Podręcznik użytkowania przestrzeni semantycznej dla języka polskiego*

Marcin Tatjewski

marcin.tatjewski@gmail.com
Instytut Podstaw Informatyki, Polska Akademia Nauk,
ul. Jana Kazimierza 5, 01-248 Warszawa

1 Wprowadzenie

1.1 Kontekst

Niniejszy podręcznik jest instrukcją użytkowania interfejsu do przetwarzania przestrzeni semantycznej dla języka polskiego. Przestrzeń tę opracowałem na podstawie zrównoważonej wersji Narodowego Korpusu Języka Polskiego [1] przy pomocy algorytmu COALS [2] zaimplementowanego w pakiecie S-Space [3]. Opisywana przestrzeń semantyczna jest modelem semantycznym dla słów z języka polskiego, który pozwala na badanie relacji podobieństwa znaczeniowego pomiędzy nimi. Prace nad przestrzenią zostały przeprowadzone w ramach projektu APPROVAL [4].

1.2 Definicje

Przestrzeń / przestrzeń semantyczna Model semantyczny dla języka polskiego lub plik zawierający ten model. Przestrzeń jest przeznaczona do obsługi przez opisywany interfejs. Aktualna wersja przestrzeni znajduje się w pliku o nazwie **„d42200w500d.sspace"**. Jej szczegółowy opis znajduje się w załączniku A.

Interfejs Interfejs użytkownika przestrzeni semantycznej będący programem o rozszerzeniu JAR służacym do badania zależności między słowami w przestrzeni. Interfejs ten jest oparty na oprogramowaniu z pakietu S-Space [3]. Najnowsza wersja znajduje się w pliku **„sspace-2.0.4-appr.jar"**.

2 Interpretacja przestrzeni semantycznej

Przestrzeń semantyczna jest statystycznym modelem zależności semantycznych w języku [5]. Jej eksploracja pozwala na badanie tych zależności pomiędzy poszczególnymi słowami zawartymi w korpusie tekstów, na którym dana przestrzeń

* Niniejsza praca powstała w ramach projektu finansowanego przez Narodowe Centrum Nauki na podstawie decyzji numer DEC-2011/03/B/HS2/02279

została wykonana. Bliskość numeryczną pary słów w przestrzeni można najprościej rozumieć jako bliskość ich znaczenia, operacjonalizowanego jako podobieństwo kontekstów, w których występują.

Przestrzeń wymieniana w niniejszej instrukcji zawiera 42200 słów najczęściej występujących w zrównoważonej wersji Narodowego Korpusu Języka Polskiego. Są to słowa, które występowały w korpusie co najmniej około 150 razy. W przestrzeni zawarte są jedynie postacie bazowe słów, otrzymane za pomocą lematyzacji i dezambiguacji.

3 Przygotowanie interfejsu oraz przestrzeni do pracy

3.1 Uwagi wstępne

Bardzo ważne jest, aby przed rozpoczęciem pracy z przestrzenią upewnić się, że komputer, z którego korzystamy, dysponuje w danej chwili co najmniej 1 GB wolnej pamięci operacyjnej RAM. W przeciwnym razie mogą wystąpić błędy opisane w rozdziale 5. Aby zwolnić część pamięci operacyjnej komputera, wystarczy na czas pracy z przestrzenią wyłączyć aplikacje, które intensywnie wykorzystują zasoby maszyny, na której pracujemy, np. przeglądarki internetowe z dużą liczbą otwartych kart.

Zalecane jest również pobranie ze strony https://www.java.com/pl/download/ i zainstalowanie najnowszej wersji darmowego oprogramowania Java.

Aby wstępnie przygotować przestrzeń oraz interfejs do użytkowania, należy umieścić plik przestrzeni oraz plik interfejsu w tym samym folderze. Szczegółowe instrukcje uruchomienia dotyczące poszczególnych systemów operacyjnych znajdują się w dalszej części tego rozdziału.

Interfejs posiada spójny zestaw komend, którego można używać w trybie interaktywnym lub poprzez wczytanie listy komend do wykonania z pliku. Komendy oraz ich interaktywny tryb użycia opisane są w sekcji 4.1. Procedura ładowania zbioru komend opisana jest w sekcji 4.2.

3.2 Przygotowanie w systemach Windows

Instalacja programu Cygwin

1. Wymagana jest instalacja programu Cygwin. W zależności od tego, czy posiadamy 32-bitową lub 64-bitową wersję systemu operacyjnego, pobieramy jeden z plików instalacyjnych:
   - 32-bitowy (http://cygwin.com/setup-x86.exe)
   - 64-bitowy (http://cygwin.com/setup-x86_64.exe)
2. Uruchamiamy plik instalacyjny i przeprowadzamy pełną instalację programu Cygwin. Zalecane jest, aby w trakcie instalacji utworzyć na pulpicie skrót do uruchamiania programu Cygwin. Odpowiednia opcja do zaznaczenia pojawi się w trakcie instalacji.

Przygotowanie i uruchomienie interfejsu za pomocą programu Cygwin

1. W folderze, w którym zainstalowaliśmy Cygwin, znajduje się podfolder „home", a z kolei w nim folder o nazwie takiej samej jak nasza nazwa użytkownika w Windows („home/<nazwa_użytkownika>"). Należy w katalogu „<nazwa_użytkownika>" umieścić przestrzeń semantyczną („d42200w500d.sspace") oraz plik interfejsu.
2. Następnie należy uruchomić program Cygwin.
3. W konsoli programu Cygwin stosujemy następującą komendę, aby uruchomić interfejs:

```
java -cp sspace-2.0.4-appr.jar edu.ucla.sspace.tools.SemanticSpaceExplorer
```

4. Teraz możemy korzystać interaktywnie z komend interfejsu zgodnie z ich opisem w sekcji 4.1. Należy pamiętać, aby wpierw załadować plik przestrzeni „d42200w500d.sspace".

3.3 Przygotowanie w systemach Linux oraz Mac OS X

Aby uruchomić interfejs interaktywny, należy wykonać następujące kroki:

1. Uruchomić program terminal lub konsola.
2. Wewnątrz terminala należy przejść do folderu, w którym zostały umieszczone przestrzeń oraz interfejs.
3. Interfejs uruchamiamy następującą komendą:

```
java -cp sspace-2.0.4-appr.jar edu.ucla.sspace.tools.SemanticSpaceExplorer
```

Po wykonaniu tych kroków pomyślne uruchomienie interfejsu rozpoznamy po obecności znaku „>", który oznacza oczekiwanie na wprowadzanie komend, opisanych w rozdziale 4.1.

Użycie wczytywania komend z pliku jest analogiczne do podobnej procedury w systemach Windows, zob. 4.2.

4 Praca z przestrzenią semantyczną

4.1 Komendy interfejsu

Pracę w interfejsie należy zacząć od wykonania komendy **„load"**, która służy do załadowania wybranej przestrzeni. Wpisywanie każdej z komend kończymy klawiszem „enter".

Każda komenda interfejsu składa się z jednego lub więcej słów połączonych łącznikiem „-". Każda z komend ma również dostępny skrót w postaci ciągłego zapisu pierwszych liter słów swojej pełnej wersji. Poniżej, dla komend wielowyrazowych podano wersje skrócone.

Wartości bliskości znaczeniowej słów podane są według miary podobieństwa cosinusowego przyjmującej wartości z przedziału [0;1] (brak jest ujemnych wartości z powodu niezerowości wartości wektorów), gdzie wartość 1 oznacza znaczenie identyczne, a 0 brak istotnego podobieństwa.

load <nazwa_pliku_przestrzeni> Komenda powoduje załadowanie przestrzeni przez interfejs, po czym przestrzeń jest gotowa do przeglądania innymi komendami. Przykład:

```
load d42200w500d.sspace
```

Wykonanie komendy „load" może potrwać do kilkunastu sekund. Próba załadowania komendą „load" nieistniejącego pliku (np. przez podanie błędnej nazwy pliku poprawnego) wywoła komunikaty o błędach i/lub zamknięcie programu. W takiej sytuacji program należy uruchomić ponownie.

gn <słowo> <liczba> Komenda powoduje wypisanie najbliższych sąsiadów zadanego słowa. Parametr <liczba> określa liczbę podanych sąsiadów. Pominięcie parametru <liczba> powoduje wypisanie 10 sąsiadów. Najbliżsi sąsiedzi są wypisani od dołu do góry. Im wyższa liczba podana obok słowa, tym większe podobieństwo znaczeniowe. Przykład:

```
gn czekolada 20
```

gs <słowo1> <słowo2> Komenda zwraca wartość podobieństwa semantycznego dla dwóch zadanych słów. Przykład:

```
gs pies kot
```

gw <ciąg_znaków> Komenda zwraca wszystkie słowa dostępne w danej przestrzeni, które zaczynają się od zadanego ciągu znaków. Nie wszystkie słowa z języka polskiego znajdują się w przestrzeni. Model zawiera wyłącznie 42200 słów najczęściej występujących w zrównoważonej wersji Narodowego Korpusu Języka Polskiego.

wcr <nazwa_pliku> <komenda> <argumenty_komendy> Komenda zapisuje wynik wykonania innej komendy do wskazanego pliku. Przykład:

```
wcr plik.txt gn czekolada 20
```

Wynik tej komendy można łatwo skopiować do arkusza kalkulacyjnego, gdzie słowa i liczby ułożą się w oddzielnych kolumnach.

help Komenda podaje pełną listę komend dostępnych w przestrzeni. Komendy są podane pełnymi słowami, lecz mogą być również zastosowane w postaci skrótów. Na przykład komenda „get-words" to „gw". Wynik tego polecenia zawiera więcej komend niż te opisane powyżej, jednak te pozostałe nie wydają się przydatne dla użytkownika przestrzeni, którą się zajmujemy w niniejszym dokumencie.

Pracę z interfejsem w trybie interaktywnym można zakończyć, używając kombinacji klawiszy „Ctrl+C".

4.2 Ładowanie pliku komend

Poza interaktywnym korzystaniem z interfejsu przestrzeni możliwe jest załadowanie do interfejsu pliku komend do wykonania i otrzymanie w rezultacie pliku zawierającego kolejno wyniki wszystkich podanych komend.

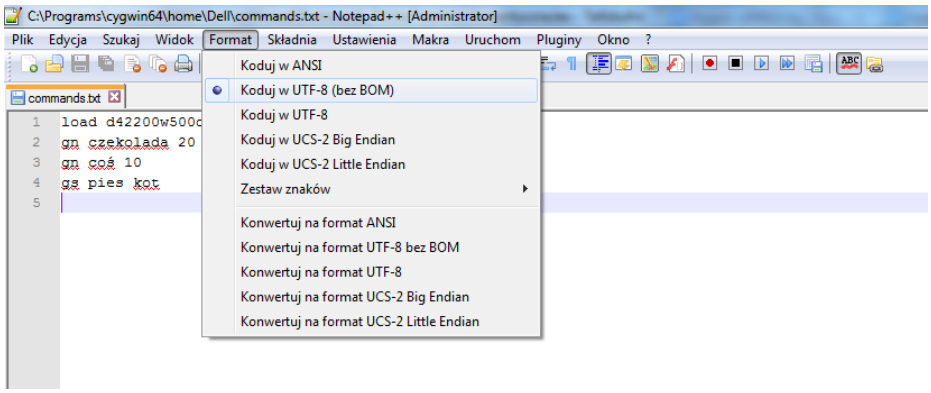
Aby wykonać tę procedurę, należy:

1. Przygotować plik z listą komend do wykonania i umieścić go w tym samym folderze co przestrzeń semantyczną. Konieczne jest, aby plik z komendami był zakodowany w formacie UTF-8 (najlepiej „UTF-8 bez BOM"). W systemie Windows prostym sposobem, aby to zapewnić, jest użycie do zapisu edytora **notepad++**, który jest dostępny bezpłatnie (dostępny pod adresem http://notepad-plus-plus.org/download/). W tym edytorze można w wygodny sposób ustawić kodowanie znaków, tak jak na rysunku 1. Pożądane komendy należy wpisać zgodnie z ich strukturą opisaną w rozdziale 4.1, rozpoczynając od komendy ładującej przestrzeń semantyczną („load").
2. Wykonać następującą komendę - na Windows w programie Cygwin, a na Linux i Mac OS X w terminalu (wykonanie może potrwać do kilkunastu sekund):

```
java -cp sspace-2.0.4-appr.jar edu.ucla.sspace.tools.SemanticSpaceExplorer -f cmd.txt > out.txt
```

gdzie
sspace-2.0.4-appr.jar to plik interfejsu,
cmd.txt to plik z komendami do wykonania,
out.txt to plik, w którym zostanie zapisany rezultat zadanych komend. Ten plik jest również zakodowany w UTF-8, więc do jego odczytu należy skorzystać ponownie z programu notepad++.

Rysunek 1. Ustawianie dodawanie znaków w notepad++



5 Najczęstsze problemy techniczne

„Exception in thread «main» java.lang.OutOfMemoryError" Błąd ten występuje, gdy komputer, z którego korzystamy, nie ma w danej chwili wystarczającej ilości wolnej pamięci operacyjnej RAM do obsłużenia przestrzeni semantycznej. Należy wtedy zwolnić część zasobów komputera, zamykając aplikacje, które intensywnie wykorzystują jego pamięć operacyjną (np. przeglądarki internetowe z wieloma otwartymi oknami).
Praca z przestrzenią wymaga około 0,5 GB RAM, zatem komputer w chwili uruchomienia przestrzeni powinien mieć co najmniej 1 GB wolnego RAM. W razie występowania problemów pomimo zamknięcia części innych programów można wspomóc się, uruchamiając interfejs przestrzeni przy użyciu poniższej komendy, która wymusza na systemie operacyjnym rezerwację odpowiedniej ilości zasobów:

```
java -Xms512m -Xmx1024m -cp sspace-2.0.4-appr.jar  edu.ucla.sspace.tools.SemanticSpaceExplorer
```

„Unknown command: ?load" Ten błąd lub podobne występujące przy wczytywaniu pliku z komendami do interfejsu może być najprawdopodobniej spowodowany niewidocznym w edytorach tekstowych znakiem BOM (byte order mark) na początku pliku komend. Najprostszy sposób na uniknięcie tego problemu to pozostawienie pierwszego wiersza w pliku komend pustego.

A Tekstowa wersja przestrzeni semantycznej

Przestrzeń znajduje się w pliku **„d42200w500d.sspace"**. Jest to wersja w formacie czytelnym dla człowieka. Jest zapisana w systemie kodowania znaków UTF-8. Zajmuje około 377 MB przestrzeni dyskowej. Jej struktura jest następująca:

```
42200 500
słowoA|0.014 0.234 ...
słowoB|0.414 0.054 ...
słowoC|0.201 0.002 ...
...
```

1. Pierwszy wiersz składa się z dwóch liczb, po kolei:
   (a) N - liczby różnych słów opisanych w przestrzeni i zarazem liczby wierszy pliku przestrzeni (pomijając ten pierwszy wiersz),
   (b) M - liczby wartości opisujących każde słowo, czyli liczby kolumn opisujących semantykę słów.
2. Każdy kolejny wiersz to:
   (a) opisywane słowo,
   (b) znak „|",
   (c) M liczb z przedziału [0;1] oddzielonych spacjami. Liczby te to wartości w wymiarach stanowiących opis semantyczny słów. Poszczególne wymiary nie mają określonej interpretacji, ponieważ są skonstruowane w wyniku redukcji wymiarów metodą SVD.

B Tworzenie przestrzeni semantycznej

Przestrzeń semantyczna, z której korzystanie jest opisane w niniejszym dokumencie, została stworzona na bazie zrównoważonej wersji Narodowego Korpusu Języka Polskiego (NKJP) [1], przy pomocy pakietu oprogramowanie S-Space Package [3]. Przestrzenie semantyczne można tworzyć na podstawie różnorodnych korpusów, również na bazie korpusów zebranych samodzielnie. Poniższe uwagi mogą być przydatne w procesie tworzenia nowej przestrzeni.

- Instrukcje na temat tworzenia nowej przestrzeni przy pomocy pakietu S-Space są dostępne pod adresem:
  https://github.com/fozziethebeat/S-Space/wiki/GettingStarted
- Zasoby Narodowego Korpusu Języka Polskiego nie są publicznie w pełni dostępne. Mogą zostać udostępnione za zgodą autorów korpusu. Dane autorów dostępne są na oficjalnej stronie Korpusu pod adresem: http://nkjp.pl/
- Narodowy Korpus Języka Polskiego zawiera również dane o swojej lematyzacji i dezambiguacji. Samodzielnej lematyzacji i dezambiguacji NKJP lub innego korpusu można dokonać za pomocą programu Pantera, dostępnego pod następującym adresem: https://code.google.com/p/pantera-tagger/
- W celu przetworzenia większości korpusów tekstów do postaci niezbędnej dla tworzenia przestrzeni za pomocą S-Space Package niezbędna jest umiejętność programowania.

Literatura

1. Adam Przepiórkowski, Mirosław Bańko, Rafał L. Górski i Barbara Lewandowska-Tomaszczyk (red.). *Narodowy Korpus Języka Polskiego*. Wydawnictwo Naukowe PWN, Warszawa, 2012.
2. Douglas L. T. Rohde, Laura M. Gonnerman, David C. Plaut. An improved model of semantic similarity based on lexical co-occurence. *COMMUNICATIONS OF THE ACM*, 8:627–633, 2006.
3. David Jurgens, Keith Stevens. The S-Space Package: an open source package for word space models. *Proceedings of the ACL 2010 System Demonstrations*, ACLDemos '10, strony 30–35, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.
4. Witryna projektu APPROVAL. `http://www.approval.uw.edu.pl/`.
5. Peter D. Turney, Patrick Pantel. From frequency to meaning: Vector space models of semantics. *J. Artif. Intell. Res. (JAIR)*, 37:141–188, 2010.