Maximum Entropy Modeling in Sparse Semantic Tagging *

Jia Cui
Center for Language and Speech Processing
Johns Hopkins University
Baltimore, MD, 21210, USA
cuijia@jhu.edu

David Guthrie
Department of Computer Science
University of Sheffield
Sheffield, S1 4DP, UK
D.Guthrie@dcs.shef.ac.uk

Abstract

In this work, we are concerned with a coarse grained semantic analysis over sparse data, which labels all nouns with a set of semantic categories. To get the benefit of unlabeled data, we propose a bootstrapping framework with Maximum Entropy modeling (MaxEnt) as the statistical learning component. During the iterative tagging process, unlabeled data is used not only for better statistical estimation, but also as a medium to integrate non-statistical knowledge into the model training. Two main issues are discussed in this paper. First, Association Rule principles are suggested to guide MaxEnt feature selections. Second, to guarantee the convergence of the bootstrapping process, three adjusting strategies are proposed to soft tag unlabeled data.

1 Introduction

Semantic analysis is an open research field in natural language processing. Two major research topics in this field are Named Entity Recognition (NER) (N. Wacholder and Choi, 1997; Cucerzan and Yarowsky, 1999) and Word Sense Disambiguation (WSD) (Yarowsky, 1995; Wilks and Stevenson, 1999). NER identifies different kinds of names such as ”person”,”location” or ”date”, while WSD distinguishes the senses of ambiguous words. For example, ”bank” can be labeled as ”financial institution” or ”edge of a river”. Our task of semantic analysis has a more general purpose, tagging all nouns with one semantic label set. Compared with NE, which only considers names, our task concerns all nouns. Unlike WSD, in which every ambiguous word has its own range of sense set, our task aims at another set of semantic labels, shared by all nouns. The motivation behind this work is that a semantic category assignment with reliable performance can contribute to a number of applications including statistical machine translation and sub-tasks of information extraction.

*This work was supported in part by NSF grant numbers IIS-0121285. Any opinions, fndings and conclusions or recommendations expressed in this material are the authors' and do not necessarily reflect those of the sponsor.

The semantic categories adopted in this paper come from the Longman Dictionary. These categories are neither parallel to nor independent of each other. One category may denote a concept which is a subset of that of another. Examples of the category structures are illustrated in Figure 1.

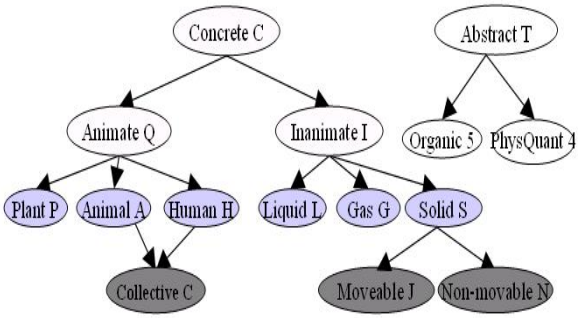

Figure 1: Structure of some semantic categories used in the Langman dictionary.

Maximum Entropy (MaxEnt) principle has been successfully applied in many classification and tagging tasks (Ratnaparkhi, 1996; K. Nigam and A.McCallum, 1999; A. McCallum and Pereira, 2000). We use MaxEnt modeling as the learning component. A major issue in MaxEnt training is how to select proper features and determine the feature targets (Berger et al., 1996; Jebara and Jaakkola, 2000). To discover useful features, we exploit the concept of Association Rules (AR) (R. Agrawal and Swami, 1993; Srikant and Agrawal, 1997), which is originally proposed in Data Mining research field to identify frequent itemsets in a large database.

Like many other classification tasks, human-annotated data for semantic analysis is expensive and limited, while a large amount of unlabeled data is easily obtained. Many researchers (Blum and Mitchell, 1998; K. Nigam and Mitchell, 2000; Corduneanu and Jaakkola, 2002) have attempted to improve performance with unlabeled data. In this paper, we also propose a framework to bootstrap with unlabeled data. Fractional counts are assigned to unlabeled instances based on current model and accessible knowledge sources. Pooled with human-annotated data, unlabeled data contributes to the next MaxEnt model.

We begin with an introduction of our bootstrapping framework and MaxEnt training. An interesting MaxEnt puzzle is presented, with its derivation showing possible directions of utilizing unlabeled data. Then the feature selection criterion guided by AR is discussed as well as the indicator selection (section 3). We discuss initialization methods for the unlabeled data. Strategies to guarantee convergence of bootstrapping process and approaches of integrating non-statistical knowledge are proposed in section 4. Finally, experimental results are presented along with conclusions and future work.

2 Bootstrapping and the MaxEnt puzzle

An instance in the corpus includes the headword, which is the noun to be labeled, and its context. To integrating the unlabeled instances in the training process, we propose a tagging method called *soft tagging*. Unlike the normal tagging, in which each instance is assigned only one label, soft tagging allows one instance to be assigned several labels with each of them being associated with a fractional credit. All credits assigned to one instance should sum up to 1. For example, a raw instance with the headword "club" can be soft tagged as (movable J:0.3, not-movable N:0.3, collective U:0.4). Once all auxiliary instances have been assigned semantic labels, we pool them together with human-annotated data to select useful features and set up feature expectations. Then a log-linear model is trained for several iterations to maximize likelihood of the whole corpora. With the updated MaxEnt model, unlabeled data will be soft tagged again. The whole process is repeated until convergence condition is satisfied. This framework is illustrated in Figure 2.

In building a mexent model, unlabeled data contributes differently to the feature selection and target estimation compared to the human-annotated data. While human-annotated instances never change, tags in unlabeled data keep updating according to the new MaxEnt model in each bootstrapping iteration, which might lead to a different feature set.

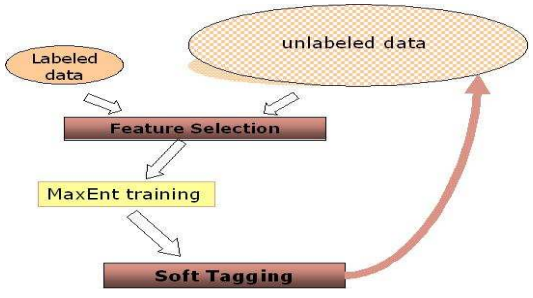

Figure 2: The bootstrapping framework

Before presenting the MaxEnt puzzle, let's first consider a regular MaxEnt formulation. Let l be a label and x be an instance. Let $k_h(l,x)$ be the h-th binary feature function which is equal to 1 if (l,x) activates the h-th feature and 0 otherwise. $P(l,x)$ is given by the model, denoting the probability of observing both l and x. $\widehat{P}(l,x)$ is the empirical frequency of (l,x) in the training data. The constraint associated with feature $k_h(l,x)$ is represented as:

$$\sum_{l,\mathbf{x}} P(l,\mathbf{x})\, k_h(l,\mathbf{x}) = \sum_{l,\mathbf{x}} \widehat{P}(l,\mathbf{x})\, k_h(l,\mathbf{x}) \qquad (1)$$

In practice, we are interested in conditional models $P(l|x)$, which assign probability mass to a label set given supporting evidence x. And we approximate $P(x)$ with $\widehat{P}(\mathbf{x})$. Consequently,

$$\sum_{\mathbf{x}} \widehat{P}(\mathbf{x}) \sum_{l} P(l|\mathbf{x})\, k_h(l,\mathbf{x}) = \sum_{l,\mathbf{x}} \widehat{P}(l,\mathbf{x})\, k_h(l,\mathbf{x}) \qquad (2)$$

Next, we show how the effect of unlabeled data disappears during the bootstrapping in a restricted situation. We make the following assumptions:

1. The feature targets are the raw frequencies in the training data. No smoothing is applied.
2. During the bootstrapping, we maintain a fixed feature set, while the feature targets might change.
3. The fractional credit assigned to label l and instance x in the $t+1$ iteration is $P^t(l|x)$.

Let N be the total number of instance($1,\cdots,N$) in labeled data and M be the total number of instance($N+1,\cdots,N+M$) in unlabeled data. Let λ be the weight assigned to unlabeled instances. The new constraint for the h-th feature function can be rewritten as:

$$\begin{aligned} & \frac{1}{N+\lambda M} \sum_{i=1}^{N} \sum_{l} P^{t+1}(l|\mathbf{x}_i)\, k_h(l,\mathbf{x}_i) \qquad (3) \\ + & \frac{\lambda}{N+\lambda M} \sum_{i=N+1}^{N+M} \sum_{l} P^{t+1}(l|\mathbf{x}_i)\, k_h(l,\mathbf{x}_i) \\ = & \frac{1}{N+\lambda M} \sum_{i=1}^{N} k_h(l_i,\mathbf{x}_i) \\ + & \frac{\lambda}{N+\lambda M} \sum_{i=N+1}^{N+M} \sum_{l} P^{t}(l|\mathbf{x}_i)\, k_h(l,\mathbf{x}_i) \end{aligned}$$

where t is the index of the bootstrap iterations. l_i is the human-annotated label attached to the ith instance.

When the procedure converges, $P^t(l|x_i) = P^{t+1}(l|x_i)$, the second parts on both sides will cancel each other. Finally, the constraint will turn out to be:

$$\sum_{i=1}^{N} \sum_{l} P^{t+1}(l|\mathbf{x}_i)\, k_h(l,\mathbf{x}_i) = \sum_{i=1}^{N} k_h(l_i,\mathbf{x}_i) \qquad (4)$$

which includes statistics only from labeled data. If the feature set stays the same, unlabeled data will make no contribution to the whole constraint set. A model which can satisfy Equation 3 must be able to satisfy Equation 4. If unlabeled instances satisfy the same distribution as labeled instances, given the same set of constraints, the model trained on only labeled data would be equivalent to the one trained on both.

The above derivation shows that with the three restrictions, the soft tagged instances make no contribution to the constraint set, which is counter-intuitive. That is why we call it a "MaxEnt Puzzle". Consequently, to utilize unlabeled data, we should break the three restrictions: to reselect feature set after each bootstrapping iteration, or adjust the soft tagging results given by the model.

3 Feature selection

Berger et al (1996) proposed an iterative procedure of adding news features to feature set driven by data. We present a simple and effective approach using some statistical heuristics for feature selection.

There are two kinds of features in MaxEnt modeling: marginal features and conditional features. Marginal features represent the priors of the semantic categories. Conditional features are composed of an indicator and a label. An indicator can be a single word, an array of words or a word cluster. Similarly, a label can be either a single semantic category or a set of categories.

One major motivation of combining unlabeled data with labeled data in the training process is that unlabeled data can provide a large pool of valuable feature candidates as well as more statistical information. An (indicator, label) pair may appear only once in labeled data, but it may occur very frequently in soft tagged data, it might be selected as a feature during the bootstrapping. Similarly, a feature selected only from human-annotated data may become relatively infrequent when the soft tagged instances are added, thus being eliminated from the feature set.

3.1 Indicators

A good indicator should have an obvious preference for definite semantic categories. For instance, "good","great" are weak indicators because they can modify almost all semantic categories; "drinkable" and "light-weighted" are usually associated with "liquid (L)" and "movable (J)" respectively, thus are strong indicators. The quality of an indicator can be measured by the entropy of the semantic categories it modifies. The lower the entropy is, the stronger preference the indicator has.

To overcome the data sparsity problem, several indicators can be clustered together to form a new single indicator. There are two possible ways of clustering. One is hard clustering, which divides all words into non-overlapping classes; The other is soft clustering, which permits one word to belong to different classes. Usually, words with similar semantic preferences are clustered together.

3.2 Association rules in feature selection

An indicator can be combined with different labels to form different features. But not all pairs including good indicators are good features. In a statistical learning algorithm, a good feature should first be statistically reliable, which is considered in most of the MaxEnt modeling implementations. However, with this single requirement, it is possible that (x, l) is a feature if both x and l are frequent, but the chance of seeing label l is tiny given the presence of indicator x. More constraints are necessary to pinpoint indicative features. We exploit Association Rule (AR) principles to put more restrictions in feature selecting.

Let x be an indicator and l be a label. $count(x, l)$ denotes the number of instances which include indicator x and are attached with label l in the training corpus. N is the total number of instances. We define three measurements to characterize the goodness of a given (x, l) pair from three different points. We put thresholds on each of them.

$support(x, l)$ denotes how frequently indicator x and label l occur together.

$$support(x, l) = \frac{count(x, l)}{N} \tag{5}$$

$confidence(x, l)$ denotes how likely it is to observe label l when indicator x is at present.

$$confidence(x, l) = \frac{count(x, l)}{count(x)} \tag{6}$$

$improvement(x, l)$ shows how much more likely to see label l when indicator x is observed relative to the overall probability of seeing l.

$$improvement(x, l) = \frac{count(x, l)/count(x)}{count(l)/N} \tag{7}$$

The concept of $improvement$ can be extended to each part of the indicators. Let S be the feature set.

$$improvement(x, l) = min_{x' \in x, (x', l) \in S} \frac{confidenc(x, l)}{confidenc(x', l)} \tag{8}$$

For instance, assume the improvement threshold is 1. Let x=x_1x_2, if $confidence(x_1, l) > confidence(x_1x_2, l)$, then (x_1x_2, l) will be ignored and only (x_1, l) is selected as a feature; otherwise, only (x_1x_2, l) remains and (x_1, l) is ignored. This idea can be applied to remove embedded features, reducing the redundancy in the feature set.

The assumption behind those criteria is that most features should be positive, i.e., given feature (x,l), the existence of indicator x always increase the odds of seeing label l, other than decrease it.

4 Soft tagging

Initialization of unlabeled instances is important for our bootstrapping framework, because its effect will be carried on from one iteration to another. Even if the auxiliary data (unlabeled data included in training) has no effect on the constraint set under some circumstances, the soft tagged auxiliary data is still a part of the training corpus, the likelihood of which is to be maximized in the MaxEnt training. On one hand, we wish the initialization to be as close as possible to the real value, of which we have no idea except for the knowledge from labeled data; on the other hand, we wish the tagging of the auxiliary data could be slightly different from the model trained only on labeled data, to avoid converging to this model itself.

As we have seen in section 2, using the model generated in the previous iteration directly to soft tag the auxiliary data

is one of the causes diluting the effect of the auxiliary data. Therefore, we propose to adjust the soft tagging results by non-statistical knowledge or through fixing some of the tags.

4.1 Initialization of unlabeled data

Inspired by the MaxEnt puzzle, we realize the importance of additional knowledge sources. To assign inital fractional credits to unlabeled instances, we look up a dictionary to narrow down choices. In particular,

1. If the headword of the instance has a unique semantic label choice according to the dictionary, it gets credit 1 for this label and 0 for others.

2. If the headword has been observed in labeled data and it has more than one possible labels in the dictionary, it is initialized as follows:

   Let w be the headword and l be a label. m is the total number of permitted labels for w according to the dictionary. Let $L(w)$ be the set of (w, l) pairs in labeled data and $D(w)$ be the set of (w, l) permitted in the dictionary.

   If $D(w) = L(w)$, then

   $$credit(w,l) = \begin{cases} \frac{count(w,l)}{count(w)} & \text{if } (w,l) \in L(w) \\ 0 & \text{otherwise} \end{cases}$$

   Otherwise, we add one and renormalize:

   $$credit(w,l) = \begin{cases} \frac{count(w,l)+1}{count(w)+m} & \text{if } (w,l) \in D(w) \\ 0 & \text{otherwise} \end{cases}$$

3. If the headword never occurs in human-annotated data, we initialize the credits with the flat distribution.

   $$credit(w,l) = \begin{cases} \frac{1}{m} & \text{if } (w,l) \in D(w) \\ 0 & \text{otherwise} \end{cases}$$

4.2 Adjust credits according to their distributions

Considering the bootstrapping process as a tug-of-war between labeled and unlabed data, with the flag denoting the model. In each bootstrapping iteration, tags of the auxiliary data are updated in accordance with the model. Unlabeled data is moving towards the labeled data while the latter stands still. One way to prevent the auxiliary data from being dragged completely to the point of the labeled data is to nail down part of the auxiliary data tags. For example, suppose an instance gets credits (0.8,0.15,0.05). Since the first category is much more likely than the others, we can fix the credits as (1,0,0) for ever. Another choice would be to remove the least probable category, i.e., changing the credits to (0.84,0.16,0). An even more radical method is to remove the whole instance from the auxiliary data if it has no strong semantic preferences.

Here are the details of these adjusting strategies:

rmlow set the least probable category with credit zero and renormalize the remaining credits. Whether to apply this action or not can depend on thresholding on the mininal credit, the ratio of the minimal credit to the maximal credit, or the entropy of the credit distribution.

kphigh assign all credit 1 to the most probable category. Possible thresholds can be set for the maximal credit, the ratio of the maximal credit to the second maximal credit, or the entropy of the credit distribution.

rmevent remove the instance if it has a flat credit distribution after several iterations. Possible thresholds can be set for the maximal credit, the ratio of the maximal credit to the minimal credit, or the entropy of the credit distribution.

The effect of those actions is permanent. That is, if a category is forbidden (set to zero) for one instance, it will be forbidden for this instance in the following iterations. Similarly, if an instance is removed, it will never be used in the training process again.

4.3 Adjust credits with non-statistical knowledge

Non-statistical knowledge is the knowlege which is not obtainable from statistics in a sparse data set, but rather from other resources like a dictionary or WordNet. This kind of knowledge may not be expressed in labeled data, therefore could not be used to form features. For instance, we can obtain easily from a dictionary that word "long-tailed" is used to modify animals. But if "long-tailed" is rarely observed in the training data, it will not not selected as a feature, thus being ignored by the model.

To take advantage of non-statistical knowledge, we use it to adjust soft tagging results. One possibility is pruning the illegal categories with a dictionary. By doing so, this kind of knowledge can affect the MaxEnt modeling indirectly. Similarly, the preference of a context word could also be used to adjust the credit distributions.

5 Experiments

5.1 Corpus

The corpus used in our experiments is provided by Sheffield University. The human-annotated data is divided into training (SHD) and test (Blind). SHD contains 197K instances and Blind contains 13K instances.

The Longman dictionary provides 36 semantic categories, among which only 21 most frequently used categories with the exception of "unknown (Z)" are used in our experiments. The distribution of the semantic categories is far from uniform. Semantic category "abstract (T)" alone forms 60% (126K) of the human-annotated instances. The histogram of the other categories is shown in Figure 3.

The inter-annotator agreement for the labeled data is 95% with exact match. Even though some semantic categories form a hierarchical structure, we always assume that human annotators chose the most specific categories. Evaluation of the classification error rate (CER) in the following experiments also uses exact match only.

5.2 Feature selection with Association Rule principles

The first group of experiments are designed to test different indicators as well as the feature selection strategies, without using unlabeled instances or bootstrapping. The indicators and feature selection thresholds which result in the best performance are used in the bootstrapping experiments.

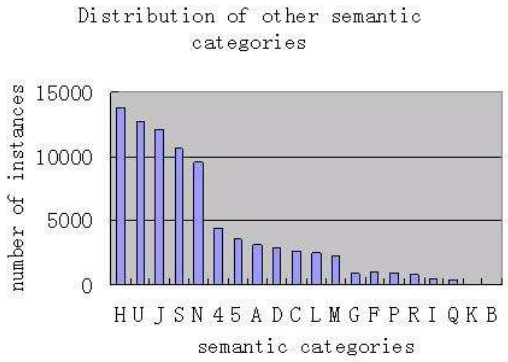

Figure 3: The number of instances labeled with semantic categories other than T

Since 96% of the headwords in Blind are also in SHD, headwords are the most effective indicators. Using headwords alone can get 9.47% classification error rate. Other indicators investigated include adjectives, which modify headwords, and their clusters. As we mentioned before, clustering can alleviate data sparseness problem. So we group headwords and adjectives. Headword clusters are soft clusters. The ambiguous headwords can belong to several clusters as long as the corresponding categories of those clusters are permitted for these headwords in the dictionary. Adjectives are hard clustered. We discard general common adjectives according to their entropies. We keep only those with strong semantic preferences and assign them to the clusters corresponding to their best preferences.

We also include compound features such as headword-adjective, headword-adjectivecluster and other combinations.

Table 1 shows the experimental results with all types of indicators which are mentioned above, and compared the results with the baseline. For different types of features, we set up different thresholds. In config-1, only the $support$ thresholds are used. From config-2 to config-5, we raise $confidence$ and $improvement$ thresholds gradually to lower the number of features. The CER results show that with proper thresholds, using $confidence$ and $improvement$ not only helps decreasing the feature set size by as much as 26%, but also improves the performance.

The last column of the table 1 shows the p-value of significnce test between each configuration and its precedence. Clearly, config 1,2,and 3 perform similar to each other, though they are significant better than using headwords only. And config 4 and 5 are significant better than config 3.

5.3 Bootstrapping

In this group of experiments, we use Blind as the auxiliary data in the bootstrapping process. Blind data is initialized with the methods mentioned in section 4. No human labels in Blind are included in the training process. The weight for the auxiliary data λ is set to 1 in our experiments. The Longman dictionary is always used in soft tagging to eliminate the illegal categories.

The bootstrapping process will stop when the constraint set does not change. Since the feature set is re-selected in each iteration, the convergence of the iterative steps is difficult to achieve. Fluctuations might be observed. With strategies proposed in 4.2, more and more auxiliary instances can be fixed with one semantic category. The whole process will stop at some point.

configuration	# of features	CER(%)	p-value
headwords only	12514	9.47	-
config-1	49861	8.18	$< 1.0e^{-12}$
config-2	44141	8.18	0.53
config-3	39136	7.94	0.034
config-4	36631	7.62	$< 1.0e^{-3}$
config-5	32214	7.64	0.65

Table 1: Classification Error Rates (CER) with different feature sets

Figure 4 plots the CER in the first 20 iterations for different setups. It shows adding Blind into the bootstrapping process can lower the error rate to 7.37%. Without any soft tagging adjustment (normal), the training process does not stop and the CER drops at first but then climbs up later.

Using entropy as thresholds is an efficient way to ensure the convergence for both **rmlow** and **kphigh**. Action **rmevent** can also keep the results from getting much worse during bootstrapping.

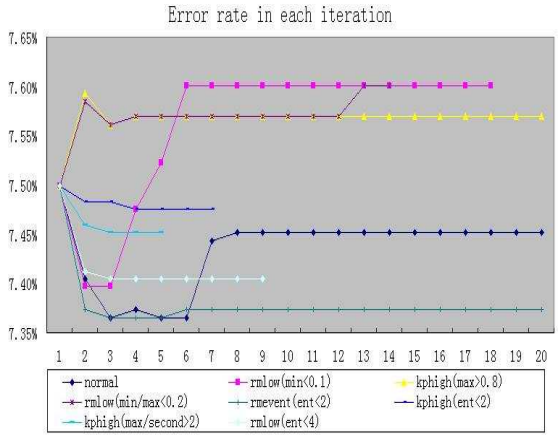

Figure 4: Error rate in each iteration for different soft tagging strategies

6 Conclusions

We introduce a general semantic analysis task, which labels all nouns with a unique set of semantic categories. In order to get benefitted from unlabeled data, we propose a bootstrapping framework with MaxEnt modeling as the main learning component. Unlabeled data, after soft tagging, can be combined with labeled data to provide evidence for MaxEnt feature selection and target estimation. Moreover, non-statistical knowledge can affect the modeling indirectly by adjusting the soft tagging results of the auxiliary data.

For MaxEnt training, we suggest using Association Rule principles to control the feature selection. The intuition behind these criteria is that the very frequent, strongly prefered (indicator,label) pairs are good enough for model training.

Using AR principles not only decreases the size of the feature set dramatically, but also improves the performances. But there is no good way to set the AR thresholds properly yet.

With the feature set changing in each iteration, the convergence of the bootstrapping is hard to guarantee. By sharpening the soft tagging results through removing the least probable label or keeping only the most probable label, we can speed up convergence.

7 Future work

At present, only headwords and adjectives are used to compose indicators. We plan to incorporate linguistically enriched features. Using parsing, we can locate verbs and use their relationships with headwords as new indicators. Another type of potential indicators are the subject codes of the headwords. Subject codes represent finer grained semantic classifications. Some of the instances in human-annotated data have been marked with subject codes. There are a total of 320 subject codes, such as "architecture","agriculture" and "business". We believe that knowing the subject codes of the headwords will help to decrease the entropy of the headword senses.

We have a very large data set BNC corpus, which comes from the same source as the labeled data we use. BNC data is composed of 1.2M unambiguous instances, 1.4M seen instances (ambiguous instances with headwords being observed in SHD) and 0.4M unseen instances (ambiguous instances with headwords never being observed in SHD). In the future, we will try bootstrapping with BNC data.

Alternative indicator clustering techniques will be explored too. We have used entropy as a heuristic in the experiments; an alternative heuristic that can be employed is mutual information.

We have provided a framework to utilize non-statistical knowledge. In addition to using a dictionary to limit the choices of unlabeled data, we can obtain plenty of information about word sense preferences from the WordNet for soft tagging adjustment.

The bootstrapping process is closely related to Expectation-Maximization procedure, in which soft tagging can be ragarded as the E-step. In many practical EM implementations, however, updating at the E-step does not use the exact theoretical value. The modification taken in the E-step can be a linear combination of the old value and the new calculated one. Similar re-estimation strategies can be applied in our work. A theoretical description of the relationship between EM and soft tagging would potentially be able to identify convergency properties of the bootstrapping framework.

Acknowledgments

The authors would like to thank Prof. Frederick Jelinek and Dr. Louise Guthrie and all other members of the Semantic group in 2003 JHU Summer Research Workshop. They are also grateful to the anonymous reviewers for their very insightful comments and suggestions.

References

D. Freitag A. McCallum and F. Pereira. 2000. Maximum entropy markov models for information extraction and segmentation. In *Proc. 17th International Conf. on Machine Learning*, pages 591–598.

A. L. Berger, S. D. Pietra, and V. J. D. Pietra. 1996. A maximum entropy approach to natural language processing. *Computational Linguistics*, 22(1):39–71.

A. Blum and T. Mitchell. 1998. Combining labeled and unlabeled data with co-training. In *COLT: Proceedings of the Workshop on Computational Learning Theory*.

A. Corduneanu and T. Jaakkola. 2002. Continuation methods for mixing heterogeneous sources. In *Proceedings of the Eighteenth Annual Conference on Uncertainty in Artificial Intelligence*.

S. Cucerzan and D. Yarowsky. 1999. Language independent named entity recognition combining morphological and contextual evidence. In *joint SIGDAT Conference on Empirical Methods in NLP and Very Large Corpora*, pages 90–99.

T. Jebara and T. Jaakkola. 2000. Feature selection and dualities in maximum entropy discrimination. In *Uncertainity In Artificial Intellegence*.

J.Lafferty K. Nigam and A.McCallum. 1999. Using maximum entropy for text classification. In *IJCAI-99 Workshop on Machine Learning for Information Filtering*, pages 61–67.

S. Thrun K. Nigam, A. K. McCallum and T. M. Mitchell. 2000. Text classification from labeled and unlabeled documents using em. *Machine Learning*, 39(2-3):103–134.

Y. Ravin N. Wacholder and M. Choi. 1997. Disambiguation of proper names in text. In *Proceedings of Fifth Conference on Applied Natural Language Processing*, pages 202–208.

T. Imielinski R. Agrawal and A. N. Swami. 1993. Mining association rules between sets of items in large databases. In Peter Buneman and Sushil Jajodia, editors, *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data*, pages 207–216, Washington, D.C., 26–28 .

A. Ratnaparkhi. 1996. A maximum entropy model for part of speech tagging. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 133–142.

R. Srikant and R. Agrawal. 1997. Mining generalized association rules. *Future Generation Computer Systems*, 13(2–3):161–180.

Y. Wilks and M. Stevenson. 1999. The grammar of sense: Using part-of-speech tags as a first step in semantic disambiguation. *Journal of Natural Language Engineering*, 4(3).

D. Yarowsky. 1995. Unsupervised word sense disambiguation rivaling supervised methods. In *Meeting of the Association for Computational Linguistics*, pages 189–196.