A Combination of Active Learning and Semi-supervised Learning Starting with Positive and Unlabeled Examples for Word Sense Disambiguation: An Empirical Study on Japanese Web Search Query

Makoto Imamura
and Yasuhiro Takayama
Information Technology R&D Center,
Mitsubishi Electric Corporation
5-1-1 Ofuna, Kamakura, Kanagawa, Japan
{Imamura.Makoto@bx,Takayama.Yasuhiro@ea}.MitsubishiElectric.co.jp

Nobuhiro Kaji, Masashi Toyoda
and Masaru Kitsuregawa
Institute of Industrial Science,
The University of Tokyo
4-6-1 Komaba, Meguro-ku Tokyo, Japan
{kaji,toyoda,kitsure}@tkl.iis.u-tokyo.ac.jp

Abstract

This paper proposes to solve the bottleneck of finding training data for word sense disambiguation (WSD) in the domain of web queries, where a complete set of ambiguous word senses are unknown. In this paper, we present a combination of active learning and semi-supervised learning method to treat the case when positive examples, which have an expected word sense in web search result, are only given. The novelty of our approach is to use "pseudo negative examples" with reliable confidence score estimated by a classifier trained with positive and unlabeled examples. We show experimentally that our proposed method achieves close enough WSD accuracy to the method with the manually prepared negative examples in several Japanese Web search data.

1 Introduction

In Web mining for sentiment or reputation analysis, it is important for reliable analysis to extract large amount of texts about certain products, shops, or persons with high accuracy. When retrieving texts from Web archive, we often suffer from word sense ambiguity and WSD system is indispensable. For instance, when we try to analyze reputation of "Loft", a name of variety store chain in Japan, we found that simple text search retrieved many unrelated texts which contain "Loft" with different senses such as an attic room, an angle of golf club face, a movie title, a name of a club with live music and so on. The words in Web search queries are often proper nouns. Then it is not trivial to discriminate these senses especially for the language like Japanese whose proper nouns are not capitalized.

To train WSD systems we need a large amount of positive and negative examples. In the real Web mining application, how to acquire training data for a various target of analysis has become a major hurdle to use supervised WSD.

Fortunately, it is not so difficult to create positive examples. We can retrieve positive examples from Web archive with high precision (but low recall) by manually augmenting queries with hypernyms or semantically related words (e.g., "Loft AND shop" or "Loft AND stationary").

On the other hand, it is often costly to create negative examples. In principle, we can create negative examples in the same way as we did to create positive ones. The problem is, however, that we are not sure of most of the senses of a target word. Because target words are often proper nouns, their word senses are rarely listed in hand-crafted lexicon. In addition, since the Web is huge and contains heterogeneous domains, we often find a large number of unexpected senses. For example, all the authors did not know the music club meaning of Loft. As the result, we often had to spend much time to find such unexpected meaning of target words.

This situation motivated us to study active learning for WSD starting with only positive examples. The previous techniques (Chan and Ng, 2007; Chen et al. 2006) require balanced positive and negative examples to estimate the score. In our problem setting, however, we have no negative examples at the initial stage. To tackle this problem, we propose a method of active learning for WSD with pseudo negative examples, which are selected from unlabeled data by a classifier trained with positive and unlabeled examples. McCallum and Nigam (1998) combined active learning and semi-supervised learning technique

Proceedings of the ACL-IJCNLP 2009 Conference Short Papers, pages 61–64,
Suntec, Singapore, 4 August 2009.

by using EM with unlabeled data integrated into active learning, but it did not treat our problem setting where only positive examples are given.

The construction of this paper is as follows; Section 2 describes a proposed learning algorithm. Section 3 shows the experimental results.

2 Learning Starting with Positive and Unlabeled Examples for WSD

We treat WSD problem as binary classification where desired texts are positive examples and other texts are negative examples. This setting is practical, because ambiguous senses other than the expected sense are difficult to know and are no concern in most Web mining applications.

2.1 Classifier

For our experiment, we use naive Bayes classifiers as learning algorithm. In performing WSD, the sense “s” is assigned to an example characterized with the probability of linguistic features $f_1,...,f_n$ so as to maximize:

$$p(\mathrm{s})\prod_{j=1}^{n} p(\mathrm{f_j} \mid \mathrm{s}) \qquad (1)$$

The sense s is positive when it is the target meaning in Web mining application, otherwise s is negative. We use the following typical linguistic features for Japanese sentence analysis, (a) Word feature within sentences, (b) Preceding word feature within bunsetsu (Japanese base phrase), (c) Backward word feature within bunsetsu, (d) Modifier bunsetsu feature and (e) Modifiee bunsetsu feature.

Using naive Bayes classifier, we can estimate the confidence score c(d, s) that the sense of a data instance “d”, whose features are f_1, f_2, ..., f_n, is predicted sense “s”.

$$\mathrm{c(d,s)} = \log p(\mathrm{s}) + \sum_{j=1}^{n} \log p(\mathrm{f_j} \mid \mathrm{s}) \qquad (2)$$

2.2 Proposed Algorithm

At the beginning of our algorithm, the system is provided with positive examples and unlabeled examples. The positive examples are collected by full text queries with hypernyms or semantically related words.

First we select positive dataset P from initial dataset by manually augmenting full text query.

At each iteration of active learning, we select pseudo negative dataset N_p (Figure 1 line 15). In selecting pseudo negative dataset, we predict word sense of each unlabeled example using the naive Bayes classifier with all the unlabeled examples as negative examples (Figure 2). In detail, if the prediction score (equation(3)) is more than τ, which means the example is very likely to be negative, it is considered as the pseudo negative example (Figure 2 line 10-12).

$$\mathrm{c(d, psdNeg)} = \mathrm{c(d, neg)} - \mathrm{c(d, pos)} \qquad (3)$$

```
01  # Definition
02  Γ (P, N): WSD system trained on P as Positive
03              examples, N as Negative examples.
04  Γ_EM(P, N, U): WSD system trained on P as
05          Positive examples, N as Negative examples,
06          U as Unlabeled examples by using EM
07          (Nigam et. all 2000)
08  # Input
09  T ← Initial unlabeled dataset which contain
10          ambiguous words
11  # Initialization
12  P ← positive training dataset by full text search on T
13  N ← φ  (initial negative training dataset)
14  repeat
15    # selecting pseudo negative examples N_p
16        by  the score of  Γ (P, T-P)   (see figure 2)
17    # building a classifier with  N_p
18    Γ_new ← Γ_EM (P,  N+N_p, T-N-P)
19    #  sampling data by using the score of  Γ_new
20    c_min ← ∞
21    foreach d ∈ (T – P – N )
22      classify d by WSD system Γ_new
23      s(d) ← word sense prediction for d using Γ_new
24      c(d, s(d)) ← the confidence of  prediction of d
25      if c(d, s(d))  < c_min  then
26         c_min ← c(d),   d_min ← d
27    end
28  end
29   provide correct sense s for d_min by human
30   if s is positive then add d_min  to P
31                   else  add d_min  to N
32  until Training dataset reaches desirable size
33  Γ_new  is the output classifier
```

Figure 1: A combination of active learning and semi-supervised learning starting with positive and unlabeled examples

Next we use Nigam’s semi-supervised learning method using EM and a naive Bayes classifier (Nigam et. all, 2000) with pseudo negative dataset N_p as negative training dataset to build the refined classifier Γ_{EM} (Figure 1 line 17).

In building training dataset by active learning, we use uncertainty sampling like (Chan and Ng, 2007) (Figure 1 line 30-31). This step selects the most uncertain example that is predicted with the lowest confidence in the refined classifier Γ_{EM}. Then, the correct sense for the most uncertain example is provided by human and added to the positive dataset P or the negative dataset N according to the sense of d.

The above steps are repeated until dataset reaches the predefined desirable size.

```
foreach d ∈ ( T – P – N )
   classify d by WSD system Γ (P, T-P)
   c(d, pos) ← the confidence score that d is
      predicted as positive defined in equation (2)
   c(d, neg) ← the confidence score that d is
      predicted as negative defined in equation (2)
   c(d, psdNeg) =  c(d, neg)  - c(d, pos)
            (the confidence score that d is
              predicted as pseudo negative)
   PN ← d ∈ ( T – P – N ) |  s(d) = neg ∧
                              c(d, psdNeg)  ≧ τ }
            (PN is pseudo negative dataset )
end
```

Figure 2: Selection of pseudo negative examples

3 Experimental Results

3.1 Data and Condition of Experiments

We select several example data sets from Japanese blog data crawled from Web. Table 1 shows the ambiguous words and each ambiguous senses.

Word	Positive sense	Other ambiguous senses
Wega	product name (TV)	Las Vegas, football team name, nickname, star, horse race, Baccarat glass, atelier, wine, game, music
Loft	store name	attic room, angle of golf club face, club with live music, movie
Honda	personal name (football player)	Personal names (actress, artists, other football players, etc.) hardware store, car company name
Tsubaki	product name (shampoo)	flower name, kimono, horse race, camellia ingredient, shop name

Table 1: Selected examples for evaluation

Table 2 shows the ambiguous words, the number of its senses, the number of its data instances, the number of feature, and the percentage of positive sense instances for each data set.

Assigning the correct labels of data instances is done by one person and 48.5% of all the labels are checked by another person. The percentage of agreement between 2 persons for the assigned labels is 99.0%. The average time of assigning labels is 35 minutes per 100 instances.

Selected instances for evaluation are randomly divided 10% test set and 90% training set. Table 3 shows the each full text search query and the number of initial positive examples and the percentage of it in the training data set.

word	No. of senses	No. of instances	No. of features	Percentage of positive sense
Wega	11	5,372	164,617	31.1%
Loft	5	1,582	38,491	39.4%
Honda	25	2,100	65,687	21.2%
Tsubaki	6	2,022	47,629	40.2%

Table 2: Selected examples for evaluation

word	Full text query for initial positive examples	No. of positive examples (percentage in trainig set)
Wega	Wega AND TV	316 (6.5%)
Loft	Loft AND (Grocery OR-Stationery)	64 (4.5%)
Honda	Honda AND Keisuke	86 (4.6%)
Tsubaki	Tsubaki AND Shiseido	380 (20.9%)

Table 3: Initial positive examples

The threshold value τ in figure 2 is set to empirically optimized value 50. Dependency on threshold value τ will be discussed in 3.3.

3.2 Comparison Results

Figure 3 shows the average WSD accuracy of the following 6 approaches.

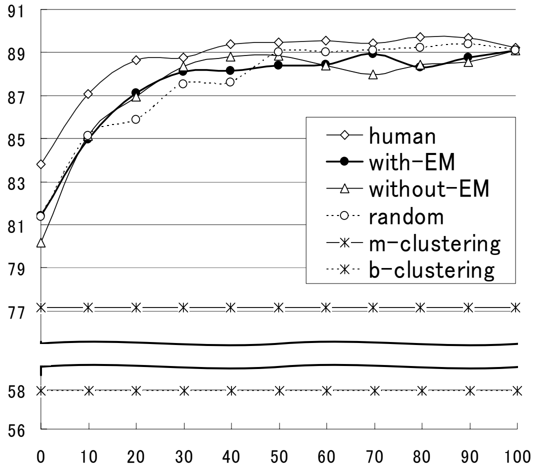

Figure 3: Average active learning process

B-clustering is a standard unsupervised WSD, a clustering using naive Bayes classifier learned with two cluster numbers via EM algorithm. The given number of the clusters are two, negative and positive datasets.

M-clustering is a variant of b-clustering where the given number of clusters are each number of ambiguous word senses in table 2.

Human labeling, abbreviated as *human*, is an active learning approach starting with human labeled negative examples. The number of human labeled negative examples in initial training data is the same as that of positive examples in figure 3. Human labeling is considered to be the upper accuracy in the variants of selecting pseudo negative examples.

Random sampling with EM, abbreviated as *with-EM*, is the variant approach where d_{min} in line 26 of figure 1 is randomly selected without using confidence score.

Uncertainty sampling without EM (Takayama et al. 2009), abbreviated as *without-EM*, is a variant approach where Γ_{EM} (P, $N+N_p$, T-N-P) in line 18 of figure 1 is replaced by Γ (P, $N+N_p$).

Uncertainty Sampling with EM, abbreviated as *uncertain*, is a proposed method described in figure 1.

The accuracy of the proposed approach *with-EM* is gradually increasing according to the percentage of added hand labeled examples.

The initial accuracy of *with-EM*, which means the accuracy with no hand labeled negative examples, is the best score 81.4% except for that of *human*. The initial WSD accuracy of *with-EM* is 23.4 and 4.2 percentage points higher than those of b-clustering (58.0%) and m-clustering (77.2%), respectively. This result shows that the proposed selecting method of pseudo negative examples is effective.

The initial WSD accuracy of *with-EM* is 1.3 percentage points higher than that of *without-EM* (80.1%). This result suggests semi-supervised learning using unlabeled examples is effective.

The accuracies of *with-EM, random* and *without-EM* are gradually increasing according to the percentage of added hand labeled examples and catch up that of *human* and converge at 30 percentage added points. This result suggests that our proposed approach can reduce the labor cost of assigning correct labels.

The curve *with-EM* are slightly upper than the curve *random* at the initial stage of active learning. At 20 percentage added point, the accuracy with-EM is 87.0 %, 1.1 percentage points higher than that of random (85.9%). This result suggests that the effectiveness of proposed uncertainty sampling method is not remarkable depending on the word distribution of target data.

There is really not much difference between the curve *with-EM* and *without-EM*. As a classifies to use the score for sampling examples in adaptation iterations, it is indifferent whether *with-EM* or *without-EM*.

Larger evaluation is the future issue to confirm if the above results could be generalized beyond the above four examples used as proper nouns.

3.3 Dependency on Threshold Value τ

Figure 4 shows the average WSD accuracies of *with-EM* at 0, 25, 50 and 75 as the values of τ. The each curve represents our proposed algorithm with threshold value τ in the parenthesis. The accuracy in the case of τ = 75 is higher than that of τ = 50 over 20 percentage data added point. This result suggests that as the number of hand labeled negative examples increasing, τ should be gradually decreasing, that is, the number of pseudo negative examples should be decreasing. Because, if sufficient number of hand labeled negative examples exist, a classifier does not need pseudo negative examples. The control of τ depending on the number of hand labeled examples during active learning iterations is a future issue.

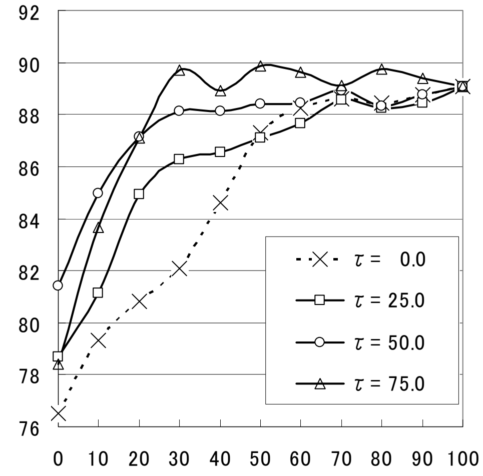

Figure 4: Dependency of threshold value τ

References

Chan, Y. S. and Ng, H. T. 2007. Domain Adaptation with Active Learning for Word Sense Disambiguation. *Proc. of ACL 2007*, 49-56.

Chen, J., Schein, A., Ungar, L., and Palmer, M. 2006. An Empirical Study of the Behavior of Active Learning for Word Sense Disambiguation, *Proc. of the main conference on Human Language Technology Conference of the North American Chapter of ACL*, pp. 120-127.

McCallum, A. and Nigam, K. 1998. Employing EM and Pool-Based Active Learning for Text Classification. *Proceedings of the Fifteenth international Conference on Machine Learning*, 350-358.

Nigam, K., McCallum, A., Thrun, S., and Mitchell, T. 2000. Text Classification from Labeled and Unlabeled Documents using EM, *Machine Learning*, 39, 103-134.

Takayama, Y., Imamura, M., Kaji N., Toyoda, M. and Kitsuregawa, M. 2009. Active Learning with Pseudo Negative Examples for Word Sense Disambiguation in Web Mining (in Japanese), Journal of IPSJ (in printing).