Computational Lexical Semantics

An Enhanced Lesk Word Sense Disambiguation algorithm through a Distributional Semantic Model

Basile et al., COLING 2014

Giulia Bonansinga
Division of Linguistics and Multilingual Studies

30-10-2014

Outline

- Lesk Algorithm and its variations
  - Simplified Lesk (Kilgarriff and Rosenzweig, 2000)
  - Adapted Lesk (Banerjee and Pedersen, 2002)
- A new approach: Lesk meets Distributional Hypothesis
- SemEval-2013 Multilingual Word Sense Disambiguation
  - Evaluation and comparison with other participants

Knowledge-based vs corpus-based approaches

- Lesk belongs to the knowledge-based approaches
- Knowledge-based methods do not perform as well as their corpus-based alternatives, but have usually larger coverage
  - They are applicable to all words in a text, while corpus-based techniques suit tasks for which a sufficient amount of annotated text is available

Lesk Algorithm

- Given two words, the algorithm selects those senses whose definitions have the maximum overlap, i.e. the highest number of common words in the definition of the senses
- Requires
  - a dictionary, with as many entries as possible meanings for each target word
    - Oxford Advanced Learner's dictionary
  - contextual information

Criticism

- Complexity
  - The number of comparisons increases combinatorially with the number of words in a text
- Definition expressiveness
  - The overlap is based only on word co-occurrences in glosses

Simplified Lesk Algorithm

- Kilgarriff and Rosenzweig, 2000
- It disambiguates one word at a time, regardless of the meaning of other words in context

(1) for each sense *i* of *W*
(2) determine *Overlap(i)*, the number of words in common between the definition of sense *i* and current sentential context
(3) find sense *i* for which *Overlap(i)* is maximized
(4) assign sense *i* to *W*

- It significantly outperforms the original Lesk algorithm (see Vasilescu et al., 2004)

Adapted Lesk Algorithm

- Banerjee and Pedersen, 2002
- It exploits relations among meanings
  - each gloss is extended by the definitions of semantically related meanings
- WordNet is adopted as semantic network and several relations are taken into account
- It outperformed plain Lesk in disambiguating nouns in SensEval-2 English task

Motivation for a new approach

- Graph approaches can disambiguate all words in a sequence at once
  - Glosses are not taken into account
- But glosses *are* descriptive of the meaning of a word!
- Even Adapted Lesk is very sensitive to the exact wording of definitions
  - The absence of a certain word can radically change the results

Distributional Semantic Spaces meet Lesk

- Instead of overlap, similarity computed on a Distributional Semantic Space (DSS) is used
- In this representation, meanings are vectors that encapsulate information about all co-occurring context words
  - grounded in Distributional Hypothesis
  - suitable for computing the overlap when no exact word matching can occur

Similarity function

- We need to compare the similarity between glosses and contexts
- Given the words g_1, g_2, ..., g_n in the gloss and the contextual words c_1, c_2, ..., c_m, their vector representations $\mathbf{g}$ and $\mathbf{c}$ are so built:

$$\mathbf{g} = \mathbf{g_1} + \mathbf{g_2} + \ldots + \mathbf{g_n}$$
$$\mathbf{c} = \mathbf{c_1} + \mathbf{c_2} \ldots + \mathbf{c_m}$$

- The cosine similarity between $\mathbf{g}$ and $\mathbf{c}$ is the score associated to the candidate meaning

Methodology

- Strenghts of Simplified and Adapted Lesk combined
- Disambiguation of one word at a time
- The sense whose gloss has the highest similarity to the context is selected
  - Different context window sizes are considered

BabelNet

- Very large multilingual semantic network built exploiting both WordNet and Wikipedia
  - Linguistic knowledge and encyclopedic concepts
  - Glosses are richer
  - Robustness for named entities
- Approach inherently multilingual and suitable for tasks such as named entity disambiguation

Algorithm

1. For each word w_i, retrieve its BabelNet synsets, first looking at WordNet
2. Build the context selecting the *l* words to the left and to the right of w_i
3. For each sense s_{ij} of w_i, expand the gloss g_{ij} to build the extended gloss g^*_{ij}
   – Using term scoring for each term in g^*_{ij}
4. Build semantic vector for each gloss g^*_{ij}
5. For each gloss g^*_{ij}, compute similarity with context c
   - Optionally, use sense distribution $p(s_{ij} \mid w_i)$ in linear combination with similarity

Gloss term scoring

- Words from the glosses of related synsets are added to the extended gloss
- Each word is weighted by a factor inversely proportional to the distance in the graph between s_{ij} and the related synsets to reflect their different origin

$$inverse\,distance = \frac{1}{1+d}$$

- To weigh more senses associated with a few words, they define the inverse gloss frequency (IGF)

$$IGF_k = 1 + log_2 \frac{|S_i|}{gf_k^*}$$

- Finally, the weight for the word w_k appearing h times in the $g*_{ij}$ is

$$weight(w_k, g_{ij}^*) = h \times IGF_k \times \frac{1}{1+d}$$

Combining sense distribution

- They run the algorithm also exploiting information on sense frequency from WordNet, based on SemCor
- They compute, for each pair <w_i, s_{ij}>, the probability that w_i is tagged with s_{ij}

$$p(s_{ij}|w_i) = \frac{t(w_i, s_{ij}) + 1}{\#w_i + |S_i|}$$

$t(w_i, s_{ij})$ = number of times the word w_i is tagged with s_{ij}

Si = number of senses of w_i

$\#w_i$ = the number of occurrences of w_i in SemCor

Getting started

- Completely developed in JAVA using BabelNet API 1.1.1
  - Software available under GNU General Public License v. 3
    - https://github.com/pippokill/lesk-wsd-dsm
- Preprocessing: tokenization with Lucene and stemming with Snowball
- The Semantic spaces are built relying on two Lucene indexes, which contain documents from British National Corpus (BNC) for English, and from Wikipedia dump for Italian
- For each language, the co-occurrences matrix M considers the 100,000 most frequent words in the corpus
  - M is reduced by Latent Semantic Analysis using the SVDLIBC tool
  - Dimension reduction is set to 200
- The algorithm uses the result of the SVD composition

Summing up

- Knowedge-based algorithm with DSM
- Language independent except for stemming and training corpus
- The gloss-context overlap is computed by using a word similarity function defined on a distributional semantic space

Evaluation

- Dataset provided for the Multilingual WSD “all-words” Task-12 of SemEval-2013
  - Systems are expected to assign the correct BabelNet synset to all occurrences of noun phrases within texts in different languages
  - Parameters:

    1) the context size (3, 5, 10, 20 and the whole text);

    2) the use of information about sense distribution

    3) the gloss term scoring function is always applied, since it provides better results.
- Simplified Lesk was implemented for comparison

  - count the common words between each g^*_{ij} and the context c, applying stemming to maximize the overlap

English evaluation

Run	*ContextSize*	*SenseDistr.*	*P*	*R*	*F*	*A*
MFS	-	-	0.656	0.656	0.656	100%
EN.LESK.1	3	N	0.525	0.525	0.525	100%
EN.LESK.6	3	Y	0.633	0.633	0.633	100%
EN.DSM.1	3	N	0.536	0.536	0.536	100%
EN.DSM.2	5	N	0.605	0.605	0.605	100%
EN.DSM.3	10	N	0.633	0.633	0.633	100%
EN.DSM.4	20	N	0.650	0.650	0.650	100%
EN.DSM.5	W	N	0.687	0.687	**0.687**	100%
EN.DSM.6	3	Y	0.669	0.669	0.669	100%
EN.DSM.7	5	Y	0.677	0.677	0.677	100%
EN.DSM.8	10	Y	0.689	0.689	0.689	100%
EN.DSM.9	20	Y	0.696	0.696	0.696	100%
EN.DSM.10	W	Y	0.715	0.715	**0.715**	100%

Table borrowed from Basile et al. (2014)

Italian evaluation

Run	*ContextSize*	*SenseDistr.*	*P*	*R*	*F*	*A*
MFS	-	-	0.572	0.572	0.572	100%
IT.LESK.2	5	N	0.531	0.530	0.530	99.71%
IT.LESK.10	W	Y	0.608	0.606	0.607	99.71%
IT.DSM.1	3	N	0.611	0.609	0.610	99.71%
IT.DSM.2	5	N	0.608	0.607	0.607	99.71%
IT.DSM.3	10	N	0.627	0.625	0.626	99.71%
IT.DSM.4	20	N	0.629	0.627	0.628	99.71%
IT.DSM.5	W	N	0.634	0.632	**0.633**	99.71%
IT.DSM.6	3	Y	0.632	0.630	0.631	99.71%
IT.DSM.7	5	Y	0.631	0.629	0.630	99.71%
IT.DSM.8	10	Y	0.636	0.634	0.635	99.71%
IT.DSM.9	20	Y	0.640	0.638	0.639	99.71%
IT.DSM.10	W	Y	0.642	0.640	**0.641**	99.71%

Table borrowed from Basile et al. (2014)

The other participants at Task 12

1. UMCC-DLSI system (Gutiìerrez et al., 2013) builds a graph using several resources: WordNet, WordNet Domains and the eXtended WordNet.

   - The best sense is selected using PageRank; prior probabilities exploit sense frequency information

2. DAEBAK system (Manion and Sainudiin, 2013) adopts a sub-graph of BabelNet generated taking into account the surrounding words of the target word

   - Uses MFS as back-off strategy

3. GETALP (Schwab et al., 2013) is inspired by the classical Lesk measure

Task 12 - Results

System	F
EN.DSM.10	**0.715**
EN.DSM.5	**0.687**
UMCC-DLSI-2	0.685
UMCC-DLSI-3	0.680
UMCC-DLSI-1	0.677
MFS	*0.656*
DAEBAK	0.604
GETALP-BN-1	0.263
GETALP-BN-2	0.266

(a) English

System	F
UMCC-DLSI-2	0.658
UMCC-DLSI-1	0.657
IT.DSM.10	**0.641**
IT.DSM.5	**0.633**
DAEBAK	0.613
MFS	*0.572*
GETALP-BN-2	0.325
GETALP-BN-1	0.324

(b) Italian

Table borrowed from Basile et al. (2014)

Future work

- Applicable to other languages
- Easy to apply to specific domains
  - It only needs a domain corpus (and, optionally, sense frequencies extracted from it)!

References

- **Satanjeev Banerjee and Ted Pedersen. 2002.** *An Adapted Lesk Algorithm for Word Sense Disambiguation Using WordNet*. In Alexander Gelbukh, editor, Computational Linguistics and Intelligent Text Processing, volume 2276 of Lecture Notes in Computer Science, pages 136–145. Springer Berlin Heidelberg.
- **Pierpaolo Basile, Annalina Caputo, Giovanni Semeraro. 2014.** *An Enhanced Lesk Word Sense Disambiguation Algorithm through a Distributional Semantic Model,* in Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, Dublin, Ireland, August 23-29, pp. 1591-160.
- **Michael Lesk. 1986.** *Automatic Sense Disambiguation Using Machine Readable Dictionaries: How to Tell a Pine Cone from an Ice Cream Cone.* In Proceedings of the 5th Annual International Conference on Systems Documentation, SIGDOC '86, pages 24–26, New York, NY, USA. ACM.
- **Rada Mihalcea**. **2006**. *Knowledge Based Methods for Word Sense Disambiguation*, book chapter in Word Sense Disambiguation: Algorithms, Applications, and Trends, Editors Phil Edmonds and Eneko Agirre, Kluwer.
- **Roberto Navigli and Simone Paolo Ponzetto. 2012.** *BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network.* Artificial Intelligence, 193:217–250.
- **Roberto Navigli, David Jurgens, and Daniele Vannella. 2013**. *SemEval-2013 Task 12: Multilingual Word Sense Disambiguation*. In Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013), pages 222–231, Atlanta, Georgia, USA, June. Association for Computational Linguistics