Semi-Automatic Construction of Polish DeriNet

Mateusz Lango

Institute of Computing Science, Poznan University of Technology

September 18, 2017

Poznań, the capital of Greater Poland

Poznań University of Technolgy

- the second university with the highest number of candidates

Resources for derivational morphology

- DeriNet for Czech – a network of $> 1M$ lemmas which are connected by derivational relation
- lack of such resources for Polish[1] (and many other languages)

[1] some information about derivation can be extracted from the Polish WordNet

Overview of the pipeline

1. Generation of frequent subsequences
2. Merging frequent subsequences into regular expressions
3. Generation of possible parents for each lemma
4. Ranking of candidate sets by machine-learned ranker

Sequential pattern mining

- one of the most important topics in frequent pattern mining
- the task is to extract all frequent subsequences with the support greater than a specified threshold
- in our case we treat lexicon as a database of sequences (words)
- we used SPADE algorithm with min. support 1% ⇒ 27K frequent patterns

Pattern	Support
n,i,e	87053
o,w,y	27099
c, z, n, o, ś, ć	7570
d, z, o, ś, ć	4792

Converting frequent patterns into regular expressions

- frequent pattern „n,i,e” ⇒ `^*n*i*e*$`
- making expressions more specific
  - delete one of the * from the expression
  - recalculate support
  - accept new expression if the support is higher than 95% of the original support
- `^*n*i*e*$` ⇒ `^nie*$`

Pattern	RegExp
n,i,e	`^nie*$`
o,w,y	`^*owy$`
c, z, n, o, ś, ć	`^*cz*nośc̈$`
d, z, o, ś, ć	`^*d*z*ośc̈$`

Filtering frequent pattens

- Problem: some regular expressions are redundant (they cover almost the same set of words)

RegExp	Support
`^*z*nośćŚ$`	7547
`^*cz*nośćŚ$`	7543

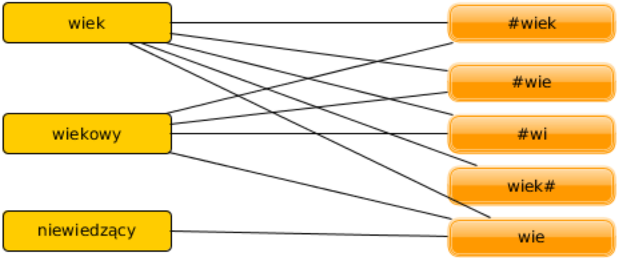
- Solution:
  - convert each regular expression to a binary feature
  - calculate phi coefficient between corresponding features
  - if ϕ is greater than 95% drop less specific regular expressions
- 27K regular expressions ⇒ 13K regular expressions

Pairwise classification

- each regular expression is used as a binary feature
- two more features: length of the common prefix and length of the common suffix
- hand-annotated training set of (derived word, base word) pairs
- the classification task: is the pair a correct one?

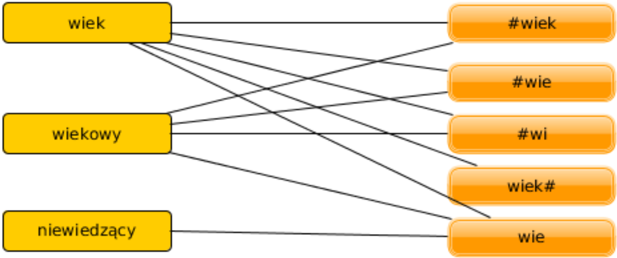
Proxinette measure

1. add a special character at the beginning and at the end of each word e.g. `#wiek#`
2. split the word into all possible substrings of length > 3 (`#wiek`, `#wie`, `#wi`, `wiek#`, `wiek`, `wie`,..)
3. create a bi-partile graph in which the words are connected to its substrings

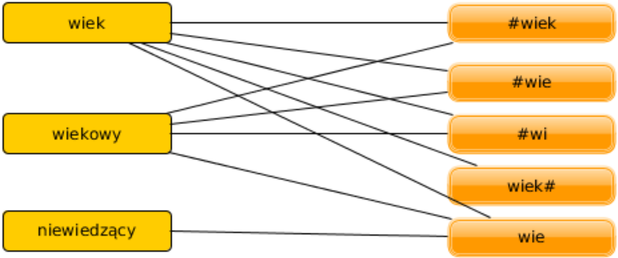


Proxinette measure

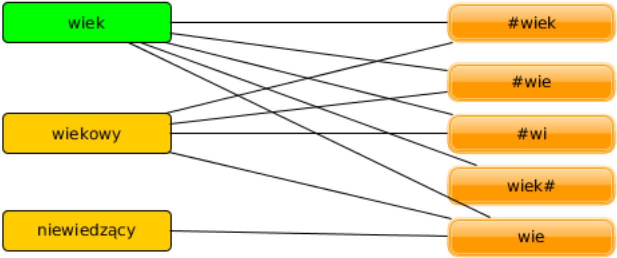
1. a weight is added to each edge which is equal to $\frac{1}{d}$ where d is the degree of the node



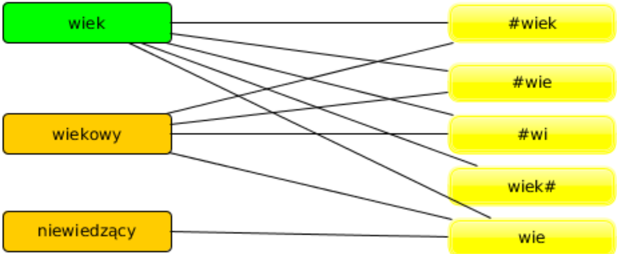
Proxinette measure



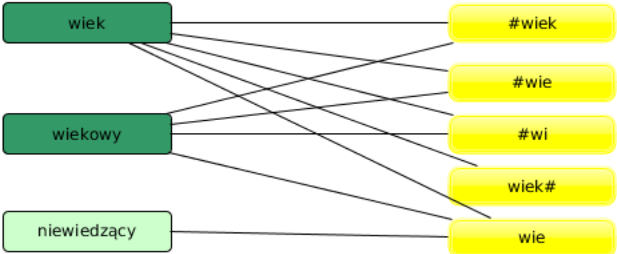
Proxinette measure



Proxinette measure



Proxinette measure



A new problem setup

- for each lemma we construct a candidate set from 100 most similar lemmas
- the problem has change: pick one (or none) from the set of candidates
- Learning to rank
    - originally proposed for ranking query results in the information retrieval systems
    - many approaches: pointwise, pairwise, listwise

Overview of the pipeline

1. Generation of frequent subsequences
2. Merging frequent subsequences into regular expressions
3. Generation of possible parents for each lemma
4. Ranking of candidate sets by machine-learned ranker

Experimental setup

- language resources
  - Morfeusz SGJP - Polish lexicon
  - Słowosieć - Polish Wordnet
- software
  - SPMF data mining library for frequent sequence mining
  - xgboost - implementation of Gradient Boosting Trees (supports learning-to-rank)
- 5-fold CV

Results

	Classification	Ranking
Precision@1	80,75%	82,33%
Avg position of correct candidate	0.64	0.49
Precision@1 with threshold	88,3%	98,8%

Polish DeriNet

- approx. 53,5 K connections were established
- 97% from 200 randomly sampled connections were correct
- we extracted 12 types of relations related with derivation from Słowosieć (the Polish WordNet) e.g. diminutives, femininity, inhabitant, derivationality
- by applying these connections to our lexicon 52K connections can be created
- finally, there is above 93,5 K connections in the network

Current work

- analysis of the inconsistencies between WordNet connections and our connections
- translation of the Czech DeriNet to Polish
- creation of a similar network for Spanish
- comparison of the structures of word-formation networks for Czech and Latin
- ...

Thank you for your attention!