

Przetwarzanie Języka Naturalnego

Lab 8 – NER

Aleksander Smywiński-Pohl

Wydział IEiT
Katedra Informatyki

11.04.2017

✦ Named Entity Recognition – NER

Rozpoznawanie jednostek referencyjnych¹ polega na określeniu, które spośród wyrażen występujących w tekście odnoszą się do specyficznych obiektów najczęściej posiadających własną nazwę oraz jaka jest kategoria semantyczna obiektów, do których odnoszą się te wyrażania.

A. Smywiński-Pohl, *Automatyczna ekstrakcja relacji semantycznych z tekstów w języku polskim*

¹W polskiej literaturze funkcjonuje również termin *rozpoznawanie jednostek nazewniczych*.

Korea Północna_[GPE] *zagroziła wystrzeleniem pocisku balistycznego w kierunku* ***USA***_[GPE].

Typy jednostek referencyjnych:

- ✘ ludzie (ang. *people*),
- ✘ organizacje (ang. *organizations*),
- ✘ miejsca (ang. *locations*),
- ✘ podmioty geopolityczne (ang. *geo-political entitites*),
- ✘ obiekty użyteczności publicznej (ang. *facilities*),
- ✘ pojazdy (ang. *vehicles*),
- ✘ etc.

- ⌘ HMM, CRF – modele statystyczne oparte na założeniu liniowości zjawisk tekstowych
- ⌘ entity linking – wykorzystanie mechanizmu ujednoznaczniania wyrażień do rozwiązania problemu NER
- ⌘ (D)RNN – wykorzystanie sieci neuronowych do modelowania odległych zależności tekstowych



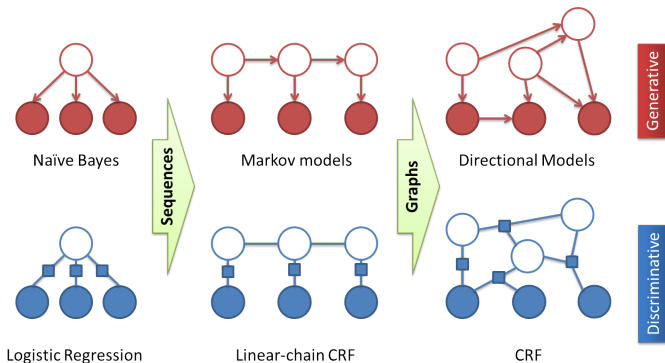
⌘ (linear chain) Conditional Random Fields – CRF

$$p(\mathbf{y}|\mathbf{x}) = \frac{1}{Z(\mathbf{x})} \prod_{C_p \in \mathcal{C}} \prod_{\Psi_c \in \mathcal{C}_p} \Psi_c(\mathbf{x}_c, \mathbf{y}_c; \theta)$$

$$\Psi_c(\mathbf{x}_c, \mathbf{y}_c; \theta_p) = \exp \left\{ \sum_{k=1}^{K(p)} \lambda_{pk} f_{pk}(\mathbf{x}_c, \mathbf{y}_c) \right\}$$

$$Z(\mathbf{x}) = \sum_{\mathbf{y}} \prod_{C_p \in \mathcal{C}} \prod_{\Psi_c \in \mathcal{C}_p} \Psi_c(\mathbf{x}_c, \mathbf{y}_c; \theta)$$

<https://www.codeproject.com/Articles/559535/Sequence-Classifiers-in-Csharp-Part-II-Hidden-Cond>



Adapted from C. Sutton, A. McCallum, "An Introduction to Conditional Random Fields", ArXiv, November 2010

<https://www.codeproject.com/Articles/559535/Sequence-Classifiers-in-Csharp-Part-II-Hidden-Cond>

✦ Wykorzystanie Wikipedii jako słownika nazw własnych

Artykuł [Dyskusja](#) Czytaj [Tekst źródłowy](#) [Historia i autorzy](#) ☆


✕
 Wiki Lubi Przyrodę - odkrywaj piękno natury, rób zdjęcia dla Wikipedii i wygrywasz nagrody!

Polska [edytuj] 📍 Na mapach: [52°N 19°E](#) [\(mapa\)](#)

 Zobacz też: [inne znaczenia](#). Na tę stronę wskazuje także [przekierowanie „RP”](#). Zobacz też: [RP](#) ([ujednoznacznienie](#))

Polska, **Rzeczpospolita Polska** (**RP**) - państwo unitarne w Europie Środkowej położone między Morzem Bałtyckim na północy a Sudetami i Karpatami na południu, w przeważającej części w dorzeczu Wisły i Odry. Powierzchnia administracyjna Polski wynosi 312 679 km²^[R1], co daje jej 70. miejsce na świecie i 9. w Europie. Zamieszkała przez prawie 38,5 miliona ludzi (2014), zajmuje pod względem liczby ludności 34. miejsce na świecie^[5], a 6. w Unii Europejskiej.

Od północy Polska graniczy z Rosją (z jej obwodem kalininingradzkim) i Litwą, od wschodu z Białorusią i Ukrainą, od południa ze Słowacją i Czechami, od zachodu z Niemcami. Większość północnej granicy Polski wyznacza wybrzeże Morza Bałtyckiego. Polska wyłączna strefa ekonomiczna na Bałtyku graniczy ze strefami Danii i Szwecji.

Rzeczpospolita Polska




Kontrola autorytatywna (obiekt geograficzny): [VIAF: 141810140](#) [|](#) [LCCN: n79131071](#) [|](#) [GND: 4046496-9](#) [|](#) [NDL: 00569130](#) [|](#) [BnF: 11880842g](#) [|](#) [SUDOC: 02658994X](#) [|](#) [WorldCat](#)

Kategorie: [Polska](#) [|](#) [Członkowie Organizacji Narodów Zjednoczonych](#) [|](#) [Państwa członkowskie Unii Europejskiej](#) [|](#) [Państwa należące do NATO](#)

Figure: Hasło *Polska* w polskiej Wikipedii.



Pokrewieństwo semantyczne dwóch haseł w Wikipedii:

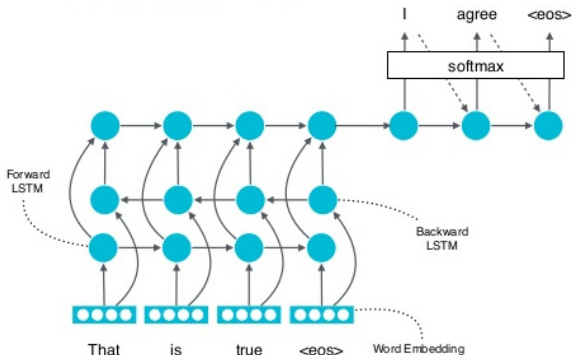
$$SR_J(\sigma_a, \sigma_b) = \begin{cases} \frac{1}{1 - \log\left(\frac{|A \cap B|}{|A \cup B|}\right)} & |A \cap B| > 0 \\ 0 & |A \cap B| = 0 \wedge a \neq b \\ 1 & |A \cap B| = 0 \wedge a = b \end{cases} \quad (1)$$

Cechy ujednoznaczniające:

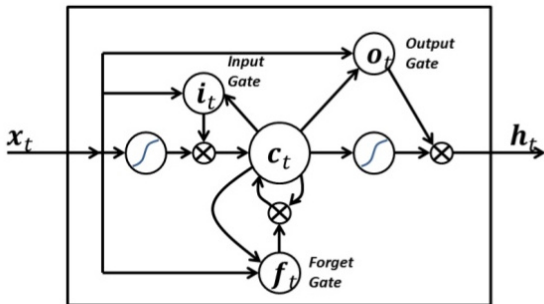
Hasło	\overline{SR}_w	P_{sense}	G	R_{SR}	R_{sense}	P_{link}	+/-
Burowie	0,32	0,93	84	0	0	0,18	+
Afrykanerzy	0,18	0,04	84	1	1	0,18	-
Burowo	0,01	0,03	84	2	1	0,18	-



- ✠ \overline{SR}_w – ważone pokrewieństwo semantyczne
- ✠ P_{sense} – prawdopodobieństwo sensu
- ✠ G – „gęstość” kontekstu semantycznego
- ✠ R_{SR} – ranga pokrewieństwa semantycznego
- ✠ R_{sense} – ranga sensu
- ✠ P_{link} – prawdopodobieństwo występowania jako odnośnik w Wikipedii
- ✠ $+/-$ – przykład pozytywny/negatywny



<https://www.slideshare.net/emorynlp/rnn-lstm-and-seq2seq-models>



<https://www.slideshare.net/eefjeopdenbuysch/machine-learning-for-robot-journalism-59993401>



Celem zadania jest utworzenie indeksu nazw osobowych i miejscowych dla pliku potop.txt.

- 1 zapoznać się ze schematami klasyfikacyjnymi oraz formatami wyjściowymi narzędzia Liner2, wybrać schemat oraz format adekwatny dla zadania, przetworzyć próbkę tekstu w oparciu o Linera (1 pkt)
- 2 stworzyć indeks nazw osobowych i miejscowych dla całego pliku potop.txt oraz narzędzie pozwalające na wyświetlenie wszystkich wystąpień określonej nazwy wraz z kontekstem (obejmującym stałą liczbę linii tekstu) wystąpienia (1 pkt)
- 3 obliczyć statystykę występowania poszczególnych nazw osobowych i miejscowych; znaleźć 10 najczęstszych nazw osobowych i 10 najczęstszych nazw miejscowych (1 pkt)

- ✘ <http://apohllo.pl/texts/lab3.tar.gz> (plik potop.txt)
- ✘ Clarin WS <http://nlp.pwr.wroc.pl/redmine/projects/nlprest2/wiki/Liner2>
- ✘ Smywiński-Pohl A. (2015). *Automatyczna ekstrakcja relacji semantycznych z tekstów w języku polskimi* (praca doktorska).
- ✘ Pohl A. (2013). *Knowledge-based Named Entity Recognition in Polish*
- ✘ Pohl A. (2012). *Improving the Wikipedia Miner Word Sense Disambiguation Algorithm.*