

Machine Translation with Cross-Lingual Information Retrieval Based Document Relevance Scores

Michael M. Wasser – mwasser@umd.edu, Advisor – Bonnie Dorr

Abstract

Current methods of statistical machine translation often disambiguate foreign words and phrases incorrectly. Words present with multiple commonly used senses can in some cases lead to translation errors. In this paper we will propose and analyze a potential method of disambiguation by considering the context of the entire document being translated. While decoding, a set of rules learned from each training document is generated. These rules are then used to translate a test document. Simplifying how the decoder functions, rules are selected based off a search algorithm that selects the set of rules that maximizes the log linear combination of a set of feature values. We created an additional feature that would contain a Cross-Lingual Information Retrieval based relevance scores. Using this method, we observed positive gains while translating with a source language of Arabic and a target language of English.

Introduction

Current methods of statistical machine translation often disambiguate foreign words and phrases incorrectly. Words with multiple commonly used senses can in some cases lead to translation errors. Words in a foreign language may not have the same set of senses as its English equivalent. For example, 辦, in Chinese means to manage or to punish. There is no English equivalent word that includes both these senses therefore a translation system must make a choice between the two English words. Even given the context of a sentence, it may be difficult for a human translator to disambiguate a particular word. One of the simplest examples, "bank" in English, can make this point clear. "I went to the bank," could mean going to the side of a river or the building that houses a financial institution. It may take the context of an entire passage for any method, machine or human, to know which sense is being referred to. While there are many techniques to help disambiguate senses such as looking at part of speech and the context near the word or phrase, these errors still occur.

In this paper we will propose and analyze a potential method of disambiguation by considering the context of the entire document being translated. Statistical machine translation involves the use a large set of training documents to create models that predict probable translations of a sentence. We will look at the potential benefit of biasing our models toward more relevant training documents. Intuitively, this method could provide more accurate translations. We hypothesized that documents that are more relevant will be similar in topic and consequently use similar syntactical structures and semantics. These similarities may mean that we can create better models for translating a specific document and ultimately lead to an improved machine translation system.

However, ranking documents based on their similarity to other documents is not a trivial task. For this, we will use a cross-lingua information retrieval (CLIR) based ranking system outlined in (Xu, Weischedel, & Nguyen, 2001). This system uses an Hidden-Markov Model (HMM) to predict how relevant a query is given a particular document is relevant. In this case, the query will be a document to be translated in the foreign language to English and the document that is relevant will be a training document in English.

We will be using a hierarchical machine translation system to be the control of our experiments. This system takes sentence alignments generated from training documents and creates a set of rules that can be used to translate phrases in a foreign language into a target language. When generating possible translations, the rules are selected based off of the log linear combination of a set of feature scores. We will augment the decoder to bias for our CLIR based document relevance scores. This will be done by adding an additional term into the aforementioned log linear combination that can account for a particular rule sources' CLIR score. Each term in this log-linear combination will be referred to as a feature score multiplied by some predetermined weight. Analysis of the experiment was completed using BLEU and TER. We will compare our results to a baseline decoder which does not include our additional feature score. While the results of these changes are specific to the translation system and documents being used for training and testing, it may provide insight into the potential benefit of such a technique with other translations systems.

This section gives an introduction to the problem of disambiguation and the possible solution using document context. The second section will go into background including how the hierarchical machine translation and the CLIR document scoring systems work. The third section will describe the specifics of our changes such as how we will implement and evaluate our system. The last section will present our conclusions and describe possible future work to follow up with what has been presented in this paper.

Background and Related Work

Overview of the Machine Translation System

The translation of a set of documents is carried out by a statistical MT system. This system uses GIZA++ with IBM models 1-4 (Brown, Pietra, Pietra, & Mercer, 1994) and a HMM based word alignment system (Vogel, Ney, & Tillmann, 1996) to build an alignment for parallel data, a hierarchical rule extraction process and decoding is done using a model similar to Chiang's Hiero (Chiang, Lopez, Madhani, Monz, Resnik, & Subotin, 2005).

Alignment

We try to solve for a "hidden" alignment a_1^J between the English source and foreign target sentences e_1^I and f_1^J in the alignment model $P(f_1^J, a_1^J | e_1^I)$. Here the alignment a_1^J describes a mapping from the source word position j to a target position a_1^J . In general, statistical models depend on a set of unknown parameters θ . This set of unknown parameters is trained using a corpus of parallel sentences (f_s, e_s) for all sentences $s = 1, \dots, S$ in the training corpus. θ is then estimated by maximizing the likely hood using the parallel training corpus:

$$\hat{\theta} = \operatorname{argmax}_{\theta} \left\{ \prod_{s=1}^S \left[\sum_a p_{\theta}(f_s, a | e_s) \right] \right\}$$

Where $p_{\theta}(f_s, a | e_s)$ is the probability that f_s maps to e_s using the alignment a with the hidden parameters θ . This can be completed using the expectation maximization algorithm. With this we can select the highest probability alignment using the Viterbi alignment:

$$\hat{a}_1^J = \operatorname{argmax}_{a_1^J} p_{\hat{\theta}}(f_1^J, a_1^J | e_1^J)$$

However, this by itself is not enough as there are limitations to this alignment model. For example, this model does not allow a source word to be aligned with more than one target word. To get around these issues, we execute alignments in the reverse direction and then combine the two alignments using the “refine method” specified in (Och & Ney, 2004).

Rule Extraction

From the resulting alignment data we pull all possible phrase translations (\hat{s}, \hat{f}) such that all words in \hat{s} are aligned to only words in \hat{f} and vice-versa. Each pair is then converted into a CFG rule $X \rightarrow (\hat{s}, \hat{f})$. We also subtract the phrase pairs from the existing rules such that rules of the form $X \rightarrow (\gamma_1 \hat{s} \gamma_2, \alpha_1 \hat{f} \alpha_2)$ are converted to a set of hierarchical rules $X \rightarrow (\gamma_1 X_i \gamma_2, \alpha_1 X_i \alpha_2)$ where i is an index not already in use in the rule. To make the rules more manageable in practice, we limit the number of possible rules created by limiting the size of phrases that can be extracted and subtracted. We also filter the rules to only include ones relevant to the current test set.

Decoding

Our decoder uses a log linear model to score each translation rule using a number of feature scores and weights as follows:

$$\log(w(X \rightarrow (\gamma, \alpha))) = \sum_i \delta_i \varphi_i(X \rightarrow (\gamma, \alpha))$$

Where φ_i are features defined for the rule and δ_i is a weight assigned to the feature. Each feature weight is determined through a method of optimization using an evaluation metric such as BLEU or TER. We used a number of features in our decoder including some specified in (Chiang, Lopez, Madnani, Monz, Resnik, & Subotin, 2005). This is of particular interest to us as we used the rule-feature mechanism to bias the translations with our CLIR scores.

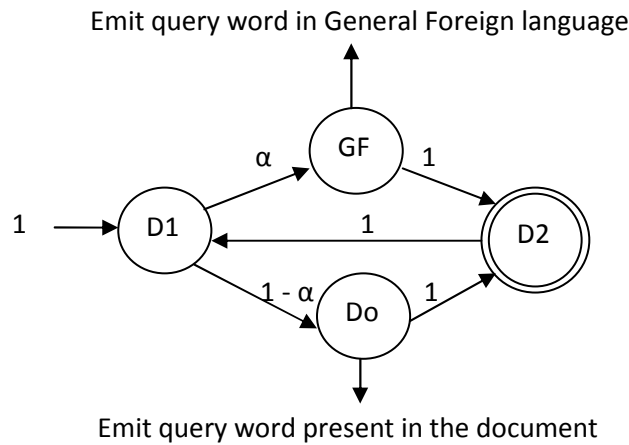
Overview of Cross-Lingua Information Retrieval Ratings

Our system will use probabilistic models for CLIR. CLIR can be used to find documents in a one language given a query in another. Information retrieval itself is defined as follows:

$$P(D|Q) = \frac{P(D) P(Q|D)}{P(Q)}$$

Where Q is a query and D is a document. $P(D)$ is the prior probability of relevance for a document. $P(Q)$ is the probability that the query Q is generated and $P(Q | D)$ is the probability that the query Q was generated given the relevant document. This version of CLIR as specified in (Xu, Weischedel, & Nguyen, 2001) assumes that $P(D)$ is constant, however, this is not the case in practice given documents our training documents are of variable length. We will address this problem in a later section. Also, since Q is a constant the term $P(Q)$ may be ignored as it will not affect the ordering of the document ranks. As such, we used $P(Q|D)$ to estimate a relevance rank of D given Q.

A Hidden Markov Model is used to simulate query generation. In particular a HMM with the following features is used:



The circles represent the possible states, the arrows between states are possible state transitions with the written transition probabilities, arrows pointing to states represent priors, and arrows point away from states represent emissions. Finally the states D1 and D2 represent dummy states while GF represents a word in a general language that is generated in the query and Do represents a word generated in the query that is in the document. α is a constant that is fixed at 0.3 based off of previous experiences. The states are cycled until all query words are generated. GF probability distribution is estimated by the following:

$$P(f|GF) = \frac{freq(f, GF)}{|GF|}$$

Where $freq(f, GF)$ is the frequency of f in the training data and $|GF|$ is the total number of tokens in the foreign language training data. In the state Do the probability distribution is computed as follows:

$$P(f|D) = \frac{freq(f, D)}{|D|}$$

Where $freq(f, D)$ is the frequency of the foreign word f in the document D and $|D|$ is the total number of tokens in the document D. Given these points, $P(Q|D)$ can be computed as follows:

$$P(Q|D) = \prod_{f \text{ in } Q} \left[\alpha P(f|GF) + (1 - \alpha) \sum_{\text{english words } e} P(e|D)P(f|e) \right]$$

Related Work

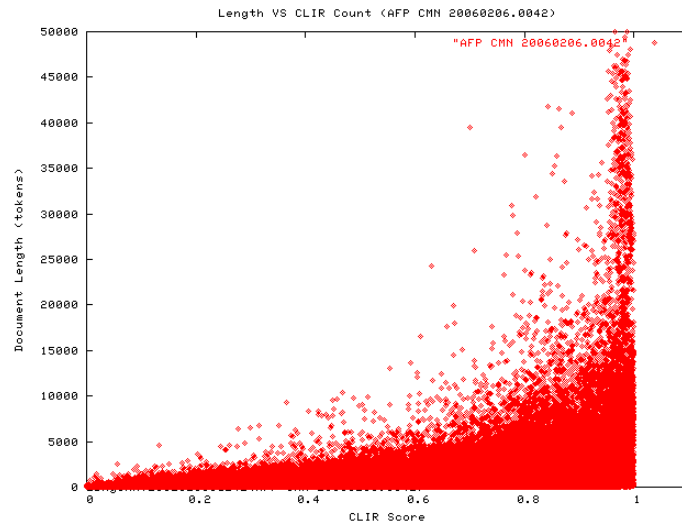
Many previous word sense disambiguation works such as (Lee & Ng, 2002) and (Vickrey, Biewald, Teysier, & Koller, 2005) have seen positive results by analyzing different aspects of the immediately surrounding context. However few have looked at larger contexts for disambiguation. Using features such as surrounding words' part of speech and single word surrounding contexts have seen some gains when compared to techniques that rely on choosing the most probable word sense. We felt we could see more accurately disambiguate words by expanding the context to the entire document using our CLIR biased translation system.

Process and Analysis

Implementing our CLIR weighted translation system involved finding training document CLIR scores, quantifying CLIR scores and formulating an appropriate method to apply the CLIR scores such that they would influence translations. We tested our CLIR augmented translation system using Chinese and Arabic as our foreign language in two different sets of tests. Our target language was English. Predefined lexicons and lists of stopwords, or tokens to ignore, were used where needed. The experiment itself was broken into several manually executed steps. We altered the structure of the training data by moving document boundaries to alter document lengths. We then ran an experiment that would generate and normalize CLIR scores among test documents. Finally we ran an experiment that would translate the test documents and analyze the results taking the CLIR scores into account.

Since each testing document is different, the range of CLIR scores for training documents was different for every testing document. Also, the distribution of scores differed between testing documents. One testing document may have had a set of scores for training documents that ranged between 0 and 350 while another may have ranged between 0 and 500. The higher the score, the more relevant the training document is to the test document. Since the decoder weighs the importance of CLIR scores in the same way regardless of the testing document, a standard way of expressing CLIR scores had to be determined. We choose two methods to do this. In the first method, we set the CLIR training document score to the value of a cumulative distribution function of all the CLIR scores for a specific test document. In the second method, we standard-normalized the CLIR score when compared to all the scores for a specific test document. Both these adjustments gave us a range of CLIR scores that would be the same regardless the test document.

Another problem was related to the assumption made by the CLIR model we used that all documents were of the same document length. The follow graph was used in an analysis confirming that these assumptions could not be ignored given our set of training documents.



Each point represents a specific training document. On the y axis is the document length and on the x axis is the CDF of the CLIR score. This chart shows that documents of longer length always have a higher average CLIR score. We concluded that we needed a way to correct for document length as it seemed to affect our results negatively. Our resolution was to chop all documents longer than a specified length into documents of a specific length. In Chinese we found a document length of 500 tokens and in Arabic a document length of 300 was appropriate given our training data. These numbers were selected by manually inspecting graphs of our training data's document lengths and selecting lengths that we felt would render the largest number of documents of about the specified document length. Another problem was documents that were shorter than this length could not easily be corrected for. In both the case of Chinese and Arabic we had collections with an average document length of less than 40 words. For our experiments we dropped these collections in both our baseline and actual CLIR experiment. Discussion of how these may be included is in the future work section.

While decoding, we needed a way to bias the importance of a training document given its relevance to the test document. We created an additional rule feature score for each rule that would contain the CLIR relevance scores. These scores were set by looking at the CLIR scores of the training documents which generated the given rule. It was common for a rule to have multiple source training documents – this necessitated a method of CLIR score combination. We ran two different sets experiments using the maximum CLIR score of the set of source documents as well as the average. This gave a total of four variants of the CLIR translation experiment from the combination of testing with CDF verses normalized CLIR scores and average verses max CLIR source score combination methods.

The results of our experiments were compared to human translated versions of the test documents. We used BLEU and TER metrics to analyze our results using case insensitive comparisons. Identifying instances of errors caused by incorrect word or phrase disambiguation proved difficult as we did not have the resources or knowledge of Chinese and Arabic to identify errors from incorrect disambiguation. As such, we relied heavily on the aforementioned metrics to gauge performance.

Results and Future Work

Our first set of experiments looked at the effects of the introduction of CLIR scores on a Chinese translation system. The first round of experiments gave us purely negative results. On closer inspection, our training documents had large variance in their document length. As stated earlier, we had previously assumed that the difference in document length was negligible. Our training document set contained approximately 385,000 documents. This set included the LDC2007E08 collection. This collection was 251,000 documents of average length of 31 words (excluding stopwords). The majority of other collections in the training set had an average length of more than 500 words. As noted earlier, our final experiment chopped up longer documents and excluded collections with short averaged document lengths. With these adaptations we received the following average results in Chinese:

Experiment (Chinese)	TER	BLEU
Baseline	55.01	34.32
Best Variant	54.97	34.17

The best available variant was CDF CLIR scoring with average combination in this case. We went through a similar process with Arabic which received the following average results:

Experiment (Arabic)	TER	BLEU
Baseline	42.01	44.76
Best Variant	42.84	45.54

Note: Instead of listing every variant, we list the best variant as data was lost for other variants due to lost experiment data for non-“best” variants.

The best available variant was standard normalized CLIR scoring with average combination. As can be seen above, the Chinese results were negative (or negligible in the case of TER) while the Arabic score were positive. This led us to several possible hypotheses and a set of topics to explore in the future. While these metrics suggest that we did have an increase in translation accuracy for Arabic, we cannot conclude that this is due to better word and phrase disambiguation as we could not confirm this.

First, we would like to explore why gains were not seen in both Chinese and Arabic. While translating, many collections in the Chinese training data were specified as “lexicon” collections and CLIR scores were not assigned to their documents. However, these same collections could be the source of a rule. In practice we just ignored these sources but the difference cause by this may have been large enough to offset the CLIR score feature weight. Another theory may be that there is something inherently different about Chinese and Arabic document relevance. However, this would be difficult to theorize about without a better understanding of Chinese and Arabic. We found that there are many limitations to analyzing results without the ability to read the test documents in the source language manually. In the future, we could also give a more complete analysis of the different experiment variants. In particular how taking the CDF of a set of CLIR scores compares to taking the standard normalization as well as comparing the maximum verses average CLIR feature score combination methods. We could also attempt to include the documents that were dropped previously due to their short length. One possible method would be to combine short documents until their document length was approximately the specified length.

Another topic of interest could involve looking at how this system could alter translation accuracy when the training lexicons come from a variety of different types of sources. Consider the extreme example; if a lexicon included the works Shakespeare (and parallel translations) and the system was translating plays from the same era, the works of Shakespeare would probably have more in common than current news articles or UN translations. A CLIR aided translation system may naturally weight towards documents of the same type and could improve the translation accuracy as a result.

Works Cited

- Brown, P. F., Pietra, S. D., Pietra, V. J., & Mercer, R. L. (1994). The Mathematics of Statistical Machine Translation: Parameter Estimation. *Computational Linguistics*, 19 (2), 263-311.
- Chiang, D., Lopez, A., Madnani, N., Monz, C., Resnik, P., & Subotin, M. (2005). The hiero machine translation system: Extensions, evaluation, and analysis. *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, 779-786.
- Lee, Y. K., & Ng, H. T. (2002, July). An Empirical Evaluation of Knowledge Sources and Learning Algorithms for Word Sense Disambiguation. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 41-48.
- Och, F. J., & Ney, H. (2004). The alignment template approach to statistical. *Computational Linguistics*, 30 (4), 417-449.
- Vickrey, D., Biewald, L., Teyssier, M., & Koller, D. (2005). Word-Sense Disambiguation for Machine Translation. *Proceedings of HLT/EMNLP*, 387-394.
- Vogel, S., Ney, H., & Tillmann, C. (1996). HMM-based word alignment. *Proceedings of the 16th conference on Computational linguistics*, 836-841.
- Xu, J., Weischedel, R., & Nguyen, C. (2001). Evaluating a Probabilistic Model for Cross-lingual. *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, 105-110.