自 然 言 語 処 理　128－11
（1998．11．5）

# A Corpus-based Bootstrapping WSD Algorithm Using a Downstairs-like Threshold

Zhenglai Gu, Naoyuki Tokuda, Yan Ye
Department of Computer Science, Utsunomiya University
{gzl,tokuda,yan}@alfin.mine.utsunomiya-u.ac.jp

## Abstract

We have developed an improved version of Yarowsky's(1995)[1] decision-list based WSD algorithm by developing a stepwise threshold function starting from a higher to lower threshold values in choosing the lists. We elucidate the validity of the threshold function because a higher threshold gives a higher accuracy at the expense of lower coverage level while noises in the initial noisy seeds will produce unreliable pieces of evidence in the decision list. A downstairs-like threshold is capable to optimize the decision list pruning processes.

# ステップ関数しきい値を利用した教師なし語意多義性解消アルゴリズム

顧 政来，徳田 尚之，葉 衍
宇都宮大学大学院情報工学研究科
{gzl,tokuda,yan}@alfin.mine.utsunomiya-u.ac.jp

## 要旨

本論文では，Yarowskyの決定リストによる語彙曖昧性解消アルゴリズムに階段状に変わるしきい関数を導入することにより，曖昧性解消の精確度で平均１．１%，決定リスト長で６４%も短くできることを示した。実験は，大型コーパスとして有名なPenntree Bankコーパスの Wall Street Journalから，Bill, Duty, Trial, Issues, Tankの５個の２義語について行った。

## 1 Introduction

The evolution of WSD algorithm is tracing the history of overcoming the knowledge acquisition bottleneck that Bar-Hillel[2] first identified over thirty years ago. Since the massive quantities of digital data ( corpus ) such as Penntree bank corpus, Brown corpus became available, many computational linguistics have started to have an interest in statistical approaches recently. Bruce and Wiebe[3](1994), Leacock and Voorhees[4](1996) approach this problem using the supervised WSD algorithm, in which ambiguous words in the examples have to be tagged by certain sense before the training phase begins. It requires obviously time-consuming hand annotations. Yarowsky[1] (1995) has been the first to propose an efficient unsupervised WSD algorithm which obviously eliminated much of the labor intensive tagging tasks of the supervised algorithms; in Yarowsky's algorithm which is based on a remarkably simplified control structure of decision lists, a bootstrapping process iteratively trains on untagged examples to progressively acquire new knowledge learning heavily from the previous disambiguation results of the decision list. A few seed collocations are all they need to start Yarowsky's bootstrapping algorithm, improving rapidly the accuracy of disambiguation if the decision list formed is reliable. We found, however, the method is quite susceptible to noises often ending up with noise-contaminated decision lists resulting in incorrectly annotated examples or unreliable pieces of evidence in the decision list. Li and Takeuchi[5] suggest to use the MDL (Minimum Description Length) principle[6] and mutual information[7] between pieces of evidence and ambiguous word to exclude noises from the decision list. However, their method has shortcomings in two respects. Firstly it is valid to noises of independent sources and secondly is used to adjust the decision list after computing the bootstrapping process managed by low threshold values. Thus many of noises generated during the bootstrapping iterative process still remain.

The purpose of our work is to report our findings that the remarkably simple and efficient Yarowsky's algorithm can still be improved in accuracy by building in a stair-like step function into the threshold function for the decision list pruning. This is remarkably effective in filtering out unnecessary noises.

We made extensive experiments on five English homographs which show that the algorithm presented perform well with promising result.

## 2 Algorithm

### 2.1 Yarowsky's Bootstrapping Algorithm

Yarowsky's original bootstrapping algorithm is given below:

Procedure Bootstrapping

(1) Manually picked small set of initial seeds

(2) Loop

(a) Annotate the sense of the homograph in the training data in accordance with the pieces of evidence in the decision list.

(b) Compute the log-likelihood ratio① of the salient words in the annotated training data

(c) Add those words with log-likelihood values above a certain threshold value $\alpha$ into the decision list.

(d) Repeat the loop. The loop will stop when the decision list does not grow.

## 2.2 Algorithm Using Threshold Function

A most critical parameter in Yarowsky's algorithm is the threshold value $\alpha$ used on which the decision whether to add or delete new or existing pieces of evidence in the decision list respectively critically depends.

At a low level of threshold values, the words are very easily added into the decision list, and so the decision list expands so rapidly with magical speed. The result is low accuracy②, because the irresponsible low level of threshold value allows many misleading uncorrelated words to be added into the decision list even in the early stages with small size of the annotated training data. However, if we choose too high a level of threshold value, on the other hands, the accuracy may increase but the algorithm has only a limited level of coverage③, because the stricter criterion makes the addition of new pieces of evidence quite difficult, eliminating even the probable candidates removed from the list. The decision list grows slowly but the accuracy of the algorithm remains high.

Determining threshold values presents a crucial trade-off problem between the high and low threshold values. It is also a balance between the accuracy and coverage. Our immediate strategy is to introduce the stair-like step function changing the threshold value $\alpha$ from a high to a low value by step function as illustrated in equation (2.2.1).

$$\text{Threshold} = \text{BeginT} - \beta * [\text{Times} / N] \quad ④ \qquad (2.2.1)$$

BeginT denotes the initial threshold used. Times denotes the number of iterations while $\beta$ is the height of each stair of the function. After N iterations, the threshold steps down by $\beta$ to the next "stairs".

# 3 Evaluation Methodology

Generally, the performance of the WSD algorithm must be evaluated from the accuracy of disambiguation and coverage point of view.

The coverage denotes the ratio of the number of the decisions made to the total number of the test data inputs. The accuracy denotes the ratio of the number of the correct decisions made for the homograph to the number of the decisions made.

In practice, a default strategy can be introduced if decision can not be made due to lack of evidence in the decision lists. This is not uncommon particularly in natural language due to the sparse data problem. In this paper, we take the majority sense of the

① Log-likelihood ratio=log $(P(w \mid s_1)/P(w \mid s_2))$, where $P(w \mid s_1)$ is the probability that the collocation word w and one sense of the homograph occur in the same window of the context, and so is $P(w \mid s_2)$.

② See chapter 3 for definition.

③ See chapter 3 for definition.

④ [Times / N] is the value taking only the integer part of the division result.

homograph as our default word sense. Thus, we define the term precision as the ratio of the number of entire correct decisions made to the number of the entire test inputs. The use of (default) dominant sense gives us the advantage of 100% coverage.

We have used the definition accuracy and precision in our experimental comparisons. Precision seems used more widely among researchers because 100% coverage is always assured.

The length of the final decision list gives an important parameter of the problem because it not only relates heavily to spatial as well as temporal complexity of the algorithm but also the length reflects directly the reliability of the evidence.

## 4 Experiments

### 4.1 Experiments Data Used

Five English noun homographs were tested through Yarowsky's and our algorithm. We extracted all the training data from the corpus of the Wall Street Journal, a machine-readable corpus collected by AC/DCI(Association of Computational Linguistics' Data Collection Initiative) in 1991. It contains 40.6 million words mainly on economy. Statistical data of the 5 homographs are given in table 1.

| | Number of Training data | Number of test data |
|---|---|---|
| Duty(obligation / tax) | 1682 | 229 |
| Issue(problem/ publish) | 4273 | 227 |
| Trial(test / law) | 3048 | 177 |
| Bill(money / file) | 7650 | 200 |
| Tank(vehicle/ container) | 770 | 95 |

Table 1. Five English homographs used

### 4.2 Experiments by Yarowsky's Algorithm

We did forty experiments for each word with different values of the parameter $\alpha$ changing from 5.5 to 1.6 at an interval of 0.1. The number of iterations for each experiment is set to 8, because we experimentally found that the length of the decision list is held constant always within 8 iterations.

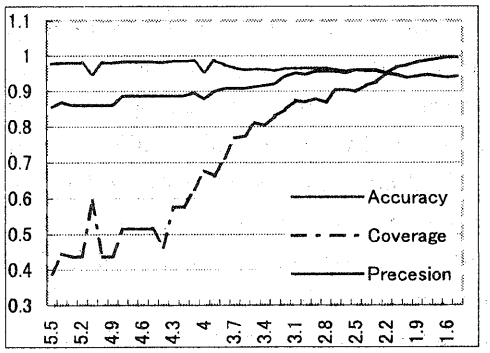


Figure 1: the accuracy, coverage and precision of the homograph "duty" by Yarowsky's algorithm

Figures 1 and 2 summarize our typical experiments.

When the threshold value is set to 5.5, the length of decision list is reduced from 20 to 10, and the accuracy is 0.978 while the coverage is lower (0.389), but on the other hand, when the threshold value is set to 1.6, the length of decision list increased rapidly

to 1018, while the accuracy is down to 0.939. From the precision point of view, the highest value is 0.961 at the threshold of 2.5 while the length of decision list is 213.

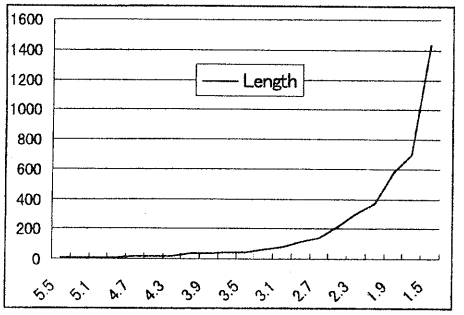


Figure 2: the Length of the decision list in the experiment of the homograph duty by Yarowsky's algorithm

### 4.3 Experiments by New Algorithm

Using the same training data, we also did forty experiments on each homograph with the different values of the two parameters BeginT and $\beta$ changing from 6.5 to 3.0 at an interval of 0.5 and 0.5 to 0.1 at an interval of 0.1, respectively. The other parameter N is set to 4. The loop is set to stop when the threshold value is less than 1.6. The result is shown in figure 3 below.

At the beginning of the process, since the annotated training sets are quite small those calculated salient log-likelihood values have low credibility. Higher thresholds are chosen at the beginning phase. As the bootstrapping process advances, the sets of the tagged training examples keep growing, and the calculated values become more and more representative and reliable. Accordingly, the level of threshold is lowered and cut down step by step. The peak precision value in this case is 0.983 at the threshold value of 3.5 while the BeginT and $\beta$ are 5.5 and 0.2, respectively.

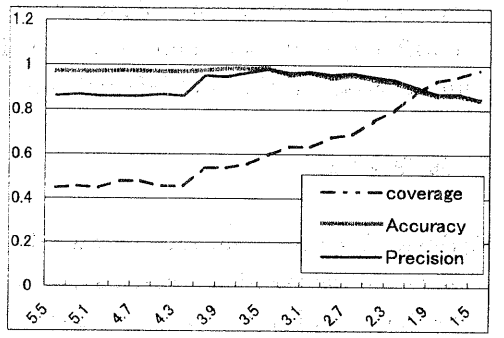


Figure3 the accuracy, coverage and precision of the homograph duty using our algorithm

| | Yarowsky | | Dowstairs | | Improvement | |
|---|---|---|---|---|---|---|
| | Prec. | Leng | Prec. | Leng | Prec. | Leng |
| duty | 96.1 | 237 | 98.2 | 48 | 2.1 | 79.7 |
| issue | 94.7 | 259 | 96.0 | 215 | 1.3 | 17.0 |
| trial | 94.4 | 330 | 95.5 | 77 | 1.1 | 76.7 |
| bill | 93.0 | 177 | 94.0 | 31 | 1.0 | 82.5 |
| tank | 94.7 | 237 | 94.7 | 74 | 0.0 | 68.8 |
| Avg. | 94.58 | 248 | 95.68 | 89 | 1.1 | 64.1 |

## 5 Summary

We have developed an improved version of Yarowsky's decision-list-based algorithm using a downstairs-like threshold function which is remarkably efficient in filtering out noises. The experiments on 5 English noun homographs show that our algorithm exhibits performance improvement of 1.1% in terms of precision and of 64.1% in terms of space for storing knowledge over Yarowsky's algorithm, indicating our algorithm is capable of generating more reliable pieces of evidence. The shorter decision list contribute to the improved complexity because the temporal complexity is of O(log(L)) where L is the

length of the decision list . This is a well-known complexity of binary search which we use. We hope our algorithm can contribute to further improvement of the unsupervised method in WSD problems.

## 6 Reference


[1] David Yarowsky. "Unsupervised word sense disambiguation rivaling supervised methods", In Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics, 189-196, 1995.

[2] Bar-Hillel Y. "The present status of automatic translation of languages", In F.L. Alt(ed), Advances in Computers. Vol.1. New York: Academic Press, 91-163,1990.

[3] Bruce Rebecca and Wiebe Janyce. "Word-sense disambiguation using decomposable models", Proceedings of *ACL-94,* 139-145,1994.

[4] Leacock Claudia, Geoffrey Towell and Ellen Voorhees. "Corpus-based statistical sense resolution", Proceedings of ARPA Human Language Technology Workshop.

[5] Hang Li and Jun-ichi Takeuchi, "Using Evidence that is both Strong and Reliable in Japanese Homograph Disambiguation", 情報処理学会研究報告 97-NL-119 53-59,1997.

[6] J.Rissanen, "Minimum-Description-Length Principle", Encyclopedia of Statistic Sciences, 523-527, 1987.

[7] James Allen, "Statistical word sense disambiguation", Natural Language Understanding, Second Edition,310-314,1995.