Symmetric is Not the Optimal Local Context Window in Chinese Word Sense Disambiguation

Gang Li, Guangzeng Kou
School of Information Management,
Wuhan University
Wuhan, China
ligang@whu.edu.cn, kouguangzeng@gmail.com

Ji Quan
Institute of Systems Engineering,
Wuhan University
Wuhan, China
quanji123@163.com

***Abstract*—Word Sense Disambiguation (WSD) is a task of classification, where the local context is the basic features to identify the sense of ambiguous word. Most systems choose optimal local context window on empirical grounds, which is usually symmetric, the distance from the ambiguous word to both sides of the window is same, such as [-1, +1] or [-2, +2]. Is symmetric window better than asymmetric window? In this paper, we take Senseval-3 Chinese data set as example. First find the optimal window estimated by cross–validation using only the training set, which is a symmetric window. Then, perform a WSD evaluation on the test data using this symmetric window for comparison with other classical symmetric window. The results show that asymmetric is better than symmetric window, and symmetric window is not always the best option.**

Keywords- Word Sense Disambiguation; local context; context window

I. INTRODUCTION

WSD is a task of classification to identify the correct meaning of ambiguous word, which can be thought of as comprising a feature set and a decision procedure. Commonly used features include surrounding words and their part of speech (POS) [1]. These features are extracted from the context. There are three types of context: local context, topical context and domain context [2]. Local context is generally considered to be some small window of words surrounding the target word, which contains location information; Topical context includes words that co-occur within a larger window usually in several sentences; Domain context is to determine discourse's domain, such as law.

This paper focuses on the local context. Most WSD systems use the local context as the primary features to identify sense of ambiguous word. The word occurring in local context window is the nearest one to ambiguous word, which can provide enough information for WSD work. However, what is the optimal local context window, how much the size of the window is the best?

Another interesting problem is symmetric window. Symmetric window means that the distance from the ambiguous word to both sides of the window is same, such as [-1, +1] or [-2, +2]. Most WSD works choose symmetric context window on empirical grounds. On the contrary, few systems use asymmetric window, for example, [-1, +3]. Symmetric and asymmetric window, which one is better?

In English WSD study, Yarowsky evaluates different local context windows by using log-likelihoods ration, and suggest that [-3, +3] and [-4, +4] is enough for local context [3]. So he just choose [-1, +1] to evaluates, which is not compared to other windows, such as asymmetric window. Leacock chooses [-3, +3] as the optimal local context window on empirical grounds, which is symmetric [4]. In Chinese WSD study, Niu et al. compare all of optimal local context window candidates through cross-validation evaluation, include [-1, +1] and [-2, +2], not considering the asymmetric window [5]. Generally speaking, WSD works usually select symmetric window as optimal window, such as [-1, +1], [-2, + 2], [-3, +3], called classical local context window.

For Chinese WSD, is symmetric window better than asymmetric window? How to choose the optimal local context window? This paper will focus on these issues. In the following, Senseval-3 Chinese lexical sample task will be taken as example. We first use cross-validation method to find the optimal local context window using only training data, which is asymmetric. Then, perform a WSD evaluation on the test data to compare the classical windows, include [-1, +1], [-2, + 2] and [-3, +3] to the asymmetric window just calculated.

II. OPTIMAL ASYMMETRIC WINDOW

In this section, in order to find the optimal local context window from candidates, we use cross-validation method to compute error of each candidate, considering the smallest one will be the optimal window. [W_{-left}, W_{+right}, $POS_{-left'}$, $POS_{+right'}$] denotes optimal local context window candidate. The local context features include:

- W_{-left}, $W_{-(left-1)} \ldots W_{right-1}$, W_{right}
- $POS_{-left'}$, $POS_{-(left'-1)i} \ldots POS_{right'-1}$, $POS_{righ'}$

A. Cross-validation

When limited examples are available, cross-validation is an important statistical method to evaluate the performance of classifier. N-fold cross-validation is most commonly used. The data set are randomly divided into n groups, and each

978-0-7695-3688-0/09 $25.00 © 2009 IEEE
DOI 10.1109/ITCS.2009.47

group will be taken as test data in turn, while the remainder as the training data. This process will be repeated by n times so that every example will be test data in just one time. Although the size of n has minimal impact on the evaluation results, 10-fold cross-validations is considered as the best choice [6].

Since the data set is divided on random, even if use same classification algorithm on same data set, it would get different results during many times. In order to eliminate the affect by random, we will use 10-time 10-fold cross-validation.

B. Naive Bayesian

Naive Bayesian algorithm is a statistical classification algorithm. WSD works always gain good performance during repeated tests. This paper use Naive Bayesian algorithm. Naive Bayesian algorithm relies on two important assumptions: the predictive attributes are conditionally independent given the class; it posits that no hidden or latent attributes influence the prediction process [7].

Let C denote the class set or sense set, and c will be a random element in C denoting a sense for an ambiguous word. In addition, let X denote example set in the form of feature vector, and x will be a random element in X denoting a case to be disambiguated [8].

$$p(C=c|X=x)=\frac{p(X=x|C=c)p(C=c)}{P(X=x)}, \text{ where}$$

$$p(X=x\,|C=c)=\prod_i p(X_i=x_i\mid C=c)$$

C. Process

Let B_i denotes optimal local context window candidate in the form of [W_{-left}, W_{+right}, $POS_{-left'}$, $POS_{+right'}$], and the collection of candidates is B that contains n elements, where n = (left +1) * (right +1) * (Left'+1) * (right' +1). Error set is E, where E_i is the error of Bi. E_{min} denotes the minimum value of E, considered as the optimal local context window. The process to determine optimal local context window is as follows:

1. Preprocessing of the data set, such as Chinese word segmentation.
2. Any B_i in B:
   a) Construct the feature vector to represent the context.
   b) Use 10-time 10-fold cross-validation to evaluate the error for B_i, called E_i.
3. Repeat step 2 for n times.

Select the smallest one in E as E_{min}, and the B_i corresponding to E_{min} will be the optimal local context window.

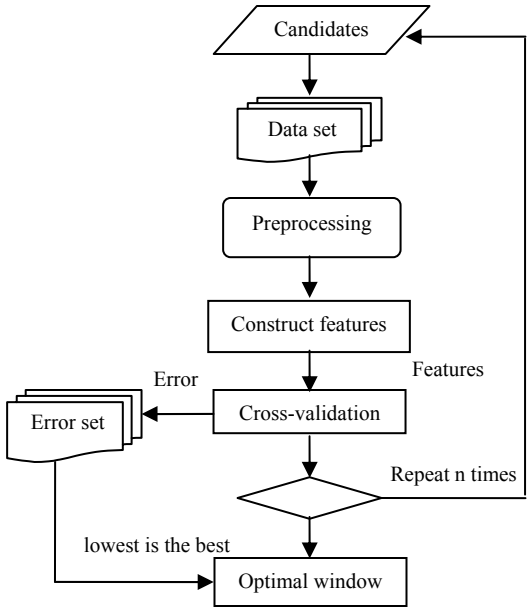

Figure 1. Flow chart of determining the optimal local context window from candidates using Cross-validation method

D. Result

We use only Senseval-3 Chinese lexical sample task training data set, a total of 20 ambiguous words and 893 examples. The interval of each element in [W_{-left}, W_{+right}, $POS_{-left'}$, $POS_{+right'}$] range from 0 to 5. There are 1296 optimal local context window candidates in all. The error indicator is root relative squared error:

$$E=\sqrt{\frac{\sum_{i=1}^{n}(P_i-T_i)^2}{\sum_{i=1}^{n}(T_i-\overline{T})^2}}, \text{ where } \overline{T}=\frac{1}{n}\sum_{i=1}^{n}T_i$$

P_i denotes predicted value; T_i denotes the true value of the example, n is the number of ambiguous word.

We choose top 5 groups from the test results ranked by error from small to large, see Table I.

TABLE I. TOP 5 GROUPS RANKED BY ERROR IN THE EXPERIMENT DETERMINING THE OPTIMAL WINDOW

Rank	Window	A/Symmetric
1	[-3, +1, -1, +1]	asymmetric
2	[-1, +1, -1, +1]	symmetric
3	[-2, +1, -1, +1]	asymmetric
4	[-3, +0, -1, +1]	asymmetric
5	[-2, +0, -1, +1]	asymmetric

We find the optimal local context window is [-3, +1, -1, +1] in the evaluation of Cross-validation, which is

asymmetric window. The symmetric window [-1, +1, -1, +1] is not the best, and other 3 groups are asymmetric window too.

III. COMPARISON TEST

Most WSD works often choose classical context window as optimal, and they are all symmetric window. However, through the evaluation of cross-validation in above, we find the optimal local context window is asymmetric. In order to verify it, we will take a comparison test on WSD between these classical windows to this asymmetric window.

This test will use same classification algorithm, same training data set and test data set. Naive Bayesian algorithm is used, and these data are all from Senseval-3 Chinese lexical sample task, for a total of 20 words, 893 examples in training data set and 380 in test data set. The performance indicator is Macro-average:

$$P_{mar} = \sum_{i=1}^{n} P_i / N, P_i = m_i / n_i$$

N is the number of ambiguous words; m_i denotes the number of labeled correctly to one target word, and n_i is the number of all test examples for this target word. See test results in Table II.

TABLE II. COMPARISON RESULT ON WSD BETWEEN ASYMMETRIC AND CLASSICAL SYMMETRIC WINDOW

Window	Accuracy	Difference
[-3, +1, -1, +1]	56.020%	
[-1, +1, -1, +1]	55.419%	+0.601%
[-2, +2, -2, +2]	55.181%	+0.839%
[-3, +3, -3, +3]	49.707%	+6.313%

Table II. shows that these classical symmetric windows are worse than the asymmetric window [-3, +1, -1, +1] on WSD for Senseval-3 Chinese lexical sample task. We will draw two conclusions: using cross-validation method to determine the optimal window is a recommend method, since the window [-3, +1, -1, +1] get best performance in this test. In addition, the performance of the symmetric window is not always the best, and sometimes the asymmetric window is better.

In practice, most WSD works choose symmetric window to label English ambiguous word, and always can get good performance. However, Chinese language is different to English. Chinese language needs word segmentation before WSD. Further, a Chinese word always contains more than words, such as "信息" composed by "信" and "息", both of which are word too. Therefore, the symmetric context window in Chinese is not always symmetric in the number of word. See an example in Fig. 2.

Ambiguous word: 路
Sentence: 由沪嘉联合投资的嘉兴中环南**路**首期工程通车。
After word segmentation:
由/p 沪/j 嘉/j 联合/v 投资/vn 的/u 嘉兴/ns 中环/j 南/j **路**/n 首/m 期/q 工程/n 通车/vn
Symmetry window [-2,+2]:
Left - [中环/j 南/j] number:3
Right - [首/m 期/q] number:2

Figure 2. Symmetric window in Chinese is not always symmetric in the number of word

IV. CONCLUSIONS

In this paper, we take Senseval-3 Chinese data set as example and use cross-validation methods to find an asymmetric window as the optimal window [-3, +1, -1, +1]. And a comparison test is preformed to compare this asymmetric window with 3 classical symmetric windows: [-1, +1, -1, +1], [-1, +1, -1, +1] and [-3, +3, -3, +3], which shows that the asymmetric is superior to the classical symmetric windows for this data set.

The performance of the symmetric window is not always the best, while the asymmetric window is sometimes better than symmetric window. In order to improve WSD performance, we can use cross-validation method or some others to determine the optimal window. We can use classical symmetric window absence of the assisted calculation, but should realize that symmetric window is not always best.

ACKNOWLEDGMENT

This paper is supported by the National Natural Science Foundation of China under grant No. 70673070.

REFERENCES

[1] Bruce R, Wiebe J., "Word-sense disambiguation using decomposable models", In *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*, Las Cruces, New Mexico, 1994, pp. 139-146

[2] Ide N., Véronis J., "Introduction to the special issue on word sense disambiguation: the state of the art", *Computational Linguistics*, 1998, pp. 1-40

[3] Yarowsky D., "Decision lists for lexical ambiguity resolution: Application to accent restoration in Spanish and French", In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics*, Las Cruces, NM, 1994, pp. 88-95

[4] Leacock C., Miller G.A., Chodorow M., "Using corpus statistics and WordNet relations for sense identification", *Computational Linguistics*, 1998.

[5] Niu Z.Y., Ji D.H., Tan C.L., "Optimizing Feature Set for Chinese Word Sense Disambiguation", In *Proceedings of the 3rd International Workshop on the Evaluation of Systems for the Semantic Analysis of Text (SENSEVAL-3)*, Barcelona, Spain, 2004.

[6] Witten I.H., Frank E., *Data Mining: Practical Machine Learning Tools and Techniques*, San Francisco: Morgan Kaufmann, 2005.

[7] Pedersen T., "A simple approach to building ensembles of Naive Bayesian classifiers for word sense disambiguation", In *Proceedings*

of the first conference on North American chapter of the Association for Computational Linguistics, 2000, pp. 63-69.

[8] George H., Langley P., "Estimating continuous distributions in Bayesian classifiers", In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, 1995, pp. 338-345