Partially Supervised Sense Disambiguation by Learning Sense Number from Tagged and Untagged Corpora

Zheng-Yu Niu, Dong-Hong Ji
Institute for Infocomm Research
21 Heng Mui Keng Terrace
119613 Singapore
{zniu, dhji}@i2r.a-star.edu.sg

Chew Lim Tan
Department of Computer Science
National University of Singapore
3 Science Drive 2
117543 Singapore
tancl@comp.nus.edu.sg

Abstract

Supervised and semi-supervised sense disambiguation methods will mis-tag the instances of a target word if the senses of these instances are not defined in sense inventories or there are no tagged instances for these senses in training data. Here we used a model order identification method to avoid the misclassification of the instances with undefined senses by discovering new senses from mixed data (tagged and untagged corpora). This algorithm tries to obtain a natural partition of the mixed data by maximizing a stability criterion defined on the classification result from an extended label propagation algorithm over all the possible values of the number of senses (or sense number, model order). Experimental results on SENSEVAL-3 data indicate that it outperforms SVM, a one-class partially supervised classification algorithm, and a clustering based model order identification algorithm when the tagged data is incomplete.

1 Introduction

In this paper, we address the problem of partially supervised word sense disambiguation, which is to disambiguate the senses of occurrences of a target word in untagged texts when given incomplete tagged corpus [1].

Word sense disambiguation can be defined as associating a target word in a text or discourse with a definition or meaning. Many corpus based methods have been proposed to deal with the sense disambiguation problem when given definition for each possible sense of a target word or a tagged corpus with the instances of each possible sense, e.g., supervised sense disambiguation (Leacock et al., 1998), and semi-supervised sense disambiguation (Yarowsky, 1995).

Supervised methods usually rely on the information from previously sense tagged corpora to determine the senses of words in unseen texts. Semi-supervised methods for WSD are characterized in terms of exploiting unlabeled data in the learning procedure with the need of predefined sense inventories for target words. The information for semi-supervised sense disambiguation is usually obtained from bilingual corpora (e.g. parallel corpora or untagged monolingual corpora in two languages) (Brown et al., 1991; Dagan and Itai, 1994), or sense-tagged seed examples (Yarowsky, 1995).

Some observations can be made on the previous supervised and semi-supervised methods. They always rely on hand-crafted lexicons (e.g., WordNet) as sense inventories. But these resources may miss domain-specific senses, which leads to incomplete sense tagged corpus. Therefore, sense taggers trained on the incomplete tagged corpus will misclassify some instances if the senses of these instances are not defined in sense inventories. For example, one performs WSD in information technology related texts using WordNet [2] as sense inventory. When disambiguating the word "boot" in the phrase "boot sector", the sense tagger will assign this instance with one of the senses of "boot" listed in WordNet. But the correct sense

[1] "incomplete tagged corpus" means that tagged corpus does not include the instances of some senses for the target word, while these senses may occur in untagged texts.

[2] Online version of WordNet is available at http://wordnet.princeton.edu/cgi-bin/webwn2.0

Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP 2006), pages 415–422,
Sydney, July 2006.

"loading operating system into memory" is not included in WordNet. Therefore, this instance will be associated with an incorrect sense.

So, in this work, we would like to study the problem of partially supervised sense disambiguation with an incomplete sense tagged corpus. Specifically, given an incomplete sense-tagged corpus and a large amount of untagged examples for a target word [3], we are interested in (1) labeling the instances in the untagged corpus with sense tags occurring in the tagged corpus; (2) trying to find undefined senses (or new senses) of the target word [4] from the untagged corpus, which will be represented by instances from the untagged corpus.

We propose an automatic method to estimate the number of senses (or sense number, model order) of a target word in mixed data (tagged corpus+untagged corpus) by maximizing a stability criterion defined on classification result over all the possible values of sense number. At the same time, we can obtain a classification of the mixed data with the optimal number of groups. If the estimated sense number in the mixed data is equal to the sense number of the target word in tagged corpus, then there is no new sense in untagged corpus. Otherwise new senses will be represented by groups in which there is no instance from the tagged corpus.

This partially supervised sense disambiguation algorithm may help enriching manually compiled lexicons by inducing new senses from untagged corpora.

This paper is organized as follows. First, a model order identification algorithm will be presented for partially supervised sense disambiguation in section 2. Section 3 will provide experimental results of this algorithm for sense disambiguation on SENSEVAL-3 data. Then related work on partially supervised classification will be summarized in section 4. Finally we will conclude our work and suggest possible improvements in section 5.

2 Partially Supervised Word Sense Disambiguation

The partially supervised sense disambiguation problem can be generalized as a model order identification problem. We try to estimate the sense number of a target word in mixed data (tagged corpus+untagged corpus) by maximizing a stability criterion defined on classification results over all the possible values of sense number. If the estimated sense number in the mixed data is equal to the sense number in the tagged corpus, then there is no new sense in the untagged corpus. Otherwise new senses will be represented by clusters in which there is no instance from the tagged corpus. The stability criterion assesses the agreement between classification results on full mixed data and sampled mixed data. A partially supervised classification algorithm is used to classify the full or sampled mixed data into a given number of classes before the stability assessment, which will be presented in section 2.1. Then we will provide the details of the model order identification procedure in section 2.2.

2.1 An Extended Label Propagation Algorithm

Table 1: Extended label propagation algorithm.

	Function: ELP(D_L, D_U, k, $Y^0_{D_L+D_U}$)
	Input: labeled examples D_L, unlabeled examples D_U, model order k, initial labeling matrix $Y^0_{D_L+D_U}$;
	Output: the labeling matrix Y_{D_U} on D_U;
1	If $k < k_{X_L}$ then
	Y_{D_U}=NULL;
2	Else if $k = k_{X_L}$ then
	Run plain label propagation algorithm on D_U with Y_{D_U} as output;
3	Else then
3.1	Estimate the size of tagged data set of new classes;
3.2	Generate tagged examples from D_U for $(k_{X_L}+1)$-th to k-th new classes;
3.3	Run plain label propagation algorithm on D_U with augmented tagged dataset as labeled data;
3.4	Y_{D_U} is the output from plain label propagation algorithm;
	End if
4	Return Y_{D_U};

Let $X_{L+U} = \{x_i\}_{i=1}^n$ be a set of contexts of occurrences of an ambiguous word w, where x_i represents the context of the i-th occurrence, and n is the total number of this word's occurrences. Let

[3] Untagged data usually includes the occurrences of all the possible senses of the target word

[4] "undefined senses" are the senses that do not appear in tagged corpus.

$S_L = \{s_j\}_{j=1}^{c}$ denote the sense tag set of w in X_L, where X_L denotes the first l examples $x_g (1 \leq g \leq l)$ that are labeled as y_g ($y_g \in S_L$). Let X_U denote other u ($l + u = n$) examples $x_h (l + 1 \leq h \leq n)$ that are unlabeled.

Let $Y^0_{X_{L+U}} \in N^{|X_{L+U}| \times |S_L|}$ represent initial soft labels attached to tagged instances, where $Y^0_{X_{L+U},ij} = 1$ if y_i is s_j and 0 otherwise. Let $Y^0_{X_L}$ be the top l rows of $Y^0_{X_{L+U}}$ and $Y^0_{X_U}$ be the remaining u rows. $Y^0_{X_L}$ is consistent with the labeling in labeled data, and the initialization of $Y^0_{X_U}$ can be arbitrary.

Let k denote the possible value of the number of senses in mixed data X_{L+U}, and k_{X_L} be the number of senses in initial tagged data X_L. Note that $k_{X_L} = |S_L|$, and $k \geq k_{X_L}$.

The classification algorithm in the order identification process should be able to accept labeled data D_L [5], unlabeled data D_U [6] and model order k as input, and assign a class label or a cluster index to each instance in D_U as output. Previous supervised or semi-supervised algorithms (e.g. SVM, label propagation algorithm (Zhu and Ghahramani, 2002)) cannot classify the examples in D_U into k groups if $k > k_{X_L}$. The semi-supervised k-means clustering algorithm (Wagstaff et al., 2001) may be used to perform clustering analysis on mixed data, but its efficiency is a problem for clustering analysis on a very large dataset since multiple restarts are usually required to avoid local optima and multiple iterations will be run in each clustering process for optimizing a clustering solution.

In this work, we propose an alternative method, an extended label propagation algorithm (ELP), which can classify the examples in D_U into k groups. If the value of k is equal to k_{X_L}, then ELP is identical with the plain label propagation algorithm (LP) (Zhu and Ghahramani, 2002). Otherwise, if the value of k is greater than k_{X_L}, we perform classification by the following steps:

(1) estimate the dataset size of each new class as $size_{new\_class}$ by identifying the examples of new classes using the "Spy" technique [7] and assuming that new classes are equally distributed;

(2) $D'_L = D_L$, $D'_U = D_U$;

(3) remove tagged examples of the m-th new class ($k_{X_L} + 1 \leq m \leq k$) from D'_L [8] and train a classifier on this labeled dataset without the m-th class;

(4) the classifier is then used to classify the examples in D'_U;

(5) the least confidently unlabeled point $x_{class\_m} \in D'_U$, together with its label m, is added to the labeled data $D'_L = D'_L + x_{class\_m}$, and $D'_U = D'_U - x_{class\_m}$;

(6) steps (3) to (5) are repeated for each new class till the augmented tagged data set is large enough (here we try to select $size_{new\_class}/4$ examples with their sense tags as tagged data for each new class);

(7) use plain LP algorithm to classify remaining unlabeled data D'_U with D'_L as labeled data.

Table 1 shows this extended label propagation algorithm.

Next we will provide the details of the plain label propagation algorithm.

Define $W_{ij} = exp(-\frac{d_{ij}^2}{\sigma^2})$ if $i \neq j$ and $W_{ii} = 0$ ($1 \leq i, j \leq |D_L + D_U|$), where d_{ij} is the distance (e.g., Euclidean distance) between the example x_i and x_j, and σ is used to control the weight W_{ij}.

Define $|D_L + D_U| \times |D_L + D_U|$ probability transition matrix $T_{ij} = P(j \rightarrow i) = \frac{W_{ij}}{\sum_{k=1}^{n} W_{kj}}$, where T_{ij} is the probability to jump from example x_j to example x_i.

Compute the row-normalized matrix $\overline{T}$ by $\overline{T}_{ij} = T_{ij} / \sum_{k=1}^{n} T_{ik}$.

The classification solution is obtained by $Y_{D_U} = (I - \overline{T}_{uu})^{-1} \overline{T}_{ul} Y^0_{D_L}$. I is $|D_U| \times |D_U|$ identity matrix. $\overline{T}_{uu}$ and $\overline{T}_{ul}$ are acquired by splitting matrix $\overline{T}$ after the $|D_L|$-th row and the $|D_L|$-th column into 4 sub-matrices.

2.2 Model Order Identification Procedure

For achieving the model order identification (or sense number estimation) ability, we use a cluster validation based criterion (Levine and Domany, 2001) to infer the optimal number of senses of w in X_{L+U}.

[5] D_L may be the dataset X_L or a subset sampled from X_L.

[6] D_U may be the dataset X_U or a subset sampled from X_U.

[7] The "Spy" technique was proposed in (Liu et al., 2003). Our re-implementation of this technique consists of three steps: (1) sample a small subset D^s_L with the size $15\% \times |D_L|$ from D_L; (2) train a classifier with tagged data $D_L - D^s_L$; (3) classify D_U and D^s_L, and then select some examples from D_U as the dataset of new classes, which have the classification confidence less than the average of that in D^s_L. Classification confidence of the example x_i is defined as the absolute value of the difference between two maximum values from the i-th row in labeling matrix.

[8] Initially there are no tagged examples for the m-th class in D'_L. Therefore we do not need to remove tagged examples for this new class, and then directly train a classifier with D'_L.

Table 2: Model order evaluation algorithm.

Function: CV(X_{L+U}, k, q, $Y^0_{X_{L+U}}$)
Input: data set X_{L+U}, model order k, and sampling frequency q;
Output: the score of the merit of k;

1 Run the extended label propagation algorithm with X_L, X_U, k and $Y^0_{X_{L+U}}$;
2 Construct connectivity matrix C_k based on above classification solution on X_U;
3 Use a random predictor ρ_k to assign uniformly drawn labels to each vector in X_U;
4 Construct connectivity matrix C_{ρ_k} using above classification solution on X_U;
5 For $\mu = 1$ to q do
5.1 Randomly sample a subset X^μ_{L+U} with the size $\alpha|X_{L+U}|$ from X_{L+U}, $0 < \alpha < 1$;
5.2 Run the extended label propagation algorithm with X^μ_L, X^μ_U, k and $Y^{0\mu}$;
5.3 Construct connectivity matrix C^μ_k using above classification solution on X^μ_U;
5.4 Use ρ_k to assign uniformly drawn labels to each vector in X^μ_U;
5.5 Construct connectivity matrix $C^\mu_{\rho_k}$ using above classification solution on X^μ_U;
Endfor
6 Evaluate the merit of k using following formula:
$M_k = \frac{1}{q}\sum_\mu (M(C^\mu_k, C_k) - M(C^\mu_{\rho_k}, C_{\rho_k}))$,
where $M(C^\mu, C)$ is given by equation (2);
7 Return M_k;

Then this model order identification procedure can be formulated as:

$$\hat{k}_{X_{L+U}} = argmax_{K_{min} \le k \le K_{max}} \{CV(X_{L+U}, k, q, Y^0_{X_{L+U}})\}. \quad (1)$$

$\hat{k}_{X_{L+U}}$ is the estimated sense number in X_{L+U}, K_{min} (or K_{max}) is the minimum (or maximum) value of sense number, and k is the possible value of sense number in X_{L+U}. Note that $k \ge k_{X_L}$. Then we set $K_{min} = k_{X_L}$. K_{max} may be set as a value greater than the possible ground-truth value. CV is a cluster validation based evaluation function. Table 2 shows the details of this function. We set q, the resampling frequency for estimation of stability score, as 20. α is set as 0.90. The random predictor assigns uniformly distributed class labels to each instance in a given dataset. We run this CV procedure for each value of k. The value of k that maximizes this function will be selected as the estimation of sense number. At the same time, we can obtain a partition of X_{L+U} with $\hat{k}_{X_{L+U}}$ groups.

The function $M(C^\mu, C)$ in Table 2 is given by (Levine and Domany, 2001):

$$M(C^\mu, C) = \frac{\sum_{i,j} 1\{C^\mu_{i,j} = C_{i,j} = 1, x_i, x_j \in X^\mu_U\}}{\sum_{i,j} 1\{C_{i,j} = 1, x_i, x_j \in X^\mu_U\}}, \quad (2)$$

where X^μ_U is the untagged data in X^μ_{L+U}, X^μ_{L+U} is a subset with the size $\alpha|X_{L+U}|$ ($0 < \alpha < 1$) sampled from X_{L+U}, C or C^μ is $|X_U| \times |X_U|$ or $|X^\mu_U| \times |X^\mu_U|$ connectivity matrix based on classification solutions computed on X_U or X^μ_U respectively. The connectivity matrix C is defined as: $C_{i,j} = 1$ if x_i and x_j belong to the same cluster, otherwise $C_{i,j} = 0$. C^μ is calculated in the same way.

$M(C^\mu, C)$ measures the proportion of example pairs in each group computed on X_U that are also assigned into the same group by the classification solution on X^μ_U. Clearly, $0 \le M \le 1$. Intuitively, if the value of k is identical with the true value of sense number, then classification results on the different subsets generated by sampling should be similar with that on the full dataset. In the other words, the classification solution with the true model order as parameter is robust against resampling, which gives rise to a local optimum of $M(C^\mu, C)$.

In this algorithm, we normalize $M(C^\mu_k, C_k)$ by the equation in step 6 of Table 2, which makes our objective function different from the figure of merit (equation (2)) proposed in (Levine and Domany, 2001). The reason to normalize $M(C^\mu_k, C_k)$ is that $M(C^\mu_k, C_k)$ tends to decrease when increasing the value of k (Lange et al., 2002). Therefore for avoiding the bias that the smaller value of k is to be selected as the model order, we use the cluster validity of a random predictor to normalize $M(C^\mu_k, C_k)$.

If $\hat{k}_{X_{L+U}}$ is equal to k_{X_L}, then there is no new sense in X_U. Otherwise ($\hat{k}_{X_{L+U}} > k_{X_L}$) new senses of w may be represented by the groups in which there is no instance from X_L.

3 Experiments and Results

3.1 Experiment Design

We evaluated the ELP based model order identification algorithm on the data in English lexical sample task of SENSEVAL-3 (including all

Table 3: Description of The percentage of official training data used as tagged data when instances with different sense sets are removed from official training data.

	The percentage of official training data used as tagged data
$S_{subset} = \{s_1\}$	42.8%
$S_{subset} = \{s_2\}$	76.7%
$S_{subset} = \{s_3\}$	89.1%
$S_{subset} = \{s_1, s_2\}$	19.6%
$S_{subset} = \{s_1, s_3\}$	32.0%
$S_{subset} = \{s_2, s_3\}$	65.9%

the 57 English words) [9], and further empirically compared it with other state of the art classification methods, including SVM [10] (the state of the art method for supervised word sense disambiguation (Mihalcea et al., 2004)), a one-class partially supervised classification algorithm (Liu et al., 2003) [11], and a semi-supervised k-means clustering based model order identification algorithm.

The data for English lexical samples task in SENSEVAL-3 consists of 7860 examples as official training data, and 3944 examples as official test data for 57 English words. The number of senses of each English word varies from 3 to 11.

We evaluated these four algorithms with different sizes of incomplete tagged data. Given official training data of the word w, we constructed incomplete tagged data X_L by removing the all the tagged instances from official training data that have sense tags from S_{subset}, where S_{subset} is a subset of the ground-truth sense set S for w, and S consists of the sense tags in official training set for w. The removed training data and official test data of w were used as X_U. Note that $S_L = S - S_{subset}$. Then we ran these four algorithm for each target word w with X_L as tagged data and X_U as untagged data, and evaluated their performance using the accuracy on official test data of all the 57 words. We conducted six experiments for each target word w by setting S_{subset} as $\{s_1\}$, $\{s_2\}$, $\{s_3\}$, $\{s_1, s_2\}$, $\{s_1, s_3\}$, or $\{s_2, s_3\}$, where s_i is the i-th most frequent sense of w. S_{subset} cannot be set as $\{s_4\}$ since some words have only three senses. Table 3 lists the percentage of official training data used as tagged data (the number of examples in incomplete tagged data divided by the number of examples in official training data) when we removed the instances with sense tags from S_{subset} for all the 57 words. If $S_{subset} = \{s_3\}$, then most of sense tagged examples are still included in tagged data. If $S_{subset} = \{s_1, s_2\}$, then there are very few tagged examples in tagged data. If no instances are removed from official training data, then the value of percentage is 100%.

Given an incomplete tagged corpus for a target word, SVM does not have the ability to find the new senses from untagged corpus. Therefore it labels all the instances in the untagged corpus with sense tags from S_L.

Given a set of positive examples for a class and a set of unlabeled examples, the one-class partially supervised classification algorithm, LPU (Learning from Positive and Unlabeled examples) (Liu et al., 2003), learns a classifier in four steps:

Step 1: Identify a small set of reliable negative examples from unlabeled examples by the use of a classifier.

Step 2: Build a classifier using positive examples and automatically selected negative examples.

Step 3: Iteratively run previous two steps until no unlabeled examples are classified as negative ones or the unlabeled set is null.

Step 4: Select a good classifier from the set of classifiers constructed above.

For comparison, LPU [12] was run to perform classification on X_U for each class in X_L. The label of each instance in X_U was determined by maximizing the classification score from LPU output for each class. If the maximum score of an instance is negative, then this instance will be labeled as a new class. Note that LPU classifies X_{L+U} into $k_{X_L} + 1$ groups in most of cases.

The clustering based partially supervised sense disambiguation algorithm was implemented by replacing ELP with a semi-supervised k-means clustering algorithm (Wagstaff et al., 2001) in the model order identification procedure. The label information in labeled data was used to guide the semi-supervised clustering on X_{L+U}. Firstly, the labeled data may be used to determine initial cluster centroids. If the cluster number is greater

[9] Available at http://www.senseval.org/senseval3

[10] we used a linear SVM^{light}, available at http://svmlight.joachims.org/.

[11] Available at http://www.cs.uic.edu/∼liub/LPU/LPU-download.html

[12] The three parameters in LPU were set as follows: "-s1 spy -s2 svm -c 1". It means that we used the spy technique for step 1 in LPU, the SVM algorithm for step 2, and selected the first or the last classifier as the final classifier. It is identical with the algorithm "Spy+SVM IS" in Liu et al. (2003).

than k_{X_L}, the initial centroids of clusters for new classes will be assigned as randomly selected instances. Secondly, in the clustering process, the instances with the same class label will stay in the same cluster, while the instances with different class labels will belong to different clusters. For better clustering solution, this clustering process will be restarted three times. Clustering process will be terminated when clustering solution converges or the number of iteration steps is more than 30. $K_{min} = k_{X_L} = |S_L|$, $K_{max} = K_{min} + m$. m is set as 4.

We used Jensen-Shannon (JS) divergence (Lin, 1991) as distance measure for semi-supervised clustering and ELP, since plain LP with JS divergence achieves better performance than that with cosine similarity on SENSEVAL-3 data (Niu et al., 2005).

For the LP process in ELP algorithm, we constructed connected graphs as follows: two instances u, v will be connected by an edge if u is among v's 10 nearest neighbors, or if v is among u's 10 nearest neighbors as measured by cosine or JS distance measure (following (Zhu and Ghahramani, 2002)).

We used three types of features to capture the information in all the contextual sentences of target words in SENSEVAL-3 data for all the four algorithms: part-of-speech of neighboring words with position information, words in topical context without position information (after removing stop words), and local collocations (as same as the feature set used in (Lee and Ng, 2002) except that we did not use syntactic relations). We removed the features with occurrence frequency (counted in both training set and test set) less than 3 times.

If the estimated sense number is more than the sense number in the initial tagged corpus X_L, then the results from order identification based methods will consist of the instances from clusters of unknown classes. When assessing the agreement between these classification results and the known results on official test set, we will encounter the problem that there is no sense tag for each instance in unknown classes. Slonim and Tishby (2000) proposed to assign documents in each cluster with the most dominant class label in that cluster, and then conducted evaluation on these labeled documents. Here we will follow their method for assigning sense tags to unknown classes from LPU, clustering based order identification process, and ELP based order identification process. We assigned the instances from unknown classes with the dominant sense tag in that cluster. The result from LPU always includes only one cluster of the unknown class. We also assigned the instances from the unknown class with the dominant sense tag in that cluster. When all instances have their sense tags, we evaluated the their results using the accuracy on official test set.

3.2 Results on Sense Disambiguation

Table 4 summarizes the accuracy of SVM, LPU, the semi-supervised k-means clustering algorithm with correct sense number $|S|$ or estimated sense number $\hat{k}_{X_{L+U}}$ as input, and the ELP algorithm with correct sense number $|S|$ or estimated sense number $\hat{k}_{X_{L+U}}$ as input using various incomplete tagged data. The last row in Table 4 lists the average accuracy of each algorithm over the six experimental settings. Using $|S|$ as input means that we do not perform order identification procedure, while using $\hat{k}_{X_{L+U}}$ as input is to perform order identification and obtain the classification results on X_U at the same time.

We can see that ELP based method outperforms clustering based method in terms of average accuracy under the same experiment setting, and these two methods outperforms SVM and LPU. Moreover, using the correct sense number as input helps to improve the overall performance of both clustering based method and ELP based method.

Comparing the performance of the same system with different sizes of tagged data (from the first experiment to the third experiment, and from the fourth experiment to the sixth experiment), we can see that the performance was improved when given more labeled data. Furthermore, ELP based method outperforms other methods in terms of accuracy when rare senses (e.g. s_3) are missing in the tagged data. It seems that ELP based method has the ability to find rare senses with the use of tagged and untagged corpora.

LPU algorithm can deal with only one-class classification problem. Therefore the labeled data of other classes cannot be used when determining the positive labeled data for current class. ELP can use the labeled data of all the known classes to determine the seeds of unknown classes. It may explain why LPU's performance is worse than ELP based sense disambiguation although LPU can correctly estimate the sense number in X_{L+U}

Table 4: This table summarizes the accuracy of SVM, LPU, the semi-supervised k-means clustering algorithm with correct sense number $|S|$ or estimated sense number $\hat{k}_{X_{L+U}}$ as input, and the ELP algorithm with correct sense number $|S|$ or estimated sense number $\hat{k}_{X_{L+U}}$ as input on the official test data of ELS task in SENSEVAL-3 when given various incomplete tagged corpora.

	SVM	LPU	Clustering algorithm with $\|S\|$ as input	ELP algorithm with $\|S\|$ as input	Clustering algorithm with $\hat{k}_{X_{L+U}}$ as input	ELP algorithm with $\hat{k}_{X_{L+U}}$ as input
$S_{subset} = \{s_1\}$	30.6%	22.3%	43.9%	47.8%	40.0%	38.7%
$S_{subset} = \{s_2\}$	59.7%	54.6%	44.0%	62.4%	48.5%	62.6%
$S_{subset} = \{s_3\}$	67.0%	53.4%	48.7%	67.2%	52.4%	69.1%
$S_{subset} = \{s_1, s_2\}$	14.6%	13.1%	44.4%	40.2%	35.6%	33.0%
$S_{subset} = \{s_1, s_3\}$	25.7%	21.1%	48.5%	37.9%	39.8%	31.0%
$S_{subset} = \{s_2, s_3\}$	56.2%	53.1%	47.3%	59.4%	46.6%	58.7%
Average accuracy	42.3%	36.3%	46.1%	52.5%	43.8%	48.9%

Table 5: These two tables provide the mean and standard deviation of absolute values of the difference between ground-truth results $|S|$ and sense numbers estimated by clustering or ELP based order identification procedure respectively.

	Clustering based method	ELP based method
$S_{subset} = \{s_1\}$	1.3±1.1	2.2±1.1
$S_{subset} = \{s_2\}$	2.4±0.9	2.4±0.9
$S_{subset} = \{s_3\}$	2.6±0.7	2.6±0.7
$S_{subset} = \{s_1, s_2\}$	1.2±0.6	1.6±0.5
$S_{subset} = \{s_1, s_3\}$	1.4±0.6	1.8±0.4
$S_{subset} = \{s_2, s_3\}$	1.8±0.5	1.8±0.5

when only one sense is missing in X_L.

When very few labeled examples are available, the noise in labeled data makes it difficult to learn the classification score (each entry in Y_{D_U}). Therefore using the classification confidence criterion may lead to poor performance of seed selection for unknown classes if the classification score is not accurate. It may explain why ELP based method does not outperform clustering based method with small labeled data (e.g., $S_{subset} = \{s_1\}$).

3.3 Results on Sense Number Estimation

Table 5 provides the mean and standard deviation of absolute difference values between ground-truth results $|S|$ and sense numbers estimated by clustering or ELP based order identification procedures respectively. For example, if the ground truth sense number of the word w is k_w, and the estimated value is $\hat{k}_w$, then the absolute value of the difference between these two values is $|k_w - \hat{k}_w|$. Therefore we can have this value for each word. Then we calculated the mean and deviation on this array of absolute values. LPU does not have the order identification capability since it always assumes that there is at least one new class in unlabeled data, and does not further differentiate the instances from these new classes. Therefore we do not provide the order identification results of LPU.

From the results in Table 5, we can see that estimated sense numbers are closer to ground truth results when given less labeled data for clustering or ELP based methods. Moreover, clustering based method performs better than ELP based method in terms of order identification when given less labeled data (e.g., $S_{subset} = \{s_1\}$). It seems that ELP is not robust to the noise in small labeled data, compared with the semi-supervised k-means clustering algorithm.

4 Related Work

The work closest to ours is partially supervised classification or building classifiers using positive examples and unlabeled examples, which has been studied in machine learning community (Denis et al., 2002; Liu et al., 2003; Manevitz and Yousef, 2001; Yu et al., 2002). However, they cannot

group negative examples into meaningful clusters. In contrast, our algorithm can find the occurrence of negative examples and further group these negative examples into a "natural" number of clusters. Semi-supervised clustering (Wagstaff et al., 2001) may be used to perform classification by the use of labeled and unlabeled examples, but it encounters the same problem of partially supervised classification that model order cannot be automatically estimated.

Levine and Domany (2001) and Lange et al. (2002) proposed cluster validation based criteria for cluster number estimation. However, they showed the application of the cluster validation method only for unsupervised learning. Our work can be considered as an extension of their methods in the setting of partially supervised learning.

In natural language processing community, the work that is closely related to ours is word sense discrimination which can induce senses by grouping occurrences of a word into clusters (Schütze, 1998). If it is considered as unsupervised methods to solve sense disambiguation problem, then our method employs partially supervised learning technique to deal with sense disambiguation problem by use of tagged and untagged texts.

5 Conclusions

In this paper, we present an order identification based partially supervised classification algorithm and investigate its application to partially supervised word sense disambiguation problem. Experimental results on SENSEVAL-3 data indicate that our ELP based model order identification algorithm achieves better performance than other state of the art classification algorithms, e.g., SVM, a one-class partially supervised algorithm (LPU), and a semi-supervised k-means clustering based model order identification algorithm.

References

Brown P., Stephen, D.P., Vincent, D.P., & Robert, Mercer.. 1991. Word Sense Disambiguation Using Statistical Methods. *Proceedings of ACL.*

Dagan, I. & Itai A.. 1994. Word Sense Disambiguation Using A Second Language Monolingual Corpus. *Computational Linguistics*, Vol. 20(4), pp. 563-596.

Denis, F., Gilleron, R., & Tommasi, M.. 2002. Text Classification from Positive and Unlabeled Examples. *Proceedings of the 9th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems*.

Lange, T., Braun, M., Roth, V., & Buhmann, J. M. 2002. Stability-Based Model Selection. *NIPS 15*.

Leacock, C., Miller, G.A. & Chodorow, M.. 1998. Using Corpus Statistics and WordNet Relations for Sense Identification. *Computational Linguistics*, 24:1, 147–165.

Lee, Y.K. & Ng, H.T.. 2002. An Empirical Evaluation of Knowledge Sources and Learning Algorithms for Word Sense Disambiguation. *Proceedings of EMNLP*, (pp. 41-48).

Levine, E., & Domany, E. 2001. Resampling Method for Unsupervised Estimation of Cluster Validity. *Neural Computation*, Vol. 13, 2573–2593.

Lin, J. 1991. Divergence Measures Based on the Shannon Entropy. *IEEE Transactions on Information Theory*, 37:1, 145–150.

Liu, B., Dai, Y., Li, X., Lee, W.S., & Yu, P.. 2003. Building Text Classifiers Using Positive and Unlabeled Examples. *Proceedings of IEEE ICDM*.

Manevitz, L.M., & Yousef, M.. 2001. One Class SVMs for Document Classification. *Journal of Machine Learning*, 2, 139-154.

Mihalcea R., Chklovski, T., & Kilgariff, A.. 2004. The SENSEVAL-3 English Lexical Sample Task. *SENSEVAL-2004*.

Niu, Z.Y., Ji, D.H., & Tan, C.L.. 2005. Word Sense Disambiguation Using Label Propagation Based Semi-Supervised Learning. *Proceedings of ACL.*

Schütze, H.. 1998. Automatic Word Sense Discrimination. *Computational Linguistics*, 24:1, 97–123.

Wagstaff, K., Cardie, C., Rogers, S., & Schroedl, S.. 2001. Constrained K-Means Clustering with Background Knowledge. *Proceedings of ICML.*

Yarowsky, D.. 1995. Unsupervised Word Sense Disambiguation Rivaling Supervised Methods. *Proceedings of ACL.*

Yu, H., Han, J., & Chang, K. C.-C.. 2002. PEBL: Positive example based learning for web page classification using SVM. *Proceedings of ACM SIGKDD*.

Zhu, X. & Ghahramani, Z.. 2002. Learning from Labeled and Unlabeled Data with Label Propagation. *CMU CALD tech report CMU-CALD-02-107*.