Comparing Similarity Measures for Original WSD Lesk Algorithm

Sulema Torres and Alexander Gelbukh

Centro de Investigación en Computación (CIC-IPN),
Unidad Profesional Adolfo-López Mateos, Av. Juan de Dios Bátiz s/n
and M. Othón de Mendizábal, Zacatenco, México, DF. 07738, Mexico
sulema7@gmail.com, www.gelbukh.com

Abstract. There are many similarity measures to determine the similarity relatedness between two words. Measures of similarity or relatedness are used in such applications as word sense disambiguation. One of the methods used to resolve WSD is the Lesk algorithm. The performance of this algorithm is connected with the similarity relatedness between all words in the text, i.e the success rate of WSD should increase as the similarity measure's performance gets better. This paper presents a comparison of several similarity measures applied to WSD using the original Lesk Algorithm**.**

Keywords: Word Sense Disambiguation, Lesk Algorithm, WordNet, Semantic Similarity.

1 Introduction

The need to determine the *degree of semantic similarity*, or *relatedness*, between two words is an important problem in Natural Language Processing (NLP). Similarity measures are used in such applications as word sense disambiguation, determining discourse structure, text summarization and annotation, information extraction and retrieval, automatic indexing, lexical selection, and automatic correction of word errors in text [4].

Human beings have an innate ability to tell if one word is more similar to a given word than another. For example, most would agree that the automotive senses of *car* and *tire* are related while *car* and *tree* are not.

There are mainly two approaches to semantic similarity [2, 17]. First approach is making use of a large corpus and gathering statistical data from this corpus to estimate a score of semantic similarity. Second approach makes use of the relations and the hierarchy of a thesaurus, which is generally a hand-crafted lexical database such as WordNet [5]. As in many other NLP studies, hybrid approaches that make benefit from both techniques also exist in semantic similarity.

There are some ways to evaluate semantic similarity measures. One is checking the correlation between the results of similarity measures and human judgments. Another one is to select an application area of semantic similarity, and compare the results of different similarity measure according to the success rates in that

Advances in Computer Science and Applications.
Research in Computing Science 43, 2009, pp. 155-166

All the other edges in the graph, including links between adjectives and adverbs, or links across different parts-of-speech, are drawn using the *lesk* measure. The results indicate that the right combination of similarity metrics can lead to a performance competing with the state-of-the-art in unsupervised word sense disambiguation.

3 WSD using the Lesk Algorithm

The Lesk algorithm [10] uses dictionary definitions (gloss) to disambiguate a polysemous word in a sentence context. The major objective of his idea is to count the number of words that are shared between two glosses. The more overlapping the words, the more related the senses are.

To disambiguate a word, the gloss of each of its senses is compared to the glosses of every other word in a phrase. A word is assigned to the sense whose gloss shares the largest number of words in common with the glosses of the other words. Figure 1 shows the graphic representation of the Lesk Algorithm.

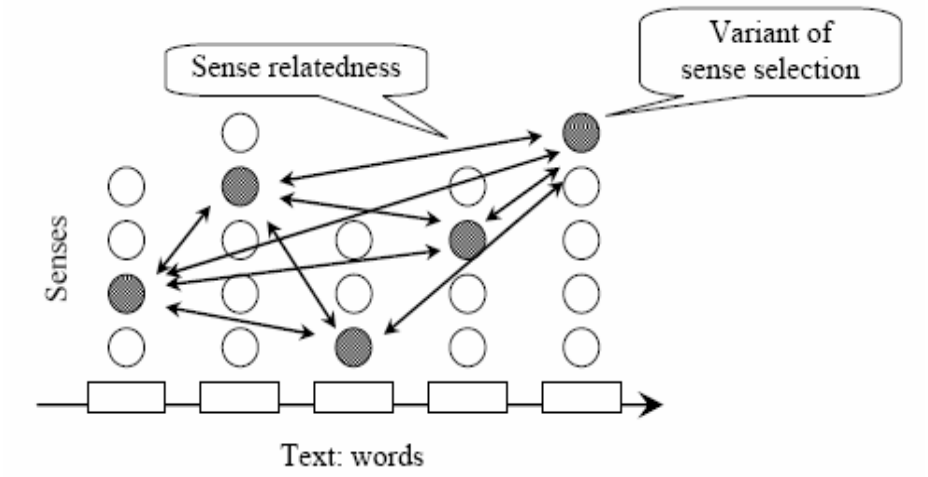

Figure 1. Graphic Representation of the Lesk Algorithm

For example: In performing disambiguation for the "pine cone" phrasal, according to the Oxford Advanced Learner's Dictionary, the word "pine" has two senses:

- sense 1: kind of evergreen tree with needle–shaped leaves,
- sense 2: waste away through sorrow or illness.

The word "cone" has three senses:

- sense 1: solid body which narrows to a point,
- sense 2: something of this shape whether solid or hollow,
- sense 3: fruit of a certain evergreen tree.

9. Leacock and M. Chodorow. Combining local context and WordNet sense similarity for word sense identification. In *WordNet, An Electronic Lexical Database*. The MIT Press, 1998.
10. M. Lesk. Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from a ice cream cone. In *Proceedings of SIGDOC '86*, 1986.
11. Lin. An information-theoretic definition of similarity. In *Proceedings of the 15th International Conference on Machine Learning*, Madison, WI, 1998.
12. M. Palmer, C. Fellbaum, S. Cotton, L. Delfs, and H. Dang. English tasks: all-words and verb lexical sample. In *Proceedings of ACL/SIGLEX Senseval-2*, Toulouse, France, 2001.
13. M. Palmer and M. Light. Introduction to the special issue on semantic tagging. *Natural Language Engineering*, 5(2): i-iv.
14. S. Patwardhan, S. Banerjee, and T. Pedersen. Using Measures of Semantic Relatedness for Word Sense Disambiguation. In *Proceedings of the Fourth International Conference on Intelligent Text Processing and Computational Linguistics*, pp. 241–257, Mexico City, Mexico, February, 2003.
15. S. Pradhan, E. Loper, D. Dligach, and M. Palmer. SemEval-2007 task 17: English lexical sample, SRL and all words. In *Proceedings of the 4th international workshop on semantic evaluations (SemEval-2007),* pp. 87–92, Prague, Czech Republic, 2007.
16. P. Resnik. Using information content to evaluate semantic similarity. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, Montreal, Canada, 1995.
17. Sebti and A. Barfroush. A new word sense similarity measure in wordnet. In *Proceedings of the IEEE International Multiconference on Computer Science and Information Technology*, October, 2008.
18. R. Sinha and R. Mihalcea. Unsupervised Graph-based Word Sense Disambiguation Using Measures of Word Semantic Similarity, In *Proceedings of the IEEE International Conference on Semantic Computing (ICSC 2007)*, Irvine, CA, September 2007.
19. B. Snyder and M. Palmer. The English all-words task. In *Proceedings of ACL/SIGLEX Senseval-3*, Barcelona, Spain, July 2004.
20. Senseval. *Evaluation Exercises for the Semantic Analysis of Text* http://www.senseval.org