

Chinese Word Sense Disambiguation Based on Lexical Semantic Ontology

Li Li¹ Qiang Zhou²

¹ School of Literature, Communication University of China, Beijing 100024, P. R. China

liliorosa@cuc.edu.cn

².Centre for Speech and Language Technologies, Research Institute of Information Technology, CSLT, Division of Technical Innovation and Development
Tsinghua National Laboratory for Information Science and Technology

Tsinghua University, Beijing 100084, P. R. China

zq-lxd@mail.tsinghua.edu.cn

Abstract

This paper describes preliminary works of word sense disambiguation on Chinese verbs using the information derived from lexical semantic ontology (LSO). In spite of sophisticated methods, simple algorithm is employed to underline the characters of the features chosen from LSO data. Several groups of tests are designed to find different effects of the features and other aspects. Some promising results are gotten from the prime tests on nine Chinese ambiguous verbs. The results show what informative features the LSO provides and the potential improving ways.

Keywords

Word sense disambiguation, lexical semantic ontology, Chinese verbs, supervised classification algorithm.

1. Introduction

Word sense disambiguation (WSD) is to assign appropriate meaning to a given ambiguous word in a context. Corpus based method is one of the successful approaches to meet this goal. There are also many different supervised learning algorithms which have been applied for WSD, for instance, example based learning (Ng and Lee, 1996),

Bayesian learning (Leacock et al., 1998), decision list (Yarowsky, 2000), maximum entropy method (Dang et al., 2002) etc.. The context of the ambiguous word is the source of the information for WSD, which is represented by feature set. Most of the information comes from the local context, such as the local collocations, part of speech information, syntax information and so on. An informative feature set will be more helpful for the classifiers to accurately disambiguate word senses than an uninformative feature set using the same classification algorithm, e.g. Niu and Ji improved their simple classifier to 60.4% through optimizing feature sets in the Chinese lexical sample task in Senseval-3 (Niu and Ji, 2004). The results of the multilingual Chinese-English lexical sample task in Semeval-2007 may represent the highest level of Chinese word sense disambiguation with the micro-average precision of 71.7% and the macro-average precision of 74.9 (Niu and Ji, 2007).

Many Chinese language resources are built to meet different kinds of research purpose, some of which can be used in WSD assignment or improve the results. Combining the information that those resources provide effectively is also one of the feasible way to get informative features for the WSD task.

In spite of comparing different classification methods to get better results, in this paper, our main purpose is to examine how the features gotten from certain resources perform on the WSD task, and choose a better feature set. We use a relatively simple supervised algorithm to try out a small-scale WSD process on Chinese verbs.

2. Resource

Our resource for WSD is lexical semantic ontology (LSO)[9]. LSO is a linguistic resource for the purpose of semantic computation. It integrates information from several resources, such as meanings and semantic categories of words from lexicons, syntactically and semantically labeled sentences from manually annotated corpus etc.

LSO hierarchy is built as a semantic lexicon, which organizes the verbs in the LSO into certain categories according to their meanings. It defines several super classes representing different kinds of situations, for example, classes initial with “E” represents general existence. Subclasses are defined in hierarchic ranks, for instance, ‘E1-1-1’ refers to the existence situation, ‘E1-2-1’ refers to an appearance situation, and they are both the subclasses of the general existence. Other super classes include “Oth” and “New”, which represent a compound category. They hold the verbs with meanings that the LSO hierarchy has not defined yet.

The verb with certain meaning is treated as a node in the hierarchy. LSO hierarchy and the LSO nodes constitute a network system. To a Chinese target verb, it provides all the possible meanings of the verb defined in the LSO hierarchy and other semantic lexicons as well, associated words and expressions annotated with syntactic function tags and semantic role tags in real text sentences and other distributional information. All of them are combined in LSO records.

A LSO record consists an ID index, target word information, sense descriptions from lexicons, lexical relation descriptions derived from annotated corpus and description of relations with other LSO records. For example, one of the LSO records of Chinese verb 'you3' has the form as described in table1.

With the LSO record above, we can get the sense descriptions of the target word, its syntactical and semantic relations with the related words and the relation with other records.

3. Method

3.1 Targets and Algorithm

Naming the ambiguous words as target words, our target words are ambiguous Chinese verbs. The definition of the LSO hierarchy is used as the criterion of the target words' meanings. The WSD context is the sentence in which the target verb is the primary predicate. We build instances based on the description of the related words in LSO records.

Naïve Bayes classifier is applied to perform our WSD task, since it is easy to build, dose not contain any process of automatic feature selection, and has satisfying performance as other sophisticated algorithms (Witten and Frank, 2005) . The Naïve Bayes classifier can calculate the possibilities of a case being each type and choose the best result, given the over all possibilities of the types could be and the feature values of the cases whose types are already known. It is founded on a hypothesis that all the features involved are independent to each other. Though it might not in that condition, it helps us to study the influence to the WSD result of every single feature we choose.

ID (of this record): 0					
Target word: you3					
Description in different lexicons and frequency information from the manually annotated corpus					
Lexicon 1		Lexicon 2		Lexicon 3	
Sense information	Frequency	Sense information	Frequency	Sense information	Frequency
E1-1-1 exist(x,L) + [L=loctim]	2967	V1.111.1 exist 存在	2952	领有, 具有, 存在, 跟 ‘无, 没’ 相对	2584
Lexical relation descriptions derived from the manually annotated corpus					
Related words:	Part of speech tags of the related words:	Semantic category of the related words:	Syntactic relation between the target word and the related words:	Semantic relation between the target word and the related words:	
Lao3Shi1	N	Human	Subject	Agent	
.....	
Relations with other LSO records: ID: 398 (Target words: you3in record 0 and you3 in record 11. Relation: record 11has tight relation to record 0, They are described in record 398.)					

Table 1. An example of a LSO record.

3.2 Feature sets

The purpose of our test is to measure what information LSO can provide to our WSD task and how it performs. Therefore, the algorithm we use does not contain any process of feature selection, and the features are chosen from the LSO data manually. We also want to find a better feature combination dealing with the WSD procedure of the target verb. Thus in addition to testing the performance with all the features, we also examine some subsets.

The data in our training set is manually annotated with correct part-of-speech tags and chunk information according to syntactic relations. We also get the head words of each chunk as related words of the target verb. The meanings of every content word in each instance are derived from Hownet¹, and organized into thirteen major categories. So every substantive in each training case is given a semantic category tag.

Feature set 1 is designed to get a baseline for our classifier before the test with features from the LSO. It comprises all the content words in the context, their part-of-speech tags and semantic category tags. The feature set simulates a word bag consists all the words in the sentence.

Feature set 2 is designed to show the performance of the features derived from the LSO. It consists of all the associated words (head words of the chunks), their syntactic function tags, part-of-speech tags and semantic category tags in the context. With feature set 2, we hope the performance will be improved greatly by adding the syntactic information of the context.

Beside the two basic feature set above, we vary feature set 2 into several subsets to look into the different influences among the features on the WSD result and try to get a better performance.

Part-of-speech tags are removed in feature set 2-1. Because most of the related words are noun, this feature is supposed to have little influence on the WSD results. The feature related words and semantic category tags can be treated as opposite sides of the semantic representation of the associated words, since the related words of the target verb are specific, and the semantic categories of the related words are more abstract and more general. Related words are removed in feature set 2-2 and semantic category tags are removed in feature set 2-3. Using semantic categories instead of related words is to try out a more abstract feature representing the semantic information

¹ Including the related words. Hownet is a lexicon. Dong Z. D, Dong Q. (2002). Hownet. <http://www.keenage.com>

of the associated words. In our test, each semantic category is very large, hence very abstract. We also hope to search the difference between the effects of these two features in the WSD results of different target verbs.

4. Tests and results

Verb	Sense Type	Number of instances	Total
You3	E1-1-1 exist(x,L)+[L=loctim]	244	583
	H1-1-1 have(x,y)+NULL	228	
	H1-3-1 contain(x,y)+NULL	41	
	Oth	70	
Mei2you3	E1-1-2 NOT_exist(x,L)+[L=loctim]	67	121
	H1-1-2 NOT_have(x,y)+NULL	54	
Gei3	H3-1-2-1 DO(x,P(x,y))_CAUSE_(NOT_have(x,y)_&_have(z,y))+[P=Give]	43	62
	H3-1-2-2 DO(x,P(x,y))_CAUSE_(NOT_have(x,y)_&_have(z,y))+[P=Provide]	19	
De2	H2-1-1 do(x,~)_CAUSE_(have(x,y)_&_NOT_have(z,y))+NULL	22	34
	Oth	12	
Fa1chu1	Oth	7	10
	E3-2-2-1 DO(x,P(x,y))_CAUSE_appear(y,L)+[P=Announce]	3	
Fa1sheng1	L1-4-1 begin(x,L)+[L=tim]	38	43
	New	5	

Shu3yu2	H1-4-1 partof(x,y)+NULL	16	23
	H1-2-1 belongto(x,y)+NULL	7	
Mei2	Oth	33	37
	E1-3-1 disappear(x,L)+[L=loctim]	4	
Fen4	Oth	20	28
	H3-1-2-7 DO(x,P(x,y))_CAUSE_(NOT_ have(x,y)_&_have(z,y))+[P=Issue]	8	

Table2. Descriptions of the target verbs.

4.1 Target verbs for tests

We choose nine Chinese verbs as target words to carry on our preliminary tests. These verbs are all ambiguous according to the LSO hierarchy. We use the definitions from the LSO hierarchy as the standard verb meanings. The representation of the meanings and the number of the instances of the verbs are given in Table 2.

Among the target verbs, ‘you3’ has the most instances and types, while the other verbs all have small number of instances. The distributions of the instances in each sense type of the verb are different. Some of them are even, such as “mei2you3” and “de2”; some of them are not, e.g. “fa1sheng1”, “mei2” and “dai4”. We assume the distributions represent the real situations of using the target verb in the real world, according to the definitions of the LSO on the verbs’ meanings.

4.2 Tests and Results

Five groups of tests are conducted on each target verb with different feature set. The test results using feature set 1 set a baseline for the following test using features from the LSO data. Then four groups of tests are carried on to give a profile of the performance using the LSO data on WSD task. 10-folds cross-validation is used to examine the practical performance of the classification. That is, divide the instance into ten parts, using nine of them as training set, while the left is used as testing set; repeat

this process ten times with letting different part as testing set and then get the average accuracy. The WSD results of each target verb are given in Table 3. 'R' represents the feature related words; 'POS' is the feature part of speech tags; 'SYN' represents the syntactic feature; 'SEM' is the feature semantic category of the related word.

The baseline drew by feature set 1 shows the WSD result with the part-of-speech and semantic information of the co-occurrence in the sentence. According to the results, using the features gotten from LSO data improves the WSD performance by about five points than the baseline on average. We assume it is mostly because of the addition of syntactic information, since the target words are verbs. The choose of the related words narrows the word bag and adds syntactic information into the analysis without decreasing the accuracy, improves it by contraries. The improvement is obvious on the verbs which have relatively more instances, such as 'you3', 'mei2you3' and 'gei3'. There are verbs whose results have little change, such as 'fa1sheng1', 'fa1chu1' and 'mei2'. They are either lack of instances or have uneven distributions of instance. Under these circumstances, the results do not rise when the classifier always classifies all the cases into the largest type despite using either of the two feature sets.

The average score of feature set 2-1 is the highest among the average results. The accuracy of each verb is also higher than (or equal to) the result using feature set 2. It indicates that the POS information does not help but disturbs the classification process. The feature POS can be viewed as a very abstract feature which describes the related word. It weakens the ability of the classifier when most of the related words are noun.

Both the feature set 2-2 and 2-3 score higher than the feature set 2 on average. The average score of using feature set 2-3 is a little bit higher, but we can not claim that the feature set 2-3 achieves better result. The results show that there are three verbs whose accuracy of feature set 2-2 is higher than the accuracy of feature set 2-3 ('you3', 'de2' and 'fa1sheng1'); four are lower ('mei2you3', 'gei3', 'shu3yu2' and 'mei2'); two are equal ('fa1chu1' and 'fen1'). To the higher ones, the results also higher than the results using feature set 2, which means the feature related word provides redundant information and affect the classification. To the lower ones, their results also lower than using the feature set 2, which indicates the features in feature set 2-2 are too general to be distinguished. It may lead to the conclusion that the generalization of the semantic representation of related word is optional but not necessary. Yet in our tests we can only say that the addition of semantic category information from the LSO can improve the result on the WSD of some particular targets.

Target verbs	Precision in each test (%)				
	F.set1: baseline	F.set2: R+POS+ SYN+SEM	F.set2-1 : R+SYN +SEM	F.set2-2: POS+ SYN+SEM	F.set2-3 : R+POS + SYN
You3	47.9	54.0	54.9	54.7	50.0
Mei2you 3	56.2	76.0	76.0	69.4	77.7
Gei3	71	77.4	82.3	74.2	80.6
De2	70.6	73.5	76.5	73.5	67.6
Fa1chu1	70.0	70.0	70.0	80.0	80.0
Fa1sheng 1	88.4	88.4	88.4	93.0	81.4
Shu3yu2	65.2	69.6	69.6	65.2	69.6
Mei2	89.2	89.2	89.2	89.2	94.6
Fen1	71.4	75.0	78.6	78.6	78.6
Average	70.0	74.8	76.2	75.3	75.6

Table 3. The results of all the tests.

To sum up, the features derived from the LSO data provide much information for our WSD task and improve the classification result remarkably than the baseline. We got the best result on the feature set comprises the related words in the context, their syntactic function tags and semantic category tags. There is no absolute answer that whether the semantic category or the related word should be used into improving the WSD result, since it differs from verb to verb. Our test results are close to other previous classifiers, but the scale is small. Thus more research has to be done to strongly support the points our primary research takes on.

5. Conclusions

While supervised classification method has been widely employed in WSD process, appropriate feature set becomes the key to improve the performance of WSD system. Certain tests are the practical ways to determine whether the features got from a resource informative enough to accomplish a task. LSO is the major resource of the features used in our WSD task. The tests examine the ability of LSO to provide the information and draw a profile of the information on Chinese verb WSD task.

Using more informative feature set derived from LSO our preliminary tests get better results than the baseline. Through different combinations of the features we improve the accuracy of each target verb and briefly analyze some characters of the results. The analysis of the results described in this paper still needs more research to support. Further tests should be carried on to get more evidence using the same algorithm, or other methods should be involved for further researches.

Acknowledgements

The research was supported by the National Natural Science Foundation of China (No. 60573185) and the National Hi-Tech Research and Development Program (863 plan) of China (No. 2007AA01Z173).

References

- Dong, Z. D, Dong Q. 2002. Hownet. <http://www.keenage.com>
- Leacock, C., Chodorow, M., & Miller G. A. 1998. *Using Corpus Statistics and WordNet Relations for Sense Identification*. Computational Linguistics, 24:1, 147–165.
- Ng, H. T., & Lee H. B. 1996. *Integrating Multiple Knowledge Sources to Disambiguate Word Sense: An Exemplar-Based Approach*. In Proc. of the 34th annual meeting on Association for Computational Linguistics, Santa Cruz, California.
- Witten, I.H., Frank, E. 2005. *Data Mining: Practical Machine Learning Tools and Techniques, Second Edition*. Morgan Kaufmann publishers. San Francisco.
- Yarowsky, D. 2000. *Hierarchical Decision Lists for Word Sense Disambiguation*.

Computers and the Humanities, 34(1-2), 179–186.

Zheng-Yu Niu, Dong-Hong Ji, Chew-Lim Tan. 2004. *Optimizing Feature Set for Chinese Word Sense Disambiguation*. In SENSEVAL-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text, Association for Computational Linguistics, Barcelona, Spain, 2004.

Zheng-Yu Niu, Dong-Hong Ji, Chew-Lim Tan. 2007. *I2R: Three Systems for Word Sense Discrimination, Chinese Word Sense Disambiguation, and English Word Sense Disambiguation*. In Proc. of the 4th International Workshop on Semantic Evaluations (SemEval-2007), Prague, 2007.

Zhou, Q. 2007. *Develop a Syntax Semantics Linking Knowledge Base for the Chinese Language*. In Proc. of the 8th Chinese Lexical Semantics Workshop, Hong Kong, May 21-23, 2007.

Qiang Zhou. 2007. *A Computational Framework to Integrate Different Semantic Resources*. To appear in TSD2008.