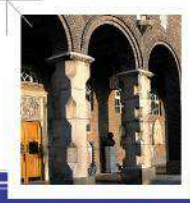
UNIVERSITAS SCIENTIARUM SZEGEDIENSIS
SZEGEDI TUDOMÁNYEGYETEM

Computing semantic similarity using large static corpora



SOFSEM 2013

András Dobó, János Csirik

Introduction

- Semantic similarity of words is crucial for many Natural Language Processing tasks
  - Information extraction
  - Spelling correction
  - Word sense disambiguation
- There are numerous existing methods for this task, but there is still room for improvement

Literature review

- Three main categories of methods
  - Using the distance of words and their gloss found in large lexical databases
  - Issuing web queries with words and processing the returned page hit counts and text snippets
  - Creating feature vectors for words based on information gathered from large static corpora, and then comparing these vectors
- Numerous researches tried to combine different types of methods in order to combine the best properties of them

Literature review

- Making use of web queries and large lexical databases have many drawbacks
  - Web queries
    - The results are estimates, they change over time,
    - Queries can have no linguistic restrictions and punctuation cannot be used,
    - Their use can be limited, time consuming, etc.
  - Lexical databases
    - Do not contain uncommon words,
    - Manually created => errors, missing entities etc.

Our methods

- Determining the part-of-speech (POS) of the test words
- Creating feature vectors based on information extracted from corpora
- Comparing words using vector similarity measurements
- Combining different methods
- Testing on two benchmark dataset

Determining the part-of-speech of the test words

- Many words can take more than one POS
- For words with different POS different features are relevant
- Assumption: each question contains words of the same POS (verb, noun, adjective or adverb)
- For a question the POS maximizing the following formula is chosen:

$$\underset{pos}{\operatorname{argmax}} \prod_{w} \ln\left(1.0001 + f_{w,pos}\right)$$

Feature extraction

- Two main approaches
  - Bag-of-words model
    - Features are the words found in a 3-word window
    - Distance-based weighting
  - Usage of grammatical relations
    - Extracting grammatical relations with the help of the C&C CCG parser
    - Features are the (grammatical relation, feature word) pairs, e.g. (subject-of, make)
- Information extracted from 3 corpora
  - BNC, Web 1T 5-gram, English Wikipedia

Creating and comparing the feature vectors

- Two main techniques
  - Modified version of the approach presented by Lin (1998)
    - Binary vectors, no weighting
    - Similarity measure defined by Lin (1998)

$$\mathrm{sim}(w_1, w_2) = \frac{2 \times I(F(w_1) \cap F(w_2))}{I(F(w_1)) + I(F(w_2))}$$

  - Usage of numerical feature vectors
    - Numerical vectors with different weighting
    - Two frequently used vector similarity measures

Weighting inside the numerical feature vectors

- Frequency (freq)
- Logarithm of the frequency (logfreq)
- Pointwise mutual information (pmi)
- Log-likelihood ratio (loglh)
- $qw(x,y) = \frac{\ln(1+c_{xy})}{\ln(1+f(y))}$
- $pw(x,y) = \ln(1+c_{xy}) \times \ln(1+I(y))$
- Rapp (2003): $A_{ij} = \log(1+f_{ij}) \cdot \left(-\sum_k p_{kj} \log(p_{kj}) \right)$

Similarity measures for the numerical feature vectors

- Cosine similarity

$$\cos(\vec{x},\vec{y}) = \frac{\vec{x}\cdot\vec{y}}{|\vec{x}||\vec{y}|} = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2}\sqrt{\sum_{i=1}^{n} y_i^2}}$$

- Dice coefficient
  - Originally only for binary vectors
  - Generalization proposed by Lin (1998)

$$\mathrm{Dice}(\vec{x},\vec{y}) = \frac{2\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2 + \sum_{i=1}^{n} y_i^2}$$

Combination of the individual methods

- In order to improve the results, the combination of the individual methods were also tested
- Combined similarity of a word pair is calculated from the similarities returned by the individual methods

$$Score_{comb} = \ln(1 + Score_1) * \ln(1 + Score_2)$$

Testing

- Testing on two benchmark dataset
  - Miller-Charles word pairs (MC)
    - 28 word pairs
    - For each pair a similarity score between 0 to 4 was assigned by 38 undergraduate students
    - Evaluation: Spearman correlation with the avg. scores
  - TOEFL synonym questions
    - 80 questions
    - In each question a test word is given with 4 alternatives, the task is to determine the most similar word to the test word
    - Evaluation: the percentage of the correct answers

Results on the MC dataset

Method	Result	Used data
Human upper bound (Resnik, 1995)	0.934	
Agirre et al. (2009)	0.92	WordNet, corpus
Patwardhan and Pedersen (2006)	0.91	WordNet
Jarmasz and Szpakowicz (2003)	0.87	Roget's Thesaurus
Lin (1998)	0.82	WordNet, corpus
bnc-bagofwords-num-cos-qw+ enwiki-parsed-num-cos-freq	**0.773**	**corpus**
bnc-bagofwords-num-cos-qw+ enwiki-parsed-num-cos-qw	**0.750**	**corpus**
Gabrilovich and Markovitch (2007)	0.72	corpus
Milne and Witten (2008)	0.70	Wikipedia links, Web search
Sahami and Heilman (2006)	0.618	Web search

Results on the TOEFL questions

Method	Result	Used data
Turney et al. (2003)	97.5%	Web search, thesaurus
Rapp (2003)	92.5%	corpus
bnc-parsed-num-cos-loglh+ enwiki-parsed-num-cos-pmi	**88.75%**	**corpus**
enwiki-bagofwords-num-cos-pmi+ enwiki-parsed-num-cos-pmi	**87.50%**	**corpus**
Tsatsaronis et al. (2010)	87.5%	WordNet
enwiki-bagofwords-num-cos-pmi	**83.75%**	**corpus**
Higgins (2005)	81.3%	Web search
Average non-English US college applicant (Landauer and Dumais, 1997)	64.5%	
Landauer and Dumais (1997)	64.3%	corpus
Lin (1998)	24.0%	WordNet, corpus

Our best methods

- Methods performing best considering both test datasets
  - *bnc-parsed-num-cos-loglh+enwiki-parsed-num-cos-pmi* (MC: 0.712, TOEFL: 88.75%)
  - the *enwiki-bagofwords-num-cos-pmi+enwiki-parsed-num-cos-pmi* (MC: 0.729, TOEFL: 87.50%)
  - *enwiki-bagofwords-num-cos-pmi+bnc-parsed-num-cos-qw* (MC: 0.737, TOEFL: 86.25%)

Summary of the results

- Considering all methods
  - Average performance on the MC dataset
  - Third best performance on the TOEFL questions
- Taking only those methods into account that solely rely on static corpora
  - Best performance on the MC dataset
  - Second best performance on the TOEFL questions

Future work

- Testing with even larger corpora
  - E.g. automatically gathered web corpus
- Combination with other types of methods
- Adaptation to languages other than English
  - E.g. Hungarian

Thank you!

The publication/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

National Development Agency
www.ujszechenyiterv.gov.hu
06 40 638 638

HUNGARY'S RENEWAL

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

ÚJ SZÉCHENYI TERV