UNIVERSITAS SCIENTIARUM SZEGEDIENSIS

Research Group on Artificial Intelligence

Wikipedia-based methods to identify noun compounds in running texts

István Nagy T.
nistvan@inf.u-szeged.hu

University of Szeged, Hungary

Multiword Expressions

- proper treatment of multiword expressions (MWEs) is essential
  - MWEs are lexical items that contain space
- subtype: Noun compounds (NCs)
  - A compound is a lexical unit that consists of two or more elements that exist on their own
- frequent in language
  - 67.3% of the sentences contain NC

Corpora used for evaluation

Corpus	Sentence	Token	NC	2	3	4≤
Wiki50	4,350	114,570	2929	2442	386	101
BNC dataset	1,000	21,631	485	436	40	9

WP based method for detecting NCs

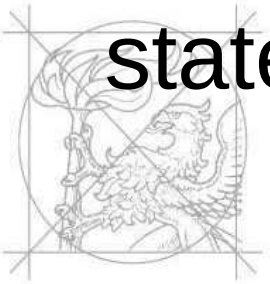
- lowercase n-grams from English Wikipedia links were collected
- three methods:
  - marked as a noun compound if it occurred in the list.
  - merge of two possible noun compounds:
    - if A B and B C both occurred in the list, A B C was also accepted as a noun compound
  - it occurred in the list and its Part of Speech (POS)-tag sequence matched one of the previously defined patterns (e.g. JJ

Results on Wiki50

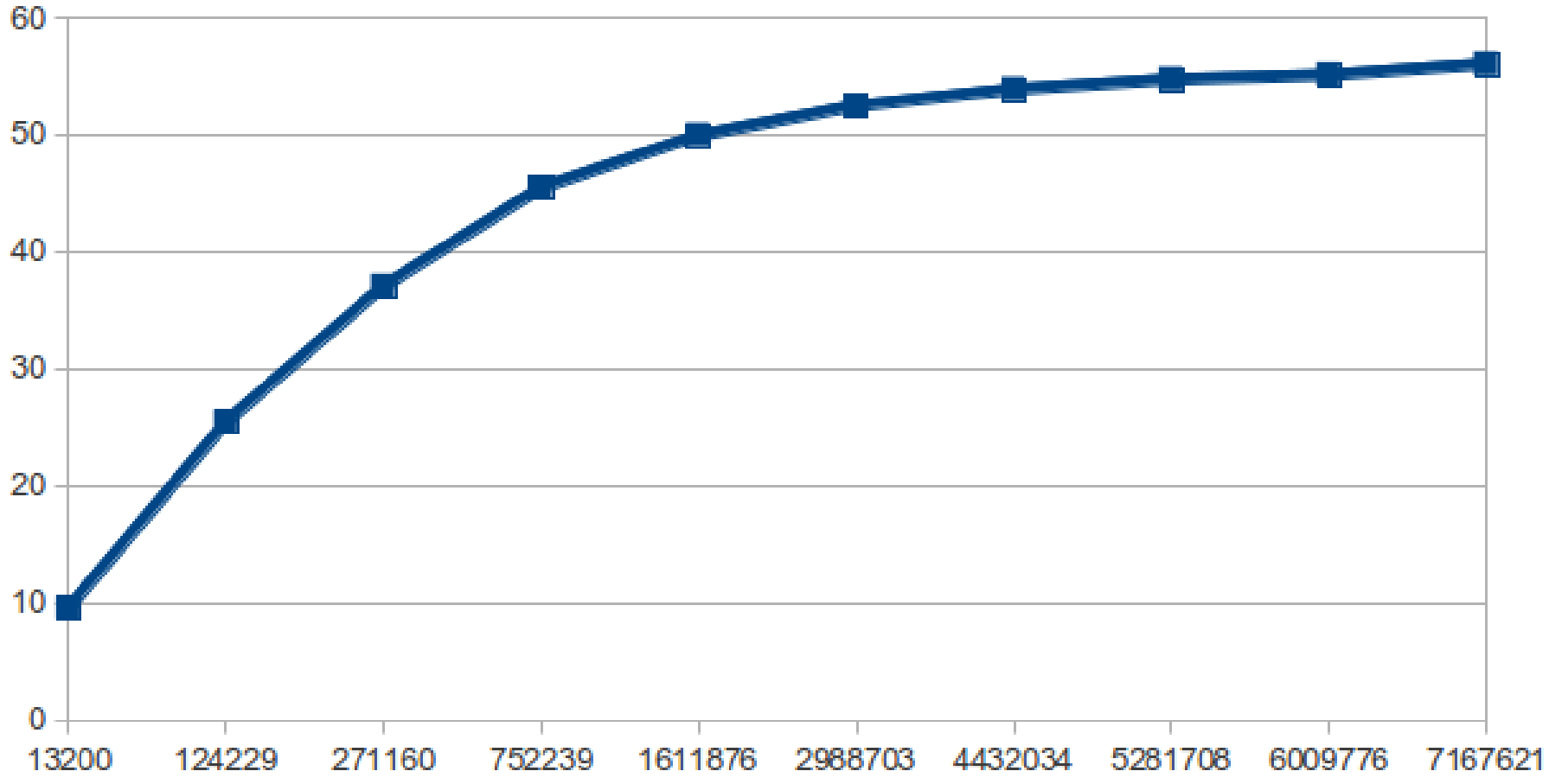
Method	Precision	Recall	F-Score
Match	37,7	54,73	44,65
Merge	40,06	57,63	47,26
POS-rules	55,56	49,98	52,62
Combined	62,66	50,69	56,04

Dictionary based method

- we investigated how the size of Wikipedia influences the results
- NC list from the actual Wikipedia status of the beginning of each year was collected
- The English Wikipedia was launched in 2001 → the first list was collected from the state of 1 January 2002.



Results of the expansion of the number of Wikipedia pages.



Machine Learning approaches

- first-order linear chain Conditional Random Fields (CRF) classifier
- basic feature set was extended with noun compound specific features.
  - Noun compound lists were added to the dictionaries.
  - The shallow linguistic features were extended with the POS-rules
  - the other entities were also specified in the sentence

Machine Learning results

- leave-one-document-out scheme on Wiki50 with 68,16 F-score
- automatically generated training database was also used
  - the training set consisted of randomly selected Wikipedia pages
  - documents were not manually annotated
  - dictionary based NC labeling was considered as the gold standard

Machine Learning results

	Recall	Precision	F-Score
LOO Wiki50	64,39	72,40	68,16
WikiTrain Wiki50	56,57	55,57	56,06
Dict. Wiki50	50,70	62,66	56,05
WikiTrain BNC	38,02	41,53	39,70
Dict BNC.	31,40	40,75	35,47

Conclusions

- dictionary vs machine learning based methods
  - heavily relied on Wikipedia
- examined the results depending on the expansion of WP over the years
  - the growth of Wikipedia can improve the results
  - the rate of improvement is reduced with time

Acknowledgement

The presentation is supported by the European Union and co-funded by the European Social Fund.

Project title: "Broadening the knowledge base and supporting the long term professional sustainability of the Research University Centre of Excellence at the University of Szeged by ensuring the rising generation of excellent scientists".

Project number: **TÁMOP-4.2.2/B-10/1-2010-0012**

NEW SZÉCHENYI PLAN

HUNGARY'S RENEWAL

The project is supported by the European Union and co-financed by the European Social Fund.

Thank you for your attention!