Probabilistic WSD in SENSEVAL-3

Judita Preiss[*]
University of Cambridge, Computer Laboratory
JJ Thomson Avenue, Cambridge CB3 0FD, UK
Judita.Preiss@cl.cam.ac.uk

Abstract

We evaluate the performance of our modular probabilistic word sense disambiguation system on the SENSEVAL-3 English all words and English lexical sample tasks. All submitted versions of the system outperform the most frequent sense baseline, the best performing system being a combination of all 26 modules. We discuss the usefulness of individual modules.

1 Introduction

The SENSEVAL evaluation exercise allows the WSD community to compare the performance of different WSD systems, by evaluating all systems on identical corpora. We evaluated our modular probabilistic word sense disambiguation system (Preiss, 2004) on the SENSEVAL-3 English all words and the English lexical sample task.

As the WSD system we employ is modular, we go on to discuss the performance of individual modules on the corpus to find the most useful modules. This investigation goes on to suggest a combination of modules which yields a higher performance than that obtained with the submitted systems. Obtaining a higher performance when a number of modules are used together shows once again that a combination system outperforms systems based on a single idea (Stevenson and Wilks, 2001).

We briefly introduce our modular probabilistic WSD system in Section 2. The results (Section 3) are followed by their discussion in Section 4, and we draw our conclusions in Section 5.

2 WSD System

Empirical studies have shown (Stevenson, 2003) that the best WSD performance is obtained

in practice by combining several different approaches. Our WSD system Preiss (2004) is based on combining 26 probabilistic modules using Bayes Rule. Many of the modules employed are based on past successful WSD approaches, such as the work of Yarowsky (2000), Mihalcea (2002), and Pedersen (2002), making it a representative supervised system.

Most of the modules within the WSD system are supervised, and there are only two modules which do not require our system to train them:[1] the *basic part of speech (PoS)* module, and the *frequency* module. The *basic PoS* module uses an HMM tagger due to Elworthy (1994) to obtain a probability distribution on PoSs. The *frequency* module uses a frequency distribution from WordNet to create a probability distribution on senses. Together these two modules generate a value for the most frequent sense baseline.

The remaining 24 modules used all require training data. Seven modules are based on part of speech information (of words one, two, or three places to the left and right, and of the word itself), three are trigram modules (the current word being the first word in the trigram, the middle word, and the last word in the trigram), seven more modules are based on the lemmas of the surrounding words (again words one, two or three places to the left and right, and the word itself). The last module is a window module containing the probability of co-occurrence of words within a window of words fifty places to the left and right of the current word. All the modules produce probability distributions on senses. For example in the *pos0* module (the part of speech of the target word module), the probability of the tag being t given that the sense of the word w is s_i is shown in Figure 1. For il-

---

[1]Note that in this work, we consider a module to be unsupervised if it is not being trained by our system.

SENSEVAL-3: Third International Workshop on the Evaluation of Systems
for the Semantic Analysis of Text, Barcelona, Spain, July 2004
Association for Computational Linguistics

System	Task	Training Data	Description
Prob0	ELS	None	Most frequent sense baseline – combination of the two unsupervised modules
Prob1	ELS	Training data provided	All 26 modules.
Prob2	ELS	Training data provided	Trigram modules, surrounding part of speech modules, head word modules, and the baseline modules.
Prob3	EAW	None	Most frequent sense baseline.
Prob4	EAW	Semcor & ELS training data (excluding verbs)	All 26 modules.
Prob5	EAW	Semcor & ELS training data (excluding verbs)	Trigram modules, surrounding part of speech modules, head word modules, and the baseline modules.

Table 1: System descriptions

lustration, we present three frequency distributions from the *pos0* module for the word *shirt* ($\mathcal{P}(\text{pos0} = \text{NN1}|\text{shirt}_i)$, $\mathcal{P}(\text{pos0} = \text{NN2}|\text{shirt}_i)$, and $\mathcal{P}(\text{pos0} = \text{VVD}|\text{shirt}_i)$) in Table 2. In this table, $f(\text{sense} \cap \text{pos})$ denotes the number of occurrences of w in the given sense with the given PoS tag, and $f(\text{sense})$ is the number of occurrences of *shirt* in the given sense.

The probability distributions produced by the modules need to be smoothed. We use Lidstone's smoothing (e.g., (Manning and Schütze, 1999)), where the optimum smoothing values are empirically determined on a development corpus by an exhaustive search. Using Lidstone's smoothing, we can make the smoothing values word and module specific, and so can make the probability distributions generated resemble uniform distributions if we are not very confident in the module for a given word.

The probability distributions produced by the 26 modules are combined using Bayes Rule:

$$\mathcal{P}(A|B) = \frac{\mathcal{P}(A \cap B)}{\mathcal{B}}$$

The prior distribution comes from the unsupervised PoS and frequency modules, and this is augmented using the remaining modules to produce the best updated estimate in the form of a posterior distribution. Combining modules using Bayes Rule is the best combination method that was tested, and outperforms the natural combination method based on Dempster–Shafer theory.

3   Results

There were three versions of the probabilistic WSD system submitted to the English lexical sample (ELS) task (Mihalcea et al., 2004), and to the English all words (EAW) task (Palmer, 2004). The descriptions of the systems and their training data can be found in Table 1. For the English all words task, the system was trained on SEMCOR 1.6 converted into 1.7.1 using (a heuristics based) automatic mapping method.

Although the output of our probabilistic system is a probability distribution on senses, this was converted a one sense assignment per instance for evaluation.[2] For the English lexical sample task, the "U" (unassignable) tag was output whenever our system gave the highest probability to none of the available senses being relevant. The system also occasionally entirely missed annotating words due to combined errors arising from the morphological decomposition component and the tagger, these were also given the "U" tag, resulting in 100% coverage. In the English all words task, the system always found an available sense. The lower coverage (97.4%) was due to the errors from the morphological and tagger components. The official system performances can be found in Table 3.

4   Discussion

Both Prob2 and Prob5 systems were investigating whether a lower number of modules would yield better performance; when a large number of modules are combined, the difference in probabilities of senses can become quite small. However, the combination chosen[3] was only optimal for the English all words tasks. Subsequent

---

[2]Note that outputting the probability distribution on senses directly would have quite a low maximal possible precision as no sense is assigned a zero probability.

[3]This choice of modules was based on a preliminary investigation with the SENSEVAL-2 English all words

$$\mathcal{P}(\text{tag} = t | \text{sense} = s_i) \quad = \quad \frac{\text{no. of occurrences of } w \text{ in sense } s_i \text{ when the PoS tag of } w \text{ is } t}{\text{no. of occurrences of } w \text{ in sense } s_i}$$

Figure 1: Probability of tag t given the sense is s_i in the *pos0* module

Sense id	PoS (t)	$f(\text{sense} \cap \text{pos})$	$f(\text{sense})$	$\mathcal{P}(t|s_i)$
shirt%1:06:00::	NN1	8	9	$\frac{8}{9}$
shirt%2:29:00::	NN1	0	1	0
shirt%1:06:00::	NN2	1	9	$\frac{1}{9}$
shirt%2:29:00::	NN2	0	1	0
shirt%1:06:00::	VVD	0	9	0
shirt%2:29:00::	VVD	1	1	1

Table 2: Part of speech distributions for the word *shirt*

System	Precision	Recall	Coverage	Precision	Recall	Coverage
ELS	Coarse-grained			Fine-grained		
Prob0	63.6%	63.6%	100.0%	54.7%	54.7%	100.0%
Prob1	71.6%	71.6%	100.0%	65.1%	65.1%	100.0%
Prob2	69.3%	69.3%	100.0%	61.9%	61.9%	100.0%
EAW	With "U"			Without "U"		
Prob3	55,1%	55.1%	100.0%	57.3%	54.7%	97.4%
Prob4	55.4%	55.4%	100.0%	57.5%	55.0%	97.4%
Prob5	57.2%	57.2%	100.0%	58.5%	56.8%	97.4%

Table 3: Results on the EAW and ELS tasks

to the evaluation taking place, we ran a search through the possible module combinations[4] resulting in a better performance with a combination of the most frequent sense modules, the trigram modules, the window module, and the root of word 3 to the left, 1 to the right and 2 to the right. The 'without U' performance for the English all words task with this module combination is 59.5% precision and 58.0% recall.

5   Conclusion

We have presented the results of our probabilistic WSD system on the SENSEVAL-3 English lexical sample and the English all words tasks. In both cases, our system outperformed the baseline for the task.[5] We have shown that the system can be further optimized, but even in its raw form it performs well.

---

task.

[4]This search was run on the English all words task.

[5]This baseline does not have access to perfect part of speech, or untagged word information.

References

D. Elworthy. 1994. Does Baum-Welch re-estimation help taggers? In *Proceedings of the 4th Conference on Applied NLP*, pages 53–58.

C. D. Manning and H. Schütze. 1999. *Foundations of Statistical Natural Language Processing*. MIT Press.

R. Mihalcea, A. Kilarriff, and T. Chklovski. 2004. English lexical sample task. In *Proceedings of SENSEVAL-3*.

R. Mihalcea. 2002. Word sense disambiguation using pattern learning and automatic feature selection. *Journal of Natural Language and Engineering*, 8(4):343–358.

M. Palmer. 2004. English all words task. In *Proceedings of SENSEVAL-3*.

T. Pedersen. 2002. Machine learning with lexical features: The Duluth approach to

Senseval-2. In J. Preiss and D. Yarowsky, editors, *Proceedings of SENSEVAL-2: Second International Workshop on Evaluating Word Sense Disambiguating Systems*, pages 139–142.

J. Preiss. 2004. Probabilistic word sense disambiguation. *Computer Speech and Language*. Forthcoming.

M. Stevenson and Y. Wilks. 2001. The interaction of knowledge sources in word sense disambiguation. *Computational Linguistics*, 27(3):321–349.

M. Stevenson. 2003. *Word Sense Disambiguation: The Case for Combining Knowledge Sources*. CSLI Publications, Stanford, CA.

D. Yarowsky. 2000. Hierarchical decision lists for word sense disambiguation. *Computers and the Humanities*, 34(1/2):179–186.