ON THE USE OF RIGHT CONTEXT IN SENSE-DISAMBIGUATING LANGUAGE MODELS

Vincent CHOW and Dekai WU

Human Language Technology Center
Department of Computer Science, HKUST
Clear Water Bay, Hong Kong
{athena,dekai}@cs.ust.hk

ABSTRACT

We investigate the utility of right-context (look-ahead information) in incremental left-to-right language models with word sense disambiguation, and discover somewhat unexpectedly that using right-context in addition to left-context (history) may actually *reduce* accuracy.

We employ word sense disambiguation as one component of a language model designed to allow hypothesis to be evaluated incrementally. In our baseline system, disambiguation is performed by a naïve-Bayes classifier that uses lexical co-occurrence features from the history.

We then augment the left-context only model with three well-motivated methods using the right-context. Perhaps surprisingly, experiment results with the three look-ahead strategies shown a 0.19% up to 10.04% *decrease* in the accuracy of disambiguating the next word.

1. INTRODUCTION

In this paper a word sense disambiguation model based on naïve-Bayes classifier will be discussed as one component of a language model designed to allow hypothesis to be evaluated incrementally. At any given time, we have a history (left-context) of words already recognized, which can be used in predicting the next word. The function of the word sense disambiguation component is to select a particular sense of each word in the history in order to improve the predictive power on the words to come. (For our speech translation project, we consider an English word to have multiple senses if it has different translations in Chinese.) Clearly, improving disambiguation accuracy leads directly to improved language model predictions.

It would seem that disambiguation accuracy could be improved if we had access to right-context words, i.e., if we could look ahead, as is common among parser architectures. To evaluate this, we constructed a variant disambiguation model, in which the left-context features are augmented by right-context features. (In this experiment, real look-ahead is available since we know the full sentence ahead of time, but in real system there are alternative ways of simulating the same effect.) Note that words in the history are already disambiguated, while words in the look-ahead have not yet been disambiguated. The evidence from look-ahead words must therefore be incorporated into the naïve-Bayes classifier in a different way than the evidence from history words. This paper presents of a comparative study on three well-motivated methods on using the right-context.

Word sense disambiguation is being a hot topic in natural language processing, as most researches focusing on monolingual senses [1, Brown et al.1991] [2, Gale et al.1992] [4, Yarowsky 1993] [3, Pedersen et al.1997]. As a component of our speech translation project, our word sense disambiguator utilizes the information from the second language to help disambiguating the word senses in the source language.

We begin with a short description of a baseline word sense disambiguation models (in section 2). Strategies on making use of the look-ahead information to help disambiguation (in section 3) will be discussed next, followed by the experimental results of disambiguation experiments on a 1,000-words English text extracted from the Wall Street Journal (in section 4). We will then give our conclusion at the end of this paper.

2. BASELINE MODEL

A bilingual word sense disambiguation model using naïve-Bayes classifier is applied as the incremental left-to-right language model in a speech translation system. For an input text $W = \{w_1, w_2, \ldots, w_m\}$ in the source language, we would like to find the best translation $S^* = \{s_1^*, s_2^*, \ldots, s_n^*\}$ in the target language, given

that each words w_i could have a set of candidate senses $\{s_{i1}, s_{i2}, \ldots, s_{ik_i}\}$ where k_i is the size of the set of candidate senses of w_i. Our goal is to find the optimal translation S^* such that

$$S^* = \arg\max_S P(S|W) \quad (1)$$

Applying an incremental left-to-right algorithm, at position i, let the disambiguated words (history) be $\{s_1, s_2, \ldots, s_{i-1}\}$ and let the undisambiguated words in the right context be $\{w_{i+1}, w_{i+2}, \ldots, w_m\}$, our goal is to find the best sense s_{ij} among the set of candidate senses $\{s_{i1}, s_{i2}, \ldots, s_{ik_i}\}$ for w_i.
In the original model using left-context only, the best sense, s_i, is estimated from

$$\arg\max_{s_{ij}} P(s_{ij}|s_1, s_2, \ldots, s_{i-1})$$

or further approximated by

$$\arg\max_{s_{ij}} \{P(s_1|s_{ij}) \ldots P(s_{i-1}|s_{ij}) P(s_{ij})\} \quad (2)$$

Note that the conditional probabilities were estimated from bigram and the last term was estimated from unigram trained on a huge monolingual corpus in the target language.
When the left context is augmented with the right context, the best sense becomes

$$\arg\max_{s_{ij}} P(s_{ij}|s_1, s_2, \ldots, s_{i-1}, w_{i+1}, \ldots, w_m)$$

or,

$$\arg\max_{s_{ij}} \{(LS)(RS)P(s_{ij})\} \quad (3)$$

which $LS = P(s_1|s_{ij})P(s_2|s_{ij}) \ldots P(s_{i-1}|s_{ij})$
and $RS = P(w_{i+1}|s_{ij})P(w_{i+2}|s_{ij}) \ldots P(w_m|s_{ij})$.
We will examine three proposed methods to estimate $P(w_n|s_{ij})$ in the next section.

3. PROPOSED STRATEGIES

3.1. Method A

Since the sense s_{ij} need not be an entry in the bilexicon (dictionary) for the word w_n, i.e. s_{ij} need not be a translation of the word w_n, we need to store the translation probabilities for every English words given every Chinese words. As we have more than 10^5 words in both the English and Chinese lexicons, the storage space needed would be of the magnitude of 10^{10}. Although today's technology allow us to store 10GB of data, but we still need a huge bilingual corpus in order to get a good maximum likelihood estimation, in which a good quality and large scale bilingual corpus is not easily obtainable.
Therefore we choosed to indirectly calculate $P(w_n|s_{ij})$ using the candidate senses of w_n as a medium. By assuming that each of the candidate senses would have the same contribution, the formula now becomes,

$$P(w_n|s_{ij}) \approx \sum_k P(w_n|s_{nk}) P(s_{nk}|s_{ij}) \quad (4)$$

Notice that we can use the same set of bigram information to calculate the latter term and a smaller bilingual corpus will be enough to pre-calculate all the possible $P(w_n|s_{nk})$.

3.2. Method B

Although we are capable to pre-calculate the probabilities $P(w_n|s_{nk})$, it requires huge intermediate storage spaces during the training and the *data sparseness problem* is still a major factor affecting the quality of the estimates. We modify the equation by further assuming that the term $P(w_n|s_{nk})$ might have even distribution w.r.t. s_{nk}. Roughly speaking, the term would have approximately constant values regardless the value of s_{nk}.

$$P(w_n|s_{ij}) \approx |k|^{-1} \sum_k P(s_{nk}|s_{ij}) \quad (5)$$

Now there is not need to calculate $P(w_n|s_{nk})$ and efficiency gain is expected. If the quality of the $P(w_n|s_{nk})$ estimation is really uniformly distributed, then we will not suffer from serious accuracy penalties.

3.3. Method C

As in the baseline model, sometimes a single sense in the context will serve as a "triggered pair" with the current candidate sense in-process. Moreover, disambiguation errors on previous steps might introduce some noise to the score of the current sense. The same argument applies to the right-context-enable mode, while we cannot guarantee that $P(s_{nk}|s_{ij})$ will not introduces noises. Therefore we go one more step further to simplify the equation, by removing the computationally expensive summation in logarithmic scale and replace it with a maximization. Note that we could ignore the term $|k|^{-1}$ because it remains constant in the maximization.

$$P(w_n|s_{ij}) \approx \max_k P(s_{nk}|s_{ij}) \quad (6)$$

With this strategy, we enjoy from more efficiency gain while it might improve accuracy.

	% decrease in accuracy	% increase in time needed
Method A	1.7374	*50.5847*
Method B	*10.0386*	41.2280
Method C	0.1930	35.9649

Table 1: Summary on experimental results

4. EXPERIMENTS

In order to evaluate the performance of the three strategies, direct comparisons on identical training and test sets were performed. Each models using the three proposed methods will be compared to the baseline model in terms of accuracy and efficiency.

4.1. Experimental Methodology

A 500-megabyte corpus of two Chinese newspapers from Hong Kong and China is acquired to train the unigram and bigram information using maximum likelihood estimation with add-one smoothing. A manually created bi-lexicon consisting of 102,331 English words, 100,071 Chinese words and 140,942 bilingual mappings (i.e. dictionary entries) is used to find the set of Chinese candidate senses for each English word in query. Test sets of size 1,000 was extracted from the Wall Street Journal which are then manually disambiguated.

The same set of compromised (i.e. approximately in the same domain) n-gram data, bi-lexicon and test sets was applied to the four models (the baseline as well as the three variations).

4.2. Experimental Results

Disambiguation accuracy and the time needed were recorded for each of the test cases on each of the four models. Each of the three variations were compared with the baseline model and a summary is given by Table 1. Experimental result suggested that, using look-ahead information, with the estimation strategies suggested in section 3, the accuracy tends to *decrease* while the efficiency significantly drops as well.

4.3. Discussion of Results

We learnt from the experimental results that using the three proposed methods as an estimate on the term $P(w_n|s_{ij})$ is not suggested. They reduce the accuracy and make the system inefficient as well. Particularly, in Method B, we observed a significant 10% decrease in accuracy. This suggested that the uniformly distributed assumption on the term $P(w_n|s_{nk})$ for all k is incorrect, i.e. the transposed lexicon does give some (if not complete) knowledges.

The experimental results shown that using more complex formula on the context part often introduced decrease in accuracy. As probabilities are small numbers, multiplying a series of small values resulted an extremely small number. In turns, the rigid unigram probability $P(s_{ij})$ became dominant in all the cases. The problem was often being introduced from improper use of approximation in the intermediate steps, or in other terms, we need to improve the quality of the bigram estimation, which is the main source of approximation when calculating the context score.

Method C gave the best result and note that it used exactly the same searching criteria on both the left and the right context, as it produced approximately the same accuracy in either cases, we can conclude that by using naïve-Bayes estimator and bigram information, left and right context gave the same effect on the scores of word senses.

5. CONCLUSION

As described in section 4, results with the three look-ahead methods respectively shown a 1.74%, 10.04%, and 0.19% *decrease* in the accuracy of disambiguating the next word. This result strengthens our belief that (1) incremental left-to-right processing is a reasonable processing paradigm even when taking word sense disambiguation into account, since the contribution of right-context is limited, and (2) using look-ahead context would likely necessitate sacrificing pure left-to-right processing in favor of non-lenear batch processing algorithms.

6. REFERENCES

[1] P. Brown, S. D. Pietra, and R. Mercer, "Word sense disambiguation using statistical methods," *Proceedings of the 29th Annual Meeting of the Association of Computational Linguistics*, pp. 264–304, 1991.

[2] W. Gale, K. Church, and D. Yarowsky, "A method for disambiguating word senses in a large corpu," in *Computers and the Humanities 26:415-439*, 1992.

[3] T. Pedersen, R. Bruce, and J. Wiebe, "Sequential model selection for word sense disambiguation," in *Proceedings of the 5th Conference on Applied Natural Language Processing (ANLP-97)*, April 1997.

[4] D. Yarowsky, "One sense per collocation," *Proceedings of the ARPA Workshop on Human Language Technology*, pp. 266–271, 1993.