Przetwarzanie Języka Naturalnego
Lab 8 – NER

Aleksander Smywiński-Pohl

Wydział IEiT
Katedra Informatyki

11.04.2017

Rozpoznawanie jednostek referencyjnych (nazewniczych)

- Named Entity Recognition – NER

Rozpoznawanie jednostek referencyjnych[1] *polega na określeniu, które spośród wyrażeń 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*

[1] W polskiej literaturze funkcjonuje również termin *rozpoznawanie jednostek nazewniczy.*

Przykład

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.

Popularne metody

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

Warunkowe pola losowe

- (linear chain) Conditional Random Fields – CRF

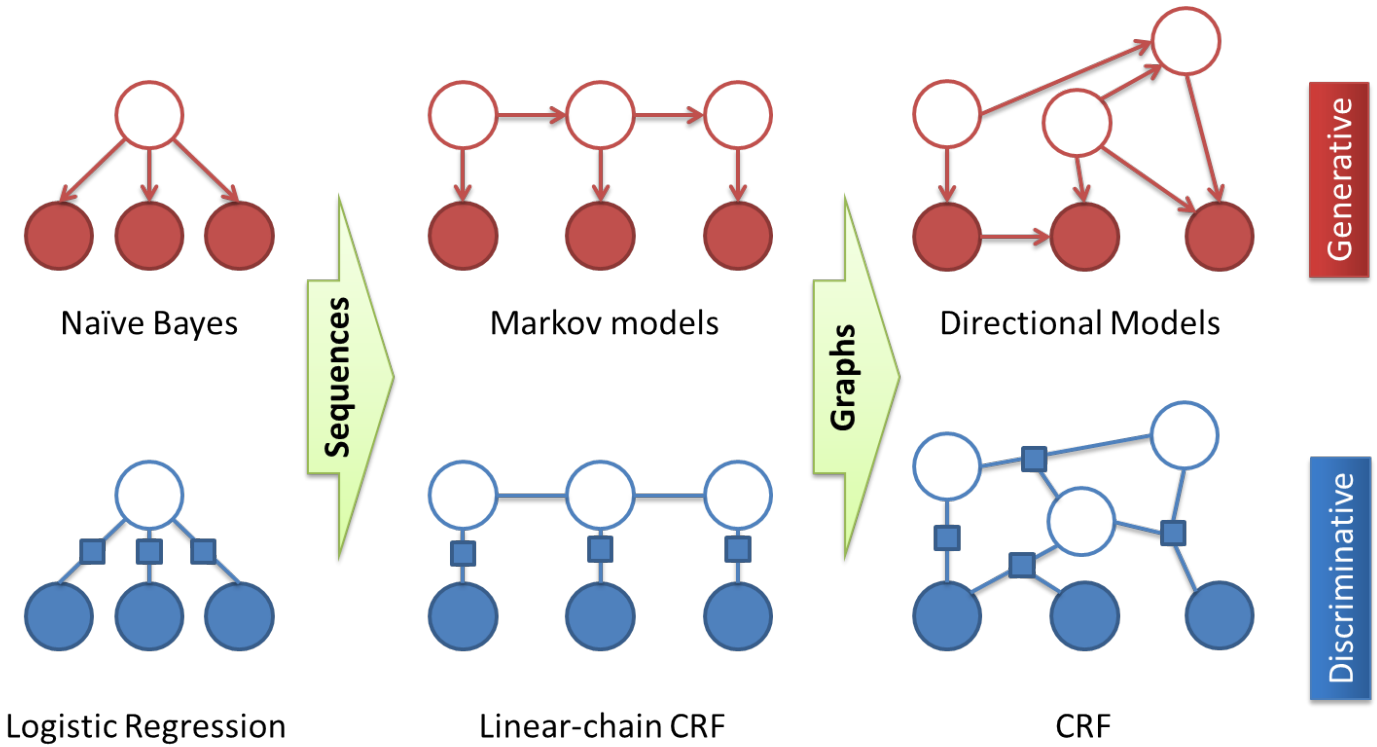
$$p(\boldsymbol{y}|\boldsymbol{x}) = \frac{1}{Z(\boldsymbol{x})} \prod_{C_p \in \mathcal{C}} \prod_{\Psi_c \in C_p} \Psi_c(\boldsymbol{x_c}, \boldsymbol{y_c};\ \theta)$$

$$\Psi_c(\boldsymbol{x_c}, \boldsymbol{y_c};\ \theta_p) = exp\left\{ \sum_{k=1}^{K(p)} \lambda_{pk} f_{pk}\ (\boldsymbol{x_c}, \boldsymbol{y_c}) \right\}$$

$$Z(\boldsymbol{x}) = \sum_{\boldsymbol{y}} \prod_{C_p \in \mathcal{C}} \prod_{\Psi_c \in C_p} \Psi_c(\boldsymbol{x_c}, \boldsymbol{y_c};\ \theta)$$

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

Modele generatywne i dyskryminacyjne



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

Entity linking

- Wykorzystanie Wikipedii jako słownika nazw własnych

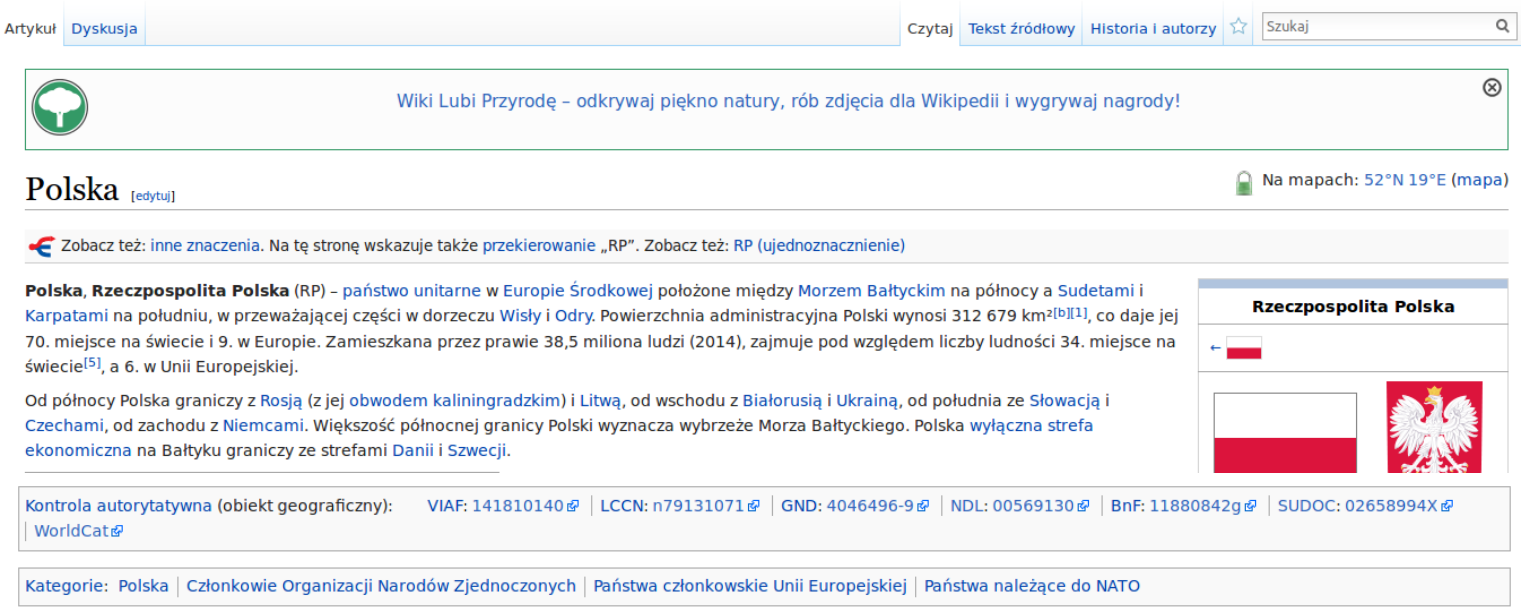


Figure: Hasło *Polska* w poslkiej Wikipedii.

Entity linking – cd.

Pokrewieństwo semantyczne dwóch haseł w Wikipedii:

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

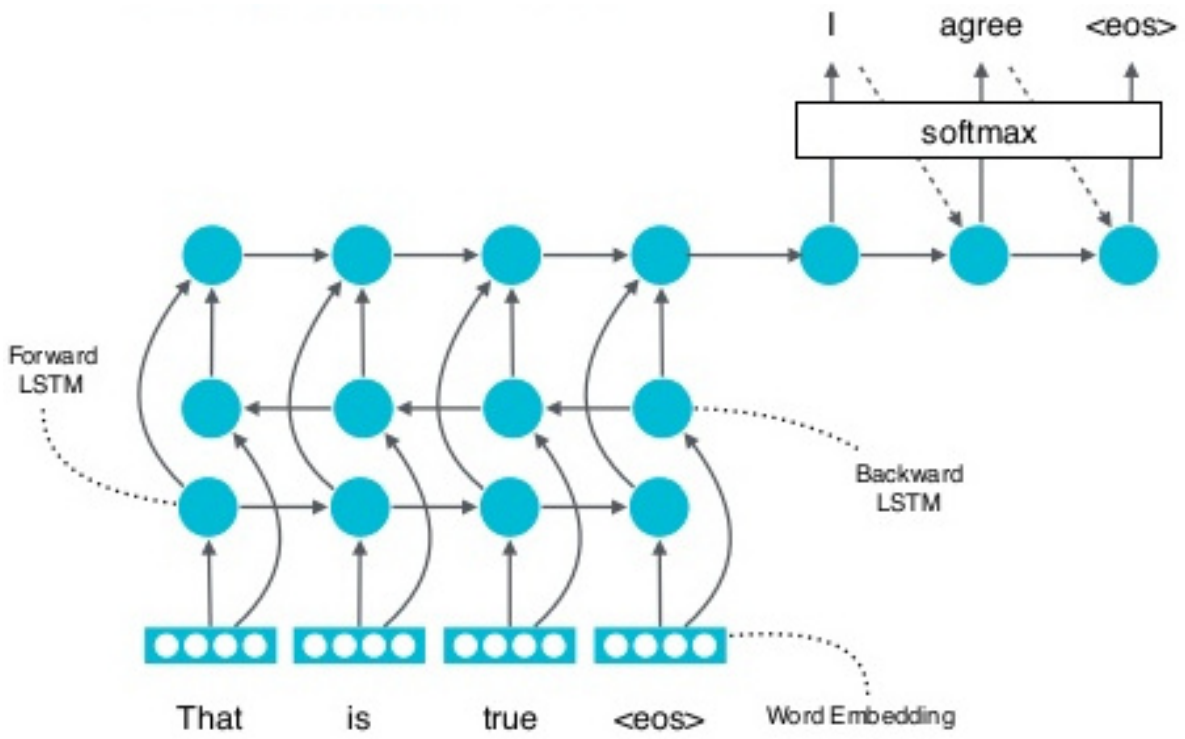
Cechy ujednoznaczniające:

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	−

Entity linking – cechy ujednoznaczniające

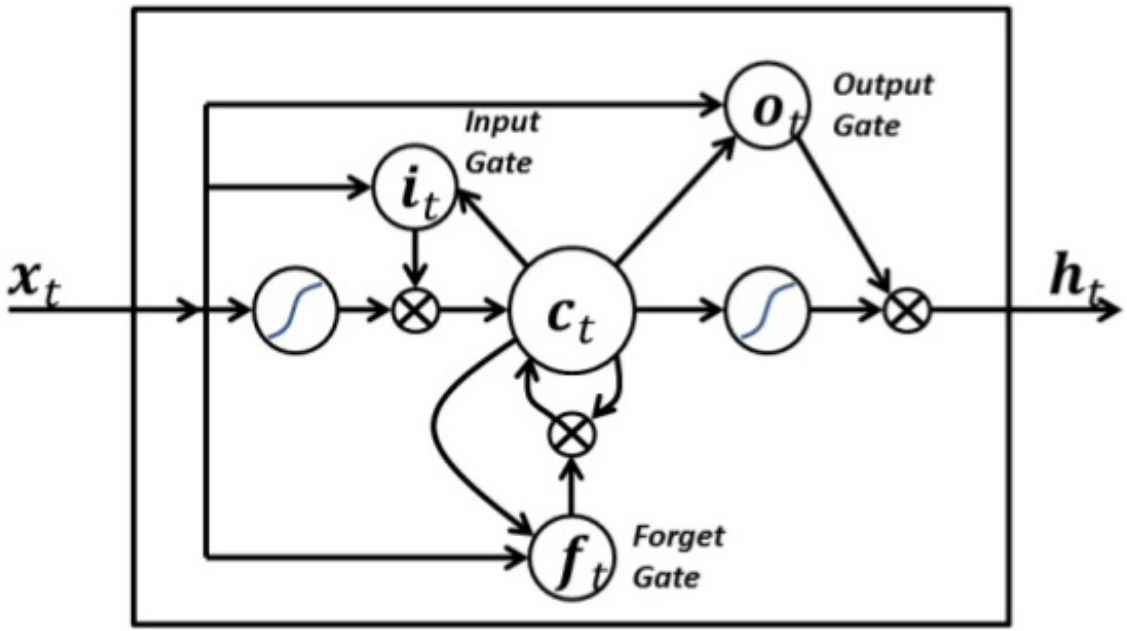
- $\overline{SR}_w$ – ważone pokrewieństwo semantyczne
- P_{sense} – prawdopodobieństwo sensu
- G – „gęstość” kontekstu semantycznego
- R_{SR} – ranga pokrewieństwa semantycznego
- R_{sense} – ranga sensu
- P_{link} – prawdopodobieństwo występowania jako odnośnik w Wikipedii
- $+/-$ – przykład pozytywny/negatywny

Rekurencyjne sieci neuronowe – seq2seq



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

Rekurencyjne sieci neuronowe – LSTM



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

Zadania

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 pozwlające 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)

Materiały

- `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 doktroska).
- Pohl A. (2013). *Knowledge-based Named Entity Recognition in Polish*
- Pohl A. (2012). *Improving the Wikipedia Miner Word Sense Disambiguation Algorithm.*