Implementation of Translation Model for Statistical Myanmar-English Translation

Thet Thet Zin
University of Computer Studies,Yangon
thetthetzin.ucsy@gmail.com

Abstract

Machine Translation is defined as the task of transforming an existing text written in a source language, into an equivalent text in a different language, the target language. The statistical machine translation approach uses two types of information: a language model and a translation model. In this paper, translation model is based on the noisy channel model. We use Bayes rule to reformulate the translation probability for translating a foreign sentence f into English sentence e. we use N-gram language model. In machine translation, reordering is requiring to reorder target phrases since different languages have different word order requirements. In this paper, we present Chunks-based reordering approach to reorder target phrases. Reordering rules are automatically generated by using bilingual corpus. The goal of this paper is to improve the translation performance for a Statistical Machine Translation of Myanmar to English.

1. Introduction

Machine translation (MT) is the task of automatically translating a text from one natural language into another. There exist different approaches to address the problem of machine translation. We will give a rough overview over these different methodologies: In rule-based system, the source language text is analyzed, e.g. using parsers and/or morphological tools, and transformed into intermediary representation. From this representation, the target language text is generated. The rules are written by human experts. As a large number of rules are required to capture the phenomena of natural language, this is a time consuming process. As the set of rules grows over time, it gets more and more complicated to extend it and ensure consistency. In the data-driven approach, bilingual and monolingual corpora are used as main knowledge source. Often, a further distinction is made between the example-based approach, where the basic idea is to do translation by analogy, and the statistical approach. In the statistical approach, MT is treated as a decision problem: given the source language sentence, we have to decide for the target language sentence that is the best translation. Then, Bayes rule and statistical decision theory are used to address this decision problem.

Statistical decision theory is a well understood area which provides a way to combine several knowledge sources into a global decision criterion with the goal of minimizing the number of errors. The model parameters are estimated from training data. As more and more training data becomes available, SMT systems get better and better.

This paper is structured as followings: a description of statistical machine translation is described in section 2. Section 3 explained Ngram based phrase extraction, translation model, Word and phrase penalty model and chunk-based reordering model. Previous works in machine translation for various languages are described in section 4. Overview of Myanmar to English translation system are presented at section 5 and the main propose of our system is presented in section 6. We end with conclusion in section 7.

2. Statistical Machine Translation

In statistical machine translation, we are given a source language sentence $f_1^J = f_1 \ldots f_j \ldots f_J$, which is to be translated into a target language sentence $e_1^I = e_1 \ldots e_i \ldots e_I$. Among all possible target language sentences, we will choose the sentence with the highest probability:

$$E^{\wedge} = argmax_e \{Pr(e_1^I \mid f_1^J)\}$$

$$= argmax_e \{\Pr(e_1^I) \cdot Pr(f_1^J \mid e_1^I)\}$$

This decomposition into two knowledge source is known as the source-channel approach to statistical machine translation. It allows an independent modeling of the target language model $Pr(e_1^I)$ and the translation model $Pr(f_1^J \mid e_1^I)$. The target language model describes the well-formedness of the target

language sentence. The translation model links the source language sentence to the target language sentence. The argmax operation denotes the search problem, i.e.., the generation of the output sentence in the target language.

An alternative to the classical source-channel approach is the direct modeling of the posterior probability $Pr(e_1^I|f_1^J)$ .Using a log-linear model we obtain:

$$Pr(E|F) = \frac{exp\{\sum_{m=1}^{M}\lambda_m h_m(E,F)\}}{\sum_{E'} exp\{\sum_{m=1}^{M}\lambda_m h_m(E',F)\}}$$

The denominator represents a normalization factor that depends only on the source sentence F. Therefore, we can omit it during the search process. As a decision rule, we obtain:

$$E^* = argmax_E\left\{\sum_{m=1}^{M}\lambda_m h_m(E,F)\right\}$$

This approach is a generalization of the source-channel approach. It has the advantage that additional model *h (.)* can be easily integrated into the overall system. In our translation model, we use source-channel approach for translation probability of Myanmar to English. We also use N-gram language model.

3. Phrase-Based Translation

3.1. Motivation

One major disadvantage of single-word based approaches is that contextual information is not taken into account. The lexicon probabilities are based only on single words. For many words, the translation depends heavily on the surrounding words. One way to the context into the translation model is to learn translations for whole phrases instead of single words [4]. A phrase is simply a sequence of words. So, the basic idea of phrase-based translation is to segment the given source sentence into phrases, then translate each phrase and finally compose the target sentence from these phrase translations. In this paper, we use heuristic learning of phrase translations from word-based alignments.

3.2. Ngram-Based Phrase Extraction

We use existing word-based alignment model for translation. This alignment model use IBM model 3, English-to-Myanmar Dictionary and Tree tagger to align the sentences. The basic idea of phrase-based translation is to segment the given source sentence into phrases, then translate each phrase and finally compose the target sentence from these phrase translations. We learn phrase pairs from word alignments generated by this alignment model. We use existing Myanmar segmenter to segment source language (Myanmar) sentence. To extract *K* phrases, we use Ngram and training corpus. The source language sentence သူမသည်စာအုပ်ကောင်းတစ်အုပ်ကိုဝယ်ခဲ့သည်။ is segmented into သူမ_သည်_စာအုပ်_ကောင်း_တစ်အုပ်_ကို_ဝယ်_ခဲ့_သည်_။ To extract phrases, we remove stop words such as (သည်) in the middle of the sentence and then we use Ngram and corpus to determine *K* phrases. We collect all aligned phrase pairs that are consistent with the word alignment. No smoothing is performed. We extract *K* phrases are:

သူမ၊ စာအုပ် ၊ကောင်း၊တစ်အုပ်၊ဝယ်ခဲ့သည်

3.3. Translation model

The phrase-based translation model is the main component of our translation system. The hypotheses are generated by concatenating target language phrases. A phrase is simply a contiguous sequence of words. The pairs of source and corresponding target phrases are extracted from the word-aligned bilingual training corpus. The phrase extraction algorithm is described in Section 3.2. We use relative frequencies to estimate the phrase translation probabilities:

$$Pr(f|e) = \frac{count(\bar{f},\ \bar{e})}{\sum_{\bar{f}} count(\bar{f},\ \bar{e})}$$

The number of co-occurrences of a phrase pair $(\tilde{f},\tilde{e})$ that are consistent with the phrase alignment is denoted as $N(\tilde{f},\tilde{e})$. If one occurrence of a target phrase $\tilde{e}$ has N > 1 possible translations, we use WSD (Word Sense Disambiguation) model to handle ambiguous words. The relative frequency estimates typically overestimate the probabilities of rare events as most of the longer phrases occur only once in the training corpus. To overcome this problem, we use word and phrase penalty model to smooth the phrase translation probabilities.

3.4. Word and phrase penalty model

We use a word penalty and a phrase penalty:

$$h_{wp}(e_1^I, s_1^k; f_1^J) = I$$
$$h_{pp}(e_1^I, s_1^k; f_1^J) = k$$

We have segmentation S_1^K of a sentence pair (f_1^J, e_1^I) into K phrase pairs. Each $s_k = (i_k; b_k, j_k)$ is a triple consisting of the last position i_k of the k^{th} target phrase $\tilde{e}_k$ and the start and end positions of the k^{th} source phrase $\tilde{f}_k$ are b_k andj_k, respectively:

$$\tilde{e}_k := e_{i_{k-1}+1} \dots e_{i_k}$$
$$\tilde{f}_k := f_{b_k} \dots f_{j_k}$$

These two models affect the average sentence and phrase lengths. The model scaling factors can be adjusted to prefer longer sentences and longer phrases. The word penalty is simply the target phrase length. If we set a negative scaling factor, longer sentences are more penalized than shorter ones, and the system will favor shorter translations. Alternatively, by using a positive scaling factor, the system will favor longer translations. Similar to the word penalty, the phrase penalty results in a constant cost per produced phrase. The phrase penalty is used to prefer either fewer and thus longer phrases or more and thus shorter phrases. Our system prefers longer sentence and longer phrases. So we use a positive scaling factor for word and phrase penalty model.

3.5. Chunk-level Reordering Model

In machine translation, reordering is one of the major problems, since different languages have different word order requirements. Some approaches have applied at the word-level, such as morphology, POS tags and word classes. They are particularly useful for the language with rich morphology for reducing the data sparseness. Other kinds of syntax reordering methods require parse trees. The parse tree is more powerful to capture the sentence structures. However, it is expensive to create tree structures and building a good quality parser is also a hard task. In our translation model, we use POS tag and chunk as the basic unit for reordering. It is not only because chunks are with more syntax than POS tags, but also they are closer to the definition of a "phrase" in phrase-based SMT and easy to use. Reordering rules are automatically generated by using bilingual corpus [2].

To generate the correct word sequence the translation system needs to have strong restricting evidence of how to rearrange the words, this is the approach taken in grammar-based systems, or it has to have weak evidence in the form of probabilities, and then test all (or at least a large number) of reordering, as is the strategy in typical phrase-based statistical translation systems.

4. Related Work

In this section, previous works in Statistical machine translation on different languages are reviewed. Various researchers have improved the quality of statistical machine translation system by using different methods on different language. (Brown et al. 1990; Brown et al. 1993), which creates probabilistic models for simulating the translation process, in the models using bilingual corpora and then decoding a test sentence by searching. Brown et al. (1993) took the translation process as a noisy-channel model. In terms of modeling, Berger et al. (1996) appended context-based information based on the Maximum Entropy principle to enrich the word-based models. In terms of training, EM algorithm (Dempster et al. 1977) dominated the parameter estimating process by taking word-level alignment of a parallel sentence pair as the latent variable. Phrase-based models emerged (Wang and Waibel 1998; Och et al. 1999; Koehn et al. 2003; Och and Ney 2004). Wang and Waibel (1998) first proposed an alignment model based on phrase structures, which were automatically acquired from parallel corpus. Beam search algorithm was used in (Och et al. 1999), which could make use of pruning strategies for balancing efficiency and accuracy. Och and Ney (2002) first introduced the log-linear model into SMT. Koehn et al. (2003) suggested using features of lexical weighting. In 2003, the famous phrase-based decoder, Pharaoh, was released to be a free SMT toolkit by Philipp Koehn and further updated to Moses (Koehn et al. 2007). These two systems are popularly used as baselines for system comparison in the SMT community. Phrase reordering is modeled in terms of offset positions at the word level (Koehn 2004a; Och and Ney 2004), making little or no direct use of syntactic information. Philipp Koehn, Franz Josef Och, Daniel Marcu used noisy channel based translation model and beam search decoder. They achieved fast decoding, while ensuring high quality. They presented experiential result on many languages (English-German, French-English, Swedish-English, and Chinese-English) [1]. They compared the performance of the three methods for phrase extraction, using the same decoder and the same trigram language model. Learning all phrases consistent with the word alignment (AP) is superior to the joint model. The performance of IBM model-4 word-based translation system is worse than both AP and Joint. Limiting the length to a maximum of only three words per phrase achieves top performance [1]. Richard Zens and Hermann Ney proposed Phrase-based Statistical Machine Translation based on log-linear model with components $h_m(.)$ and scaling factors λ_m. They solve search problem using dynamic programming and beam search with three pruning methods. A comparison with Moses showed that the presented decoder is significantly faster at the same level of translation quality [4].

5. Overview of the Statistical Machine Translation of Myanmar to English

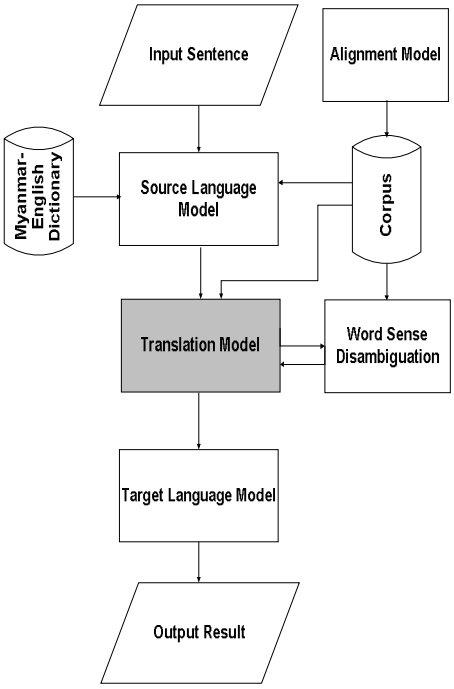


Figure 1.Machine translation system of Myanmar to English

In this system input sentence (source language) is Myanmar language. We need to segment input sentence. So we use existing Myanmar Word Segmenter. Source Language Model based on Ngram. Alignment model use Myanmar-English Dictionary and corpus to generate word alignment. Translation model use segmented sentence, generate phrases by using ngram source language model and corpus, looking up these phrases/words at corpus for relevant target translation phrases/word. One source phrase/word have more than one Target words, we use existing Myanmar Word Sense Disambiguation (WSD) model. Myanmar and English word orders are different. So we use chunk-based reordering method in translation model. Target language model checks sentence patterns and grammar patterns of target language sentence. In this Myanmar to English machine translation system, we focus on Translation model. Translation model are central components of any statistical machine translation system.

6. Proposed Translation Model

In my proposed system, we need segmented words as an input. But there is no boundary to determine different words in Myanmar Language. Thus, we use existing Myanmar Word Segmenter. Then, Preprocessing includes removing stopping words (သည်, မြီး,မျှ) depend on input words. We use source language model (Ngram) and corpus to extract phrases for input words. After extracting phrases for input sentence, translation model that assigns a probability $P\ (f|e)$ to any pair of English and Myanmar strings. Our processing procedure of the system is shown in Figure 2.

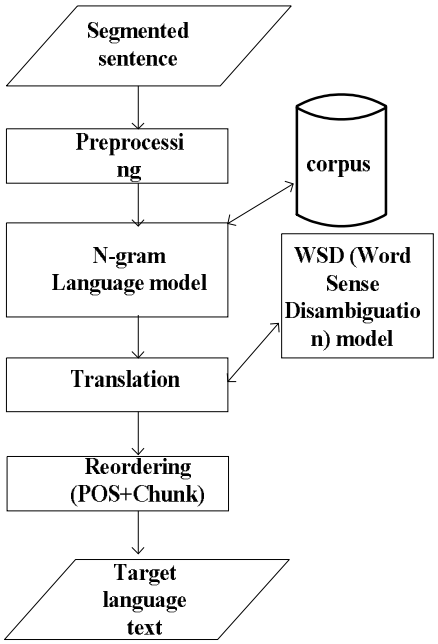


Figure 2.System Architecture of Translation Model

In our system, we estimate the phrase translation probability distribution by relative frequency:

$$Pr(f|e) = \frac{count(\bar{f},\ \bar{e})}{\sum_{\bar{f}} count(\bar{f},\ \bar{e})}$$

No smoothing is performed. If one occurrence of a target phrase $\tilde{e}$ has more than one possible translation, we use WSD (Word Sense Disambiguation) model to handle ambiguous words. We multiply its phrase translation probability with the language model probability for the generated English phrases. We use ngram language model to compute $p(e)$. Myanmar and English languages have different word order. Therefore, we need to reorder the target phrases. In our proposed system, reordering is based on POS tagging and English Chunk rules. These reordering rules are automatically extracted by using corpus and English chunk rules.

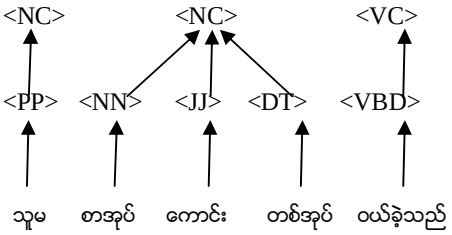
Example: Phrases for this sentence (သူမသည်စာအုပ်ကောင်းတစ်အုပ်ကိုဝယ်ခဲ့သည်။) are

သူမ၊စာအုပ်၊ကောင်း၊တစ်အုပ်၊ဝယ်ခဲ့သည်

and alignment data is

[0]သူမ/[0]she[PP]
[1]စာအုပ်/[2]book[NN]
[2]ကောင်း/[2]good[JJ]
[3]တစ်အုပ်/[3]a[DT]
[4]ဝယ်ခဲ့သည်/[1]bought[VBD]

A First bracket [0] is the index of source word; second bracket [0] is the index of English word. We get tags of source language by using tags of target language. A source sentence is firstly parsed into chunks



These chunks will be reordered by some rules which are automatically extracted by using corpus and reordering rules file. Myanmar chunks <NC><NC><VC> are reorder according to the English chunk rule.

<NC><NC><VC> ⟶ <NC><VC><NC>

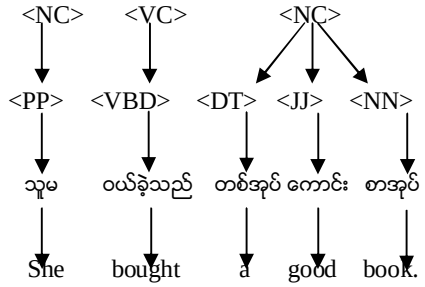
Differences in word order can be local or global. Local reordering is for example the swapping of adjective and noun in language pairs like Myanmar and English. For this global reordering rule, we have local reorder rules for English are

<PP> ⟶ <NC>
<DT><JJ><NN> ⟶ <NC>
<VBD> ⟶ <VC>

By using local reordering rules and global reordering rule, we get two sentences such as

သူမ<NC>ဝယ်ခဲ့သည်<VC> တစ်အုပ် ကောင်း စာအုပ်<NC>
တစ်အုပ်ကောင်း စာအုပ်<NC> ဝယ်ခဲ့သည် <VC>သူမ<NC>

We compute probability of position of <NC> noun chunk by using corpus. Target sentence is



7. Conclusion

In this paper, we have presented a translation model based on n-gram language model, noisy channel model, WSD model and chunk-based reordering method. This system depends on training corpus. If we get larger corpus size, we can get the best translation result. We implement this system to improve translation quality based on statistical approach.

References

[1]Philipp Koehn, Franz Josef Och, Daniel Marcu,"Statistical Phrase-Based Translation", Presentation at DARPA IAO Machine Translation Workshop, July22-23,2002, Santa Monica,CA.

[2]Yuqi Zhang, Richard Zens and Hermann Ney, "Improved Chunk-level Reordering for Statistical Machine Translation".

[3] Xianchao WU,"Statistical Machine Translation Using Large-scale Lexicon and Deep Syntactic Structures", Submitted to the Graduate School of the University of Tokyo on December 15th,2009 in Partial Fulfillment of the Requirements for the Degree of Doctor of Information Science and Technology.

[4]Richard Zens and Hermann Ney, "Improvements in Phrase-Based Statistical Machine Translation".

[5]Daniel Marcu and William Wong," A Phrase-Based, Joint Probability Model for Statistical Machine Translation"

[6]Ulrich Germann, Michael Jahr, Kevin Knight,Daniel Marcu, and Kenji Yamada, "Fast Decoding and Optimal Decoding for Machine Translation "

[7]Michel Galley and Christopher D.Manning, "A Simple and Effective Hierarchical Phrase Reordering Model", Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing, pages 848-856, Honolulu, October 2008.

[8]Yuqi Zhang, Richard Zens and Hermann Ney, " Chunk-Level Reordering of Source Language Sentences with Automatically Learned Rules for Statistical Machine Translation.

[9]Yu Zheng-tao, Deng Bin, Hou Bo, Han Lu and Guo Jian-yi, " Word Sense Disambiguation Based on Bayes Model and Information Gain", International Journal of Advanced Science and Technology vol.3, February, 2009.

[10]George Foster, Roland Kuhn and Howard Johnson, " Phrasetable Smoothing for Statistical Machine Translation".