Adaptive Word Sense Tagging on Chinese Corpus

Sue-Jin Ker
Department of Computer& Information Science
Soochow University
Taipei, Taiwan
ksj@cis.scu.edu.tw

Jen-Nan Chen
Department of Computer Science &
Information Engineering
Ming Chuan University
Taipei, Taiwan
jnchen@mcu.edu.tw

Abstract

This study describes a general framework for adaptive word sense disambiguation. The proposed framework begins with knowledge acquisition from the relatively easy context of a corpus. The proposed framework heavily relies on the adaptive step that enriches the initial knowledge base with knowledge gleaned from the partially disambiguated text. Once adjusted to fit the text at hand, the knowledge base is applied to the text again to finalize the disambiguation decision. The effectiveness of this approach was examined through sentences from the Sinica corpus. Experimental results indicated that adaptation significantly improved the performance of WSD. Moreover, the adaptive approach, achieved an applicability improvement from 33.0% up to 74.9% with a comparable precision.

1 Introduction

Word sense disambiguation is a long-standing problem in natural language understanding. Statistically acquiring sufficient knowledge about a language to build a robust WSD system is extremely difficult. For such a system to be efficient, a large mass of balanced materials must be gathered to cover many idiosyncratic facets of the language. Three issues must be addressed in a lexicalized statistical word sense disambiguation (WSD) model: data sparseness, lack of abstraction, and static learning. First, a word-based model has a multiplicity of parameters that are difficult to measure consistently, even with an extremely large corpus. Under-trained models lead to low precision. Second, word-based models lack a crucial degree of abstraction for a broad coverage system. Third, a static WSD model is probably neither robust nor portable, since it is difficult to construct a model relevant to a broad range of unrestricted texts. Several WSD systems have been created that apply word-based models to a specific domain to disambiguate senses appearing in generally easy contexts with a large number of typically salient words. In an unrestricted text, however, the context is usually diverse and difficult to capture with a lexicalized model; therefore, a corpus-trained system is unlikely to transfer suitably to a new domain.

Generality and adaptability are, therefore, essential to a robust and portable WSD system. An adaptive system, armed with an initial knowledge base extracted from defined words, is superior in two ways to static word-based models trained on a corpus. First, the initial knowledge is sufficiently rich and unbiased for a large portion of text to be disambiguated correctly. Second, based on the initial disambiguation, an adaptation step can then be implemented render the knowledge base more relevant to the task, thus resulting in broader and more precise WSD.

This study explores in detail whether word-based knowledge provides a general solution for disambiguating contexts of unrestricted texts. This method assumes that a major part of a given text is *easy* or prototypical and, therefore, understandable using general knowledge. Adapting contextual representation of word senses to those in the easy context, will hopefully allow us to

interpret the other part, which is normally considered a *hard* context. Adaptation makes the knowledge base more relevant to the text and, therefore, more effective for WSD in a hard context. Experimental results demonstrate the feasibility of this adaptive WSD approach.

The rest of this paper is organized as follows. Section 2 reviews recent WSD literature from the viewpoints of various contextual knowledge types and different representation systems. Section 3 then describes the strategy of using the adapted knowledge base and default is described. Next, Section 4 provides a detailed account of experiments conducted to evaluate the effectiveness of the adaptive approach, including the experiment setup, results and evaluation. Conclusions are finally made in Section 5.

2 Previous Works

Using a machine to select the chosen sense of a polysemous word in a specific context has received increasing interest. Various methods of WSD have been proposed recently in natural language processing literature, and newer ones have rapidly superseded old ideas. Central to these efforts are the contextual knowledge encoded and the way this knowledge is described. This section reviews recent WSD literature from the viewpoints of different forms of contextual knowledge and their representational schemes.

Any scheme for gaining contextual information on word sense must begin with means of identifying the word sense, since word sense is an abstract concept, unclear on the surface. With this completed, the surrounding words to construct a contextual representation of the word sense for WSD. Three approaches are available to divide word senses. First, human means can be used to derive a hand-tagged corpus of word senses. Earlier WSD works adopted this approach and hand tagged the intended sense of each polysemous word in the training corpus (Kelly and Stone 1975; Hearst 1991). Second, the numbered sense entries readily available in a machine-readable dictionary can be taken, with their definitions and examples treated as contextual information (Lesk 1986; Veronis and Ide 1990; Wilks et al. 1990; Guthrie et al. 1991). The third way of eliciting word sense uses linguistic constraints. For instance, three linguistic constraints can be exploited for successful sense tagging and WSD.

One sense per discourse The senses of all instances of a polysemous word are highly consistent within any given document.

One sense per collocation Words in close proximity offer strong and consistent clues to the sense of a target word, conditional on the relative distance, order, and syntactic relationship.

One sense per translation Translations in a bilingual corpus can represent the senses of words.

To exemplify the first constraint, consider the word *suit*. The constraint captures the intuition that if the first occurrence of *suit* is a LAWSUIT sense, then later occurrences in the same discourse are also likely to refer to LAWSUIT (Gale, Church and Yarowsky 1992a). The second constraint reveals that most works on statistical disambiguation have assumed that word sense is closely correlated with particular contextual features, like occurrence of particular words in a window around the ambiguous word. However, Yarowsky (1995) proposed that strong collocations should be identified for WSD. In a bilingual corpus, differences in translations of the polysemous word allowed one to identify the intended sense, particularly in contrasting polysemy. Gale, Church and Yarowsky (1992b) used French translations in parallel texts to disambiguate some polysemous words in English. For instance, the senses of *duty* were typically translated as two different French words, *droit* and *devoir*, respectively, representing

the senses *tax* and *obligation*. Thus, a number of *tax* sense examples of *duty* could be collected by extracting instances of *duty* that were translated as *droit,* and the same could be done for *obligation* sense examples of *duty.*

Once word senses are identified, the context of a particular word sense can then be gathered and encoded in some way for use in the following disambiguation step. At least two ways are available to encode contextual knowledge. The obvious way, the lexicalized representation, is a surface scheme that keeps a weighted list of words occurring in the context of a particular sense. Conversely, the conceptual representation encodes the categories of words that might appear in the context.

3 An Adaptive Method for WSD

The first step of this study was to construct an initial knowledge from training corpus, then to describe how the knowledge was employed to resolve ambiguity for polysemous words in a context.

3.1 Construct an Initial Knowledge from Training Set

To avoid accumulating extraneous information in the knowledge acquisition, sense information was only adopt from the easy text in context during the acquisition phase. First, a preparatory segmentation was made for the training material. Next, target words in each sentence were labelled associated senses. A knowledge base was then construct based on the occurrence frequency of the surrounding target words in each sentence. Finally, the target words' co-occurrence probability was computed and the above descriptive outline of the procedure was summed up as Algorithm 1.

Algorithm 1:

Step 1: Let D_w denote a set of sentence collections consisting of a target word w in training material.

Step 2: For each sense division s of word w, let count(c, w, s) denote the frequency count that word c occurs in all word occurrences for word w in sense s.

Step 3: For each target word w, and its sense division s, the co-occurrence probability Pr(c, w, s) of possible context word c and w is calculated as the following:

$$\Pr(c,w,s) = \frac{count(c,w,s)+a}{\sum_{s' \in senses\ of\ w}(count(c,w,s')+a)} \qquad \text{Eq. 1}$$

3.2 Adaptive Sense Disambiguation

This section took advantage of an adaptive approach to automatically resolve ambiguity for polysemous words in context. Naïve Bayes is a simple but effective text classification algorithm for learning from labelled data alone (Lewis, 1998; McCallum and Nigam, 1998). The parameterization given by Naïve Bayes defines an underlying generative model assumed by the classifier. Considering the sense-tagging task as a classification problem this model assumes each word in a sentence was generated separately from the others, given the word sense.

The adaptive process began with an initial collection of labelled sentences L and one of unlabeled sentences set U. For all instances of polysemous word w in U and each sense s' of w, the sense-conditional probabilities score(s'| w, S) was computed. Next, the ratio R(w, S) was computed for all instances of w in U, and the most secure instance S was picked from U. Following that, the instance was labelled to sense s^1 and S added to L. Next, the same process

was performed on sentences in *U* until they had been disambiguated. Algorithm 2 which gives a formal and detailed description of adaptive WSD, is shown as follows:

Algorithm 2:

Step 1: If un-labeled test set *U* is an empty set, stop.

Step 2: For all instances of polysemous word *w* in *U* and each sense *s'* of *w*, compute the sense-conditional probabilities score(*s'*| *w*, *S*).

$$\text{score}(s'|w,S) = P(s')\prod_{k=1}^{|S|}\Pr(c_k,w,s') \qquad \text{Eq. 2}$$

Step 3: Compute R(*w*, *S*) for all instances of polysemous word *w* in *U*.

$$\text{R}(w,S) = \frac{score(s^1|w,S)}{score(s^2|w,S)} \qquad \text{Eq. 3}$$

where $s^1 = \arg\max_{s \in \text{senses of } w} score(s|w,S)$, and $s^2 = \arg\max_{s \in \text{senses of } w, and\ s \neq s^1} \text{score}(s|w,S)$.

Step 4: Pick the most secure instance *S* with the largest value of R(*w*, *S*) and R(*w*, *S*) is greater than a preset threshold, θ. Then, label sense s^1 to *S* and add it to *L*.

Step 5: Go to Step 1.

4 Experiment and Evaluation

4.1 Experiment

Two experiments were conducted to assess the effectiveness of the proposed method: a WSD experiment with adaptive process and an experiment without adaptive process.

The experimental setup is described in a number of steps as follows. (1) A set of 20 polysemous words was chosen as the target for disambiguation and evaluation. Table 1 lists these words. The senses number from 2 to 8 of these words and their average sense numbers are 3.2. (2) For each polysemous word, a sense division was established based on the Chinese WordNet (Miller, 1990; Fellbaum, 1998; CKIP, 2003). (3) Tests were performed on the sentences from the Sinica corpus (CKIP, 1995). The ambiguity of these testing words in our experiment is shown as Table 1.

4.2 Evaluation

To assess the performance, two human judges were asked to give a sense label to each example of these twenty words in the testing set. The results of running the two programs on the testing set were compared against those of human assessors. The number of test instances and correct assignments in these two experiments were tallied to produce the precision rate for each experiment. Tables 2 and Table 3are summarized the experimental results. Based on the results, the adapting approach was reasonably helpful for WSD, achieving an applicability improvement from 33.0% up to 74.9% with comparable precision. Table 4 described the experimental precision and applicability for each run.

5 Conclusion

This study presented an adaptive approach to word sense disambiguation. Under this novel learning strategy, an initial knowledge set for WSD was first built based on the sense definition in training data. These disambiguated texts can be used to adjust the fundamental knowledge in

an adaptive fashion so to improve disambiguation precision. We have demonstrated that this approach can outperform established static approaches based on direct comparison of experimental results. This level of performance is achieved without lengthy training or the use of a very large training corpus.

Table 1 Ambiguities of experimental testing data sets.

Word	Pos	# of senses	Word	Pos	# of senses	Word	Pos	# of senses
報紙	Na	2	後	Ncd	2	股	Na	4
傳真	Na	2	回	Nf	2	後	Nes	4
臉色	Na	2	節	Nf	2	面	Na	5
靈魂	Na	2	同志	Na	3	頭	Nf	5
信	Na	2	命	Na	3	頭	Na	6
感	Na	2	故	Na	3	股	Nf	8
前	Ncd	2	嫌	Na	3			

Table 2 Experimental results without adaptive approach.

Word	Pos	# of instance	#of tagged	correct	Precision (%)	Applicability (%)
報紙	Na	244	102	94	92.1	41.8
傳真	Na	10	5	5	100	50
臉色	Na	16	8	7	87.5	50
靈魂	Na	53	21	21	100	39.6
信	Na	258	152	151	99.3	58.9
感	Na	5	1	1	100	20
前	Ncd	936	287	265	92.3	30.6
後	Ncd	182	64	63	98.4	35.1
回	Nf	319	124	123	99.1	38.8
節	Nf	25	16	16	100	64
同志	Na	26	8	6	75	30.7
命	Na	108	39	39	100	36.1
故	Na	11	2	2	100	18.1
嫌	Na	14	5	5	100	35.7
股	Na	17	13	13	100	76.4
後	Nes	62	3	3	100	4.8
面	Na	239	35	34	97.1	14.6
頭	Nf	106	44	42	95.4	41.5
頭	Na	303	96	88	91.6	31.6
股	Nf	381	68	66	97	17.8
Total		3315	1093	1044	95.5	33.0

Table 3 Experimental results with adaptive approach.

Word	Pos	# of instance	#of tagged	correct	Precision (%)	Applicability (%)
報紙	Na	244	232	216	93.1	95.1
傳真	Na	10	7	7	100.0	70.0
臉色	Na	16	8	7	87.5	50.0
靈魂	Na	53	30	30	100.0	56.6
信	Na	258	228	225	98.7	88.4
感	Na	5	2	2	100.0	40.0
前	Ncd	936	916	849	92.7	97.9
後	Ncd	182	162	160	98.8	89.0
回	Nf	319	144	128	88.9	45.1
節	Nf	25	16	16	100.0	64.0
同志	Na	26	10	6	60.0	38.5
命	Na	108	55	53	96.4	50.9
故	Na	11	2	2	100.0	18.2
嫌	Na	14	5	5	100.0	35.7
股	Na	17	16	15	93.8	94.1
後	Nes	62	5	4	80.0	8.1
面	Na	239	117	99	84.6	49.0
頭	Nf	106	89	84	94.4	84.0
頭	Na	303	291	250	85.9	96.0
股	Nf	381	149	143	96.0	39.1
Total		3315	2484	2301	92.6	74.9

Table 4 Experimental results for each runs.

Runs	tagged	correct	Precision (%)	Applicability (%)
1	1093	1044	95.5	33.0
2	2185	2043	93.5	65.9
3	2440	2265	92.8	73.6
4	2484	2301	92.6	74.9

Acknowledgements

The authors would like to thank the National Science Council of the Republic of China for partially supporting this research under Contract No. NSC 92-2411-H-031-014-ME.

References

CKIP. 1995. The Content and Illustration of Sinica Corpus of Academia Sinica, Technical Report No. 95-02.

CKIP. 2003. The sense and semantic of Chinese Word, Technical Report No. 03-01, 03-02.

Fellbaum C. 1998. *WordNet: An Electronic Lexical Database*, The MIT Press.

Gale, W. A., K. W. Church, and D. Yarowsky. 1992b. Using Bilingual Materials to Develop Word Sense Disambiguation Methods. In *Proceedings of the 4th International Conference on Theoretical and Methodological Issues in Machine Translation*, 101-112.

Gale, W. K., W. Church and D. Yarowsky. 1992a. One Sense Per Discourse, In *Proceedings of the Speech and Natural Language Workshop*, 233-237.

Guthrie, J., L. Guthrie, Y. Wilks and H. Aidinejad. 1991. Subject-dependent Co-occurrence and Word Sense Disambiguation. In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, 146-152.

Hearst, M. 1991. Noun Homonym Disambiguation using Local Context in Large Text Corpora. In *Proceedings of the 7th International Conference on of UW Centre for the New OED and Text Research: Using Corpora*, pages 1-22.

Kelly, E. and P. Stone. 1975. *Computer Recognition of English Word Senses*, North-Holland, Amsterdam.

Lesk., M. E. 1986. Automatic Sense Disambiguation using Machine Readable Dictionaries: How to Tell a Pine Cone from an Ice Cream Cone. In *Proceedings of the ACM SIGDOC Conference,* 24-26, Toronto, Ontario.

Lewis, D. D. 1998. Naïve Bayes at forty: The independence assumption retrieval. In *Proceedings of ECML-98.*

McCallum, A. and K. Nigam. 1998. A comparison of event models for Naïve Bayes Text Classification. In *AAAI-98 Workshop on Learning for Text Classification*.

Miller, G. A., 1990. "WordNet: An Online Lexical Database. In *Special Issue of International Journal of Lexicography*, 3(4). Veronis J. and N. Ide. 1990. Word Sense Disambiguation with Very Large Neural Networks Extracted from Machine Readable Dictionaries. In *Proceedings of the 13th International Conference on Computational Linguistics*, 389-394.

Veronis J. and N. Ide. 1990. Word Sense Disambiguation with Very Large Neural Networks Extracted from Machine Readable Dictionaries. In *Proceedings of the 13th International Conference on Computational Linguistics*, 389-394.

Wilks Y. A., D. C. Fass, C. M. Guo, J. E. McDonald, T. Plate and B. M. Slator. 1990. Providing Tractable Dictionary Tools. *Machine Translation*, 5, 99-154.

Yarowsky. D. 1995. Unsupervised Word Sense Disambiguation Rivaling Supervised Methods. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics*, 189-196.