Benchmarking word sense disambiguation systems for Swedish

Luis Nieto Piña, Richard Johansson

University of Gothenburg
{luis.nieto.pina, richard.johansson}@gu.se

Abstract

We compare several word sense disambiguation systems for Swedish and evaluate them on seven different sense-annotated corpora. Our results show that unsupervised systems beat a random baseline, but generally do not outperform a first-sense baseline considerably. On a lexical-sample dataset that allows us to train a supervised system, the unsupervised disambiguators are strongly outperformed by the supervised one.

1. Introduction

Word sense disambiguation (WSD) is a difficult task for automatic systems. The most accurate WSD systems build on supervised learning models trained on annotated corpora (Taghipour and Ng, 2015), but because of the difficulty of the sense annotation task and the amount of training data required, the luxury of supervised training for wide-coverage WSD is available for just a few languages.

An approach that circumvents the lack of annotated corpora is to take advantage of the information available in lexical knowledge bases (LKBs) like WordNet (Fellbaum, 1998). This kind of resource encodes word sense lexicons as graphs connecting lexically and semantically related concepts. Several methods are available that use LKBs for WSD (Navigli and Lapata, 2007; Agirre and Soroa, 2009; Johansson and Nieto Piña, 2015a; Nieto Piña and Johansson, 2016). Different approaches range from embedded representations of the LKB to relatively complex analyses of its underlying graph.

In this paper, we present a comparison of several such unsupervised WSD systems evaluated on a series of Swedish language datasets containing sense annotations. For the purpose of illustrating the difference in performance between unsupervised and supervised systems, a supervised model is also included in our experiments.

2. Models

Personalized PageRank (UKB). The UKB system (Agirre and Soroa, 2009) is a graph-based WSD system that carries out disambiguation by applying the Personalized PageRank algorithm to the sense graph.

Post-processed word sense vectors (PP). The system introduced by Johansson and Nieto Piña (2015a) takes advantage of the geometrical interpretation of semantic relatedness in word vector spaces to score the senses of ambiguous words based on their context. A post-processing algorithm (Johansson and Nieto Piña, 2015b) is applied on a word vector space to embed an LKB which describes relationships between word senses. This results in a word *sense* vector space, in which related senses are expected to be located near each other. The WSD task is then tackled by scoring the senses of an ambiguous word in relation to their distance to the words in its context: the highest scoring sense is chosen to disambiguate each instance.

Random walk-based word sense vectors (RW). The method presented by Nieto Piña and Johansson (2016) is closely related to the one above: geometrical distances between concepts in a word sense vector space are used to score the senses of ambiguous words. In this case, however, the word sense vectors are directly extracted from the LKB without the need of a preexistent word vector space. Random walks (RW) on the LKB's underlying graph are used to generate a synthetic corpus: starting a RW on an concept's node in the graph generates a series of related concepts which are treated as the context for the corresponding sense; by repeating this process several times over the LKB's collection of concepts, a collection of synthetic sentences is created which is then used as a corpus to train vector representations. The length of RWs is controlled by a stop probability that parameterizes the model. Once the vectors are obtained, they are used in a scoring scheme similar to the one described above. From the two versions presented in the original paper, the unweighted-graph approach is used here for the purpose of comparison with the UKB model, which is also applied on an unweighted graph.

Supervised word experts (Sup). Supervised systems typically achieve high accuracies in evaluations for languages where sufficient training data is available, which makes it useful to include a system of this type as well. We used a "word expert" approach: one separate SVM classifier was trained for each ambiguous lemma type. The SVMs used a bag-of-words feature representation of a context window, where the words were weighted by their proximity to the target word to disambiguate.

3. Experiments

The WSD systems listed in the previous section were tested on a series of Swedish annotated corpora in which the ambiguous words have been manually disambiguated according to the SALDO lexicon (Borin et al., 2013); random and first-sense baselines are also given for comparison.

3.1 The SALDO Lexicon

SALDO is the largest freely available LKB for Swedish: the version used in this paper contains roughly 125,000 entries defining word senses in terms of semantic network.

The unsupervised systems rely on the structure of the SALDO network, which is defined in terms of semantic *de-*

scriptors. A descriptor of a sense is another sense used to define its meaning. The most important descriptor is called the *primary* descriptor (PD), and since every sense in SALDO (except an abstract root sense) has a unique PD, the PD subgraph of SALDO forms a tree. A sense is typically related to its PD through hyponymy or synonymy, but other relations are also possible.

3.2 Evaluation Corpora

Our first two evaluation datasets, the *SALDO examples* (SALDO-ex) and *Swedish FrameNet examples* (SweFN-ex) consist of 2,364 sentences selected by lexicographers to exemplify the senses (Johansson and Nieto Piña, 2015a). An additional four collections contain 4,811 sentences and are taken from the *Koala* annotation project (Johansson et al., 2016); each collection is sampled from text in one of four domains: blogs, novels, Wikipedia, and European Parliament proceedings. Our final corpus comes from the Swedish Senseval-2 lexical sample (Kokkinakis et al., 2001). It uses a different sense inventory, which we mapped manually to SALDO senses. After removing instances of a few lemmas that were unambiguous in SALDO, we ended up with 7,052 training and 1,246 testing instances. Because this dataset has more instances per lemma type – there are just 33 different types – it is the only one where the supervised system is applicable. We preprocessed the sentences in all seven corpora to tokenize, compound-split, and lemmatize the texts, and to determine the set of possible senses in a given context. We used content words only: nouns, verbs, adjectives, and adverbs.

3.3 Evaluation

We parameterized the models as follows. The stop probability for the RWs on SALDO was 0.25. The word sense vectors were trained on the resulting synthetic corpus using 10 iterations of the training algorithm. Sense vectors used by both the RW and the PP systems had a dimensionality of 200. We used version 2.0 of UKB, run in the *word-by-word* mode, using an unweighted graph based on the PD tree.

The disambiguation mechanism of each model introduced in Section 2 is applied to sentences containing one ambiguous word. A score is then calculated for each of the senses of an ambiguous target word in a context window of size 10 (to each side of the target word) and the highest scoring sense is selected to disambiguate the entry. The accuracy of the method is then obtained by comparing these selections with the annotations of the test datasets.

The results on the test corpora are shown in Table 1. Along the tested models are also included a uniformly random (Rand.) and first-sense (S1) baselines. As we can see, the S1 baseline scores much higher in running-text corpora than in datasets selected by lexicographers.

The result for the Senseval lexical sample dataset clearly illustrates the superiority of supervised systems in such circumstances: even being a relatively simple model, the supervised approach stands out with the highest accuracy. Among the unsupervised models, the post-processing approach performs generally best; all of them beat a random baseline, but it is worth observing how the first sense is still a hard baseline for these models. Notice also the clear influence of the different corpora on the WSD results: while the Koala datasets are sampled from running text, the SALDO and SweFN examples and the Senseval dataset are built to have a good coverage of the sense variation, which results in an overall decrease in accuracy.

Corpus	RW	UKB	PP	Sup.	S1	Rand
SALDO-ex	52.1	55.5	**64.0**	–	53.2	39.3
SweFN-ex	51.0	53.7	**64.2**	–	54.3	40.3
Blogs	49.8	70.0	**74.9**	–	72.4	40.8
Europarl	55.7	67.6	**74.3**	–	67.9	42.3
Novels	56.6	70.1	**78.3**	–	77.2	40.1
Wikipedia	60.4	69.5	**80.4**	–	76.8	41.2
Senseval	38.5	45.0	54.1	**77.9**	50.8	35.7

Table 1: WSD accuracies on the test sets.

4. Conclusion

In the comparison presented in this paper we have shown that unsupervised WSD systems are able to beat a random and, in some cases, a first-sense baseline on WSD tasks. However, these approaches are still far from performing on par with a supervised disambiguator. Nevertheless, the results obtained by some of the systems tested in this paper are promising; given their relative lack of sophistication, they leave ample room for improvements in leveraging existing LKBs and large unlabeled corpora. Thus, we argue that further research on unsupervised systems might be beneficial due to their independence from annotated resources, which are scarce and expensive to produce.

References

Eneko Agirre and Aitor Soroa. 2009. Personalizing PageRank for word sense disambiguation. In *EACL*.

Lars Borin, Markus Forsberg, and Lennart Lönngren. 2013. SALDO: a touch of yin to WordNet's yang. *Language Resources and Evaluation*, 47:1191–1211.

Christiane Fellbaum, editor. 1998. *WordNet: An electronic lexical database*. MIT Press.

Richard Johansson and Luis Nieto Piña. 2015a. Combining relational and distributional knowledge for word sense disambiguation. In *NODALIDA*, pages 69–78.

Richard Johansson and Luis Nieto Piña. 2015b. Embedding a semantic network in a word space. In *Proc. NAACL*, pages 1428–1433.

Richard Johansson, Yvonne Adesam, Gerlof Bouma, and Karin Hedberg. 2016. A multi-domain corpus of Swedish word sense annotation. In *Proc. LREC*.

Dimitrios Kokkinakis, Jerker Järborg, and Yvonne Cederholm. 2001. SENSEVAL-2: The Swedish framework. In *Proc. SENSEVAL-2*, pages 45–48.

Roberto Navigli and Mirella Lapata. 2007. Graph connectivity measures for unsupervised word sense disambiguation. In *Proc. IJCAI*, pages 1683–1688.

Luis Nieto Piña and Richard Johansson. 2016. Embedding senses for efficient graph-based word sense disambiguation. In *Proc. TextGraphs*, pages 2710–2715.

Kaveh Taghipour and Hwee Tou Ng. 2015. One million sense-tagged instances for word sense disambiguation and induction. In *Proc. CoNLL*, pages 338–344.