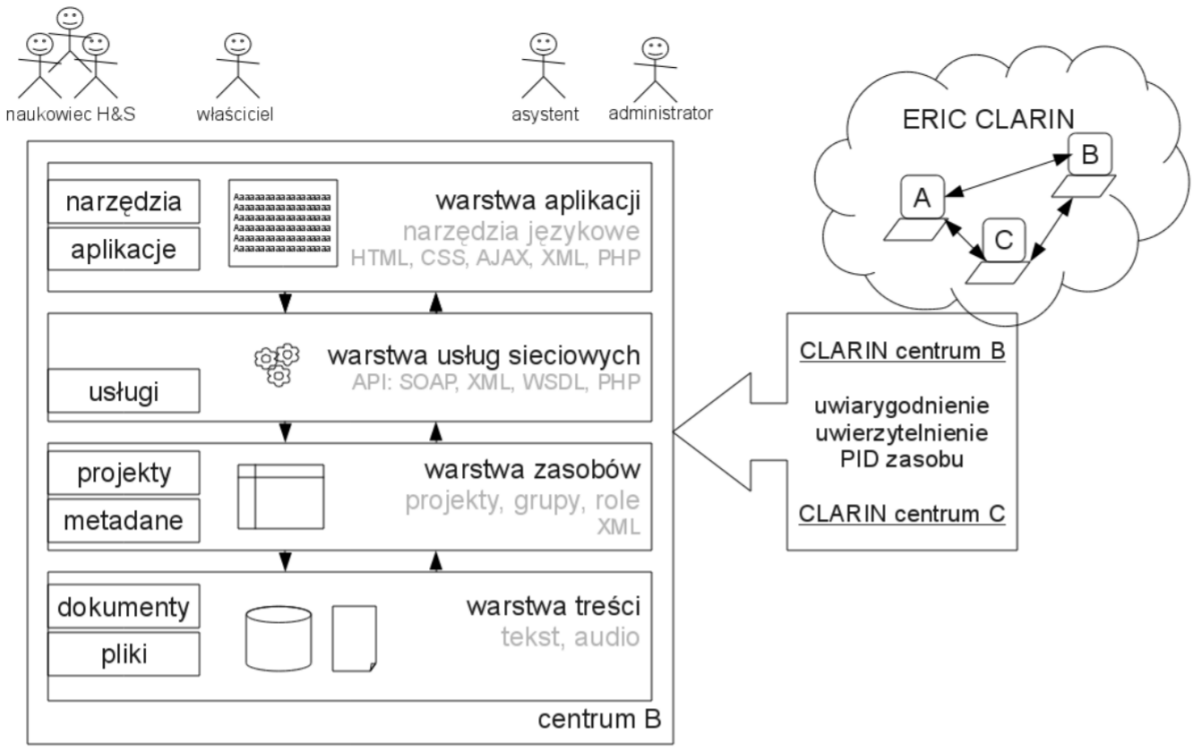
3.2 CLARIN-PL

CLARIN-PL to polska część europejskiej infrastruktury CLARIN. CLARIN-PL tworzą polskie uniwersytety i instytuty badawcze, w których powstają zarówno repozytoria tekstów pisanych i mówionych w języku polskim oraz tekstów równoległych w innych językach, jak i narzędzia do pracy z tekstami. System ten ma ułatwić pracę na tekstach źródłowych nie tylko naukowcom w Polsce, ale także w całej Europie [26].

W Polsce głównym podmiotem odpowiedzialnym za rozwój polskiej części systemu jest Politechnika Wrocławska, która jest centrum typu B. Głównymi zadaniami centrum jest [27]:

- budowa repozytorium z zasobami i narzędziami,
- dbanie o spójność techniczńa powstających zasobów,

- kontrola przestrzegania praw dotyczących własności intelektualnej, licencji i zasad etycznych powstających zasobów,
- ustanawianie polityki bezpieczeństwa.



Rysunek 1: Schemat architektury dla polskiego centrum typu B [27].

3.3 O Słowosieci

3.3.1 Definicja

Słowosieć wchodzi w skład zasobów i narzędzi CLARIN-PL i jest to polsko-angielska sieć leksykalno-semantyczna, w której znaczenie danej jednostki leksykalnej jest opisywane poprzez jej sąsiedztwo w sieci, które wyraża relacje znaczeniowe, w jakie wchodzi ona z innymi jednostkami [3].

Słowosieć tworzona jest w sposób półautomatyczny: manualną pracę lingwistów wspomagają narzędzia informatyczne, analizujące duże korpusy tekstów [15]. Oznacza to, że Słowosieć cechuje się stosunkowo wysoką jakością dostarczanych informacji, które zostały sprawdzone przez specjalistów w dziedzinie języka.

Obecnie Słowosieć opisuje 178000 rzeczowników, czasowników, przymiotników i przysłówków, zawiera niemal 259000 unikatowych znaczeń i ponad 600000 instancji relacji. Jest największym wordnetem na świecie, większym nawet od Princeton WordNet. Decyzją władz Politechniki Wrocławskiej Słowosieć jest udostępniana nieodpłatnie do wszelkich zastosowań, także komercyjnych, w oparciu o licencję wzorowaną na licencji Princeton WordNet [15].

3.3.2 Struktura

Struktura Słowosieci jest zbliżona do struktury swojego protoplasty, czyli Princeton WordNet. Podstawą struktury Słowosieci jest synset. Jest to zbiór jednostek leksykalnych, które reprezentują te same części mowy i wchodzą w takie same relacje semantyczne z innymi jednostkami leksykalnymi [15]. Samo słowo synset jest zbitką wyrazową angielskich słów synonym i set, które oznaczają odpowiednio synonim oraz zbiór. Znaczy to, że jest to zbiór słów mających to samo znaczenie, a jednostki leksykalne mające więcej niż jedno znaczenie należą do więcej niż jednego synsetu. Określa się je mianem jednostek polisemicznych.

Relacje leksykalno-semantyczne w poszczególnych wordnetach różnią się z zależności od języka. Biorąc pod uwagę rozbudowaną morfologię języka polskiego w Słowosieci, rozszerzono zestaw podstawowych relacji z WordNetu. Miało to na celu bardziej precyzyjne określanie znaczeń poszczególnych słów [11]. Same relacje może podzielić ze względu na to, czy łączą one jednostki (mają charakter derywacyjny), czy synsety (mają charakter semantyczny).

Relacja	**Opis**	**Przykłady**
Synonimia	łączy jednostki leksykalne w synsety	lokum i dom
Antonimia	wskazuje na każdą znaczeniową przeciwstawność między jednostkami leksykalnymi	zwierzę i roślina kręgowiec i bezkręgowiec mąż i żona
Synonimia międzyparadygmatyczna	łączy: rzeczownik odczasownikowy z czasownikiem, czasownik z przymiotnikiem, rzeczownik z przymiotnikiem	myślenie i myśleć żyć i życiowy beztroska i beztroski
Nacechowanie	dla wyrazów deminutywnych, wyrazów augmentatywnych / ekspresywnych oraz istot młodych	chłopak i chłopczyk dziewczyna i dziewucha kot i kocię
Żeńskość	dla nazw żeńskich wyprowadzanych z rzeczowników męskich	aktor i aktorka nauczyciel i nauczycielka
Rola	gdy rzeczownik jest derywatem odczasownikowym lub odrzeczownikowym	kłamca i kłamać skrzypek i skrzypce
Zawieranie roli	gdy czasownik jest derywatem odrzeczownikowym	dyrektorować i dyrektor monitorować i monitor
Nosiciel stanu/cechy	gdy rzeczownik jest derywatem odprzymiotnikowym; relacją odwrotną jest zawieranie stanu/cechy	ślepiec i ślepy głupek i głupi

Derywacyjność	łączy derywaty synchroniczne reprezentujące procesy mniej regularne	ołowica i ołów
Fuzzynimia	relacja nieokreślona	tort i urodziny

Tabela 1: Relacje jednostek w Słowosieci dla rzeczownika [11]

Najistotniejszymi relacjami w sieci są synonimia oraz hiperonimia i hiponimia, ponieważ synset uznaje się za zdefiniowany dopiero wtedy, gdy posiada on wyraz nadrzędny (hiperonim) lub jest częścią pewnej całości (jest meronimem). W przypadku nazw własnych to są one wprowadzane do Słowosieci tylko wtedy, kiedy funkcjonują wyrazy pospolite powstałe od nich np. Polska i polski.

Relacja	**Opis**	**Przykłady**
Hiperonimia / hiponimia	relacja nadrzędności i podrzędności; jest ona odwrotna	skoro owoc jest hiperonimem jabłka, to jabłko jest hiponimem owocu
Holonimia / meronimia	relacja cząstkowości / całościowości; nie zawsze jest ona odwracalna	las jest holonimem drzewa, drzewo zaś nie zawsze jest meronimem lasu
Bliskoznaczność	łączy leksemy o różnych rejestrach stylistycznych	słonina i szpyrka szkoła i buda oraz sztuba
Mieszkaniec	relacja o charakterze derywacyjnym; określa bycie mieszkańcem / obywatelem jakiegoś kraju, regionu, miasta	Polak i Polska wrocławianin i Wrocław

Tabela 2: Relacje synsetów w Słowosieci dla rzeczownika [11]

4 Linked Open Data

4.1 Definicja

Już w roku 1997 w globalnej sieci znajdowało się około 55 mln stron internetowych [19]. Dynamiczny rozwój Internetu, jaki postępował w zasadzie od początku jego wypuszczenia z rąk wojska sprawił, że od początkowego serwowania statycznych informacji, czytelnych jedynie na człowieka przeszedł do udostępniania interaktywnych treści, które z powodzeniem mogą być eksplorowane przez automaty. W ostatnich latach mówimy już o sieci WEB 3.0. zwaną również Semantic Web, która strukturą przypomina ogromną bazę danych, a Link Open Data jest jej realizacją.

4.2 Generacje sieci WWW

4.2.1 Web 1.0

WWW lub Web 1.0 to system powiązanych dokumentów hipertekstowych dostępnych za pośrednictwem internetu. Pierwsza implementacja sieci reprezentuje właśnie Web 1.0, którą według Bernersa-Lee można uznać za „sieć tylko do odczytu". Innymi słowy, wczesna sieć pozwalała nam jedynie na wyszukiwanie informacji. Interakcja z użytkownikiem istniała jedynie w szczątkowym stopniu. Celem strony internetowej było tylko i aż udostępnianie informacje online w dowolnym momencie każdemu, kto był nią zainteresowany [8].

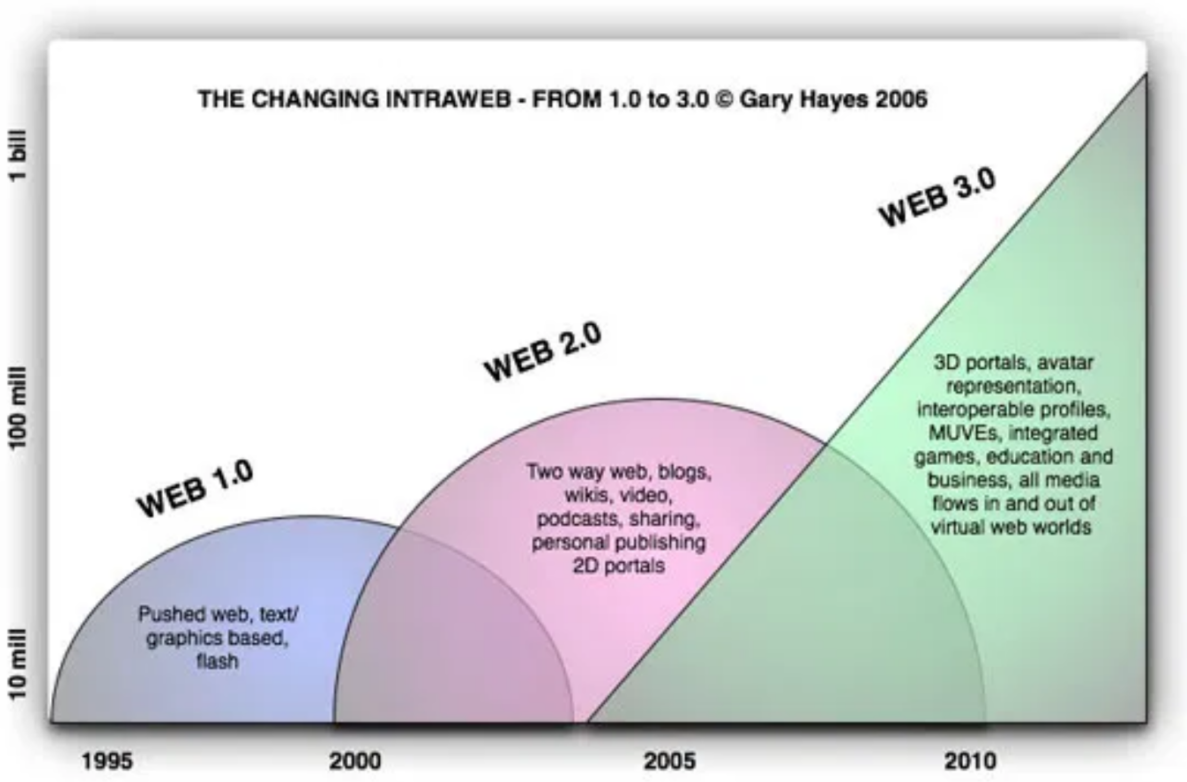
4.2.2 Web 2.0

Web 2.0 jest naturalną ewolucją sieci WWW, która koncentruje się na użytkownikach i ich interakcjach ze stroną. Termin Web 2.0 jest powszechnie kojarzony z aplikacjami internetowymi, które ułatwiają interaktywne dzielenie się informacjami, interoperacyjność i są zorientowane na użytkownika. Witryna Web 2.0 daje użytkownikom swobodę wyboru w zakresie interakcji i współpracy w ramach sieci społecznościowej. Umożliwia dialog medialny między twórcami treści (prosumentami) a odbiorcami (konsumentami). Użytkownicy nie są już ograniczeni jedynie do pasywnego przyjmowania treści, które zostały dla nich stworzone. Przykładami sieci 2.0 są serwisy społecznościowe, blogi, strony wiki czy witryny do udostępniania filmów [13].

4.2.3 Web 3.0

Web 3.0 opisuje ewolucję sieci w kierunku bazy danych. Tim Berners-Lee wyjaśnienia, że Web 3.0 byłby czymś podobnym do sieci typu „odczyt-zapis". Byłby on definiowany jako tworzenie wysokiej jakości treści i usług. Web 3.0. miałby na celu przekształcenie internetu w bazę danych, która mogłaby być eksplorowana przez aplikacje inne niż przeglądarki internetowe i byłaby ona wykorzystywana np. przez sztuczną inteligencję. Web 3.0 to sieć, w której zanika koncepcja witryny lub strony, gdzie dane nie są własnością, ale są udostępniane, a usługi prezentują je inaczej w zależności od potrzeb aplikacji lub użytkownika. Web 3.0 jest nazywana

także siecią semantyczną (ang. The Semantic Web), w której kluczowymi pojęciami są metadane i struktura [20][5].



Rysunek 2: Ewolucja sieci Web [10].

4.3 Link Data i Link Open Data

Linked Data to jeden z głównych filarów sieci semantycznej (Web 3.0), który polega na tworzeniu połączeń między zestawami danych, zrozumiałych nie tylko dla ludzi, ale także dla maszyn. Link Data można rozumieć jako zestaw dobrych praktyk projektowania danych połączonych dostosowanych do eksploracji przez aplikacje. Ów dobre praktyki zostały okerśone przez Tima Bernersa-Lee i mają one na celu budowanie łatwej do ekspolracji, spójnej i skalowalnej sieci [31]:

- Identyfikatory URI jako nazwy.

  Uniform Resource Identifier (URI) to powszecgnie stosowany system identyfikacji służący do nadawania unikatowych nazw przedniotom i zjawiskom z najróżniejszych dziedzin - od treści cyfrowych dostępnych w Internecie po rzeczywiste obiekty i abstrakcyjne pojęcia. Za pomocą identyfikatorów URI można rozróżniać różne rzeczy lub stwierdzać, że rzecz z jednego zestawu danych jest taka sama jak inna w innym zbiorze danych.

- Identyfikatory URI HTTP, aby ludzie mogli wyszukać nazwy.

  Ponieważ protokół HTTP zapewnia prosty i powszechny mechanizm pozyskiwania zasobów, a identyfikatory URI pomagają nazywać zasoby to w rezultacie w łatwy sposób

można odszukiwać zasoby w sieci. Przyspiesza to publikowanie wszelkiego rodzaju danych i dodawanie ich do globalnej przestrzeni danych.

- Poszukiwanie identyfikatorów URI, korzystając z ogólnoprzyjętych standardów (RDF, SPARQL).

  RDF i SPARQL pozwalają na efektywne wyszukiwanie zasobów korzystając z URI.

  RDF to oparty na grafie format reprezentacji do publikowania i wymiany danych w sieci opracowany przez W3C. Jest również wykorzystywany w bazach danych z grafami semantycznymi (znanych również jako triplestores RDF) - technologia opracowana do przechowywania połączonych danych i wnioskowania nowych faktów z istniejących.

  SPARQL z kolei jest znormalizowanym językiem zapytań W3C do wyszukiwania i manipulowania danymi przechowywanymi w formacie RDF. Jako taki pozwala nam przeszukiwać sieć danych (lub dowolną bazę danych) i identyfikować relacje.

- Łączenie identyfikatorów URI, w celu dostarczenia dodatowych informacji

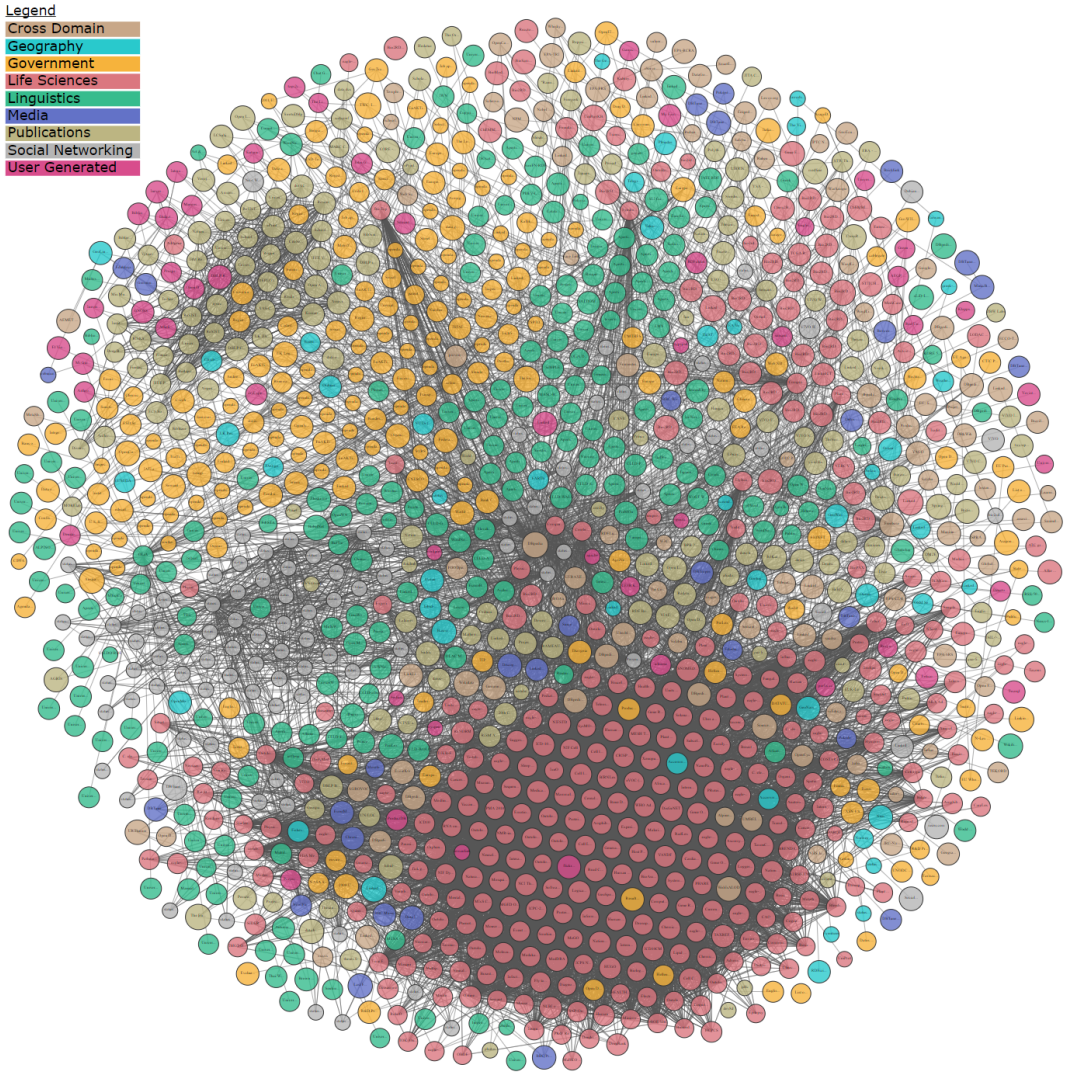
  Podobnie jak w sieci hipertekstowej, linki do innych identyfikatorów URI sprawiają, że dane są ze sobą połączone, przez co proces przeszukiwania danych jest prostszy i szybszy. Łącząc nowe informacje z istniejącymi zasobami, maksymalizujemy ponowne użycie i łączenie istniejących danych oraz wzbogacamy sieć o dodatkowe informacje istotne z punktu widzenia przetwarzania maszynowego.

Otwarte dane połączone będące wolnodostępnymi danymi połączonymi na stan z marca 2019 roku zawierają 1239 zbiorów danych z 16 147 linkami [28].

4.4 DBpedia

Centralnym punktem LOD jest DBpedia, do której łączy się znaczna część zbiorów. Ma ona na celu usystematyzowanie i powiązanie ze sobą danych z Wikipedii oraz udostępnianie ich w łatwy i przejrzysty sposób. Projekt ten został zainicjalizowany przez Wolny Uniwersytet Berlina oraz Uniwersytetu w Lipsku, a pierwsze zbiory informacji zostały opublikowane już w 2007 roku. Cała DBpedia jest udostępniana na licencji wolnego oprogramowania zarówno do zastosowań prywatnych, jak i komercyjnych. Sama DBpedia na stan z kwietnia 2016 roku opisywała ponad 6 milionów encji, z czego 5,2 miliona było sklasyfikowanych w spójnej ontologi, w tym 1,5 miliona osób, 810 tysięcy miejsc, 135 tysięcy albumów muzycznych, 106 tysięcy filmów, 20 tysięcy gier wideo, 275 tysięcy organizacji, 301 tysięcy gatunków roślin i zwierząt oraz 5 tysięcy chorób [29].

Do udostępniania danych DBpedia korzysta z RDF (Resource Description Framework). Jest to język służący do opisywania zasobów, ze składnią opartą na XMLu, opracowany przez W3C. W RDF zasoby są opisywane za pomocą trójek (ang. triples) podmiot - orzeczenie - obiekt (ang. subject – predicate – object) np. Quo vadis - autor - Henryk Sienkiewicz. W 2014 roku DBpedia zawierała 3 miliardy trójek, z czego 580 milionów było wyekstrahowanych z anglojęzycznej Wikipedii [30].



Rysunek 3: Chmura Link Open Data [28].

4.5 Wikidata

Podobnie jak DBpedia stanowi silnie połączony z innymi bazami punkt Link Open Data. Wikidata to baza wiedzy tworzona na podstawie Wikipedii. Posiada prawie 74 milionów pojęć [32]. Co więcej jest centralną platformą do zarządzania Wikipedią oraz podobnie jak DBpedia udostępnia swoje zasoby za pośrednictwem RFD.

5 Realizacja pracy

5.1 Opis metodologii mapowania danych

Mapowanie danych ze Słowosieci w taksonomię Link Open Data było przeprowadzane za pomocą programu komputerowego, który przyjął formę aplikacji konsolowej. Uruchomiona aplikacja miała za zadanie dla każdego pojęcia z zadanego tekstu pojąć próbę przyporządkowania strony w DBpedii, która jest częścią LOD, i w której to zawarte są informacje powiązane z danym pojęciem. Tekst do analizy mógł być podany aplikacji bezpośrednio w komendzie uruchamiającej program lub też dane mogły zostać pobrane z pliku tekstowego.

Dane wynikowe były zapisywane do plików w formacie json lub xml, przez co mogły być z powodzeniem przetwarzane wzrokowo. Ponadto wymienione formaty plików, ze względu na prostotę budowy i popularność, mogą z łatwością posłużyć jako pliki wsadowe do innych programów.

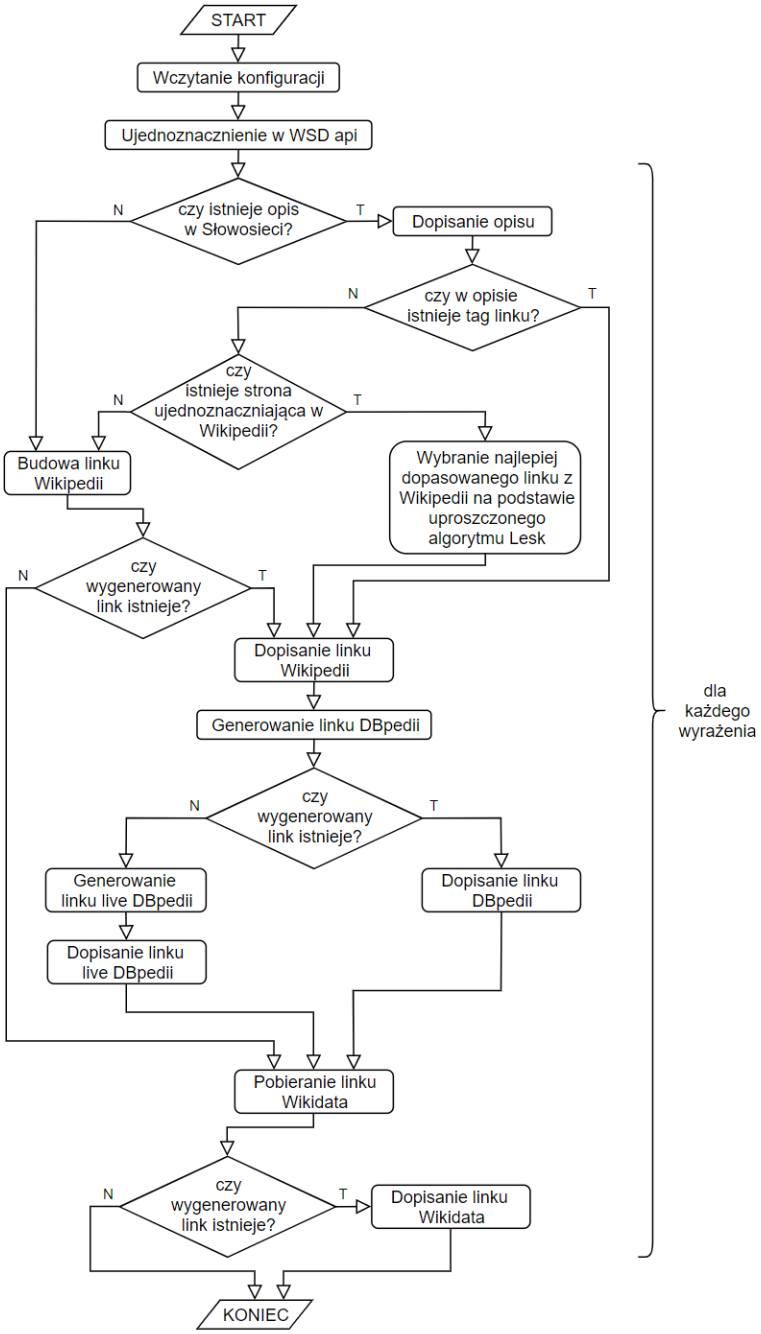
5.2 Wykorzystane technologie

Aplikacja w całości powstała przy pomocy Javy w wersji 10. Budowanie paczki odbywało się za pomocą Mavena. Paczka była budowana do postaci jara ze względu na możliwość łatwego uruchamiania aplikacji na różnych platformach.

5.3 Parametry konfiguracyjne programu

- -t <tekst> - podanie tekstu do analizy,
- -f <nazwa_pliku> - podanie nazwy pliku z tekstem do analizy (posiada niższy priorytet niż opcja -t),
- -s - włączenie trybu cichego, wyłączenie logowania do konsoli,
- -j <nazwa_pliku> - podanie nazwy pliku, do którego zostanie zapisany rezultat działania programu w postaci jsona,
- -ji <liczba_wcięć> - ustawienie wielkości wcięć w pliku json,
- -x <nazwa_pliku> - podanie nazwy pliku, do którego zostanie zapisany rezultat działania programu w postaci xmla,
- -xi <liczba_wcięć> - ustawienie wielkości wcięć w pliku xml,

5.4 Fazy działania programu



Rysunek 4: Schemat działania aplikacji.

5.4.1 Wczytanie konfiguracji

Konfiguracja programu jest wczytywana z linii poleceń podczas uruchamiania aplikacji. W ten sposób z łatwością można zmieniać parametry wywołania programu przy każdorazowym jego uruchomieniu, bez potrzeby modyfikacji plików konfiguracyjnych. Zgodnie z parametrami opisanymi w punkcie 5.3 konfiguracji podlega zarówno sposób wczytywania danych, jak i

logowania rezultatów pracy programu.

5.4.2 Ujednoznacznienie w CLARIN api

W celu jednoznacznego połączenia leksemów z zadanego tekstu z synsetami ze Słowosieci wykorzystano interfejs programistyczny udostępniany przez infrastrukturę CLARIN. Pozwala on na przypisanie wyrażeniu znaczenia ze Słowosieci. Wywołania do interfejsu zostały skonfigurowane tak, aby możliwe było wykrywanie wyrażeń wielowyrazowych.

5.4.3 Ekstrakcja danych ze Słowosieci

Ze względu na niestabilność zasobów udostępnianych przez CLARIN wszelka analiza danych ze Słowosieci była przeprowadzana offline, na wersji kompatybilnej z wersją interfejsu ujednoznaczniającego. Sama Słowosieć zawiera dla każdego pojęcia szereg informacji takich jak definicja czy przykłady użycia. W niektórych przypadkach w opisie w Słowosieci znajduje się także link kierujący bezpośrednio do definicji danego pojęcia w Wikipedii.

Ze względu na to, że informacje zostały zweryfikowane przez językoznawców, w pracy uznano je za najbardziej wiarygodne i to informacje pozyskiwane w tym kroku są priorytetowo uznawane za poprawne dla danego pojęcia.

5.4.4 Ujednoznacznienie za pośrednictwem Wikipedii

Dla pojęć, które posiadają więcej niż jedno znaczenie w Wikipedii, powstają tzw. strony ujednoznaczniające. Adresy URL takich stron przyjmują postać:

https://pl.wikipedia.org/wiki/<pojęcie>_ujednoznacznienie

w przypadku, gdy jedno ze znaczeń słowa jest dominujące. W przypadku, w którym wszystkie znaczenia są jednakowo często występujące adres URL ma postać:

https://pl.wikipedia.org/wiki/<pojęcie>

Samo ujednoznacznienie opierało się o uproszczony algorytm Lesk. Polega on na zliczeniu słów wspólnych z definicją pojęcia [1]. Jako tekst reprezentujący wybrano pierwszy akapit strony z Wikipedii danego pojęcia, natomiast odniesieniem był kontekst analizowanego tekstu. Co więcej, aby zwiększyć trafność ujednoznacznienia, wykluczono z obu źródeł słowa będące na stop liście słów dla języka polskiego. Stop lista to zbiór słów nieistotnych z punktu dostarczanych informacji.

Jako że informacje uzyskiwane w tym procesie są wynikiem pewnego rodzaju ujednoznacznienia, można je uznać za stosunkowo wiarygodne. Ich stopień wiarygodności nie jest tak wysoki, jak stopień danych zweryfikowanych przez językoznawców w Słowosieci, ale nie są też to dane będące wynikiem przypadkowych operacji na ciągach znaków.

5.4.5 Budowa linku Wikipedii

Każdy adres URL Wikipedii poprzedzony jest protokołem oraz nazwą domeny:

https://pl.wikipedia.org/wiki

Po nich następuje nazwa pojęcia, jakie opisuje dana strona. Jeżeli na Wikipedii występuje więcej niż jedno pojęcie o takiej samej nazwie, dodatkowo dopisany jest przyrostek doprecyzowujący, który określa jakiej dziedziny wiedzy dane zagadnienie dotyczy.

W tym przypadku wiarygodność tych danych jest stosunkowo niska i nie można mówić o żadnym ich zweryfikowaniu. Opierają się one o przekształcenia ciągów znaków i nie są w żaden sposób weryfikowane.

5.4.6 Mapowanie do LOD

Przyrostki adresów URL Wikipedii i DBpedii są takie same, więc posiadając adres do strony z Wikipedii można w łatwy sposób przemapować go do adresu strony DBpedii. Co więcej, aby doprecyzować mapowanie i uzupełnić zbiór informacji z DBpedii uzupełniono mapowanie o odniesienie do Wikidaty. W ten sposób jest zapewniony dodadkowy punkt zaczepienia do Danych Połączonych.

5.4.7 Logowanie rezultatów

Prezentacja wyników, w zależności od konfiguracji programu, mogła być wykonywana na trzy sposoby:

- w konsoli programu,
- do pliku json,
- do pliku xml.

W przypadku logowania do plików dla celów czytelności przez człowieka możliwe jest ustawienie wielkości wcięć.

5.5 Struktura danych wynikowych

Dane wynikowe, ze względu na swoją obszerność, podzielone są na mniejsze kawałki. Pierwszym stopniem gradacji są frazy (ang. chunks). Każda fraza podzielona jest na zdanie (ang. sentence). W skład zdania oprócz listy tokenów (ang. token), czyli słów wchodzą tagi z metadanymi, określające np. które z kolei jest dane zdanie. W pojedynczych tokenach zapisane są informacje o typie, części mowy, jaką reprezentują czy z jakim znaczeniem ze Słowosieci są powiązane. Pojęcia najczęściej są jednowyrazowe, choć mogą składać się z większej liczby słów. Dodatkowo odrębnymi pojęciami są znaki interpunkcyjne zawarte w analizowanym tekście. Powyższa struktura wynika ze struktury danych wynikowych z API ujednoznaczniającego.

Do węzłów tokenów są wkładane dane będące rezultatem działania programu. Zawierają one informacje takie jak identyfikator jednostki w Słowosieci, opis pojęcia ze Słowosieci, link do Wikipedii oraz linki z odnośnikami do Link Open Data, czyli DBpedii i Wikidaty.


```
{"chunkList": {"chunk": [
  {
    "sentence": {
      "tok": [
        {
          "orth": "zamek",
          "prop": [
            {
              "key": "sense:ukb:syns_id",
              "content": 43594
            },
            {
              "key": "sense:ukb:syns_rank",
              "content": "43594/302.3605796316 4189/226.6504355617 ...
            },
            {
              "key": "sense:ukb:unitsstr",
              "content": "zamek.4(3:wytw)"
            }
          ],
          "units": [{
            "unit": "63884",
            "wiki": "https://pl.wikipedia.org/wiki/Zamek_(broń)",
            "details": "mechanizm broni palnej służący do zamykania na czas ...
            "dbpedia": "http://dbpedia.org/page/Lock_(firearm)",
            "wikidata": "https://www.wikidata.org/wiki/Special:EntityPage/Q1134386",
            "desc": "##K: książk. ##D: mechanizm broni palnej służący do ...
          }],
          "lex": {
            "ctag": "subst:sg:nom:m3",
            "disamb": 1,
            "base": "zamek"
          }
        },
        ...
      ],
      "ns": [
        "",
        ""
      ],
      "id": "s1"
    },
    "id": "ch1"
  },
  ...
}
```


Rysunek 5: Przykładowa struktura danych zwracanych przez program.

5.6 Testowanie aplikacji

Testowanie polegało na ręcznym przyporządkowaniu pojęciom z zadanego tekstu linków z Wikipedii oraz porównaniu rezultatów prac programu z danymi referencyjnymi. Przyporządkowywane były linki z Wikipedii, a nie linki z Link Open Data, ponieważ ręczne przyporządkowanie strony z Wikipedii było szybsze i redukowało szansę na wykonanie błędnego mapowania przez człowieka, a przejście pomiędzy Wikipedią a LOD nie jest obarczone błędem.

Samo testowanie opierało się na sprawdzaniu poprawności mapowania na dwóch podstawowych płaszczyznach:

- pojedynczych pojęć w kontekście
- całych zdań

Taki podział kierunków testowania daje, po pierwsze, możliwość weryfikacji i oceny najbardziej zawodnych elementów aplikacji w przypadkach, w których mapowanie powinno być teoretycznie proste i jednoznaczne. Po drugie natomiast długie teksty, w których kontekst znaczeniowy nie jest taki oczywisty, daje możliwość sprawdzenia, jak aplikacja radzi sobie w sytuacjach, kiedy wyznaczenie poprawnego mapowania nie jest tak oczywiste.

5.6.1 Testowanie pojedynczych pojęć w kontekście

W przeprowadzanych testach kontekst znaczeniowy wypowiedzi był rozumiany jako otoczenie słowa, które stanowi zbiór słów towarzyszących testowanemu pojęciu, a który to zbiór nierozerwalnie łączy się z jednym konkretnym znaczeniem. Sama niejednoznaczność słów jest zjawiskiem powszechnym w każdym języku i można ją podzielić na dwa podstawowe rodzaje polisemię i homonimię [6]. Polisemia polega na wyrażaniu różnych treści za pomocą jednego środka językowego, czyli słowa np. „powód” jako przyczyna i termin prawny oznaczający osobę pozywającą do sądu. Homonimia natomiast próbuje przekazywać różne znaczenia pojęć za pomocą takich samych form językowych np. „dam” jako forma czasownika „dać” i rzeczownika „dama”.

Testy opierały się na budowaniu zdań, w których kontekst wskazywał jednoznacznie na dane znaczenie wybranego pojęcia oraz porównywaniu czy wskazane przez program pojęcie jest zgodne z pojęciem wskazanym przez człowieka. Wskazanie odbywało się poprzez użycie w zdaniach wyrazów nierozerwalnie łączących się z jednym znaczeniem danego słowa. Testowane były następujące klasy problemów mapowania pojęć:

- pojęcia jednowyrazowe:
  - wieloznaczne
  - nazwy własne
  - eponimy

- wielowyrazowe:
  - jednoznaczne
  - nazwy własne

5.6.2 Testowanie całych zdań

Przed przystąpieniem do testów wymagane było ręczne przygotowanie poprawnie zmapowanych pojęć będących odniesieniem. Następnie pojęcia jedno lub wielowyrazowe podlegały mapowaniu przez program, którego poprawność była weryfikowana.

W przypadku tak przeprowadzanych testów ryzyko popełnienia błędu podczas ręcznego mapowania kilkuset wyrazowych tekstów jest stosunkowo wysokie. Z tego względu należy mieć na uwadze to, że wyniki mogą być w nieznacznym stopniu obarczone dodatkowym błędem, który jednak nie powinien wypaczyć wniosków.

Przykłady wybrane do testów miały za zadanie sprawdzić jak aplikacja radzi sobie z tekstami średniorozległymi oraz jak poziom skomplikowania języka tekstu wpływa na jakość mapowania.

6 Wyniki prac

6.1 Tesowanie pojęć jednowyrazowych

6.1.1 Jednowyrazowe pojęcia wieloznaczne

Testy mapowania jednowyrazowych pojęć wieloznacznych zostały przeprowadzone, aby potwierdzić, że program radzi sobie w ogólności z pojęciami, których znaczenie nie jest oczywiste. W tym celu zbudowanych zostało kilka zdań, w których testowane pojęcie jest pojęciem wieloznacznym, a kontekst zdania jednoznacznie wskazuje na jedno wybrane znaczenie testowanego pojęcia. Tabela 3. przedstawia szczegółowe wyniki dla słowa „zamek", natomiast tabela 4. ogólne wyniki dla różnych słów wieloznacznych.

Analizowana fraza	**Testowane pojęcie**	**Otrzymane znaczenie**	**Pochodzenie mapowania**	**Przyczyna błędu**
Zamek z fosą i wieżą	zamek (budowla)	zamek (broń)	Słowosieć	API WSD
Zamek sięgający do nieba	zamek (budowla)	zamek (broń)	Słowosieć	API WSD
Miasto z zamkiem	zamek (budowla)	zamek (broń)	Słowosieć	API WSD
Król mieszka w zamku	zamek (budowla)	zamek (broń)	Słowosieć	API WSD
Drzwi z zamkiem	zamek (mechanizm)	zamek (broń)	Słowosieć	API WSD
Sejf z zamkiem szyfrowym	zamek (mechanizm)	zamek (broń)	Słowosieć	API WSD
Klucz do zamka	zamek (mechanizm)	zamek (broń)	Słowosieć	API WSD
Zamek magnetyczny	zamek (mechanizm)	zamek (broń)	Słowosieć	API WSD
Zamek służył do zapalenia prochu	zamek (broń)	zamek (broń)	Słowosieć	API WSD

Tabela 3: Szczegółowe rezultaty testowania pojedynczych pojęć jednowyrazowych na przykładzie słowa „zamek"

Analizowana fraza	**Testowane pojęcie**	**Otrzymane znaczenie**	**Pochodzenie mapowania**	**Przyczyna błędu**
Golf to sport uprawiany na trawiastych terenach	golf (sport)	golf (sport)	Słowosieć	-

Golf jest dobrze izolującym ubraniem	golf (ubiór)	golf (ubiór)	Slowosiec	-
Rakieta to pojazd latający lub pocisk	rakieta (pojazd)	rakieta (pocisk)	Wikipedia ujednoznacz.	API WSD
Rakieta to przyrząd do gry w tenisa	rakieta (tenis)	rakieta (pocisk)	Slowosiec	API WSD
Klucz służy do otwierania zamków i kłódek	klucz (do zamka)	klucz (do zamka)	Wikipedia ujednoznacz.	API WSD
Klucz to szczytowy kliniec łuku lub niektórych typów sklepienia	klucz (architektura)	klucz (architektura)	Wikipedia ujednoznacz.	-

Tabela 4: Ogólne rezultaty testowania pojedynczych wieloznacznych pojęć jednowyrazowych

Rezultaty jednoznacznie pokazują, że API ujednoznaczniające CLARIN nie radzi sobie najlepiej z rozpoznawaniem różnych znaczeń słów. W przypadku słowa „zamek" każdokrotnie, niezależnie od kontekstu wypowiedzi, przypisuje mu znaczenie jako części broni. W sytuacji, w której istnieje link do Wikipedii w znaczeniu przypisanym przez API, program nie próbuje korzystać z algorytmu ujednoznaczniającego bazującego na stronach ujednoznaczniających Wikipedii.

6.1.2 Jednowyrazowe eponimy

Następnie aplikacja została sprawdzona pod kątem rozpoznawania eponimów, czyli słów utworzonych od nazw własnych. Eponimy zostały wybrane tak, aby możliwe było ich odwzorowanie w taksonomię Link Open Data, to znaczy, aby istniały dla każdego z nich powiązane strony w Wikipedii.

Analizowana fraza	**Testowane pojęcie**	**Otrzymane znaczenie**	**Pochodzenie mapowania**	**Przyczyna błędu**
Colt był znany jako kontruktor i producent broni	Colt (osoba)	Colt (osoba)	Wikipedia ujednoznacz.	-
Colt ma prostą budowę	colt (broń)	colt (broń)	Wikipedia ujednoznacz.	-

Skrótem od jedostki mach jest Ma lub M	mach (jednostka)	mach (jednostka)	Wikipedia ujednoznacz.	-
Mach to austriacki fizyk i filozof	Mach (osoba)	Mach (osoba)	Wikipedia ujednoznacz.	-
Nobel jest wysoce aktywny	Nobel (pierwiastek)	nobel (pierwiastek)	Słowosieć	API WSD
Nobel wynalazł dynami	nobel (osoba)	nobel (pierwiastek)	Słowosieć	-
Tantal należy do metali przejściowych	tantal (pierwiastek)	tantal (pierwiastek)	Słowosieć	-
Tantal był synem Zeusa	Tantal (postać)	tantal (pierwiastek)	Słowosieć	API WSD
Tesla jest jednostką indukcji magnetycznej w układzie SI	tesla (jednostka)	tesla (jednostka)	Wikipedia ujednoznacz.	-
Nikola Tesla zmarł w USA	Tesla (osoba)	Tesla (osoba)	Wikipedia ujednoznacz.	-

Tabela 5: Rezultaty testów eponimów

Ponownie w przypadkach, w których mapowanie zostało przeprowadzone na podstawie istniejących linków ze Słowosieci nie były rozpoznawane różne znaczenia pojęć. W przypadkach, w których mapowanie dokonywane było za pomocą stron ujednoznaczniających w Wikipedii, wyniki są satysfakcjonujące, nawet przy zastosowaniu tak prostego algorytmu, jak uproszczony algorytm Lesk. Należy mieć na uwadze fakt, że teksty zostały dobrane tak, aby algorytm nie miał problemów ze zliczaniem słów w różnych formach, gdyż testowane siągi słów nie były poddawane stemmingowi.

6.1.3 Jednowyrazowe nazwy własne

Kolejne próby miały za zadanie pokazać jak aplikacja zmapuje nazwy własne, które posiadają swoje definicje w Wikipedii. Dla uogólnienia wyników wbrane zostały znane pojęcia z różnych dziedzin nauki i życia.

Analizowana fraza	**Testowane pojęcie**	**Otrzymane znaczenie**	**Pochodzenie mapowania**	**Przyczyna błędu**
Titanic zatonął w 1912 roku	Titanic	Titanic	Wikipedia ujednoznacz.	-

Mercedes ma w swojej kolekcji 550 pojazdów	Mercedes	Mercedes	Wikipedia ujednoznacz.	-
Nokia znana jest z produkcji telefonów komórkowych	Nokia	Nokia	Wikipedia ujednoznacz.	-
Aktualnie Warszawa jest stolicą Polski i największym miastem kraju	Warszawa	Warszawa	Słowosieć	-
Japonia jest ojczyzną sushi	Japonia	Japonia	Słowosieć	-

Tabela 6: Rezultaty testów jednowyrazowych nazw własnych

W tym przypadku zarówno ujednoznacznie za pośrednictwem stron ujednoznaczniających Wikipedii, jak i WSD API dało zadowalający rezultat. WSD API spisało się dobrze, ponieważ w tym przypadku nie miało miejsce mapowanie pojęć wieloznacznych, nie była konieczna analiza kontekstu tekstu i wybór poprawnego znaczenia.

6.2 Tesowanie pojęć wielowyrazowych

6.2.1 Wielowyrazowe pojęcia jednoznaczne

Ze względu na to, że wywołanie WSD API było skonfigurowane tak, aby rozpoznawać pojęcia wielowyrazowe przeprowadzono testy również tej funkcjonalności. W tym celu testowane były zarówno proste pojęcia wielowyrazowe jak i wielowyrazowe nazwy własne posiadające swoje strony w Wikipedii. Rezultaty zostały przedstawione w tabelach 7. i 8.

Analizowana fraza	**Testowane pojęcie**	**Otrzymane znaczenie**	**Pochodzenie mapowania**	**Przyczyna błędu**
Związek frazeologiczny to połączenie dwóch lub więcej wyrazów	Związek frazeologiczny	Związek frazeologiczny	Wikipedia ujednoznacz.	-
Barakuda wielka to drapieżna ryba morska	Barakuda wielka	-	-	-

Kościół Adwentystów Odpocznienia Sabatu grupuje około dwóch tysięcy członków	Kościół Adwentystów Odpocznienia Sabatu	-	-	-
Drugą częścią Biblii chrześcijańskiej jest Nowy Testament	Nowy Testament	-	-	-

Tabela 7: Rezultaty testów jednoznaczych pojęć wielowyrazowych

6.2.2 Wielowyrazowe nazwy własne

Analizowana fraza	**Testowane pojęcie**	**Otrzymane znaczenie**	**Pochodzenie mapowania**	**Przyczyna błędu**
Statua Wolności wznosi się na wyspie u ujścia rzeki Hudson	Statua Wolności	-	-	-
Tomasz Kot grał w sztukach teatralnych i filmowych	Tomasz Kot	-	-	-
Alicja w Krainie czarów została napisana w 1865 roku	Alicja w Krainie czarów	-	-	-
Mont Blanc jest najwyższym szczytem Alp	Mont Blanc	-	-	-

Tabela 8: Rezultaty testów wielowyrazowych nazw własnych

Daje się zauważyć, że mapowanie pomimo wyboru opcji rozpoznawania pojęć wielowyrazowych w ogóle nie radzi sobie z ich ujednoznacznianiem. Pojęcia wielowyrazowe w ogóle nie są rozpoznawane, API zwraca znaczenia dla poszczególnych wyrazów, nie rozpatrując całego pojęcia.

6.3 Testowanie całych zdań

Miało na celu sprawdzenie jak aplikacja będzie radzić sobie z mapowaniem pojęć z tekstu ciągłego. Proces testowania podzielony był na dwa niezależne etapy, które poddawały mapowaniu teksty:

- zawierające liczne pojęcia wieloznaczne,
- o rosnącym poziomie skomplikowania językowego,

token	otrzymany link	źródło mapowania
ogromny	-	generowanie
zamek	https://pl.wikipedia.org/wiki/Zamek_(broń)	Slowosiec
z	https://pl.wikipedia.org/wiki/Impedancja	lesk
fosa	https://pl.wikipedia.org/wiki/Fosa	Slowosiec
?	https://pl.wikipedia.org/wiki/1870	lesk
potrafić	-	generowanie
zrobić	-	generowanie
?	https://pl.wikipedia.org/wiki/1870	lesk
wra	-	generowanie
?	https://pl.wikipedia.org/wiki/1870	lesk
enie	-	generowanie
,	https://pl.wikipedia.org/wiki/,	generowanie
jednak	-	generowanie
to	https://pl.wikipedia.org/wiki/To_(kana)	lesk
zamek	https://pl.wikipedia.org/wiki/Zamek_(broń)	Slowosiec
w	https://pl.wikipedia.org/wiki/Wolfram	lesk
kurtka	https://pl.wikipedia.org/wiki/Kurtka	generowanie
chronić	-	generowanie
my	https://pl.wikipedia.org/wiki/My_(litera)	lesk
bardziej	-	generowanie
przed	-	generowanie
zimno	https://pl.wikipedia.org/wiki/Zimno_(Polska)	lesk
,	https://pl.wikipedia.org/wiki/,	generowanie
ni	https://pl.wikipedia.org/wiki/Ny_(litera)	lesk
?	https://pl.wikipedia.org/wiki/1870	lesk
zamek	https://pl.wikipedia.org/wiki/Zamek_(broń)	Slowosiec
w	https://pl.wikipedia.org/wiki/Wolfram	lesk
broń	https://pl.wikipedia.org/wiki/Bro%C5%84_palna	Slowosiec
palny	-	generowanie
chronić	-	generowanie
my	https://pl.wikipedia.org/wiki/My_(litera)	lesk
przed	-	generowanie
?	https://pl.wikipedia.org/wiki/1870	lesk
ar	https://pl.wikipedia.org/wiki/Argon	Slowosiec
.	-	generowanie

Rysunek 6: Wizualizacja rezultatów testów całych zdań.

6.3.1 Testowanie mapowania tekstów zawierających pojęcia wieloznaczne

Do realizacji zadania skonstruowano zdania, w których występowały liczne słowa wieloznaczne. Przykład reprezentacji wyników działania programu dla zdania:

> *Ogromny zamek z fosą potrafi zrobić wrażenie, jednak to zamek w kurtce chroni nas bardziej przed zimnem, niż zamek w broni palnej chroni nas przed żarem.*

został zaprezentowany na rysunku 6.

Jak daje się zauważyć, wyniki działania programu są mało czytelne i niepoprawne. Błędy w wynikach działania programu wynikają bezpośrednio z niepoprawnego działania API ujednoznaczniającego. Interfejs programistyczny wykorzystywany przez aplikację miał problemy z interpretacją polskich znaków diakrytycznych. CLARIN WSD z nieznanych przyczyn dzieliło wyrazy na tokeny, rozpoznając polskie litery jako znaki zapytania i w wyniku tych błędów sama aplikacja podejmowała próbę mapowania tokenów, które w kontekście wypowiedzi nie miały żadnego sensu.

Dla potwierdzenia, że wina nie leży po stronie kodowania danych w żądaniu do API, sprawdzono również efekty ujednoznaczniania w demo dostępnym w sieci. Ów demo korzysta z tych samych zasobów i funkcji co udostępnione API, które wykorzystuje aplikacja oraz dostępne jest pod adresem:

> *https://ws.clarin-pl.eu/wsd.shtml*

Testy potwierdziły, że to API ma problemy z przetwarzaniem polskich znaków, a przykładowe efekty analizy dla pojęcia „wrażenie” prezentuje rysunek 7.

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE chunkList SYSTEM "ccl.dtd">
<chunkList>
 <chunk id="ch1" type="p">
  <sentence id="s1">
   <tok>
    <orth>wra</orth>
    <lex disamb="1"><base>wra</base><ctag>ign</ctag></lex>
   </tok>
   <ns/>
   <tok>
    <orth>?</orth>
    <lex disamb="1"><base>?</base><ctag>interp</ctag></lex>
   </tok>
   <ns/>
   <tok>
    <orth>enie</orth>
    <lex disamb="1"><base>enie</base><ctag>ign</ctag></lex>
   </tok>
  </sentence>
 </chunk>
</chunkList>
```

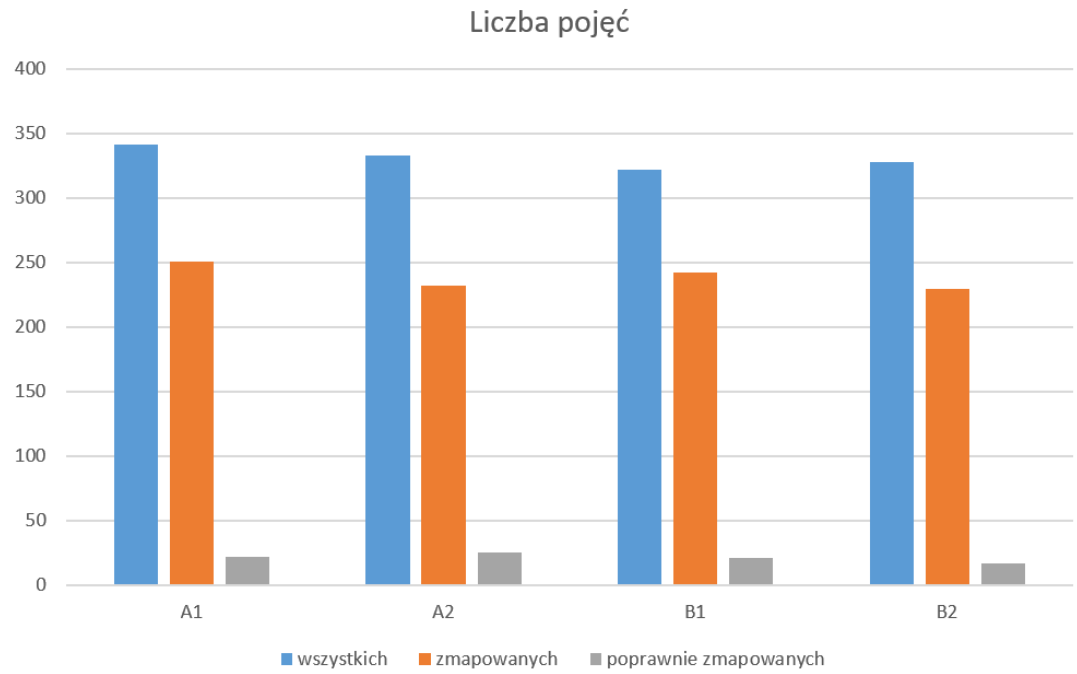
Rysunek 7: Efekt działania aplikacji ujednoznaczniającej będącej częścią CLARIN

6.3.2 Testowanie mapowania tekstów o rosnącym poziomie skomplikowania

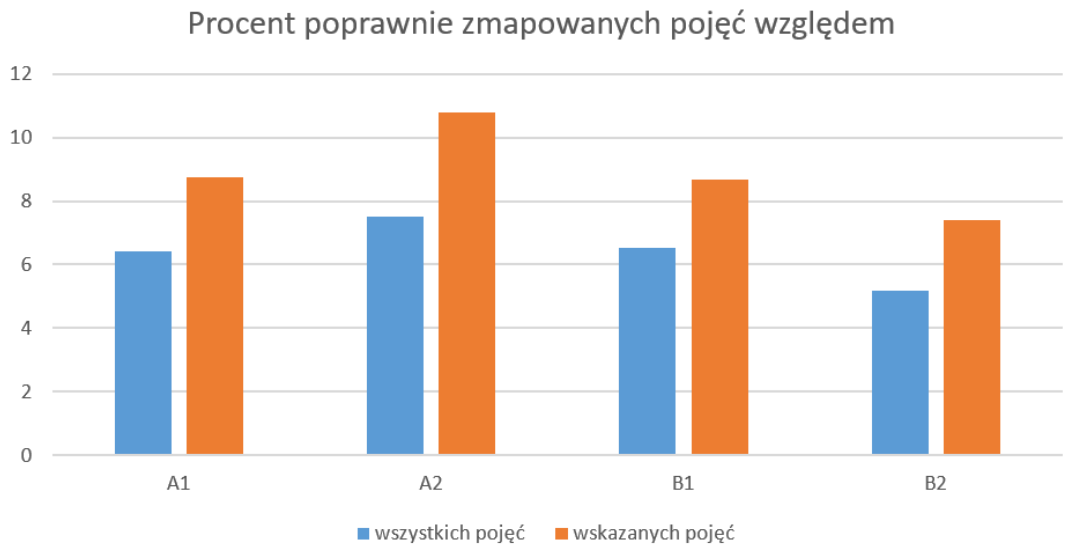
Pomimo błędów w działaniu API przeprowadzono testy dla różnych poziomów języka polskiego, aby wyłapać ogólne tendencje i trendy przy mapowaniach. Próby przeprowadzano dla poziomów języka polskiego od A1 do B2. Teksty wybrano arbitralnie, tak aby opierały się słownictwu ogólnym. Wyniki zaprezentowane są na wykresach 8 i 9.

Nietrudno zauważyć, że błędy wynikające z niepoprawnego podziału tekstu na tokeny skutkują niezwykle niską poprawnością mapowania, która waha się od około 7 do około 11 procent. Taki stosunek wszystkich pojęć do pojęć poprawnie zmapowanych pokazuje, że stosowanie aplikacji w aktualnym stanie API jest bezcelowe, gdyż wyniki są zaszumione, nieczytelne i praktycznie bezużyteczne.

Na podstawie wyników można także przypuszczać, że skomplikowanie na poziom językowym w niewielkim stopniu wpływa na procent zmapowanych pojęć. Różnice dla poszczególnych poziomów mieszczą się w zakresie błędu statystycznego. Zastosowane teksty opierają się o słownictwo ogólne i można wnioskować, że dla słownictwa specjalistycznego lub dziedzinowego procent poprawnie zmapowanych pojęć byłby jeszcze niższy. ‘



Rysunek 8: Liczba pojęć rozpoznanych przez aplikację dla tekstów na różnym poziomie językowym



Rysunek 9: Wyniki procentowe realizacji mapowania dla różnych poziomów języka

7 Podsumowanie

Praca miała na celu podjęcie próby mapowania pojęć w taksonomię Link Open Data. Samą próbę można uznać za umiarkowanie udaną.

Ogniwem programu, które zawodziło najczęściej był interfejs programistyczny do ujednoznaczniania w jednostki ze Słowosieci należący do infrastruktury CLARIN. Interfejs ten nie radził sobie najlepiej z ujednoznacznianiem pojęć, nawet jednowyrazowych. Często dla danego słowa, niezależnie od kontekstu, w jakim występowało, wskazywał na to samo znaczenie. Przykładem takiego zachowania może być słowo „zamek", któremu API każdokrotnie przypisywało znaczenie zamka jako części broni palnej.

Ponadto interfejs nie radził sobie z ujednoznacznianiem pojęć wielowyrazowych. Testy wykazały, że były one rozumiane jedynie w nielicznych przypadkach i nie można w tym przypadku mówić o żadnej powtarzalności czy poprawności.

Co najgorsze testy wykazały, że infrastruktura ma problemy implementacyjne. Zarówno od strony REST API, jak i od strony dema aplikacji udostępnionego w internecie występowały problemy z polskimi znakami. Aplikacja internetowa należąca do infrastruktury CARIN w niektórych przypadkach dla tekstów posiadających polskie znaki nie rozpoznawała polskich znaków diakrytycznych i zwracała dla nich niezrozumiałe i niepoprawne rezultaty.

Nie mniej jednak interfejs ujednoznaczniający ma także swoje mocne strony. Jego niewątpliwą zaletą jest dostępność. Przez cały czas trwania testów nie wystąpił problem z wyłączeniem usługi czy też jej chwilową niedostępnością. Ponadto API wystawione do użytku cechuje się stosunkowo wysoką niezawodnością. Wewnętrzne błędy serwera podczas przetwarzania tekstu zdarzały się sporadycznie i nie zaburzały procesu testowania.

W przypadku ujednoznacznienia korzystającego ze stron ujednoznaczniających Wikipedii oraz uproszczonego algorytmu Lesk pomimo wykorzystania tak prostego algorytmu wyniki były zadowalające, choć mocno zależne od form poszczególnych składowych analizowanego tekstu. W znacznej większości przypadków ujednoznacznianie na podstawie jednego akapitu było zadowalające i nie wymagało poszerzania analizowanego tekstu na większą liczbę akapitów opisu z Wikipedii.

Jeżeli chodzi o mapowanie przeprowadzane poprzez generowanie linków z form bazowych pojęć, było ono stosowane dosyć rzadko, gdyż w większości mapowań kroki 1 i 2, czy odpowiednio ekstrakcja linku ze Słowosieci oraz ujednoznacznienie oparte o algorytm Lesk, znajdowały mapowanie, jakie zwracała aplikacja. Jest do duży plus działania aplikacji, gdyż oba poprzedzające kroki opierają się na statystykach, natomiast generowanie odbywa się poprzez operacje na znakach, przez co charakteryzuje się niskim poziomem ufności.

8 Kierunki rozwoju

Ze względu na to, że rezultaty otrzymane na bazie testów nie są w pełni zgodne z oczekiwaniami wskazać można w aplikacji conajmniej kilka obszarów, których usprawnienie powinno poprawić jakość otrzymywanych mapowań. Można do nich zaliczyć:

- implementację własnego narzędzia ujednoznaczniającego.

  Jako że API będące częścią infrastruktury CLARIN odpowiedzialne za sprowadzanie pojęć do węzłów Słowosieci zawodziło w pracy aplikacji najczęściej, to naturalnym ruchem wydaje się wyeliminowanie tego słabego ogniwa. Narzędzie powinno celniej rozpoznawać znaczenie słów oraz, jako że mogłoby korzystać w pełni z zasobów lokalnych jego szybkość i niezawodność byłaby dużo wyższa niż zasobów udostępnianych w sieci. Dodatkowo można by pokusić się o implementację rozpoznawania pojęć wieloznacznych, czego brakowało w wykorzystanych funkcjach infrastruktury CLARIN.

- bardziej zaawansowany i dokładniejszy algorytm ujednoznaczniania poprzez Wikipedię.

  Z racji tego, że wykorzystany algorytm jest algorytmem stosunkowo prostym nie jest on odporny na znaczne różnice analizowanego tekstu z definicją z Wikipedii. Należało by wykorzystać bardziej złożone algorytmy ujednoznaczniania, które obarczałyby wynik mniejszym błędem. Dodatkowo słusznym wydaje się użycie stemera, który uodporniłby algorytm na zmieniające się formy słów.

- weryfikacja wygenerowanego linku

  Samo generowanie linku jest operacją przeprowadzaną na znakach. W związku z tym jej wiarygodność stoi na stosunkowo niskim poziomie. Co prawna udział mapowań otrzymanych w kroku generowania linku nie jest zbyt wysoki, ale mimo to otrzymane w ten sposób mapowania powinny być poddane weryfikacji.

Literatura

[1] Aliwy A. H., Abbas A. R. *Improvement WSD dictionary using annotated corpus and testing it with simplified Lesk algorithm*. Department of Computer Science, University Of Technology, Baghdad 2015

[2] Berners-Lee T. *Hearing on the "Digital Future of the United States: Part I – The Future of the World Wide Web"*. 2007

[3] Dziob A., Łazarewicz P. *Słowosieć jako narzędzie wspomagające pracę tłumacza*. Rocznik Kognitywistyczny V/2011 DOI 10.4467/20843895RK.12.004.0408

[4] Erxleben F., Günther, M., Krötzsch M. Mendez, J., Vrandecic D. *Introducing Wikidata to the Linked Data Web*. Comm. ACM, 2014

[5] Fay R. *Linked: Semantic Web and metadata*. 2010

[6] Gaustad T. *Linguistic Knowledge and Word Sense Disambiguation*. University of Groningen 2004

[7] Hinrichs, E., Krauwer S. *The CLARIN Research Infrastructure: Resources and Tools for E-Humanities Scholars*. Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014), May 2014, 1525–31.

[8] Ivanowa M., Ivanowa T. *Web 2.0 and 3.0 Environments: Possibilities for Authoring and Knowledge Representation*. Revista de Informatica Sociala, VII, 7-21.

[9] Kędzia P., Piasecki M., Rudnicka E., Przybycień K. *Automatic Prompt System in the Process of Mapping plWordNet on Princeton WordNet*. Wrocław University of Technology

[10] Loureiro A., Bettencourt T. *Building Knowledge in the Virtual World – Influence of Real Life Relationships*. Journal of Virtual Worlds Research, February 2010

[11] Maziarz M., Piasecki M., Szpakowicz S., Rabiega-Wiśniewska J. *Semantic Relations among Nouns in Polish WordNet Grounded in Lexicographic and Semantic Tradition*. Cognitive Studies 11, s. 161–182

[12] Miller, G. A. *WordNet: An on-line lexical database*. International Journal of Lexicography 3, 4 (Winter 1990)

[13] Naik U., Shivalingaiah D. *Technological March from Web 1.0 to Web 3.0: A Comparative Study*. Revista de Informatica Sociala, VII, 7-21.

[14] Ostuni, V. C., Di Noia T. *Top-N Recommendations from Implicit Feedback leveraging Linked Open Data*. 7th ACM conference on Recommender systems, pages 85-92, Hong Kong 2013

[15] Piasecki M., Szpakowicz S., Broda B. *A Wordnet from the Ground Up*. Wrocław 2009: Wydawnictwo Politechniki Wrocławskiej

[16] Roszkowski M. *Linked Data – model danych powiązanych w Semantic Web*. Zagadnienia Informacji Naukowej, 2010 nr 2, s. 52-68.

[17] Rudnicka E., Piasecki M., Piotrowski T. Grabowski Ł., Bond F. *Mapping wordnets from the perspective of inter-lingual equivalence*. Wrocław University of Technology

[18] Rudnicka E., Witkowski W., Kaliński M. *Towards the methodology for extending Princeton WordNet*. Wrocław University of Technology 2015 r.

[19] Stokowska A., Korulska E. *Historia Internetu*. Centrum Edukacji Obywatelskiej, 2014

[20] Waqar A. *Third Generation of the Web: Libraries, Librarians and Web 3.0*. Library Hi Tech News, Vol. 32 Iss 4

[21] Wright R. *The man who invented the web*. Time, May 19, 1997 vol. 149 no. 20

[22] Váradi, T., Wittenburg P., Krauwer S, Wynne M., Koskenniemi K. *CLARIN: Common Language Resources and Technology Infrastructure*. 2008

[23] Vrandecic D., Krötzsch, M. *Wikidata: A free collaborative knowledge base*. Comm. ACM, 2014

[24] *https://www.esfri.eu/about*. styczeń 2020

[25] *https://www.clarin.eu/content/about-eric*. styczeń 2020

[26] *http://clarin-pl.eu/pl/o-projekcie*. styczeń 2020

[27] *http://clarin-pl.eu/pl/o-projekcie/clarin-pl*. styczeń 2020

[28] *https://lod-cloud.net/*. styczeń 2020

[29] *https://blog.dbpedia.org/2016/10/19/yeah-we-did-it-again-new-2016-04-dbpedia-release/*. styczeń 2020

[30] *https://wiki.dbpedia.org/about/facts-figures*. styczeń 2020

[31] *https://www.ontotext.com/knowledgehub/fundamentals/linked-data-linked-open-data/*. styczeń 2020

[32] *https://www.wikidata.org/wiki/Wikidata:Statistics*. styczeń 2020

Wykaz skrótów

- PWN (ang. Princeton WordNet) – pierwsza leksykalna baza danych języka angielskiego,
- URL (ang. Uniform Resource Locator) -– ujednolicony format adresowania zasobów,
- WWW (ang. World Wide Web) – światowa rozległa sieć komputerowa,
- HTML (ang. HyperText Markup Language) – hipertekstowy język znaczników,
- LD (ang. Linked Data) – dane połączone
- LOD (ang. Linked Open Data) – wolne dane połączone
- CLARIN (ang. Common Language Resources and Technology Infrastructure) – ogólnoeuropejska infrastruktura naukowa
- CLARIN-PL – polska część infrastruktury CLARIN
- ERIC (ang. European Research Infrastructure Consortium) – organ zarządzający i koordynujący CLARIN
- ESFRI (ang. European Strategy Forum on Research Infrastructures) – Europejskie Forum Strategiczne ds. Infrastruktur Badawczych

Spis rysunków

Spis tabel