Lesk algorithm

From Wikipedia, the free encyclopedia

The **Lesk algorithm** is a classical algorithm for word sense disambiguation introduced by Michael E. Lesk in 1986. [1]

Contents

Overview

The Lesk algorithm is based on the assumption that words in a given "neighborhood" (section of text) will tend to share a common topic. A simplified version of the Lesk algorithm is to compare the dictionary definition of an ambiguous word with the terms contained in its neighborhood. Versions have been adapted to use WordNet.[2] An implementation might look like this:

1. for every sense of the word being disambiguated one should count the amount of words that are in both neighborhood of that word and in the definition of each sense in a dictionary
2. the sense that is to be chosen is the sense which has the biggest number of this count

A frequently used example illustrating this algorithm is for the context "pine cone". The following dictionary definitions are used:

```
PINE
1. kinds of evergreen tree with needle-shaped leaves
2. waste away through sorrow or illness
```

```
CONE
1. solid body which narrows to a point
2. something of this shape whether solid or hollow
3. fruit of certain evergreen trees
```

As can be seen, the best intersection is Pine #1 ∩ Cone #3 = 2.

Simplified Lesk algorithm

In Simplified Lesk algorithm,[3] the correct meaning of each word in a given context is determined individually by locating the sense that overlaps the most between its dictionary definition and the given context. Rather than simultaneously determining the meanings of all words in a given context, this approach tackles each word individually, independent of the meaning of the other words occurring in the same context.

"A comparative evaluation performed by Vasileseu et al. (2004)[4] has shown that the simplified Lesk algorithm can significantly outperform the original definition of the algorithm, both in terms of precision and efficiency. By evaluating the disambiguation algorithms on the Senseval-2 English all words data, they measure a 58% precision using the simplified Lesk algorithm compared to the only 42% under the original algorithm.

Note: Vasileseu et al. implementation considers a back-off strategy for words not covered by the algorithm, consisting of the most frequent sense defined in WordNet. This means that words for which all their possible meanings lead to zero overlap with current context or with other word definitions are by default assigned sense number one in WordNet."[5]

Simplified LESK Algorithm with smart default word sense (Vasilescu et al., 2004)[6]

```
function SIMPLIFIED LESK(word,sentence) returns best sense of word

    best-sense <- most frequent sense for word
    max-overlap <- 0
    context <- set of words in sentence
    for each sense in senses of word do

        signature <- set of words in the gloss and examples of sense
        overlap <- COMPUTEOVERLAP (signature,context)
        if overlap > max-overlap then

                max-overlap <- overlap
                best-sense <- sense

end return (best-sense)
```

The COMPUTEOVERLAP function returns the number of words in common between two sets, ignoring function words or other words on a stop list. The original Lesk algorithm defines the context in a more complex way.

Criticisms and other Lesk-based methods

Unfortunately, Lesk's approach is very sensitive to the exact wording of definitions, so the absence of a certain word can radically change the results. Further, the algorithm determines overlaps only among the glosses of the senses being considered. This is a significant limitation in that dictionary glosses tend to be fairly short and do not provide sufficient vocabulary to relate fine-grained sense distinctions.

Recently, a lot of works appeared which offer different modifications of this algorithm. These works uses other resources for analysis (thesauruses, synonyms dictionaries or morphological and syntactic models): for instance, it may use such information as synonyms, different derivatives, or words from definitions of words from definitions.[7]

There are a lot of studies concerning Lesk and its extensions:[8]

- Kwong, 2001;
- Nastase and Szpakowicz, 2001;
- Wilks and Stevenson, 1998, 1999;
- Mahesh et al., 1997;
- Cowie et al., 1992;
- Yarowsky, 1992;
- Pook and Catlett, 1988;
- Kilgarriff & Rosensweig, 2000,
- Alexander Gelbukh, Grigori Sidorov, 2004.

Accuracy

The original method achieved 50–70% accuracy (depending on the word) on *Pride and Prejudice* and selected papers of the Associated Press.

Lesk Variants

- Original Lesk (Lesk, 1986)
- Adapted/Extended Lesk (Banerjee and Pederson, 2002/2003)

See Also

- Word Sense Disambiguation

Reference

1. ^ Lesk, M. (1986). Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone (http://portal.acm.org/citation.cfm?id=318728&dl=GUIDE,ACM&coll=GUIDE&CFID=103485667&CFTOKEN=64768709). In SIGDOC '86: Proceedings of the 5th annual international conference on Systems documentation, pages 24-26, New York, NY, USA. ACM.
2. ^ Satanjeev Banerjee and Ted Pedersen. *An Adapted Lesk Algorithm for Word Sense Disambiguation Using WordNet (http://www.cs.cmu.edu/~banerjee/Publications/cicling2002.ps.gz)*, Lecture Notes In Computer Science; Vol. 2276, Pages: 136 - 145, 2002. ISBN 3-540-43219-1
3. ^ Kilgarriff and J. Rosenzweig. 2000. English SENSEVAL:Report and Results. In Proceedings of the 2nd International Conference on Language Resourcesand Evaluation, LREC, Athens, Greece.
4. ^ Florentina Vasilescu, Philippe Langlais, and Guy Lapalme. 2004. Evaluating Variants of the Lesk Approach for Disambiguating Words. LREC, Portugal.
5. ^ Agirre, Eneko & Philip Edmonds (eds.). 2006. Word Sense Disambiguation: Algorithms and Applications

5. ^ Agirre, Eneko & Philip Edmonds (eds.). 2006. Word Sense Disambiguation: Algorithms and Applications. Dordrecht: Springer. www.wsdbook.org
6. ^ Florentina Vasilescu, Philippe Langlais, and Guy Lapalme. 2004. Evaluating Variants of the Lesk Approach for Disambiguating Words. LREC, Portugal.
7. ^ Alexander Gelbukh, Grigori Sidorov. Automatic resolution of ambiguity of word senses in dictionary definitions (in Russian). J. Nauchno-Tehnicheskaya Informaciya (NTI), ISSN 0548-0027, ser. 2, N 3, 2004, pp. 10–15.
8. ^ Roberto Navigli. *Word Sense Disambiguation: A Survey* (http://www.dsi.uniroma1.it/~navigli/pubs/ACM_Survey_2009_Navigli.pdf), ACM Computing Surveys, 41(2), 2009, pp. 1–69.

Retrieved from "http://en.wikipedia.org/w/index.php?title=Lesk_algorithm&oldid=595760625"
Categories: Natural language processing | Semantics | Computational linguistics | Word-sense disambiguation

- This page was last modified on 16 February 2014 at 18:33.
-