

Instrukcja do narzędzia WoSeDon

Najnowsza wersja narzędzia wraz z zasobami znajduje się na repozytorium:
git clone git@156.17.135.23:wosedon

- Zasoby umieszczone zostały w katalogu *wosedon/wosedon/resources/*
- Plik konfiguracyjny potrzebny do uruchomienia narzędzia umieszczony został w katalogu *wosedon/wosedon/cfg/*

Charakterystyka narzędzia WoSeDon:

Narzędzie przeznaczone do ujednoznaczniania znaczeń leksykalnych słów (ang. Word Sense Disambiguation). Algorytm, który został zaimplementowany do rozstrzygania niejednoznaczności to algorytm PageRank [1]. Narzędzie WoSeDon zawiera różne modyfikacje owego algorytmu oraz umożliwia manipulację wieloma parametrami mającymi szczególny wpływ na jakość ujednoznaczniania znaczeń leksykalnych.

Instalacja narzędzia WoSeDon:

1. Narzędzie WoSeDon wymaga zainstalowania narzędzia Corpus2, którego opis oraz instrukcja instalacji dostępne są na stronie <http://nlp.pwr.wroc.pl/redmine/projects/corpus2/wiki>
2. Narzędzie WoSeDon wymaga zainstalowania narzędzia graph-tool, które dostępne jest na stronie <http://graph-tool.skewed.de/>. W przypadku problemów z instalacją sprawdź wersję gcc czy na pewno jest zgodna z tą wymaganą, jak i g++.
3. Następnie należy zainstalować narzędzie PLWNGraphBuilder, które znajduje się w katalogu z pobranym z repozytorium projektem WoSeDon. W katalogu *wosedon/PLWNGraphBuilder/* znajduje się skrypt instalacyjny *setup.py*
4. Ostatnim krokiem jest instalacja narzędzia WoSeDon. W katalogu *wosedon/wosedon/* znajduje się skrypt instalacyjny *setup.py*

Opis pliku konfiguracyjnego:

[wosedon]

Nazwa parametru	Możliwe wartości dla parametru	Opis parametru
context	Document	Parametr, który umożliwia ustawienie kontekstu dla ujednoznacznianego słowa.
	Sentence	
gbuilders	SynsetGraphBuilder	Odpowiednia wartość przypisana do parametru specyfikuje rodzaj grafu, na którym zostanie uruchomiony algorytm PageRank. W przypadku jeżeli zamierzamy połączyć dwa grafy podajemy wartości parametru po spacji np. gbuilders = SynsetGraphBuilder MSRGraphBuilder
	LexicalUnitGraphBuilder	
	MSRGraphBuilder	
	BidirectionalMSRGraphBuilder	
	WNDGraphBuilder	
	SUMOGraphBuilder	
mergers	SynsetsLUMerger	Odpowiednia wartość przypisana do

	SynsetsLUMerger2	parametru specyfikuje rodzaj łączenia dwóch grafów w jeden. Nazwa dla każdej metody łączenia grafu, podana w kolumnie obok, sugeruje w jakiej kolejności powinny być podane grafy w poprzednim parametrze. Jeżeli ustawimy np. mergers = SynsetsLUMerger2 to należy ustawić również gbuilders = SynsetGraphBuilder LexicalUnitGraphBuilder
	SynsetsSUMOMerger	
	SynsetsWNDMerger	
	SynsetsMSRMerger	
wsdalgorithm	GTPersonalizedPR	Odpowiednia wartość przypisana do parametru specyfikuje rodzaj algorytmu PageRank, jaki zostanie wykorzystany w procesie ujednoznaczniania słów.
	GTPersonalizedPRNorm	
	GTPersonalizedPRNorm2	
	GTPersonalizedPRNormIt	
	GTPersonalizedPRNormReduction	
	GTPersonalizedW2WPR	
	GTPersPRNormItModV	
	GTPersPRNormItModVRankNorm	
	GTPersPRNormModV	
	GTStaticPR	
rerankers	LemmaRankingNormalizer	Odpowiednia wartość przypisana do parametru specyfikuje czy wartości rankingu dla każdego synsetu mają zostać znormalizowane w obrębie ujednoznacznianego słowa.
	NodeDegreeRanker	

[wosedon:resources]

Nazwa parametru	Opis parametru
sumo_graph_file	Ścieżka do pliku z grafem SUMO.
mapping_sumo_file	Ścieżka do pliku z rzutowaniem Słowsieci na ontologię SUMO.
wnd_graph_file	Ścieżka do pliku z grafem WordNet Domains.
mapping_wnd_file	Ścieżka do pliku z rzutowaniem Słowsieci na WordNet Domains.
msr_file	Ścieżka do pliku z MSR.
plwn_graph_file	Ścieżka do pliku zawierającego spakowane grafy, a mianowicie graf, którego wierzchołkami są wyłącznie synsety oraz graf, którego wierzchołkami są wyłącznie jednostki leksykalne. Grafy można zbudować za pomocą narzędzia PLWNGraphBuilder
tagset	Tagset.

[wosedon:build_options]

Nazwa parametru	Opis parametru
unique_edges	Jeżeli ustawiono wartość parametru na True to oznacza, iż z grafu zostaną usunięte powtarzające się krawędzie.
directed_graphs	Jeżeli ustawiono wartość parametru na True to oznacza, iż graf będzie grafem skierowanym.
syn_rel_ids	Wartościami parametru są identyfikatory relacji pomiędzy synsetami w Słownosieci, podawane kolejno po spacji. Jeżeli zostaną podane identyfikatory relacji, to z grafu (synsetów) zostaną usunięte wszystkie relacje, których identyfikator jest różny od podanego (podanych). Jeżeli nie zostaną podane identyfikatory, to w grafie pozostaną wszystkie relacje.
lu_rel_ids	Wartościami parametru są identyfikatory relacji pomiędzy jednostkami leksykalnymi w Słownosieci, podawane kolejno po spacji. Jeżeli zostaną podane identyfikatory relacji, to z grafu (jednostek leksykalnych) zostaną usunięte wszystkie relacje, których identyfikator jest różny od podanego (podanych). Jeżeli nie zostaną podane identyfikatory, to w grafie pozostaną wszystkie relacje.
accept_pos	Wartościami parametru są identyfikatory części mowy, podawane kolejno po spacji. Jeżeli zostaną podane identyfikatory części mowy, to z grafu zostaną usunięte wszystkie wierzchołki, których identyfikator części mowy jest różny od podanego (podanych). Jeżeli nie zostaną podane identyfikatory, to w grafie pozostaną wszystkie wierzchołki.
syn_syn_rel_weight	Wartością parametru jest lista relacji wraz z przypisanymi do nich wagami w formacie identyfikator relacji:waga dla relacji Wartości powinny być podawane po spacji. Jeżeli nie zostaną ustawione wagi dla relacji to relacje otrzymają wagę 0.
lu_lu_rel_weight	Wartością parametru jest lista relacji wraz z przypisanymi do nich wagami w formacie identyfikator relacji:waga dla relacji Wartości powinny być podawane po spacji. Jeżeli nie zostaną ustawione wagi dla relacji to relacje otrzymają wagę 0.
sumo_sumo_rel_weight	Wartością parametru jest lista relacji wraz z przypisanymi do nich wagami w formacie nazwa relacji:waga dla relacji Wartości powinny być podawane po spacji. Jeżeli nie zostaną ustawione wagi dla relacji to relacje otrzymają wagę 0.
wnd_wnd_rel_weight	Wartością parametru jest lista relacji wraz z przypisanymi do nich wagami w formacie nazwa relacji:waga dla relacji Wartości powinny być podawane po spacji. Jeżeli nie zostaną ustawione wagi dla relacji to relacje otrzymają wagę 0.

[wosedon:merge_options]

Nazwa parametru	Opis parametru
-----------------	----------------

syn_lu_rel_weight	Wartością parametru jest waga dla relacji pomiędzy wierzchołkiem synsetem, a wierzchołkiem jednostką leksykalną. Ma to znaczenie w przypadku łączenia dwóch grafów. Jeżeli nie ustawiono wagi to otrzyma ona wartość 0.
syn_msr_rel_weight	Wartością parametru jest waga dla relacji pomiędzy wierzchołkiem synsetem, a wierzchołkiem msr. Ma to znaczenie w przypadku łączenia dwóch grafów. Jeżeli nie ustawiono wagi to otrzyma ona wartość 0.
syn_sumo_rel_weight	Wartością parametru jest waga dla relacji pomiędzy wierzchołkiem synsetem, a wierzchołkiem z pojęciem SUMO. Ma to znaczenie w przypadku łączenia dwóch grafów. Jeżeli nie ustawiono wagi to otrzyma ona wartość 0.
syn_wnd_rel_weight	Wartością parametru jest waga dla relacji pomiędzy wierzchołkiem synsetem, a wierzchołkiem z dziedziną WordNet. Ma to znaczenie w przypadku łączenia dwóch grafów. Jeżeli nie ustawiono wagi to otrzyma ona wartość 0.

[wosedon:wsd_alg]

Nazwa parametru	Opis parametru
dumping_factor	Współczynnik tłumienia, najczęściej ustawiany na 0.85
max_iter	Maksymalna liczba iteracji algorytmu PageRank.
edge_weight	Jeżeli ustawiono na True, oznacza to, iż zostaną wykorzystane wagi przypisane do relacji, w przeciwnym wypadku wagi nie będą brane pod uwagę.

Źródła

[1] Agirre, E., de Lacalle, O. L., and Soroa, A. (2014). Random walks for knowledge-based word sense disambiguation. Computational Linguistics, 40(1):57–84.