Automatyczna ekstrakcja i klasyfikacja semantyczna wielosegmentowych jednostek leksykalnych języka naturalnego

Paweł Chrząszcz

17 lutego 2016

1 Wstęp

Przetwarzanie języka naturalnego wymaga użycia algorytmów ekstrakcji cech z tekstu. Najprostsze cechy to po prostu miary statystyczne. Aby uzyskać wyższą skuteczność, konieczne jest skorzystanie z informacji zależnej od języka, czyli cech morfologicznych i syntaktycznych – jest to szczególnie istotne w przypadku języków o rozbudowanej fleksji, takich jak j. polski. Przykładowo, analizując segment „psem" w zdaniu „Wyszedłem z psem na spacer" możemy stwierdzić, że słowo to jest rzeczownikiem rodzaju męskiego żywotnego nieosobowego w narzędniku liczby pojedynczej. Ekstrakcję takiej informacji może umożliwić narzędzie wyposażone w słownik fleksyjny, np. Słownik Fleksyjny Języka Polskiego – SFJP z biblioteką CLP [11, 8], Morfeusz [25] czy też Morfologik [14]. Zasoby te zawierają setki tysięcy wyrazów wraz z ich formami gramatycznymi, jednak w dalszym ciągu istnieją wyrazy występujące rzadko, których w tych słownikach nie ma. W przypadku takich słów pomocne mogą się okazać tagery, które oznaczają tekst cechami gramatycznymi. Narzędzia te wykorzystują statystyczne algorytmy uczenia z nadzorem (supervised learning), takie jak SVM, HMM czy też CRF. Są one trenowane na korpusie tekstu otagowanym wzorcowymi znacznikami i dla języka polskiego osiągają dokładność znakowania cechami syntaktycznymi na poziomie 90% [22, 17]. Narzędzia te są również przydatne do ujednoznaczniania ekstrahowanych cech, np. segment "miał" może być formą czasownika lub rzeczownika.

Opisane cechy syntaktyczne mogą okazać się niewystarczające, np. jeżeli szukamy zdań związanych ze słowem „zwierzę", nie znajdziemy zdania „Wyszedłem z psem na spacer", ponieważ powiązanie słów „zwierzę" i „pies" ma charakter semantyczny, czyli dotyczy znaczenia słów. Ekstrakcja cech semantycznych z tekstu jest zagadnieniem bardziej skomplikowanym i zwykle rozwiązuje się je tworząc w sposób ręczny złożone taksonomie oparte na relacjach paradygmatycznych takich jak np. hiponimia i synonimia. Przykładami takich zasobów są WordNet [13] i ontologie, np. CYC [12]. Główną wadą tych sieci taksonomicznych jest brak relacji syntagmatycznych, czyli zależności obrazujących związki między wyrazami występującymi w konkretnym zdaniu, pełniącymi określone role semantyczne. Zasoby zawierające takie relacje to np. FrameNet [20] dla j. angielskiego. Rozwijany jest też Słownik Semantyczny Języka Polskiego, lecz jest on w dalszym ciągu daleki od ukończenia.

Głównym rodzajem słów, których nie odnajdziemy w wyżej wymienionych zasobach, są wielosegmentowe jednostki leksykalne (inaczej: **wyrazy wielosegmentowe**, dalej zwane WW), czyli wyrażenia składające się z kilku segmentów, które posiadają własne, odrębne znaczenie. Przykładami takich wyrazów są terminy („tlenek węgla"), idiomy („panna młoda", „mówić trzy po trzy"), nazwy własne („Polski Związek Wędkarski") czy też nazwy osób („Lech Wałęsa"). Znaczenie WW jest często inne niż suma znaczeń poszczególnych segmentów, np. słowa „panna" i „młoda" nie są semantycznie powiązane ze ślubem, a całe wyrażenie już jest. Powoduje to konieczność dołączenia wyrazów wielosegmentowych do słowników, sieci semantycznych i innych zasobów językowych.

Potrzebujemy więc zasobów językowych zawierających WW oraz metod ich ekstrakcji z tekstu. Dodatkowo przydatna byłaby płytka klasyfikacja semantyczna, ograniczająca się do przydzielenia wyrazowi jedynie prostej etykiety semantycznej – np. słowu „pies" przydzielimy etykietę „zwierzę". Pozwoli to na przynajmniej częściowy opis znaczenia, a jeżeli etykiety same znajdą się w sieci semantycznej, wówczas będziemy mogli z nią powiązać również etykietowane słowo (np. jeżeli słowo „Cessna" otrzyma etykietę „samolot" znajdującą się w sieci semantycznej, będziemy mogli je powiązać z tematyką lotniczą).

1.1 Analiza problemu

Najprostsze metody wykrywania wystąpień WW w tekście polegają na używaniu statystycznych miar współwystępowania słów, jednak uzyskiwane wyniki są niskie [18, 27, 15, 19]. Do podniesienia skuteczności potrzebne są leksykony WW i korpusy treningowe, zawierające oznaczone wystąpienia WW [5]. W przypadku języka polskiego problem polega na tym, że zasoby te nie są dostępne – niniejsza praca ma dopiero umożliwić ich tworzenie. Widzimy więc, że badania nad nowymi metodami realizującymi nakreślone cele są **w pełni uzasadnione**, a niniejsza praca ma w dużej mierze charakter eksploracyjny, ponieważ nie istnieją poprzednie wyniki będące punktem odniesienia. Jednym z założeń pracy jest niekorzystanie z ręcznie tworzonych reguł i zbiorów treningowych – pozwala to stwierdzić, z jaką dokładnością można ekstrahować wyrazy wielosegmentowe z nieuporządkowanego tekstu polskiego bez użycia otagowanych zbiorów treningowych, ręcznie tworzonych reguł i korzystających z nich klasyfikatorów i tagerów. Badania takie nie były jeszcze prowadzone, a ich efektem jest nie tylko wyznaczenie punktu odniesienia (*baseline*) dla dalszych prac, ale również stworzenie brakujących zasobów zawierających WW dla języka polskiego.

Okazuje się, że obecnie coraz częściej zasoby językowe takie jak WordNet zastępowane są Wikipedią, co niejednokrotnie pozwala podnieść skuteczność różnych algorytmów ekstrakcji informacji z tekstu, np. [7]. Zawartość Wikipedii może posłużyć do ekstrakcji wyrazów w tym wielosegmentowych (hasła), etykiet semantycznych (definicje), relacji semantycznych (przekierowania, linki, kategorie) oraz do trenowania algorytmów statystycznych (treść jako korpus). Podjęto więc decyzję o wykorzystaniu polskiej Wikipedii [23] jako głównego zasobu używanego do ekstrakcji WW.

1.2 Tezy

Podstawowym celem niniejszej pracy jest umożliwienie ekstrakcji wyrazów wielosegmentowych dla języka polskiego – pozwala to sformułować pierwszą tezę pracy.

> **TEZA 1**
> *Możliwe jest opracowanie algorytmu ekstrahującego w sposób automatyczny wyrazy wielosegmentowe z tekstu w języku polskim, wykorzystującego jako źródła danych słownik fleksyjny i Wikipedię.*

Algorytm ekstrakcji może działać samodzielnie, jednak przede wszystkim może on zostać użyty do stworzenia słownika WW. Dlatego też w niniejszej pracy wykazana zostanie również prawdziwość poniższej tezy.

> **TEZA 2**
> *Możliwe jest utworzenie w sposób automatyczny słownika wyrazów wielosegmentowych z haseł Wikipedii oraz wyrazów wielosegmentowych wyekstrahowanych przy pomocy algorytmu opisanego w Tezie 1.*

Niniejsza praca opisuje więc głównie badania nad **ekstrakcją** wyrazów wielosegmentowych. Odnośnie **klasyfikacji semantycznej** tych wyrazów, prace ograniczono do dopracowania wcześniejszego algorytmu wyznaczającego etykiety semantyczne haseł Wikipedii [4] oraz wstępnych eksperymentów dotyczących wyznaczania takich etykiet dla nowo wyekstrahowanych wyrazów. Dalsze badania ujęte są w planach przyszłych prac.

2 Definicja wyrazów wielosegmentowych

Problem automatycznej ekstrakcji wyrażeń wielosegmentowych z tekstu jest rozważany co najmniej od kilkunastu lat – w literaturze anglojęzycznej funkcjonuje pojęcie *multiword expressions* (MWE), które w pracy Saga i in. [21] zdefiniowano jako "idiosynkratyczne interpretacje przekraczające granice słów". W pracy tej wyróżniono 4 kategorie takich wyrażeń dla języka angielskiego. Poniżej przedstawiono ich najbliższe polskie odpowiedniki:

1. Wyrażenia nieodmienne – mają stałe, odrębne znaczenie, są nieodmienne i semantycznie niedekomponowalne. Przykłady: "ad hoc", "mimo wszystko", "ani mru-mru".
2. Wyrażenia o ustalonej strukturze – mają stałe, odrębne znaczenie, funkcjonują jako jedna jednostka słownikowa odmieniająca się przez odpowiednie formy gramatyczne. Przykłady: "panna młoda", "biały kruk", "mówić trzy po trzy".

3. Wyrażenia o swobodnej strukturze – jak wyżej, lecz dopuszczają dodawanie lub zamianę niektórych segmentów, a także rozbijanie na części oddzielone innymi segmentami, co nie prowadzi do utraty znaczenia, np. “działać jak płachta na byka”, “gotów na czyjeś każde skinienie”, “popełnić błąd”.
4. Utarte wyrażenia – nie posiadają odrębnego znaczenia (znaczenie całego wyrażenia jest sumą znaczeń segmentów), np. “czyste powietrze”, “dookoła świata”, “ciężka praca”.

W niniejszej pracy ograniczono się do **drugiej** kategorii z powyższej listy. Ponadto zdecydowano, że ekstrahowane będą jedynie wyrażenia pełniące rolę rzeczownikową. Ograniczenia te pozwalają uniknąć trudnych decyzji odnośnie tego, czy dany wyraz jest WW [15] oraz problemów z nieciągłością wyrażeń [9, 10]. Wyrazy wielosegmentowe w kontekście niniejszej pracy można zdefiniować jako wyrażenia odmienne, o dokładnie zdefiniowanej, ustalonej strukturze, odmieniające się w całości jak rzeczowniki, pełniące w tekście rolę rzeczowników i posiadające określone, stałe znaczenie. Przykłady takich wyrażeń znajdują się w tabeli 1. WW zdefiniowane w ten sposób mają dobrze zdefiniowaną strukturę gramatyczną – jest to ciąg co najmniej dwóch segmentów, z których każdy należy do jednej z poniższych kategorii, przy czym przynajmniej jeden z segmentów musi być odmienny.

- **Segmenty odmienne** tworzą główną część WW. Mogą nimi być rzeczowniki, przymiotniki, liczebniki lub imiesłowy przymiotnikowe. Segmenty te odmieniają się wraz z całym wyrazem przez przypadki i liczby. W formie podstawowej wszystkie segmenty odmienne występują, podobnie jak cały wyraz, w mianowniku liczby pojedynczej (wyjątkiem są wyrazy wielosegmentowe *pluralia tantum*). Segmenty odmienne nie muszą mieć takiego samego rodzaju, np. “kobieta kot”, jednak nie mogą zmieniać rodzaju podczas odmiany.
- **Segmenty nieodmienne** to wszelkie pozostałe segmenty, których forma nie zmienia się niezależnie od formy gramatycznej całego wyrazu. Mogą to być wyrazy odmienne (rzeczowniki, przymiotniki, czasowniki itp.), wyrazy nieodmienne (np. partykuły, spójniki lub wyrazy obcojęzyczne), znaki interpunkcyjne (przecinek, myślnik, kropka, cudzysłów itp.), liczby arabskie bądź rzymskie czy też inne segmenty (np. K2).

Tabela 1: Przykłady wyrazów wielosegmentowych, których ekstrakcja jest przedmiotem pracy. Segmenty odmienne podkreślono.

Typ wyrazu	**Przykłady**
Nazwy osób	Józef Piłsudski, Allen Vigneron, Szymon z Wilkowa
Inne nazwy własne	Lazurowa Grota, Polski Związek Wędkarski
Wyrażenia zawierające nazwę	rzeka Carron, jezioro Michigan, premier Polski
Wyrazy pospolite semantycznie niedekomponowalne	panna młoda, świnka morska, czarna dziura
Wyrazy pospolite semantycznie dekomponowalne	chlorek sodu, baza wojskowa, lampa naftowa, zaimek względny

3 Metody ekstrakcji wyrazów wielosegmentowych

Schemat działania zaimplementowanego systemu przedstawiono na rys. 1. Pierwszym krokiem jest wyekstrahowanie danych z Wikipedii. Wykorzystano w tym celu ogólnodostępne zrzuty bazy danych projektów fundacji Wikimedia[1]. Ekstrahowane dane to treści stron, przekierowania, linki między artykułami, szablony i kategorie. Badano również przydatność Wikisłownika [24], lecz okazało się się, że podczas gdy wśród haseł Wikipedii odnaleziono 973 tys. wyrazów wielosegmentowych, w Wikisłowniku było ich jedynie 1118.

Przetwarzanie języka naturalnego wymaga użycia zasobów słownikowych. Podstawowym słownikiem wykorzystywanym w niniejszej pracy jest Słownik Fleksyjny Języka Polskiego (SFJP) [11], a konkretnie biblioteka CLP. Podczas prac nad ekstrakcją wyrazów oraz ich etykiet semantycznych z Wikipedii okazało się, że znaczący odsetek błędnych wyników był spowodowany brakiem pewnych wyrazów w SFJP – podjęto więc decyzję o rozszerzeniu danych SFJP o dane zasobów Morfeusz [25] i Morfologik [14]. Cechą odróżniającą te zasoby od biblioteki CLP jest całkowicie odmienny format danych, wykorzystujący **znaczniki morfosyntaktyczne** – dokonano więc scalenia danych, a rezultat zapisano w nowym formacie CLPM, będącym rozszerzeniem CLP. Ponieważ czas dostępu do słownika ma znaczenie krytyczne dla systemu, dane zapisano w wysoko zoptymalizowanej na czas odczytu bazie danych LMDB. Jako przykład działania słownika przeanalizujmy **znacznik słownikowy** zwrócony dla napotkanego w tekście segmentu "wole":

$$\{(\texttt{ADA-wola}, \{1\}), (\texttt{AEA-wole}, \{2, 8, 11, 14\}), (\texttt{CC-woli}, \{15, 21\})\}$$

Rozpoznanie jest niejednoznaczne – są trzy możliwe jednostki słownikowe: `ADA-wola` (rzecz., r. żeński), `AEA-wole` (rzecz., r. nijaki) i `CC-woli` (przymiotnik). Każda z nich może wystąpić w różnych formach, np. zapis $\{2, 8, 11, 14\}$ oznacza dopełniacz l.p. lub mianownik, biernik albo wołacz l.mn.[2]

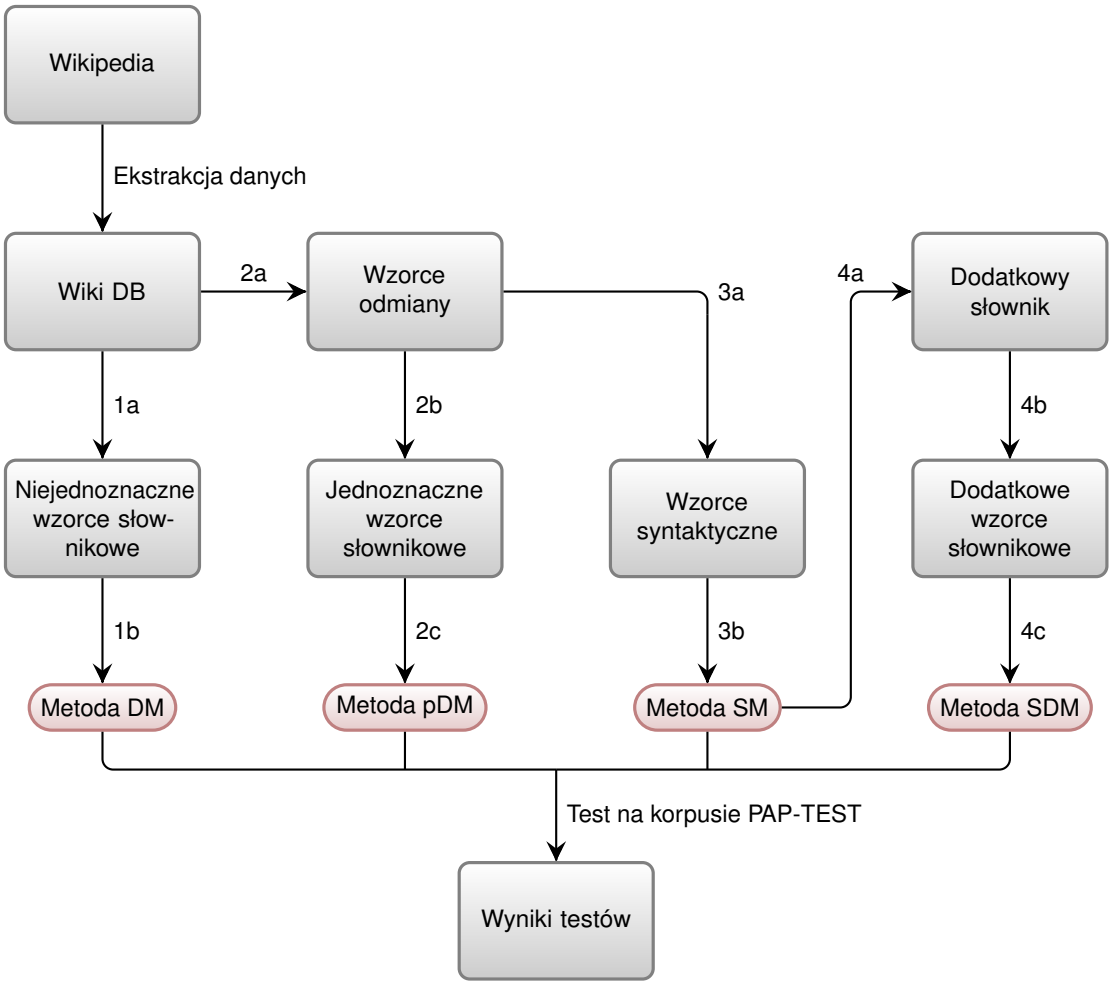
Wyekstrahowane z Wikipedii dane są następnie używane przez metody ekstrakcji wyrazów wielosegmentowych. W niniejszej pracy przygotowano i przetestowano kilka różnych algorytmów ekstrakcji.

3.1 Metoda DM

Hasła Wikipedii można potraktować jako słownik wyrazów wielosegmentowych. Jest to oczywiście duże uproszczenie: nie wiadomo, które segmenty są odmienne, niektóre z nich mogą być niejednoznaczne, a część haseł nie będzie w ogóle wyrazami wielosegmentowymi. Ponadto zakres Wikipedii jest ograniczony. Mimo to z całą pewnością taka prosta metoda może posłużyć jako punkt wyjścia, a zarazem odniesienia (*baseline*) dla metod bardziej zaawansowanych oraz jako składnik ewentualnych metod złożonych. Aby dane wyrażenie mogło zostać rozpoznane w tekście, potrzebny jest algorytm rozpoznający. Zastosowane podejście polega na utworzeniu na podstawie haseł **wzorców słownikowych** (rys. 1, operacja 1a), które są później rozpoznawane w tekście. Wzorce te mogą być niejednoznaczne, ponieważ trzeba wziąć pod

[1] `http://dumps.wikimedia.org`

[2] Niejednoznaczności można eliminować, korzystając ze statystycznych tagerów lub parserów regułowych, jednak wprowadza to duży odsetek błędów przenoszący się do kolejnych etapów przetwarzania danych.



Rysunek 1: Schemat działania systemu ekstrakcji wyrazów wielosegmentowych.

uwagę wszystkie możliwe warianty odmiany danego wyrażenia. Jako przykład rozważmy hasło “Droga wojewódzka nr 485”. Występują tu następujące niejednoznaczności:

- Segment “Droga” może być pisany wielką bądź małą literą – nie możemy tego stwierdzić, ponieważ hasła Wikipedii zaczynają się zawsze od wielkiej litery.
- Segment “Droga” może być odmienny lub nieodmienny. Analogicznie, segment “wojewódzka” może być odmienny lub nieodmienny. Wiemy jedynie, że co najmniej jeden z nich musi być odmienny, by wyrażenie było WW.
- Segment “Droga” może on być rzeczownikiem lub przymiotnikiem. Jeżeli jest on odmienny, będzie to miało wpływ na sposób odmiany.

Utworzono prosty tekstowy format zapisu wszystkich możliwych wariantów, a następnie powstałe wzorce posłużyły do skonstruowania **automatu Moore’a**[3] (rys. 1, operacja 1b) rozpoznającego je w tekście. Ponieważ opisywany problem dotyczy nie tylko rozpoznawania wyrażeń w tekście, ale także ich

[3]Wybrano ten rodzaj automatu, ponieważ pozwala on na wypisywanie w każdym stanie bieżąco rozpoznanego wzorca, a zatem umożliwia rozpoznanie wielu częściowo pokrywających się wzorców jednocześnie.

ekstrakcji, dla każdego rozpoznanego wyrażenia zapisywane są w bazie danych wszystkie możliwości jego odmiany. Przykładowo, w zdaniu "Rozpoczął się remont drogi wojewódzkiej nr 485." uda się rozwiązać wszystkie powyższe niejednoznaczności, ale zdanie "Droga wojewódzka nr 485 rozpoczyna się w Gdańsku." nie pozwoli na to. Ponadto algorytm wspiera rozpoznawanie wzorców pokrywających się częściowo lub całkowicie – dzięki temu można dokonać późniejszej analizy i ewaluacji wszystkich możliwości. Ten algorytm ekstrakcji WW nazwano **DM** (*Dictionary Matching*).

3.2 Metoda pDM

Po analizie metody DM w ramach eksperymentu podjęto próbę zastosowania heurystycznego algorytmu ujednoznaczniającego wzorce słownikowe, co spowodowało zmniejszenie niejednoznaczności wyników ekstrakcji. W niniejszej pracy dążymy jednak do tego, by unikać metod, które wprowadzają ograniczenia strukturalne rozpoznawanych wyrazów. W związku z tym potrzebna jest metoda automatycznego wyznaczenia **wzorców odmiany** haseł Wikipedii (rys. 1, operacja 2a). Pomysł polega na tym, by wykorzystać linki przychodzące do artykułów. Linki zawierają hasło w różnych formach fleksyjnych, np. do hasła "Czarna dziura" może prowadzić link "czarnej dziury". Pozwala to na stwierdzenie, które segmenty są odmienne. Powinno to również umożliwić ujednoznacznienie wielu niejednoznacznych segmentów, a także podjęcie decyzji co do tego, czy hasło pisane jest wielką czy małą literą. Kolejną zaletą jest fakt, że hasła, do których nie prowadzą linki, często nie są wyrazami wielosegmentowymi[4], więc możemy je przy okazji odfiltrować. Wada metody polega jednak na tym, że treść linku jest czasami błędna. Powoduje to konieczność zastosowania dość złożonego algorytmu:

1. W pierwszej kolejności tworzona jest statystyka linków przychodzących.
2. Następnie dla każdego linku wyznaczane są odpowiadające mu wzorce odmiany hasła.
3. Kolejnym krokiem jest próba korekty pisowni pierwszej litery hasła.
4. Kolejny etap to wyznaczanie zbioru linków o maksymalnej liczności, dla którego nie ma sprzeczności we wzorcach odmiany.
5. Następnie do bazy danych zapisywany jest nowy wiersz odpowiadający wzorcowi odmiany.

W przypadku tych haseł, dla których udało się utworzyć jednoznaczne wzorce odmiany, tworzone są wzorce słownikowe, a następnie konstruowany jest automat analogiczny jak dla metody DM (rys. 1, operacje 2b i 2c). Ten wariant nazwano **pDM**.

3.3 Metoda SM

Dotychczas opisane metody ekstrakcji wyrazów wielosegmentowych, DM i pDM, dokonywały jedynie rozpoznawania wyrazów będących hasłami Wikipedii. Aby pokonać to ograniczenie, konieczne jest wprowadzenie pewnych reguł lub wzorców, które mogłyby posłużyć do ekstrakcji nowych wyrazów.

[4]Obserwacja opiera się przeglądaniu kilkuset losowo wybranych haseł zarówno z linkami przychodzącymi jak i bez nich.

Wzorce takie zwykle są definiowane ręcznie [1, 26, 2, 16, 19]. Okazuje się jednak, że wiele można osiągnąć wykorzystując opisaną w poprzednim podrozdziale metodę automatycznego wyznaczania wzorców odmiany haseł Wikipedii – skoro dla danego hasła znamy wzorzec odmiany, można wykorzystać jego budowę do znajdowania w tekście wyrazów o podobnej strukturze. Np. dla wyrazów wielosegmentowych "tlenek węgla", "siarczan miedzi", "wodorotlenek sodu" pierwszy segment to odmienny rzeczownik r. męskiego, a drugi – nieodmienny rzeczownik w dopełniaczu. Dodatkowo wzorzec może uwzględniać kontekst, w którym występuje wyraz wielosegmentowy[5], np. wymienione związki chemiczne występują często w podobnych wyrażeniach, np. "…zawartość *tlenku węgla* w …", "…reakcja *siarczanu miedzi* z …", "…nadmiar *wodorotlenku sodu* w …".

W oparciu o powyższe obserwacje utworzono algorytm, który w oparciu o wzorce odmiany z metody pDM oraz analizę kontekstu wystąpień linków tworzy **wzorce syntaktyczne** opisujące strukturę składniową samego WW, a także kontekstu, w którym występuje (rys. 1, operacja 3a). Rozważano różne poziomy szczegółowości wzorców i wybrano wariant, w którym zapisywane są następujące informacje:

- Część mowy i odmienność każdego z segmentów hasła, a także rodzaj i liczba dla segmentów odmiennych oraz przypadek dla nieodmiennych.
- Kontekst ograniczony jest do jednego segmentu po lewej i po prawej stronie. Dla segmentów kontekstu zapisywana jest informacja zbliżona do tej dla nieodmiennych segmentów hasła.

Przykładowo dla wyrażenia "centralnej **czarnej dziury**." zapiszemy wzorzec `cc16, cc17, cc20 *cc15 *ad1_p`. Oznacza on przymiotnik w dopełniaczu, celowniku lub miejscowniku l.poj. r. żeńskiego, po którym występuje WW składający się z dwóch odmiennych segmentów w rodzaju żeńskim: przymiotnika i rzeczownika. Prawy kontekst to znak interpunkcyjny. Razem z wzorcem zapisywana jest forma, w której wystąpił – tutaj dopełniacz l.p. W taki sposób tworzymy statystykę wzorców wraz z formami, w których wystąpiły. Następnie konstruowany jest automat podobny jak dla DM i pDM (rys. 1, operacja 3b), który służy do rozpoznawania wzorców. Powstałą metodę nazwano **SM**. W przeciwieństwie do metod słownikowych daje ona wyniki silnie niejednoznaczne – dane wyrażenie może pasować do wielu wzorców. Wybór właściwego wyniku wymaga wprowadzenia funkcji oceniającej wynik. W tym przypadku zdecydowano się na miarę ilościową, sumującą wystąpienia danego wzorca w Wikipedii w konkretnej formie gramatycznej. Wprowadzono parametr rs_{min} umożliwiający odcięcie wyników poniżej pewnej wartości tej miary[6].

3.4 Metoda SDM

Wynik działania metody SM na pewnym korpusie tekstów można przekształcić do postaci słownikowej (rys. 1, operacja 4a) – w ten sposób uzyskamy dodatkowy zasób słownikowy, który następnie może zwiększyć skuteczność rozpoznawania i ekstrakcji WW z tekstu. Zdecydowano wykorzystać do tej operacji dwa korpusy tekstowe:

[5] Zauważono to też np. w pracy [6].

[6] W przyszłości można wprowadzić tutaj metody uczenia maszynowego z nadzorem, wymagają one jednak dużego wysiłku poświęconego na tworzenie zbiorów treningowych.

- PAP-TRAIN – korpus notatek prasowych PAP liczący ok. 3.6 mln segmentów.
- WIKI – korpus zawierający treść wszystkich artykułów Wikipedii, liczący 202.7 mln segmentów.

Słownik utworzony z korpusu WIKI poddano szczegółowej analizie. Jego dokładność zależy od wybranej wartości progu rs_{min}. Przykładowo, jeżeli próg ten ustalimy tak, że słownik ma 1 milion haseł, ponad 75% z nich będzie poprawnymi WW. Po utworzeniu słownika należy podobnie jak dla metody pDM utworzyć wzorce słownikowe, a następnie automat je rozpoznający (rys. 1, operacje 4b i 4c). Powstałą metodę nazywamy **SDM**.

4 Testy metod ekstrakcji

Aby zweryfikować prawdziwość Tezy 1, trzeba ocenić jakość wyników generowanych przez algorytmy ekstrahujące WW z tekstu. W tym celu przetestowano działanie algorytmów na losowo wybranej próbce 100 notatek prasowych z korpusu PAP, w której ręcznie oznakowane zostały wyrazy wielosegmentowe. Tagowanie przeprowadzane było przez dwie osoby (autor i promotor pracy). Powstały korpus oznaczmy przez PAP-TEST[7]. Fragment otagowanej notatki pokazano poniżej:

```
Zdaniem prezes {{*** Narodowego Banku Polskiego}} {{*--- Hanny Gronkiewicz-Waltz}}
{{** Jarosław Bauc}} jest odpowiednim kandydatem na {{*- ministra finansów}}.
```

Podwójne nawiasy klamrowe oznaczają miejsca wystąpień WW, a segmenty odmienne i nieodmienne oznaczamy odpowiednio przez $*$ i -.

Test polega na wyborze co najmniej jednej spośród dostępnych metod (**DM**, **pDM**, **SM** i **SDM**), ustaleniu wartości ich parametrów liczbowych (np. rs_{min} dla metody SM)[8] oraz wykonaniu tagowania na korpusie PAP-TEST pozbawionym tagów wybranymi metodami – w przypadku wyboru kilku metod należy określić ich priorytety. W wyniku tagowania otrzymujemy otagowany korpus wynikowy PAP-WW. Porównując go z PAP-TEST możemy wyznaczyć cztery zbiory wyrażeń:

- T_i – zbiór poprawnie rozpoznanych wyrażeń z prawidłowo zidentyfikowanymi segmentami odmiennymi.
- T_d – zbiór poprawnie rozpoznanych wyrażeń z nieprawidłowo zidentyfikowanymi segmentami odmiennymi.
- F_n – zbiór wyrażeń, które powinny być rozpoznane, lecz nie zostały rozpoznane.
- F_p – zbiór wyrażeń, które nie powinny być rozpoznane, lecz zostały rozpoznane.

Wprowadzono dwa rodzaje testu w zależności od sposobu traktowania wyrażeń ze zbioru T_d: **test rozpoznawania** uznaje je za poprawne, natomiast **test ekstrakcji** uznaje je za błędne – podział ten

[7] Należy tutaj podkreślić, że wybrane notatki zostały wykluczone z korpusu treningowego PAP-TRAIN.

[8] Wartości optymalnych parametrów były walidowane krzyżowo: korpus PAP-TEST dzielono na pół, po czym jedną z połówek używano do optymalizacji, a druga do testu.

wynika z faktu, że o ile elementy T_d są poprawnie rozpoznane, to jednak nie można ich uznać za w pełni wyekstrahowane WW, ponieważ posiadają błędny wzorzec odmiany.

Wyniki działania algorytmów rozpoznawania i ekstrakcji informacji z tekstu tradycyjnie podaje się w postaci wartości wskaźników **precyzji** (*precision*, P) i **pełności** (*recall*, R). Precyzja określa, jaka część rozpoznanych wyników jest poprawna, natomiast pełność – jaką część oczekiwanych wyników rozpoznano poprawnie. Dla testu rozpoznawania wskaźniki te wyrażają się wzorami:

$$P_{rec} = \frac{|T_i \cup T_d|}{|T_i \cup T_d \cup F_p|} \qquad R_{rec} = \frac{|T_i \cup T_d|}{|T_i \cup T_d \cup F_n|}$$

Z kolei dla testu ekstrakcji obowiązują wzory:

$$P_{ext} = \frac{|T_i|}{|T_i \cup T_d \cup F_p|} \qquad R_{ext} = \frac{|T_i|}{|T_i \cup T_d \cup F_n|}$$

Dla obu metod wyznaczamy jeszcze miarę F_1 (F-measure) będącą ich średnią harmoniczną: $F_1 = \frac{2PR}{P+R}$. Jest to popularnie stosowana miara łącząca precyzję i pełność. Współczynniki F_1 dla obu testów oznaczymy odpowiednio przez F_{rec} i F_{ext}.

4.1 Wyniki testów

Wyniki testów wszystkich metod zebrano w poniższej tabeli 2. Najwyższą precyzję osiąga metoda pDM, ponieważ ekstrahuje ona wyłącznie hasła Wikipedii, które dodatkowo zostały przefiltrowane podczas wyznaczania wzorców odmiany. Widać też wyraźną poprawę P_{ext} dla pDM w stosunku do DM. Metoda SM co prawda sama osiąga niezbyt wysokie wyniki, lecz pozwala ona na skonstruowanie słownika, z którego korzysta metoda SDM osiągająca wysoką pełność. W ostatnim wierszu przedstawiono metodę łączoną, wykorzystującą kolejno pDM, SDM i SM. Dzięki takiej kolejności zostaje zachowana w dużym stopniu precyzja pDM, natomiast SDM i SM zwiększają wartość pełności. Metoda ta osiąga najlepsze

Tabela 2: Wyniki testów rozpoznawania i ekstrakcji wyrazów wielosegmentowych różnymi metodami. **Wyróżniono** najlepszy wynik w każdej z kolumn.

	Test rozpoznawania			**Test ekstrakcji**		
Metoda	P_{rec}	R_{rec}	F_{rec}	P_{ext}	R_{ext}	F_{ext}
DM	80.97	42.54	55.78	58.71	30.85	40.44
pDM	**90.12**	38.64	54.09	**86.96**	37.29	52.19
SM	50.46	64.75	56.72	47.82	61.36	53.75
SDM	62.83	64.75	63.77	60.86	62.71	61.77
pDM + SDM + SM	72.27	**70.14**	**71.19**	69.23	**67.19**	**68.19**

rezultaty, jednak istnieje też znaczna liczba błędnych wyników – wśród przyczyn błędów dominują:

- Długa, nietypowa struktura wyrażeń, np. zamiast "V Liceum Ogólnokształcące im. Augusta Witkowskiego" rozpoznano osobno "Liceum Ogólnokształcące" i "Augusta Witkowskiego". W tym przypadku jeden błąd spowodował zwiększenie F_n o jeden element i F_d o dwa.
- Brak obcojęzycznych nazw i nazwisk w CLPM, np. "Pete Sampras".
- Błędy ortograficzne, np. "W.Brytania" (brak spacji po kropce), "Białego Domy".
- Nadmiarowe wyrażenia z Wikipedii, np. "stycznia 1921", "grudniu 1981".

Podsumowując możemy stwierdzić, że rezultat liczbowy dość dobrze odzwierciedla rzeczywistą jakość wyników, chociaż może on być zaniżony. Istnieją możliwości dalszej poprawy.

5 Etykiety semantyczne wyrazów wielosegmentowych

Algorytm ekstrakcji etykiet semantycznych został zaprojektowany i zaimplementowany w ramach pracy magisterskiej Autora [3], natomiast później już w ramach przygotowań do pracy doktorskiej został on dopracowany i dostosowany do nowej struktury bazy danych. Ulepszona wersja została opisana w publikacji [4], po czym jeszcze została ona zmodyfikowana tak, by korzystała ze słownika CLPM.

Celem działania algorytmu jest wyznaczenie **etykiety semantycznej** – krótkiej definicji składającej się z kilku słów, np. dla słowa "Kraków" etykieta powinna brzmieć "miasto", a dla "Karol Bielecki" – "piłkarz ręczny". Etykieta zawiera **rzeczownik główny** oraz inne opcjonalne rzeczowniki lub przymiotniki, jednak powinna być krótka i zwięzła. Czasami trudno jest podać definicję przy pomocy rzeczownika i potrzebne są dodatkowe **operatory**, np. "część samochodu", "rasa kota", "grupa ludzi", które powinny zostać dołączone do etykiety.

Jako źródłowy zasób danych ponownie wykorzystano Wikipedię, a konkretnie – wyekstrahowane z niej początkowe akapity każdego z artykułów. Problem polega na przydzieleniu każdemu wyrazowi wielosegmentowemu z Wikipedii etykiety, która jest ekstrahowana z pierwszych zdań artykułu. Algorytm opiera się na spostrzeżeniach odnośnie struktury typowej definicji encyklopedycznej haseł i składa się z kilku etapów.

1. Usunięcie powtórzonego hasła z początkowego akapitu.
2. Podział artykułu na zdania i ich fragmenty, uporządkowane według rozpoczynającego je segmentu, np. fragment zdania zaczynający się od znaku "—" będzie prawdopodobnie zawierał definicję.
3. Wyszukiwanie rzeczownika głównego we fragmentach zdań z uwzględnieniem operatorów.
4. Uzupełnianie definicji o dodatkowe elementy.

Algorytm korzystający z CLPM generuje słownik zawierający 94.3% poprawnych etykiet semantycznych[9], co jest poprawą o ok. 2% w stosunku do poprzedniej wersji wykorzystującej bibliotekę CLP.

[9] Test wykonano na próbce 500 haseł.

Oprócz przydzielenia etykiet hasłom Wikipedii istnieje potrzeba ekstrakcji etykiet semantycznych dla dowolnych wyrazów wielosegmentowych wyekstrahowanych z tekstu. Jest to problem złożony, ponieważ w tekście nie znajdziemy bezpośredniej informacji na temat znaczenia danego wyrazu. Podjęto próbę zbadania, czy można wyznaczyć etykietę nowo wyekstrahowanego WW na podstawie etykiet haseł, z których wygenerowano wzorce syntaktyczne (metoda SM), jednak okazało się, że podejście to daje niskie wyniki – wstępne testy pokazały dokładność poniżej 25% dla 100 przypadkowo wybranych wyrazów z automatycznie utworzonego słownika liczącego 171 tys. wyrazów. Nie pomogła również próba użycia WordNetu do znalezienia wspólnego hiperonimu w przypadku kilku konfliktujących etykiet. Powodem jest drobnoziarnistość etykiet oraz brak bezpośredniej implikacji miedzy syntaktyką a semantyką. W przyszłości należy dopracować istniejące etykiety tak, by mogły posłużyć za zbiór treningowy i użyć uczenia maszynowego z nadzorem do ekstrakcji etykiet dla nowych wyrazów.

6 Podsumowanie

Przeprowadzone badania wykazują prawdziwość przedstawionych tez. **Teza 1** została udowodniona przez wyniki uzyskane przez metody SM, SDM oraz metodę łączoną. Zaprezentowane rezultaty pokazują, że istnieje możliwość automatycznej ekstrakcji wyrazów wielosegmentowych z tekstu przy pomocy słownika fleksyjnego i artykułów Wikipedii bez wykorzystania dodatkowych reguł i zbiorów treningowych. Metoda łączona (pDM + SDM + SM) uzyskała w teście rozpoznawania wyrazów wielosegmentowych wartość F_1 przekraczającą 71%, a w teście ekstrakcji – 68%, co pozwala stwierdzić, że teza ta została potwierdzona. Prawdziwość **Tezy 2** wykazują z kolei przedstawione metody tworzenia słownika WW z Wikipedii (metody DM i pDM) i z wyników działania algorytmu SM.

Literatura

[1] Božo Bekavac i Marko Tadic. *A generic method for multi word extraction from Wikipedia*. 30th International Conference on Information Technology Interfaces (ITI), str. 663–668. IEEE, 2008.

[2] Aleksander Buczyński i Adam Przepiórkowski. *Spejd: A shallow processing and morphological disambiguation tool*. Human Language Technology. Challenges of the Information Society, str. 131–141. Springer, 2009.

[3] Paweł Chrząszcz. *Automatyczne rozpoznawanie i klasyfikacja nazw wielosegmentowych na podstawie analizy haseł encyklopedycznych*. Praca magisterska, Akademia Górniczo-Hutnicza im. Stanisława Staszica w Krakowie, 2009.

[4] Paweł Chrząszcz. *Enrichment of inflection dictionaries: automatic extraction of semantic labels from encyclopedic definitions*. Proceedings of the 9th International Workshop on Natural Language Processing and Cognitive Science (NLPCS, w połączeniu z ICEIS), str. 106–119. SciTePress, 2012.

[5] Matthieu Constant i Anthony Sigogne. *MWU-aware part-of-speech tagging with a CRF model and lexical resources*. Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World, str. 49–56. Association for Computational Linguistics, 2011.

[6] Meghdad Farahmand i Ronaldo Martins. *A supervised model for extraction of multiword expressions based on statistical context features*. Proceedings of the 10th Workshop on Multiword Expressions (MWE, w połączeniu z EACL), str. 10–16. Association for Computational Linguistics, 2014.

[7] Evgeniy Gabrilovich i Shaul Markovitch. *Computing semantic relatedness using Wikipedia-based explicit semantic analysis.* Proceedings of the 20th International Joint Conference on Artifical Intelligence (IJCAI), tom 7, str. 1606–1611. Morgan Kaufmann Publishers Inc., 2007.

[8] Marek Gajęcki. *Słownik fleksyjny jako biblioteka języka C*. Słowniki komputerowe i automatyczna ekstrakcja informacji z tekstu (pod redakcją Wiesława Lubaszewskiego). Wydawnictwa AGH, Kraków, 2009.

[9] Filip Graliński, Agata Savary, Monika Czerepowicka i Filip Makowiecki. *Computational lexicography of multi-word units: how efficient can it be?* Proceedings of the Workshop on Multiword Expressions: from Theory to Applications (MWE), str. 1–9. Association for Computational Linguistics, 2010.

[10] Roman Kurc, Maciej Piasecki i Bartosz Broda. *Constraint based description of Polish multiword expressions*. Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC), str. 2408–2413. European Language Resources Association, 2012.

[11] Wiesław Lubaszewski, H. Wróbel, M. Gajęcki, B. Moskal, A. Orzechowska, P. Pietras, P. Pisarek i T. Rokicka. *Słownik Fleksyjny Języka Polskiego*. Grupa Lingwistyki Komputerowej, Katedra Informatyki AGH i Katedra Lingwistyki Komputerowej UJ, Kraków, 2001.

[12] Cynthia Matuszek, John Cabral, Michael J. Witbrock i John DeOliveira. *An introduction to the syntax and content of Cyc*. AAAI Spring Symposium: Formalizing and Compiling Background Knowledge and Its Applications to Knowledge Representation and Question Answering, str. 44–49. 2006.

[13] Marek Maziarz, Maciej Piasecki i Stanisław Szpakowicz. *Approaching plWordNet 2.0*. Proceedings of the 6th Global Wordnet Conference. Global WordNet Association, 2012.

[14] *Morfologik. Analizator morfologiczny + słownik morfologiczny + korektor gramatyczny + biblioteki*. Dostępny 8 maja 2015.
`http://morfologik.blogspot.com`

[15] Pavel Pecina. *A machine learning approach to multiword expression extraction*. Proceedings of the LREC Workshop – Towards a Shared Task for Multiword Expressions (MWE), str. 54–61. European Language Resources Association, 2008.

[16] Jakub Piskorski, Peter Homola, Małgorzata Marciniak, Agnieszka Mykowiecka, Adam Przepiórkowski i Marcin Woliński. *Information extraction for Polish using the SProUT platform*. Intelligent Information Processing and Web Mining, tom 25 z serii *Advances in Soft Computing*, str. 227–236. Springer Berlin Heidelberg, 2004.

[17] Aleksander Pohl i Bartosz Ziółko. *A comparison of Polish taggers in the application for automatic speech recognition*. Proceedings of the 6th Language and Technology Conference (LTC), str. 294–298. 2013.

[18] Carlos Ramisch, Paulo Schreiner, Marco Idiart i Aline Villavicencio. *An evaluation of methods for the extraction of multiword expressions*. Proceedings of the LREC Workshop – Towards a Shared Task for Multiword Expressions (MWE), str. 50–53. European Language Resources Association, 2008.

[19] Carlos Ramisch, Aline Villavicencio i Christian Boitet. *MWEToolkit: a framework for multiword expression identification*. Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC), str. 662–669. European Language Resources Association, 2010.

[20] Josef Ruppenhofer, Michael Ellsworth, Miriam R.L. Petruck, Christopher R. Johnson i Jan Scheffczyk. *FrameNet II: Extended theory and practice*. International Computer Science Institute, Berkeley, CA, 2006.

[21] Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann Copestake i Dan Flickinger. *Multiword expressions: a pain in the neck for NLP*. Computational Linguistics and Intelligent Text Processing, tom 2276 z serii *Lecture Notes in Computer Science*, str. 1–15. Springer Berlin Heidelberg, 2002.

[22] Jakub Waszczuk. *Harnessing the CRF complexity with domain-specific constraints. The case of morphosyntactic tagging of a highly inflected language.* Proceedings of the 24th International Conference on Computational Linguistics (COLING), str. 2789–2804. 2012.

[23] *Wikipedia. Wolna encyklopedia*. Dostępny 8 maja 2015.
`https://pl.wikipedia.org`

[24] *Wikisłownik. Wolny, wielojęzyczny słownik*. Dostępny 23 maja 2015.
`http://pl.wiktionary.org`

[25] Marcin Woliński. *Morfeusz — a practical tool for the morphological analysis of Polish*. Advances in Soft Computing, 26(6), str. 503–512, 2006.

[26] Michał Woźniak. *Automatic extraction of multiword lexical units from Polish text*. 5th Language and Technology Conference (LTC). 2011.

[27] Yi Zhang, Valia Kordoni, Aline Villavicencio i Marco Idiart. *Automated multiword expression prediction for grammar engineering*. Proceedings of the Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties, str. 36–44. Association for Computational Linguistics, 2006.