Annotating Words Using WordNet Semantic Glosses

Julian Szymański[1,⋆] and Włodzisław Duch[2,3]

[1] Department of Computer Systems Architecture, Gdańsk University of Technology, Poland
julian.szymanski@eti.pg.gda.pl
[2] Department of Informatics, Nicolaus Copernicus University, Toruń, Poland
[3] School of Computer Engineering, Nanyang Technological University, Singapore
Google: W. Duch

Abstract. An approach to the word sense disambiguation (WSD) relaying on the WordNet synsets is proposed. The method uses semantically tagged glosses to perform a process similar to the spreading activation in semantic network, creating ranking of the most probable meanings for word annotation. Preliminary evaluation shows quite promising results. Comparison with the state-of-the-art WSD methods indicates that the use of WordNet relations and semantically tagged glosses should enhance accuracy of word disambiguation methods.

Keywords: Word Sense Disambiguation, WSD, WordNet, Wikipedia, NLP.

1 Introduction

Ambiguity of natural language is the source of many problems in automatic text processing. It is quite evident for example in classification or clustering of documents represented by features derived from word frequencies. Automatic semantic annotation is still a great challenge, requiring solution to the word sense disambiguation (WSD) problem. To realize the promise of semantic Internet, with machine-enabled interpretation, words in the text should be annotated with specific senses. Adding elementary semantic information during the initial processing phase greatly facilitates text annotation and interpretation. It can be achieved by creating text representation based on word senses instead of string tokens extracted from the content of the text [5]. Other types of annotations, not discussed here, include syntactic parts of speech tagging, grammatical, anaphoric, prosodic and affective annotations.

In analogy to the NP-complete problems in computational complexity theory [7] the word disambiguation and many other Natural Language Processing (NLP) problems are defined as AI-complete [14], meaning that their solution is as hard as passing the Turing Test [17]. The problem of words disambiguation implies several issues that should be taken into consideration.

First, how to distinguish and represent word meanings? The level of granularity of senses and how they relate to each other needs to be defined. The use of synonyms and/or homonyms must be considered. Many approaches have been developed to acquire word senses in an automatic way [18] eg. using Latent Semantic Indexing based

⋆ Corresponding author.

T. Huang et al. (Eds.): ICONIP 2012, Part IV, LNCS 7666, pp. 180–187, 2012.

on Singular Value Decomposition applied to statistics of words occurrences. Still the most successful results are achieved creating word senses by a hand.

Second, how to encode lexical knowledge to enable effective interpretation similar to the way people use natural language? Many approaches have been devised here, using for example manually crafted ontologies, ex. Sumo/Milo [12], or semantic networks [16], such as manually created WordNet [11], or semi-automatically created ConceptNet [8] and MindNet [13]. Other approaches try to acquire linguistic knowledge using statistical methods applied to a large collections of data eg. HAL [9] or ADIOS [15].

One fruitful approach to semantic annotation of texts is to move beyond bag-of-words representation, using atoms of lexical knowledge to represent the elementary word meanings (senses), and converting the text into a graph linking senses rather than words. We shall focus here only on the word sense disambiguation during initial text processing phase, mapping words form texts to the structures that carry elementary meanings that may be treated as semantic atoms (senses). WordNet synsets are well-suited for that purpose, grouping words into sets of synonyms related to word definitions, providing sense identifiers and recording semantic relations between synsets. Different people rarely use the same words describing the same object, scene or situation. The use of synsets helps to capture similarities of texts that contain different words but have similar meaning. Employing synsets allows for using WordNet semantic network formed by relations between synsets. Text annotated at a higher abstraction level is can be clustered in a better way because similarities between texts are more clear.

Enhancing document representation with superordinate categories works even better for clustering [4], simulating spreading neural activation responsible for simple inference processes in the reader's mind. New features expose content of the text in more obvious way, simplifying conceptual processing. This elementary representations of words meanings has been already used in our projects aimed at constructing algorithms for automatic analysis of texts[1]. The approach to the word sense disambiguation introduced here is based on contextual information obtained from synsets related to a given synset by exploiting its definition. Description of the used algorithm, examples of the disambiguations and evaluation on the set of several polysemous test words is given below.

2 Disambiguation with WordNet Semantic Glossses

Semantic Glosses (SG) approach employs relations between synsets, or more precisely relations obtained from references between synsets that are related to their definitions (gloss tags)[2]. This idea differs from one of the most popular approaches – adapted Lesk algorithm [1] that uses structural information and traverses the hypernym hierarchy formed as a tree of senses. A graph of related synsets is used here to strengthen mutual associations of synsets. It resembles the spreading activation process [2], automatic activation of related concepts during sentence comprehension, formation of patterns of activations related to word meanings. Activations of the network of synsets may serve

[1] `http://kask.eti.pg.gda.pl/CompWiki/`
[2] `http://wordnet.princeton.edu/glosstag.shtml`

Algorithm 1. Pseudocode of Semantic Glosses algorithm

```
1: function SEMANTICGLOSS(text)
2:     wordSenses ← empty set
3:     Nwords ← number of adjacent words included
4:     for word in text do
5:         for sense in word.def.senses do
6:             wordSenses[sense] ← 0;
7:         end for
8:     end for
9:     for word in text do
10:        Refs ← Nwords words adjacent to the word
11:        for reference in Refs do
12:            for wordSense in word.def.senses do
13:                if sense ∈ reference.def.senses then
14:                    scores[sense] + +          ▷ Add a point to the score
15:                end if
16:            end for
17:        end for
18:    end for
19:    result ← an empty set
20:    for word in wordSenses do
21:        result ← a sense for word with the highest score
22:    end for
23:    return result
24: end function
```

as an approximation of semantic representation [3], where the already active network constrains selection of the next synset.

SG approach for disambiguation of word meanings employs relations between synsets. WordNet not only lists various word meanings, representing them by synsets, but also provides several types of relations between them. Many structural relations, such as hypernyms, troponyms, or meronyms, are defined, usually for a particular part of speech. Starting with the version 3.0 WordNet also provides *semantically annotated disambiguated gloss corpus*. Glosses are short definitions providing proper meanings of words and thus whole synsets. The gloss annotations cover also concepts, collocations (multi-word forms), tagging discontiguous spans of text, for example converting "personal or business relationship" to "personal_relationship", "business_relationship". Glosses have been linked manually to the context-appropriate sense in WordNet, disambiguating the corpus. Tagging includes part of speech, potential lemma forms, a few semantic classes (acronym, number, year, currency, etc). This information creates many new opportunities, but in this paper only associations between synset definitions are used. With semantically annotated gloss corpus each synsets is loosely coupled with several others, related to its definition. In this way additional contextual relations are provided and these relations are not restricted to the one part of speech, as is the case with most structural relations.

The main steps in the disambiguation process are as follows:

- Disambiguated word W is mapped on its possible meanings (synsets) $\{Ts(W)\}$.
- For each synset from $\{Ts(W)\}$ set retrieve all synsets Tgs that may be derived from its glosses.
- Rank all Ts synset according to the number of relations with glosses in Tgs.

The details of this algorithm are described in the pseudocode Algorithm 1. Wordnet glosses may be extended by Wikipedia articles that may be linked directly to the appropriate synsets.

3 Results

Examples of disambiguations performed using relations between synset glosses are presented below. For illustration different texts using different meanings of a test word *horse* have been selected. Disambiguation results for different meanings of the word are presented in Tables 2–6. To compare the results achieved with this approach, denoted as SG (*Semantic Glosses*), results obtained by the Stanford Parser[3] (SP) are also given. Probabilities in percentages indicating the most suitable sense are presented. Results are presented in the form $\boxed{\text{word}_x^y}$, where x is the sense number and y is a Greek letter used to enumerate consecutive words that appear in the text.

WordNet offers following senses of the word *horse*:

1. **horse, Equus caballus** – solid-hoofed herbivorous quadruped domesticated since prehistoric times.
2. **horse, gymnastic horse** – a padded gymnastic apparatus on legs.
3. **cavalry, horse cavalry, horse** – troops trained to fight on horseback; 500 horse led the attack.
4. **sawhorse, horse, sawbuck, buck** – a framework for holding wood that is being sawed.
5. **knight, horse** – a chessman shaped to resemble the head of a horse; can move two squares horizontally and one vertically (or vice versa).

The Wikipedia articles about these different meaning of *horse*, with each appearance labeled, are shown below, followed by tables showing results of the Stanford Parser (SP) and our Semantic Glosses (SG) method.

1) The $\boxed{\text{horse}_1^\alpha}$ is a hooved (ungulate) mammal, a subspecies of the family Equidae. $\boxed{\text{Horses}_1^\beta}$ and humans interact in a wide variety of sport competitions and non-competitive recreational pursuits, as well as in working activities such as police work, agriculture, entertainment, and therapy. $\boxed{\text{Horses}_1^\gamma}$ were historically used in warfare, from which a wide variety of riding and driving techniques developed, using many different styles of equipment and methods of control. Humans provide domesticated

[3] `http://nlp.stanford.edu/software/lex-parser.shtml`

$\boxed{\text{horses}_1^{\delta}}$ with food, water and shelter, as well as attention from specialists such as veterinarians and farriers. The results of disambiguation of the word "horse" in the first sense are shown in Table 1. Scores are given in percents, and bold face shows the highest score for a given method and a given word position. SG has slight preference for wrong sense 3 and 5 over correct 1.

Table 1. Results of disambiguating four occurrences of word "horse" in sense 1 (horse_1)

	horse_1^{α}		horses_1^{β}		horses_1^{γ}		horses_1^{δ}	
	SP	SG	SP	SG	SP	SG	SP	SG
#1	**48%**	22%	**40%**	22%	**41%**	**22%**	**62%**	**22%**
#2	26%	18%	17%	19%	11%	19%	7%	19%
#3	18%	**24%**	18%	**23%**	10%	21%	17%	21%
#4	8%	15%	24%	15%	0%	16%	9%	16%
#5	1%	21%	1%	21%	38%	**22%**	4%	**22%**

2) The $\boxed{\text{horse}_2^{\alpha}}$ is an artistic gymnastics apparatus. It is used by only male gymnasts, due to intense strength requirements. Originally made of a metal frame with a wooden body and a leather cover, modern pommel $\boxed{\text{horses}_2^{\beta}}$.

Both parsers found for the second occurrence collocations '*pommel horse*', defined as "a metal body covered with foam rubber and leather, with plastic handles (or pommels)", therefore the second occurrence of the word has not been tagged. The results of disambiguation of the word "horse" in that sense are shown in Table 2. SG shows very stron preference for correct meaning, while SP fails here.

Table 2. Results of disambiguating sense horse_2

	horse_2^{α}		horses_2^{β}	
	SP	SG	SP	SG
#1	**57%**	0%	n/a	n/a
#2	42%	**98%**	n/a	n/a
#3	1%	2%	n/a	n/a
#4	0%	0%	n/a	n/a
#5	0%	0%	n/a	n/a

Table 3. Results of disambiguating sense horse_3

	horse_3^{α}	
	SP	SG
#1	31%	n/d
#2	6%	n/d
#3	**63%**	n/d
#4	0%	n/d
#5	0%	n/d

3) $\boxed{\text{Horse}_3^{\alpha}}$ cavalry were soldiers or warriors who fought mounted on horseback. Cavalry were historically the third oldest (after infantry and chariotry) and the most mobile of the combat arms. A soldier in the cavalry is known by a number of designations such as cavalryman or trooper.

In this case SG approach found collocation '*horse cavalry*' with score 69% and annotated it as '*an army unit mounted on horseback*', and with score 31% '*troops trained to fight on horseback*'. Although this is essentially correct single word has not been annotated, hence n/d in Table 3.

4) A $\boxed{\text{horse}_4^{\alpha}}$ is a beam with four legs used to support a board or plank for sawing. The sawhorse may be designed to fold for storage. A sawhorse with a wide top is particularly useful to support a board for sawing or as a field workbench, and is more useful as a single, but also more difficult to store. The results of disambiguation of word horse in that sense have been shown in Table 4. Here SG has very stron correct preference while SP fails.

Table 4. Results of disambiguating sense horse_4

	horse_4^{α}	
	SP	SG
#1	32%	0%
#2	19%	10%
#3	10%	0%
#4	2%	**90%**
#5	**37%**	0%

Table 5. Results of disambiguating sense horse_5

	horse_5^{α}		horse_5^{β}	
	SP	SG	SP	SG
#1	28%	7%	26%	11%
#2	8%	6%	19%	9%
#3	27%	36%	**36%**	33%
#4	0%	5%	3%	8%
#5	**36%**	**47%**	15%	**38%**

5) The $\boxed{\text{horse}_5^{\alpha}}$ is a piece in the game of chess, representing a knight (armored cavalry). It is normally represented by a $\boxed{\text{horse}_5^{\beta}}$'s head and neck. Each player starts with two knights, which start on the rank closest to the player, one square from the corner. The results of disambiguation of the word horse in that sense are shown in Table 5. SG is correct by a wide margin over other senses, SP fails for the second position in favor of sense 3.

The evaluation of the SG approach has been performed on a test set of eight multi-sense words. For different senses of these words 51 test texts have been prepared, and evaluation of disambiguation performed in the same way as in the case of a word *horse*.

The results of disambiguation for 8 test words are shown in Table 6. The percentage values describes the fraction of proper disambiguations achieved for each word aggregated for all senses.

Table 6. Accuracy of disambiguating 8 test words

	SP	SG
Horse	66%	75%
King	30%	53%
Road	100%	66%
Computer	100%	50%
Grass	60%	100%
Kernel	33%	100%
Shell	20%	55%
Root	50%	80%

Table 7. Aggregated graded scores of disambiguating 8 test words (see text)

	SP	SG
Horse	63%	81%
King	38%	50%
Road	83%	66%
Computer	100%	50%
Grass	30%	100%
Kernel	33%	100%
Shell	20%	50%
Root	41%	100%

Results presented in Table 6 describe only the precision of disambiguation in terms of binary correct-incorrect decisions. However, sometimes the difference between two

top-scored synsets may be very small, or even zero, as for example in Table 1 where the horse$_1^\gamma$ Semantic Glosses algorithm equally score sense #1 and #5. A graded evaluation that values bigger differences in scores gives more useful information. If the proper sense of the word has been determined with high confidence (more than 10% difference from the next sense) the graded score is 1. If this difference is in the range $3 - 10\%$ it obtains 0.75, and for differences within $\pm 3\%$ the graded score is 0.5 (even if the correct synset is cored below the winner). Finally, if the proper meaning fells below the 3% of the winner, or for some reason could not be evaluated the score is 0.

Summing all graded scores and dividing this sum by the number of performed disambiguations expressed in percentages allows for measuring results including some estimation of disambiguation confidence. These results are presented in Table 7.

4 Discussion and Future Directions

The algorithm that employs semantically annotated glosses provides quite promising results. So far it has been evaluated only on a small test set of 8 multi sense words (51 different meanings). As the preliminary results are promising the method is now being tested on a larger scale, and some improvements will be introduced.

The approach can run into problems while disambiguating different meanings of the same word in one sentence eg. *„Turtle's shells provide protection to parts of the animal body, like egg shell protects birds' embryo."* The first 'shell' is related to the turtle shell the second to the egg shell. The task for disambiguating such cases is relatively easy for humans because using semantic memory collocations are easily discovered and require much smaller context for proper sense classification. Experiments with variable context length dependent on the number of identical words with different meanings in one sentence will be performed to check how to deal with such difficulties.

Some WordNet synsets are larger and have more relations than others, the distribution is very uneven. This causes preference for larger synsets that may confuse many algorithms degrading results for meanings that correspond to synsets with small number of relations. To simulate effects of spreading activation weighed relations between synsets may be introduced, describing patterns of more and less important activations. One should also explore the use of WordNet structural information given in predefined relations that extends the network of relations between synsets. Also it is possible to use references between glosses obtained from higher order relations that should have smaller weights.

It should be fruitful to employ additional relations from mining Wikipedia hyperreferences [10] to introduce more relations between synsets. This task requires first a mapping between WordNet synsets and Wikipedia articles. Results of the semi-automatic approach [6] to perform such mapping are quite good. Another aspect is the use of negative knowledge about the words present in glosses that do not appear in the wider context.

To perform experiments presented in this article the application for testing different methods of word sense disambiguation has been created. Using it one can enter the sentence and obtain the text with semantic annotations. This application integrates selected parsers and allows for experimentation displaying results in the user-friendly form. The source code of this application is freely available for

academic use at the address: http://kask.eti.pg.gda.pl/semagloss/annotations.zip. This project resulted also in development of API in C# and Java for WordNet semantically annotated gloss corpus. The API is available for download at the address http://kask.eti.pg.gda.pl/semagloss/index.html.

Acknowledgments. The work has been supported by the Polish Ministry of Science and Higher Education under research grant N N 516 432338. Authors would like also to thank Adam Kuśmierz for his contribution to the application development.

References

1. Banerjee, S., Pedersen, T.: An Adapted Lesk Algorithm for Word Sense Disambiguation Using WordNet. In: Gelbukh, A. (ed.) CICLing 2002. LNCS, vol. 2276, pp. 136–145. Springer, Heidelberg (2002)
2. Collins, A., Loftus, E.: A Spreading-Activation Theory of Semantic Processing. Psychol. Rev. 82, 407 (1975)
3. Duch, W., Matykiewicz, P., Pestian, J.: Towards Understanding of Natural Language: Neurocognitive Inspirations. In: de Sá, J.M., Alexandre, L.A., Duch, W., Mandic, D.P. (eds.) ICANN 2007. LNCS, vol. 4669, pp. 953–962. Springer, Heidelberg (2007)
4. Duch, W., Matykiewicz, P., Pestian, J.: Neurolinguistic Approach to Natural Language Processing with Applications to Medical Text Analysis. Neural Netw. 21, 1500–1510 (2008)
5. Kehagias, A., Petridis, V., Kaburlasos, V.G., Fragkou, P.: A Comparison of Word and Sense-based Text Categorization Using Several Classification Algorithms. J. Intell. Inf. Syst. 21, 227–247 (2003)
6. Korytkowski, R., Szymański, J.: Collaborative Approach to WordNet and Wikipedia Integration. In: The Second International Conference on Advanced Collaborative Networks, Systems and Applications, COLLA 2012, pp. 23–28 (2012)
7. Kubale, M.: Introduction to Computational Complexity and Algorithmic Graph Coloring, Gdańskie Towarzystwo Naukowe, Poland (1998)
8. Liu, H., Singh, P.: ConceptNet: A Practical Commonsense Reasoning Tool-Kit. BT Technol. J. 22, 211–226 (2004)
9. Lund, K., Burgess, C.: Producing High-Dimensional Semantic Spaces from Lexical Cooccurrence. Behav. Res. Methods 28, 203–208 (1996)
10. Medelyan, O., Milne, D., Legg, C., Witten, I.: Mining Meaning from Wikipedia. Int. J. Hum-Comput. St. 67, 716–754 (2009)
11. Miller, G.A., Beckitch, R., Fellbaum, C., Gross, D., Miller, K.: Introduction toWordNet: An On-line Lexical Database. Princeton University Press, New Jersey (1993)
12. Niles, I., Pease, A.: Towards a Standard Upper Ontology. In: Proceedings of the International Conference on Formal Ontology in Information Systems, pp. 2–9. ACM (2001)
13. Richardson, S., Dolan, W., Vanderwende, L.: MindNet: Acquiring and Structuring Semantic Information from Text. In: Proceedings of the 17th International Conference on Computational Linguistics, vol. 2, pp. 1098–1102. Association for Computational Linguistics (1998)
14. Shahaf, D., Amir, E.: Towards a Theory of AI Completeness. In: AAAI Spring Symposium: Logical Formalizations of Commonsense Reasoning, pp. 150–155 (2007)
15. Solan, Z., Horn, D., Ruppin, E., Edelman, S.: Unsupervised Learning of Natural Languages. Proceedings of the National Academy of Sciences of the USA 102(33), 11629 (2005)
16. Sowa, J.: Principles of Semantic Networks. Morgan Kaufmann Series in Representation and Reasoning. Morgan Kaufmann, San Mateo (1991)
17. Turing, A.: Computing Machinery and Intelligence. Mind 59, 433–460 (1950)
18. Turney, P., Pantel, P.: From Frequency to Meaning: Vector Space Models of Semantics. J. Artif. Intell. Res. 37, 141–188 (2010)