A New Approach to Pseudoword Generation

Lubomir Otrusina, Pavel Smrz

Brno University of Technology
Bozetechova 2, Brno, Czech Republic
{iotrusina,smrz}@fit.vutbr.cz

Abstract

Sense-tagged corpora are used to evaluate word sense disambiguation (WSD) systems. Manual creation of such resources is often prohibitively expensive. That is why the concept of pseudowords – conflations of two or more unambiguous words – has been integrated into WSD evaluation experiments. This paper presents a new method of pseudoword generation which takes into account semantic-relatedness of the candidate words forming parts of the pseudowords to the particular senses of the word to be disambiguated. We compare the new approach to its alternatives and show that the results on pseudowords, that are more similar to real ambiguous words, better correspond to the actual results. Two techniques assessing the similarity are studied – the first one takes advantage of manually created dictionaries (wordnets), the second one builds on the automatically computed statistical data obtained from large corpora. Pros and cons of the two techniques are discussed and the results on a standard task are demonstrated. Even though the wordnet-based method improves the modelling accuracy of the generated pseudowords (and one of the automatic corpus-based procedures provides results close to it), the key observation is that the results of WSD on the Senseval-like data is highly unstable and that the comparison of the WSD systems should be application-specific.

1. Introduction

WSD refers to the natural language processing problem of determining which sense of a word is activated by the use of the word in a particular context (Agirre and Edmonds, 2006). The most successful systems apply supervised machine learning algorithms and classify an occurrence of the word in context into one or more of its sense classes defined by a dictionary. To be able to evaluate WSD systems, one needs a "gold-standard" corpus annotated with the target senses. Either a sample of the words is processed or all (known) words in a piece of running text are disambiguated. The test corpora are usually hand-annotated. Producing such a resource is very expensive.

Pseudowords offer an alternative to the manual sense tagging. Two or more words are chosen (e. g., *dentist* and *spaceship*) and their individual occurrences are replaced by their conflation (i. e., *dentist/spaceship*). The task of the WSD system is then to identify the part of the pseudoword that corresponds to the original word in each individual context. Unfortunately, the results of the evaluation on pseudowords significantly differ from the real ones. The contexts of the pseudoword individual components are often much more distinctive than those of different senses corresponding to a real word. Therefore, pseudowords are accepted as an upper bound of the true WSD accuracy only (Nakov and Hearst, 2003).

Two main factors play a key role in the unequal performance – the selection process of pseudoword components and the way the results are compared. The original pseudoword-forming procedures (Gale et al., 1992; Schütze, 1992; Yarowsky, 1993) expected a random selection of the constituent (unambiguous) words. Consequently, they were likely to combine semantically distinct words that could perhaps model homography but not much more frequent polysemy. Nakov and Hearst also mention that the results produced using such pseudowords were difficult to characterize in terms of the types of ambiguity they model (Nakov and Hearst, 2003). Later approaches employed semantic similarity/relatedness derived from manually created resources — general-purpose wordnet-style networks such as Chinese CUP-Dic (Lu et al., 2006) or domain-specific lexical hierarchies, e. g. MeSH (Nakov and Hearst, 2003).

Pseudoword constituent selection is also related to the target application of WSD. For example, Ide and Wilks (Ide and Wilks, 2006) argue that the WSD research should focus on broad-discrimination tasks that most NLP applications require. Also, the primary motivation of the research reported in this paper lies in WSD for text mining purposes that seldom demand fine-grained division of senses. If the "ground truth" sense inventory is very fine-grained (such as in wordnet used in our experiments), it is questionable whether the pseudoword constituents need to be necessarily unambiguous words (wrt the base dictionary). Intuitively, it seems more appropriate (and the results of our experiments confirm this assumption) to lose the constraint and to allow combinations of unambiguous and ambiguous constituents that better model the real ambiguity that a particular application needs to deal with. For example, if the task aims at finding documents/sentences/contexts in which word *Apple* refers to the company (and does not care about the distinction between senses "a fruit" and "a tree"), *Google/Pear* could become the pseudoword even though *Pear* presents the same metonymic relation (can also mean "a fruit" or " a tree").

If the pseudoword building process focuses on the correct modelling of the underlying ambiguity that needs to be taken away in a particular task, not only individual words but also multi-word expressions should be considered as pseudoword constituents. This is especially true in the specialised domains where it can be difficult to find single-word candidates similar to particular senses of a modelled word. However, one needs to be careful when dealing with the multi-word expressions that do not correspond to headwords in the used dictionary (e.g., noun phrases from definitions or multi-word keywords automatically identified as similar to a given sense of the disambiguated word). The multi-word expression itself can be ambiguous and it can be difficult to automatically identify this situation.

Public WSD evaluation campaigns (Senseval-[1-3], SemEval-1) showed that the tasks aiming at fine-grained sense discrimination, where human annotators agreed in only 85 % of word occurrence and where the baseline accuracy (always choosing the most frequent sense) is 50–60 %, cannot expect better accuracy than 70 %. Unfortunately, per-sense accuracies are not always reported. Then, it is difficult to estimate the performance of the state-of-the-art WSD methods on tasks that deal with the non-dominate senses only. An example could be pre-filtering of documents to identify occurrences of a low-frequent sense and then to apply specialised information extraction on the candidate context. Our experiments show that the results for this kind of data are extremely unstable. Even intuitively it is clear that when one trains on two or three examples of a category and tests on another two or three only, the precision can easily go down to 0 % (no testing context covered). The evaluation should therefore explicitly construct confusion and cost/benefit matrix for the senses and measure the results with respect to the task-specific conditions.

Previous pseudoword approaches took into account the frequency of the constituents and conflated words with the frequency ratio corresponding to that of particular senses of the original word. However, there is no evidence that this selection scheme actually helps to build pseudowords that better model the sense variety. Especially in the case of the above-mentioned low-frequency senses, it is obvious that a correct evaluation needs to run many experiments randomly sampling from the potentially large population of the contexts. If one aims at the best modelling of the underlying ambiguity, it is crucial not only to correctly choose the pseudoword constituents, but also the sampling procedure to identify contexts most similar to those of the original senses.

The ability to measure context similarities brings also an additional value extending the primary use in the pseudoword construction discussed in this paper. As mentioned above, the targeted WSD application field of our work is a user-driven information extraction system (Schmidt and Smrz, 2009). The system typically starts with one or two annotated examples, identifies all the ambiguous occurrences in the available data and asks the user to disambiguate additional data. Assessing the context similarities helps to identify the contexts that will maximise the gain of the active learning process.

The experiments described in the following sections apply context similarity measurement to the pseudoword forming. We investigate two approaches. The first one takes advantage of pre-existing lexical hierarchies such as wordnets (that are available for few languages and few application domains only). The other one is based on the data automatically derived from large corpora. The training/testing data sets are taken from Senseval-3 Task 6 (English lexical sample) which is directly mapped to the Princeton Wordnet. We compare the results to the baseline and to the real accuracy values on the original contexts. In contrast to the previous approaches, the new methods construct pseudoword that better model real sense distinctions.

The rest of the paper is organized as follows. The next section introduces the proposed methods and discusses the details of the implemented similarity measures. Also, the basic statistics of the training/test data sets are presented. Section 3 summarizes results of the experiments and compares them to alternative approaches. The following section then relates the presented approach to the previous works in the field. We conclude with future directions of our research.

2. Pseudoword Construction Procedures

The ultimate goal of the pseudoword building process is to find constituents (individual words or multi-word expressions) that occur in similar contexts as corresponding senses of the real ambiguous word. The pool of constituent candidates is derived either directly from the Princeton wordnet, or by means of automatic term similarity techniques.

To understand the difficulties of the selection, let us first characterize the English lexical sample we deal with. In Senseval-3 Task 6, the data is divided into train and test sets per word sense. The work reported in this paper deals with 20 nouns included in the dataset. Rarely, more than one sense or "none of the listed" labels are associated with the evaluation word occurrence. These cases account for less than 1–5 % in particular categories. For the sake of clarity, we do not consider them in the described experiments. Note, however, that the presented methods can model unknown/multi-label annotation.

Word	S	Train set	Test set
argument	5	32:102:8:4:23	20:47:1:4:18
arm	5	19:201:6:29:3	8:108:5:8:1
atmosphere	4	14:2:23:74	4:1:14:38
audience	3	9:122:36	6:60:26
bank	6	7:2:10:163:18:39	5:1:3:86:11:17
degree	6	61:136:3:3:3:18	29:66:6:2:1:11
difference	5	38:26:27:25:83	23:14:15:11:35
difficulty	4	3:11:12:16	2:8:8:4
disc	4	41:18:72:40	19:10:38:24
image	5	60:20:52:3:4	26:11:27:1:1
interest	7	13:69:12:47:7:30:2	11:38:3:24:2:11:2
judgment	7	4:3:15:20:4:1:15	2:2:5:9:2:1:11
organization	3	85:18:5	40:7:6
paper	7	24:10:8:34:2:53:38	19:1:7:12:1:25:29
party	5	16:145:26:4:16	8:71:14:2:7
performance	5	28:42:38:13:31	18:26:21:4:13
plan	3	23:16:103	7:5:57
shelter	3	31:53:80	23:24:33
sort	4	4:13:110:22	2:11:50:18
source	5	11:27:3:2:4	4:18:1:3:3

Table 1: Frequency of train and test examples for individual senses of ambiguous words in the evaluation data set

Table 1 presents the numbers of training/test examples for individual words and their senses. The dataset is anchored in the Princeton wordnet. The "informed" wordnet-based method takes advantage of the sense keys referring to particular synsets in the wordnet database (e. g., (*arm%1:06:00::*). Although not all wordnet senses for the target words are covered in the dataset, the resulting sense inventory is far from being coarse-grained. For example, one of the words – *atmosphere* – does not consider the latter of the two wordnet related senses:

- the mass of air surrounding the Earth
- the envelope of gases surrounding any celestial body

but still distinguishes other two senses with close definitions and overlapping contextual use examples:

- a particular environment or surrounding influence
- a distinctive but intangible quality surrounding a person or thing.

The wordnet-based method simply takes "surroundings" of particular word senses and generates candidates from unambiguous words or short phrases. To determine which words could be taken as unambiguous for the current task in hand (see the discussion on the metonymy of word *Pear* above), we analyse the sense definitions, extract genus words and look for correspondences. For the comparison with the Senseval-3 datasets, however, we do not join the individual senses and search for literals that appear only once in the wordnet.

As other literals appearing in the synset with ambiguous words often refer to ambiguous words as well, the method extends the search to direct siblings (hyponyms of the direct hypernym), the genus phrase from the definition and direct hyponyms. We take English GigaWord Corpus (Graff and Cieri, 2003) as the source of pseudoword constituent contexts. If the identified candidates do not occur frequently enough (to match the numbers of the training/test examples for the particular sense), even larger word vicinity is considered – hypernyms, other relations (such as meronyms) and the second level siblings.

Two methods that do not need manually created dictionaries to generate pseudowords have been tested. First, we employed a latent variable technique called Explicit Semantic Analysis (ESA) (Gabrilovich and Markovitch, 2007). ESA maps each word to the space of Wikipedia articles. It computes word similarity in the dimensions defined by the articles. For example, a word can obtain vector $(1.2, 0.1, 3.5, \cdots)$ where the similarity between the word and the first article in Wikipedia is 1.2, the similarity between the word and the second article is 0.1 and so on. For more details on ESA, see (Gabrilovich and Markovitch, 2007).

The cosine of the angle between the vectors is used to measure similarity of individual words. To identify words appearing in the contexts similar to those of a particular sense of a given ambiguous word, all the words in contexts are lemmatized. We rank the lemmata by $tf_{i,j} \cdot idf_i$ where $tf_{i,j} = \frac{n_{i,j}}{\sum_k m_{k,j}}$, where $n_{i,j}$ is the number of occurrences of term t_i in context of sense c_j and $idf_i = \frac{|C|}{|\{c:t_i \in c\}|}$, where $|C|$ is the total number of senses and $\{c : t_i \in c\}$ is a set containing all the senses whose contexts contain term t_i. The vector representing the whole context is computed as a normalized weighted sum of individual words appearing in a window of a given size. The weight combines the tf.idf score and the distance from the word in focus (the absolute value of the difference in word positions).

The other context relatedness method employs Lin's similarity measure (Lin, 1998). From a general perspective, the similarity between two objects is defined as the amount of information that the objects have in common divided by the amount of information given by each of the objects individually. Let $T(w)$ be a set of all features of word w. Let $I(w, f)$ be the mutual information of word w and feature f. The similarity between words w_1 and w_2 can then be computed as $sim_{Lin}(w_1, w_2) = \frac{\sum_{f \in T(w_1) \cap T(w_2)} (I(w_1,f)+I(w_2,f))}{\sum_{f \in T(w_1)} I(w_1,f)+\sum_{f \in T(w_2)} I(w_2,f)}$.

3. Experiment Settings and Results

A standard implementation of Support Vector Machines – SVMlight (Joachims, 1999) – is used as the classifier in all the reported experiments. The context is defined as a window of size ± 15, i. e., lemmata of 15 preceeding and 15 following words are taken into account. Stop-list filtering is applied (excluding words *a, the, and, or*, etc.).

Classification on the original Senseval-3 English lexical sample runs on the contexts from this dataset. For the experiments with pseudowords, we extract relevant contexts from the English GigaWord corpus. To guarantee the same conditions, the numbers of train and test examples for pseudowords exactly match those of the individual senses corresponding to the constituents.

Word	Bas	S3	S3 avg.	FineGr
argument	52.2	53.3	52.3±2.2	52.6±2.3
arm	83.0	84.6	84.7±1.1	83.1±1.3
atmosphere	66.7	61.4	62.4±3.4	65.1±3.1
audience	65.2	75.0	74.8±2.6	68.8±5.5
bank	69.9	78.1	78.9±2.3	71.7±2.1
degree	57.4	67.8	64.2±3.0	59.8±2.9
difference	35.7	50.0	47.8±4.5	43.6±10.8
difficulty	18.2	40.9	29.5±8.8	37.7±7.9
disc	41.7	68.1	67.5±3.3	59.0±7.0
image	39.4	59.1	67.6±4.3	62.0±5.4
interest	41.8	48.3	43.7±4.2	42.6±3.6
judgment	28.1	40.6	44.8±5.5	41.0±8.5
organization	75.5	77.4	78.8±2.3	76.5±2.5
paper	26.6	39.4	47.4±3.6	41.8±8.5
party	69.6	74.5	73.3±1.6	72.2±3.7
performance	31.7	41.5	43.3±4.4	41.5±4.7
plan	82.6	87.0	84.8±3.0	81.4±2.6
shelter	41.3	48.8	51.4±4.0	49.5±4.6
sort	61.7	55.6	57.0±2.8	66.0±1.9
source	62.1	65.5	49.7±6.3	56.2±6.7

Table 2: Results on pseudowords with semantically close constituents

Word	S3	Rand	WN	ESA	Lin
argument	52.3±2.2	59.9±6.4	56.0±3.2	59.8±3.2	57.7±2.5
arm	84.7±1.1	84.9±1.4	83.2±1.3	83.0±6.7	84.3±1.1
atmosphere	62.4±3.4	75.2±5.6	73.7±3.9	67.9±3.7	66.7±2.5
audience	74.8±2.6	78.3±6.3	72.3±3.4	80.8±2.9	72.4±2.2
bank	78.9±2.3	72.2±2.0	71.9±1.8	67.3±3.2	77.2±2.1
degree	64.2±3.0	67.8±5.3	59.5±2.0	73.2±2.6	69.2±1.9
difference	47.8±4.5	51.7±7.1	41.2±6.3	45.5±5.3	59.7±4.8
difficulty	29.5±8.8	46.6±12.0	34.1±10.6	38.2±8.4	40.9±9.0
disc	67.5±3.3	72.9±5.0	61.6±4.9	58.6±3.3	76.8±8.1
image	67.6±4.3	76.4±3.9	65.0±6.8	67.1±5.6	62.6±8.1
interest	43.7±4.2	48.8±11.8	43.2±5.0	51.1±5.3	39.3±5.4
judgment	44.8±5.5	59.7±8.1	49.5±6.7	44.1±8.2	43.3±12.0
organization	78.8±2.3	78.0±3.9	77.4±1.2	75.5±2.5	78.5±2.4
paper	47.4±3.6	48.1±3.0	54.2±3.0	41.5±5.2	56.3±5.5
party	73.3±1.6	72.3±1.5	73.7±2.8	70.3±2.0	69.1±1.8
performance	43.3±4.4	51.7±8.1	53.2±4.1	52.6±5.7	40.5±6.4
plan	84.8±3.0	83.7±2.6	82.2±2.8	84.2±1.5	81.9±2.6
shelter	51.4±4.0	63.8±6.5	62.4±5.0	55.8±2.8	61.0±5.1
sort	57.0±2.8	64.2±4.0	64.0±2.6	66.5±2.6	60.4±1.9
source	49.7±6.3	63.0±6.3	53.8±5.8	56.9±6.0	56.2±6.2
Sum of diffs	-	134.4	98.6	112.9	101.2

Table 3: Comparison of pseudoword creation methods, differences between the accuracies on real words and the pseudowords

The first simple experiment shows that the methods computing the relatedness of contexts allow building pseudowords from constituents that are close enough to model any fine-grained sense distinctions. Table 2 summarizes the results. Column *Bas* corresponds to the baseline method that assigns the most frequent label to each test example. *S3* presents the accuracy of the classifier on the original dataset. *S3 avg* shows the average accuracy and its standard deviation for 10 runs on the same data (the same proportion of test and train examples), where the train set is randomly chosen from all the examples (and the rest is used for testing). The last column brings the accuracy values on the pseudowords constructed from unambiguous words that are as close as the most similar senses of the target ambiguous word.

The poor results reported on the pseudowords modelling very fine-grained sense distinctions prove that there is no need to consider pseudowords as an upper bound of the true WSD accuracy only. On contrary, the special construction can set a kind of a lower bound which pinpoints the potential low performance of the disambiguation if the real senses overlap or are extremely difficult to distinguish.

The key question tackled in this paper consists in determining the method that constructs pseudowords and selects their contexts on which the WSD accuracy values match those of the real ambiguous words. Table 3 presents the results of the methods discussed above. Column *S3* repeats the accuracy of the classifier on the original dataset. The remaining columns give the relative difference of the particular method to *S3*. *Rand* refers to the random selection of unambiguous words as pseudoword constituents. The selection of the unambiguous words by means of the wordnet similarity, ESA and Lin's methods are reported in columns *WN*, *ESA* and *Lin* respectively.

The last row of the table shows that the sum of differences of the particular average accuracies to S3. The method based on handcrafted (wordnet) similarities provides the best approximation to the target value. The last column proves that even an automatically derived data (from large corpus) can offer a reasonable base for pseudoword generation. Note, however, that the variation in the S3 column itself is rather high. In some cases (e. g., words *source* and *difficulty*), it is questionable whether the methods should really model the particular contexts tested (or the words should be excluded as outliers).

4. Related Work

(Nakov and Hearst, 2003) pointed out that the standard method of the random selection of pseudoword constituents does not produce a suitable model for real ambiguous words. The paper proposed a lexical category-based method that builds on a medical term hierarchy extracted from MeSH (Medical Subject Headings). The authors favour candidate words that have similar frequencies to the individual senses of the modelled ambiguous words. However, for the experiments on the constructed pseudowords, they focus on the pseudowords with evenly distributed senses only. Even though the paper correctly states that in common texts, the more frequent sense for two-sense words is reported to occur 92 % of the time in average, the results are reported for the "difficult settings" only. Consequently, the demonstrated accuracy of the WSD algorithms is lower than that of randomly chosen candidates.

(Gaustad, 2001) compared results of WSD on the randomly generated pseudowords to the real ambiguous words on the

Senseval-1 dataset. Similarly to our experiments, she set the same value for all parameters and used the same WSD algorithm in order to achieve the most objective results. The paper discovered that the accuracy of the algorithm tested on the pseudowords-based data is much higher (8–20 %) than the accuracy for the same algorithm tested on real ambiguous words.

(Lu et al., 2006) proposed another method for creation of pseudowords. Authors used CUP-Dic – a wordnet-style Chinese semantic lexicon. They searched for an alternative unambiguous word for each sense of real ambiguous words in the same synset only. The method defined a new evaluation data set for Chinese. Unfortunately, no comparison to the randomly selected pseudoword constituents was given.

5. Conclusions and Future Directions

This paper compared two approaches to pseudoword construction. The best results have been obtained by the wordnet-based method that identifies candidates for pseudoword constituents from wordnet lexical relations. It also showed that similar results can be achieved by means of automatic methods computing word relatedness from large corpora.

A very high variability on the Senseval-3 standard dataset has also been observed. This casts doubt upon the stability and reliability of the general comparisons of WSD systems. As mentioned in the text, some applications focus on the non-dominant sense identification. It should be taken into account in the future WSD challenges and a detailed analysis of the performance (per sense), supplemented by the whole set of sense confusion matrices, should be published.

Our future work will deal with advanced context-similarity techniques. Particular attention will be paid to the additional value brought by (shallow) parsing of the word contexts. We will also focus on user feedback integration in the active learning settings.

Acknowledgement

This work was partially supported by the EC's Seventh Framework Programme (FP7/2007-2013) under grant agreement No. 247829, and by the BUT FIT grant FIT-10-S-2 and the research plan MSM0021630528.

6. References

E. Agirre and P. G. Edmonds. 2006. *Word Sense Disambiguation: Algorithms and Applications*. Springer.

I. Dagan, L. Lee, and F. Pereira. 1997. Similarity-based methods for word sense disambiguation. In *Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics*, page 63. Association for Computational Linguistics.

E. Gabrilovich and S. Markovitch. 2007. Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence*, pages 6–12.

W. Gale, K. W. Church, and D. Yarowsky. 1992. Work on statistical methods for word sense disambiguation. In *Probabilistic Approaches to Natural Language: Papers from the 1992 AAAI Fall Symposium*, pages 23–25.

T. Gaustad. 2001. Statistical Corpus-Based Word Sense Disambiguation: Pseudowords vs. Real Ambiguous Words. In *Proceedings of 39th Annual Meeting of ACL (ACL/EACL 2001), Student Research Workshop*. Citeseer.

D. Graff and C. Cieri. 2003. English GigaWord, linguistic data consortium.

N. Ide and Y. Wilks. 2006. Making Sense About Sense. In *Word Sense Disambiguation: Algorithms And Applications*, chapter 3. Springer, Dordrecht.

T. Joachims. 1999. Making large scale SVM learning practical.

D. Lin. 1998. Automatic retrieval and clustering of similar words. In *Annual Meeting-Association for Computational Linguistics*, volume 36, pages 768–774. Association for Computational Linguistics.

Z. Lu, H. Wang, J. Yao, T. Liu, and S. Li. 2006. An equivalent pseudoword solution to Chinese word sense disambiguation. In *Proceedings of the 21st International Conference on Computational Linguisticsand 44th Annual Meeting of the ACL*, page 457.

P. I. Nakov and M. A. Hearst. 2003. Category-based pseudowords. In *NAACL '03: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, pages 67–69, Morristown, NJ, USA. Association for Computational Linguistics.

M. Schmidt and P. Smrz. 2009. Information extraction in semantic wikis. In *Proceedings of the Fourth Workshop on Semantic Wikis*, volume 2009, pages 17–29.

H. Schütze. 1992. Dimensions of meaning. In *Proceedings of the 1992 ACM/IEEE conference on Supercomputing*, pages 787–796. IEEE Computer Society Press Los Alamitos, CA, USA.

D. Yarowsky. 1993. One sense per collocation. In *HLT'93: Proceedings of the workshop on Human Language Technology*, pages 266–271, Morristown, NJ, USA. Association for Computational Linguistics.