

A Concept-based Adaptive Approach to Word Sense Disambiguation

Jen Nan Chen
Department of Computer Science
National Tsing Hua University
Hsinchu 30043, Taiwan
jnchen@mcu.edu.tw

Jason S. Chang
Department of Computer Science
National Tsing Hua University
Hsinchu 30043, Taiwan
jschang@cs.nthu.edu.tw

Abstract

Word sense disambiguation for unrestricted text is one of the most difficult tasks in the fields of computational linguistics. The crux of the problem is to discover a model that relates the intended sense of a word with its context. This paper describes a general framework for adaptive conceptual word sense disambiguation. Central to this WSD framework is the sense division and semantic relations based on topical analysis of dictionary sense definitions. The process begins with an initial disambiguation step using an MRD-derived knowledge base. An adaptation step follows to combine the initial knowledge base with knowledge gleaned from the partial disambiguated text. Once the knowledge base is adjusted to suit the text at hand, it is then applied to the text again to finalize the disambiguation result. Definitions and example sentences from LDOCE are employed as training materials for WSD, while passages from the Brown corpus and Wall Street Journal are used for testing. We report on several experiments illustrating effectiveness of the adaptive approach.

1 Introduction

Word sense disambiguation for unrestricted text is one of the most difficult tasks in the fields of computational linguistics. The crux of the problem is to discover a model that relates the intended sense of a word with its context. It seems to be very difficult, if not impossible, to statistically acquire enough word-based knowledge about a language necessary to build a robust system capable of automatically

disambiguating senses in unrestricted text. For such a system to be effective, a great deal of balanced materials must be assembled in order to cover many idiosyncratic aspects of the language. There exist three issues in a lexicalized statistical word sense disambiguation (WSD) model - data sparseness, lack of a level of abstraction, and static learning strategy. First, word-based models have a plethora of parameters that are difficult to estimate reliably even with a very large corpus. Under-trained models lead to low precision. Second, word-based models lack a degree of abstraction that is crucial for a broad coverage system. Third, a static WSD model is unlikely to be robust and portable, since it is very difficult to make a single static model relevant to a wide variety of unrestricted texts. Recent WSD systems have been developed using word-based model for specific limited domain to disambiguate senses appearing in usually easy context (Leacock, Towell, and Voorlees 1996) with a lot of typical salient words. For unrestricted text, however, the context tends to be very diverse and difficult to capture with a lexicalized model, therefore a corpus-trained system is unlikely to port to new domains and run off the shelf.

Generality and adaptiveness are therefore key to a robust and portable WSD system. A concept-based model for WSD requires less parameter and has an element of generality built in (Liddy and Paik 1993). Conceptual classes make it possible to generalize from word-specific context in order to disambiguate a word sense appearing in a particularly unfamiliar context in term of word recurrences. An adaptive system armed with an initial lexical and conceptual knowledge base extracted from machine-readable dictionaries (MRDs), has two strong advantages over static lexicalized models trained using a corpus. First, the initial

knowledge is rich and unbiased such that a substantial portion of text can be disambiguated precisely. Second, based on the result of initial

disambiguated text. Subsequently, the knowledge base is adjusted to suit the text at hand. The adjusted knowledge base is then

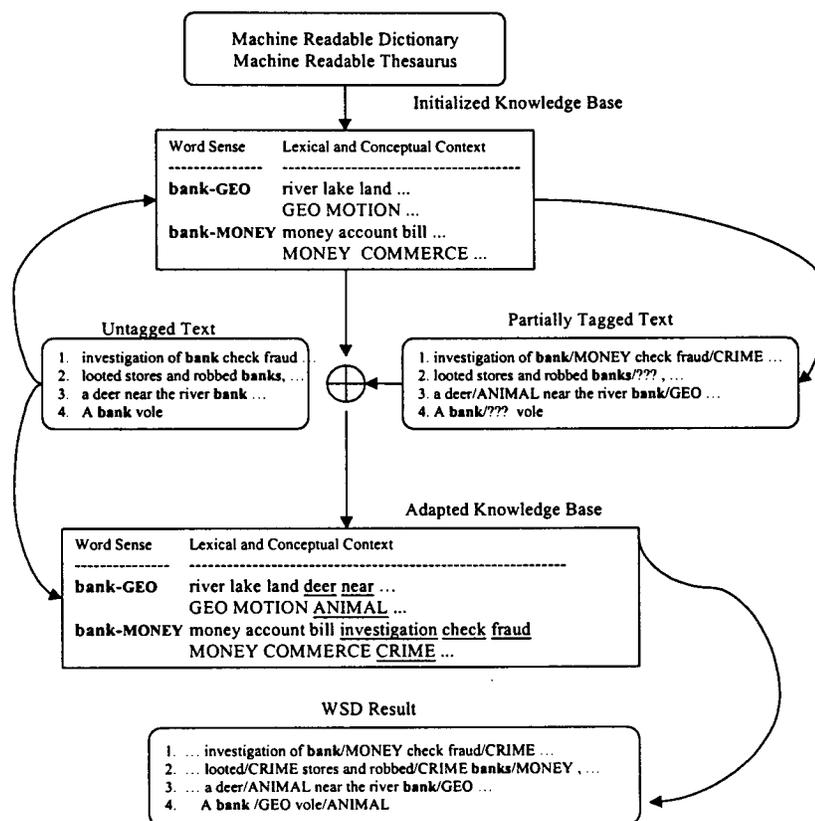


Figure 1 General framework for WSD using MRD.

disambiguation, an adaptation step is taken to make the knowledge base more relevant to the task at hand, leading to broader and more precise WSD.

Figure 1 lays out the general framework for an adaptive conceptual WSD approach, under which this research is being carried out. The learning process described here begins with a step of knowledge acquisition from MRDs. With the acquired knowledge, the system reads the input text and starts the step of initial disambiguation. Adaptive step follows to combine the initial knowledge base with knowledge gleaned from the partially

applied to the text again to finalize the disambiguation result. For instance, Figure 1 shows the initial contextual representation (CR) extracted from the Longman Dictionary of Contemporary English (Protor 1978, LDOCE) for the GEO-bank sense contained both lexical and conceptual information: $\{land, river, lake, \dots\} \cup \{GEO, MOTION, \dots\}$. The initial CR is informative enough to disambiguate a passage containing *a deer near the river bank* in the input text. The initial disambiguation step produces sense tagging of *deer/ANIMAL* and *bank/GEOGRAPHY*, but certain instances of *bank* are left untagged for lack of relevant WSD

knowledge. For instance, the GEO-bank sense in the context of *vole* is unresolved since there is no information linking ANIMAL context to GEOGRAPHY sense of *bank*. The adaptation step adds *deer* and ANIMAL to the contextual representation for GEO-bank. The enriched CR therefore contains information capable of disambiguating the instance of *bank* in the context of *vole* to produce final disambiguation result.

2 Acquiring Conceptual Knowledge from MRD

In this section we apply a so-called *TopSense* algorithm (Chen and Chang 1998) to acquire CR for MRD senses. The current implementation of *TopSense* uses the topical information in Longman Lexicon of Contemporary English (McArthur 1992, LLOCE) to represent WSD knowledge for LDOCE senses. In the following subsections we describe how that is done.

2.1 Contextual Representation from MRDs

Dictionary is a text whose subject matter is a language. The purpose of dictionary is to provide definitions of word senses, and in the process it supply knowledge not just about the language, but the world (Wilks et al. 1990). A good-sized dictionary usually has a large vocabulary and good coverage of word senses useful for WSD. However, short MRD definitions and examples per se lack a level of abstraction to function effectively as a contextual representation of word sense. On the other hand, the thesaurus organizes word senses into a fixed set of coarse semantic categories and thus could potentially be useful as the basis of a conceptual CR of word sense. To get the best of both worlds of dictionary and thesaurus, we propose to link an MRD sense to thesaurus categories to produce conceptual representation of its context. Content words extracted directly from the definition sentence of a word sense can be put to use as the word-level contextual representation of that particular word sense.

One way of producing such conceptual CR is to link MRD senses to their relevant thesaurus

senses and categories. These links furnish the MRD senses with information necessary for building a conceptual CR. We will describe one such approach under which each MRD sense is linked to a relevant thesaurus sense according to its defining words. The linked thesaurus sense, unlike the isolated MDR sense, falls within a certain semantic category. Consequently, we can establish relations between defining words and semantic category that eventually lead to conceptual CR.

With the word lists in a thesaurus category cast as a document representing a certain subject matter or topic, the task of constructing conceptual representation of context for a certain MRD sense bears a striking resemblance to the document retrieval task in information retrieval (IR) research. Relatively well-established IR techniques of weighting terms and ranking documents are applied to build a list of topics that are most relevant to the definition of each MRD sense. This list of ranked topics, for a particular word sense, forms a vectorized conceptual representation of context in the space of all possible topics.

2.2 Illustrative Example

One example is given in this subsection to illustrate how *TopSense* works.

Example 1. Conceptual representation of an LDOCE sense

crane.1.n.1. a machine for lifting and moving heavy objects by means of a very strong rope or wire fastened to a movable arm (JIB).

For the most relevant topics to fine-grained sense, we get the following ranked list **Hd** (EQUIPMENT), **Ha** (MATERIALS), **Ma** (MOVING).

Furthermore, the definition and examples of a particular sense on the surface level seldom are information sufficient to represent context of the sense. For instance, the words *machine*, *lift*, *move*, *heavy*, *object*, *strong*, *rope*, *wire*, *fasten*, *movable*, *arm*, *jib* in the definition of the sense, **crane.1.n.1**, are hardly enough contextual information to resolve a **crane.1.n.1** instance in the Brown corpus shown below:

Unsinkable slowed and stopped, hundreds of brilliant white flares swayed eerily down from

the black, the air raid sirens ashore rose in a keening shriek, the anti-aircraft guns coughed and chattered- and above it all *motors* roared and the bombs came whispering and wailing and crashing down among the ships at anchor at Bari. They had come from airports in the Balkans, these hundred-odd Junkers 88's. They had winged over the Adriatic, they had taken Bari by complete surprise and now they were battering her, attacking with deadly skill. They had ruined the radar warning system with their window, they had made themselves invisible above their *flares*. And they also had the lights of the city, the port wall *lanterns*, and a shore crane's spotlight to guide on.

However, with a level of abstraction made possible by using a thesaurus, it is not difficult to build a conceptual CR of word sense, which is intuitively more effective for WSD. For instance, based on LLOCE topics, the conceptual CR (EQUIPMENT, MATERIALS, MOVING) derived from the definition of **crane.1.n.1**, is general enough to characterize many salient words appearing in the context of the **crane.1.n.1** instance, including *motor* (EQUIPMENT), *lantern* (EQUIPMENT), and *flare* (EQUIPMENT, MATERIALS).

3 The Adaptive WSD Algorithm

We sum up the above descriptions and outline the procedure for the algorithm in this section. In what follows an adaptive disambiguation algorithm based on class-based approach will be described. Next, we give an illustrative example to show how the proposed algorithm works for unrestricted text.

3.1 The algorithm

The proposed algorithm starts with the step of initial disambiguation using the contextual representation $CR(W, S)$ derived from the MRD for the sense S of the head entry W . A step of adaptation followed to produce a knowledge base from the partially disambiguated text. Finally, the undisambiguated part is disambiguated according to the newly acquired knowledge base. The following algorithm gives a formal and detailed description of adaptive WSD.

Algorithm *AdaptSense*

Step 1: Preprocess the context and produce a list of lemmatized content words $CON(W)$ in a polysemous word W 's context.

Step 2: For each sense S of W , compute the similarity between the context representation $CR(W, S)$ and topical context $CON(W)$.

$Sim(CR(W, S), CON(W))$

$$= \frac{\sum_{t \in M} (W_{t,s} + W_t)}{\sum_{t \in CR(W,S)} W_{t,s} + \sum_{t \in CON(W)} W_t}, \text{ where}$$

$M = CR(W, S) \cap CON(W)$,

$W_{t,s}$ = weight of a contextual word t
with sense S in $CR(W, S)$,

$$W_t = \text{weight of } t \text{ in } CON(W) = \frac{1}{\sqrt{|X_t|}}$$

X_t = distance from t to W in number of words.

Step 3: For each word W , choose a relevant sense S_w if passes a preset threshold then construct triples $T = \{(W, S, CON(W))\}$.

Step 4: Compute a new set of contextual representation $CR(W, S) = \{u \mid u \in CON(W) \text{ and } (W, S, CON(W)) \in T\}$

Step 5: Infer remaining less relevant sense for W in CON

3.2 Illustrative Example

Consider the following passage from the Brown corpus:

... Of cattle in a pasture without throwin' 'em together for the purpose was called a "pasture count". The counters rode through the pasture countin' each bunch of grazin' cattle, and drifted it back so that it didn't get mixed with the uncounted cattle ahead. This method of countin' was usually done at the request, and in the presence, of a representative of the **bank** that held the papers against the herd. The notes and mortgages were spoken of as "cattle paper". A "book count" was the sellin' of cattle by the books, commonly resorted to in the early days, sometimes much to the profit of the seller. This led to the famous sayin' in the Northwest of the "books won't freeze". This became a common byword durin' the ...

In our experiment, we observed that *hold* and *paper* are related to both MONEY and ROAD sense in the initial knowledge base.

Thus, this instance of *bank* is left unresolved in the initial disambiguation step. The adaptation step discovers that both *hold* and *paper* co-occur with some MONEY-bank instances in the partially disambiguated text. Therefore, the system is able to correctly resolve this bank instance to MONEY sense.

4 Experiments and Discussions

4.1 Experiment

In our experiment, we use the materials of text windows of 50 words to the left and 50 words to the right of thirteen polysemous words in the Brown corpus and a sample of Wall Street Journal articles. All instances of these thirteen words are first disambiguated by two human judges. For these thirteen words under investigation, only nominal senses are considered. The experimental results show that the adaptive algorithm disambiguated correctly 71% and 77% of these test cases in the Brown corpus and the WSJ sample. Table 1 provides further details. However, there are still room for improvement in the area of precision. Evidence have shown that by exploiting the constraint of so-called “one sense per discourse,” (Gale, Church and Yarowsky 1992b) and the strategy of bootstrapping (Yarowsky 1995), it is possible to boost coverage, while maintaining about the same level of precision.

4.2 Discussions

Although it is often difficult to compare studies on different text domain, genre and experimental setup, the approach presented here seems to compare favorably with the experimental results reported in previous WSD research. Luk (1995) experiments with the same words we use except the word *bank* and reports that there are totally 616 instances of these words in the Brown corpus, (slightly less than the 749 instances we have experimented on). The author reports that 60% of instances are resolved correctly using the definition-based concept co-occurrence (DBCC) approach. Leacock et al. (1996) report that precision rate of 76% for disambiguating the word *line* in a sample of

WSJ articles.

One of the limiting factors of this approach is the quality of sense definition in the MRD. Short and vague definitions tend to lead to inclusion of inappropriate topics in the contextual representation. Using inferior CR, it is not possible to produce enough and precise samples in the initial step for subsequent adaptation.

Table 1(a) Disambiguation results for thirteen ambiguous words in Brown corpus.

Word	# of senses	# of instances	Without adaptation	With adaptation
			# of correct	
bank	8	97	68	71
bass	2	16	16	16
bow	5	12	3	3
cone	2	14	14	14
duty	2	75	67	69
galley	3	4	4	4
interest	4	346	213	228
issue	4	141	67	88
mole	2	4	2	2
sentence	2	32	30	30
slug	5	8	4	6
star	6	46	28	29
taste	3	51	36	36
Total		846	552	596
Precision			65.2%	70.5%

Table 1(b) Disambiguation results for thirteen ambiguous words in Wall Street Journal articles.

Word	# of senses	# of instances	Without adaptation	With adaptation
			# of correct	
bank	8	370	350	353
bass	2	2	2	2
bow	5	-	-	-
one	2	-	-	-
duty	2	25	19	22
galley	3	-	-	-
interest	4	221	123	127
issue	4	260	181	177
mole	2	-	-	-
sentence	2	12	11	12
slug	5	-	-	-
star	6	7	3	2
taste	3	6	3	3
Total		903	692	698
Precision			76.6%	77.3%

The experiment and evaluation shows that adaptation is most effective when a high-frequency word with topically contrasting senses is involved. For low-frequency senses such as EARTH, ROW, and ROAD senses of *bank*, the approach does not seem to be very effective. For instance the following passage containing an instance of *bank* has the ROW sense but our algorithm fails to disambiguate it.

... They slept- Mynheer with a marvelously high-pitched snoring, the damn seahorse ivory teeth watching him from a bedside table. In the ballroom below, the dark had given way to moonlight coming in through the bank of french windows. it was a delayed moon, but now the sky had cleared of scudding black and the stars sugared the silver-gray sky. Martha Schuyler, old, slow, careful of foot, came down the great staircase, dressed in her best lace-drawn black silk, her jeweled shoe buckles held forward.

Non-topical sense like ROW-bank can appeared in many situations, thus are very difficult to captured using a topical contextual representation. Local contextual representation might be more effective.

Infrequent and non-topical senses are problematic due to data sparseness. However, that is not specific to the adaptive approach, all other approaches in the literature suffer the same predicament. Even with a static knowledge acquired from a very large corpus, these senses were disambiguated at a considerably lower rate.

5 Related approaches

In this section, we review recent WSD literature from the prospective of types of contextual knowledge and different representational schemes.

5.1 Topical vs. Local Representation of Context

5.1.1 Topical Context

With topical representation of context, the context of a given sense is reviewed as a bag of words without structure. Gale, Church and Yarowsky (1992a) experiment on acquiring topical context from substantial bilingual training corpus and report good results.

5.1.2 Local Context

Local context includes the structured information on word order, distance, and syntactic feature. For instance, the local content of *a line from* does not suggest the same sense for the word *line* as *a line for* does. Brown et al. (1990) use the trigram model as a way of resolving sense ambiguity for lexical selection in statistical machine translation. This model makes the assumption that only the previous two words have any effect on the translation, thus word sense, of the next word. The model attacks the problem of lexical ambiguity and produces satisfactory results, under some strong assumption. A major problem with trigram model is that of long distance dependency. Dagan and Itai (1994) indicate that two languages are more informative than one; an English corpus is very helpful in disambiguating polysemous words in Hebrew text. Local context in the form of lexical relations are identified in a very large corpus. Brown, et al. (1991) describe a statistical algorithm for partitioning word senses into two groups. The authors use mutual information to find a contextual feature that most reliably indicates which of the senses of the French ambiguous word is used. The authors report a 20% improvement in the performance of a machine translation system when the words are first disambiguated this way.

5.2 Static vs. Adaptive Strategy

Of the recent WSD systems proposed in the literature, almost all have the property that the knowledge is fixed when the system completes the training phase. That means the acquired knowledge never expands during the course of disambiguation. Gale, et al. (1992a) report that if one had obtained a set of training materials with errors no more than twenty to thirty percent, one could iterate training materials selection just once or twice and have training sets that had less than ten percent errors. The adaptive approach is somehow similar to their idea of incremental learning and to the bootstrap approach proposed by Yarowsky (1995). However, both approaches are still considered static models which are changed only in the training phase.

6 Conclusions

We have described a new adaptive approach to word sense disambiguation. Under this learning strategy, first contextual representation for each word sense is built from the sense definition in MRD and represented as a weighted-vector of concepts represented as word lists in a thesaurus. Then the knowledge base is applied to the text for WSD in an adaptive fashion to improve on disambiguation precision. We have demonstrated that this approach has the potential of outperforming established static approaches. This performance is achieved despite the fact no lengthy training time or a very large corpus is required. It is evident that the WSD algorithms proposed herein are simple, take up little time and space, and most importantly, require no human intervention in all phases of WSD. Sense tagging of training material, knowledge acquisition from training data, and disambiguation all are done automatically.

Acknowledgements

This work is partially supported by ROC NSC grants 84-2213-E-007-023 and NSC 85-2213-E-007-042. We are grateful to Betty Teng and Nora Liu from Longman Asia Limited for the permission to use their lexicographical resources for research purpose. Finally, we would like to thank the anonymous reviewers for many constructive and insightful suggestions.

References

- Brown, P. F., S. A. Della Pietra, V. J. Della Pietra, and R. L. Mercer (1991). *Word-sense disambiguation using statistical methods*. In Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics, pp 264-270.
- Chen, J. N. and J. S. Chang (1998). *Topical clustering of MRD senses based on information retrieval techniques*. Special Issue on Word Sense Disambiguation, Computational Linguistics, 24(1), pp 61-95.
- Dagan, I. and A. Itai (1994). *Word Sense Disambiguation Using a second language monolingual corpus*. Computational Linguistics, 20(4), pp 563-596.
- Gale, W. A., K. W. Church, and D. Yarowsky (1992a). *Using bilingual materials to develop word sense disambiguation methods*. In Proceedings of the 4th International Conference on Theoretical and Methodological Issues in Machine Translation, pp 101-112.
- Gale, W. A., K. W. Church and D. Yarowsky (1992b). *One sense per discourse*. In Proceedings of the Speech and Natural Language Workshop, pp 233-237.
- Leacock, C., G. Towell, and E. M. Voorlees (1996). *Towards building contextual representations of word senses using statistical models*. In B. Boguraev and J. Pustejovsky, editor, *Corpus Processing for Lexical Acquisition*. MIT Press, Cambridge, MA.
- Liddy, E. D. and W. Paik (1993). *Document filtering using semantic information from a machine readable dictionary*. In Proceedings of the Workshop on Very Large Corpora, pp 20-29.
- Luk, A. K. (1995). *Statistical sense disambiguation with relatively small corpora using dictionary definitions*. In Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics, pp 181-188.
- McArthur, T. (1992). *Longman lexicon of contemporary English*. Longman Group (Far East) Ltd., Hong Kong.
- Proctor, P. (ed.) (1978). *Longman dictionary of contemporary English*. Harlow: Longman Group.
- 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, pp 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, pp 189-196.