A Phrase-Based Approach Based on Morphological Information for Japanese-Uighur Statistical Machine Translation System

Maimitili Nimaiti[†] and Izumi Yamamoto[††],

Nagoya Institute of Technology

Summary
The statistical translation approach of Japanese Uighur language in machine translation system is a blank. This paper analyses the approach of statistical machine translation system in Uighur language, discusses how to establishing of dictionary and parallel corpus and phrase based statistical machine translation system based on linguistic rules for Uighur language, and it presents the method of statistical machine translation system based on morphological information of Uighur, the rule base and the dictionary.

Key words:
Statistical machine translation (SMT), Uighur lexicon

1. Introduction

Along with the broad methods of Japanese rule based translation in the field of machine translation, the Uighur Japanese machine translation were established and improved by recent researchers (Muhtar) (Maimitili). But the statistical machine translation in the Uighur language is a blank to be fulfilled. A lot of errors arising in suffix and verb translation in rule based methods and translation rules and translation dictionary need to be increased manually work. Therefore, it is a task that must be undertaken.

Significant progress has been made by various research groups towards the goal of getting reliable statistical translation results. Researchers focus their efforts on enhancing the way different tasks of MT are performed. Some researchers focus on innovating better models for word based, phrases based, and syntax based Statistical Machine Translation (SMT). Other researchers consider the development of better decoding algorithms.

There are very few researches on machine translation in Uighur. They are implemented machine translation system systems between English to Uighur [1], Chinese to Uighur and Japanese to Uighur [2][3][4][5]. They uses the rule based approach, whose all knowledge from linguists is externalized as a set of inference rules [6][7]. In these work, a translation system is implemented that works on word by word translation. And case suffixes are considered only for both languages. Actually Both Japanese and Uighur include lots of suffixes. The harmonization about Uighur language is not explained clearly. These approaches have related several drawbacks to time consumption and rule conflict.

In this work we present a Japanese-Uighur statistical machine translation methods by incorporating morphological information to enhance the translation model by better utilizing the source languages. In contrast to the usual word-based and phrase-based approaches that concentrates morpheme and dictionary features on target languages to improve translation models.

2. Related Works

Our initial experiments with statistical machine translation [9] into Uighur showed that when Japanese – Uighur parallel data were aligned at the word level, a Uighur word typically have to align with a complete phrase on the Japanese side, And that sometimes theses phrases on the Japanese side could be discontinuous, and suggested that exploiting sub-lexical structure would be a fruitful avenue to pursue.

The default standard model that for phrase-based statistical machine translation [8] is only conditioned on movement distance and nothing else. However, some phrases are reordered more frequently than others. Hence, we want to consider a lexicalized reordering model that conditions reordering on the actual phrases. One concern, of course, is the problem of sparse data. A particular phrase pair may occur only a few times in the training data, making it hard to estimate reliable probability distributions from these statistics.

Supposing we want to translate a source language sentence S_1^N =S1…SN into a target language sentenceE_1^M=E1…EM, we can follow a noisy-channel approach regarding the translation process as a channel, which distorts the target sentence and outputs the source sentence defining SMT as the optimization problem expressed by M

$$\hat{e} = \text{argmax Pr } (E_1^M / S_1^N)$$

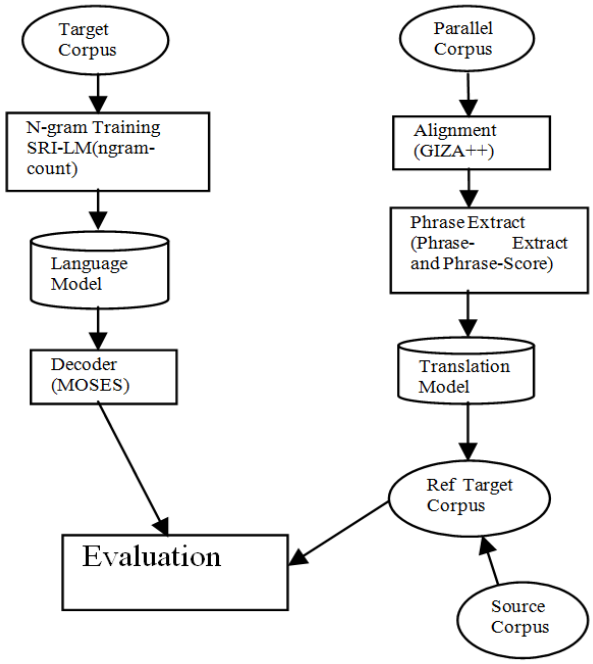
Typically, Bayes rule is applied, obtaining the following expression

$$\hat{e} = \text{argmax Pr } (E_1^M) \text{ Pr } (S_1^N / E_1^M)$$

Manuscript received January 5, 2014
Manuscript revised January 20, 2014

This way, translating S_1^N becomes the problem of detecting which E_1^M among all possible target sentences scores best given the product of two models: Pr(E_1^M) forms the target language model (The Pr(E_1^M) is typically the standard n-gram language model), and Pr (S_1^N/E_1^M)) forms the most important are the phrase-based translation model.

Figure 1 shows that The phrase-based model captures the basic idea of phrase-based translation to segment source sentence into phrases, then translate each phrase and finally compose the target sentence from phrase translations.



Figur1. Phrase based SMT approach

The standard implementation of a decoder is essentially a beam search algorithm. The current state of the art decoder is the factored decoder implemented in the Moses toolkit. As name suggests, this decoder is capable of considering multiple information sources (called factors) in implementing the argmax search (searches for the best according to a linear combination of models). We can get the language model from a monolingual corpus (in the target language) and use it to check how fluent the target language is.

The translation model is obtained by using an aligned bilingual corpus and used to check how the output (in the target language) matches the input (in the source language). We start from a sentence-aligned parallel training corpus and generate word alignments with the GIZA++ toolkit based on IBM Model 1-5 and hidden Markov model.

The phrase translation table is learnt from parallel corpus. It is word-aligned bi-directionally and using various heuristics phrase correspondence is obtained. From the phrase pairs, the phrase translation probability is calculated by relative frequency.

3. Proposed Approach

3.1 Phrase based SMT based on morphological information

Our parallel data consists mainly from translation of story book and Japanese text book. Figure 2 shows our proposed approach. We process these as follows:

i) : We segment the words in our Uighur corpus into lexical morphemes whereby differences in the surface representations of morphemes due to word-internal phenomena are abstracted out to improve statistics during alignment. Note that as with many similar languages, the segmentation of a surface word is generally ambiguous, we first generate a representation using morphological analyzer (CHASEN) that contains both the lexical segments and the morphological features encoded for all possible segmentations and interpretations of the word and perform morphological disambiguation using morphological features. Once the contextually salient morphological interpretation is selected, we discard the features leaving behind the lexical morphemes making up a word.

ii) : We tag the Japanese side using Tree Tagger [10], which provides a lemma and a part-of-speech for each word. We then remove any tags which do not imply an explicit morpheme or an exceptional form. So for instance, if the word book gets tagged as +NN, we keep book in the text, but remove +NN. For books tagged as +NNS or booking tagged as +VVG, we keep book and +NNS, and book and +VVG. A word like went is replaced by go +VVD.

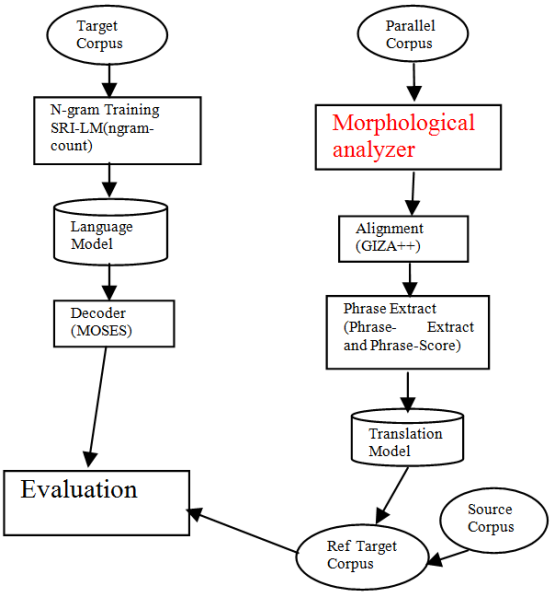
iii) : From these morphologically segmented corpora, we also extract for each sentence, the sequence of roots for open class content words (nouns, adjectives, adverbs, and verbs). For Uighur, this corresponds to removing all morphemes and any roots for closed classes. For Japanese, this corresponds to removing all words tagged as closed class words along with the tags such as +VVG above that signal a morpheme on an open class content word. We use this to augment the training corpus and bias content word alignments, with the hope that such roots may get a better chance to align without any additional “noise” from morphemes and other function words.

Table 1 presents various statistical information about this parallel corpus. One can note that Uighur has many more distinct word forms (about twice as many as Japanese), but

has much less number of distinct content words than Japanese .For language models in decoding and n-best list rescoring, we use, in addition to the training data, a monolingual Uighur text of about 1582 sentences (in a segmented and disambiguated form).

Uighur		
	Sentences	Words
Train	1582	13272
Test	100	1721
Japanese		
	Sentences	Morphems
Train	1582	23097
Test	100	2817

Table1. Statistics on Uighur and Japanese Training and Test Data



Figur2. Proposed Approach

3.1 Lexical Processing

Uigur belongs to Altaic language branch of the Turkic language, Uigur is make up of 32 alphabetic. They are spelling characters as figure 3.

Uyghur Arabic Yeziqi (**UAY**): is primarily written in a script based on the Arabic abjad which start from right site.

Uyghur Latin Yeziqi (**ULY**): is an auxiliary alphabet for the Uyghur language based on the Latin alphabet which starts from left.

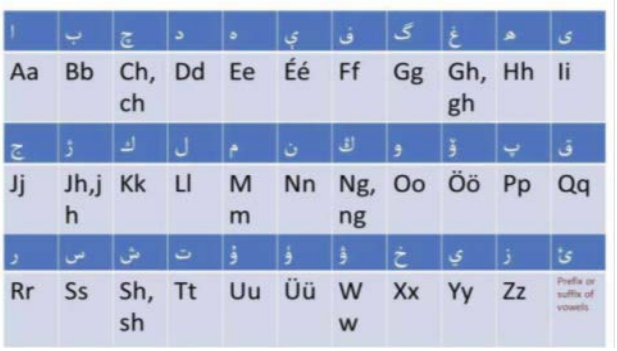


Figure 3. UAY and ULY

4. Experiments and Evaluation

4.1 Tools Employed

We employed the phrase-based statistical machine translation framework [11], and use the Moses toolkit [12], and the SRILM language modeling toolkit [13], and Japanese morphological analyzer[14] using a single reference translation.

We conducted two experiments and performed four sets of experiments employing different morphological representations on the Uighur sentences and adjusting the Japanese representation accordingly wherever needed. Also we applied three evaluation system (BLEU individual score, BLEU cumulative score, NIST individual score) to compare between two approaches.

4.2 Data Preparation and Sentence Translation

As in table 2, total 180sentence were selected in this experiment. Include 80 be verb sentence, 50 do verb sentence and 50 complex sentence (some of similar sentence as a to be or to do part of simple sentence included) were written in ULA.

	System with morphological information	System without morphological information
To be(80)	**39**	**23**
To do(50)	**19**	**11**
Complex(50)	**11**	**6**

Table2. Statistics of correct translated sentences

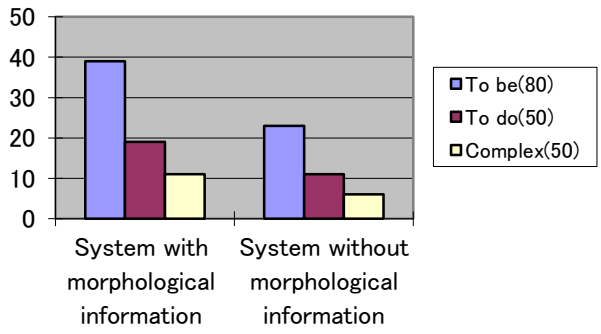
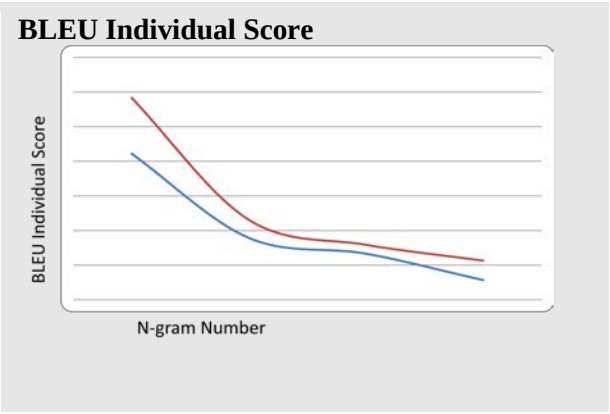


Figure 4. Statistics of correct translated sentences

4.3 Experiment Result

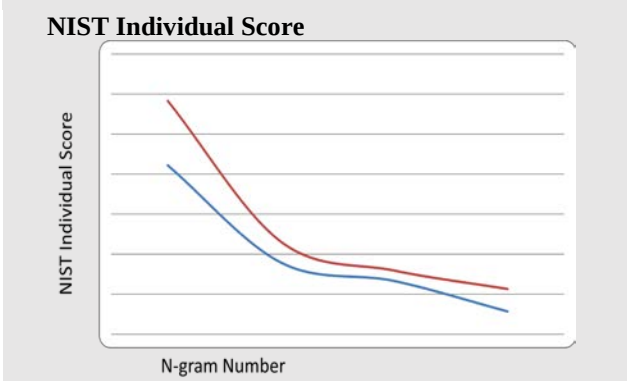
As Figure 4 shows that when we applied morphological information to SMT approach, there are few improves in the translation of to be sentence. For to be sentence, there are some improvement after using morphological information which have better translation of modified verb. However, translation of complex sentences has disconnection between translation and original sentences. There are still having rooms for improvement.

Figure5, Figure6 and Figure 7 shows that various methods of analysis and transformation can be used before obtaining the final result. Along with these statistical approaches may be augmented generating hybrid systems. The methods which are chosen and the emphasis depends largely on the design of the system,



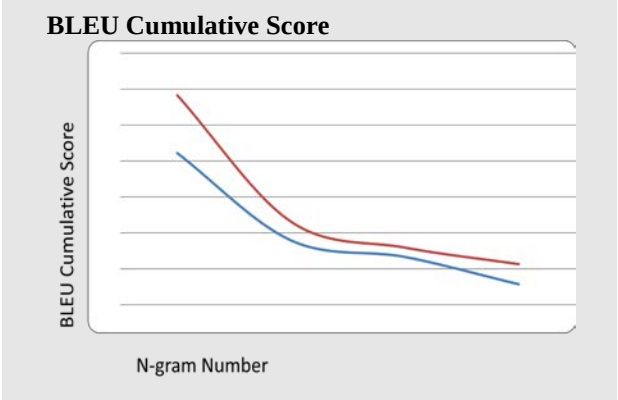
N-gram	1	2	3	4
Without M	0.422	0.179	0.133	0.057
With M	0.583	0.231	0.159	0.113

Figure 5.BLEU Individual Score



N-gram	1	2	3	4
Without M	2.491	0.161	0.031	0
With M	2.921	0.268	0.047	0.011

Figure 6. NIST Individual Score



N-gram	1	2	3	4
Without M	0.422	0.273	0.155	0.132
With M	0.583	0.289	0.178	0.151

Figure 7. BLEU Cumulative Score

4.4 Error Analysis

Phrase-based Systems based on morphological information for Machine Translation model has a visible future for low source minority languages. But some lexical translation doesn't help to the original meaning.

If you would like to itemize some parts of your manuscript, please make use of the specified style “itemize” from the drop-down menu of style categories.

In the case that you would like to paragraph your manuscript, please make use of the specified style “paragraph” from the d

5. Conclusion

In this paper , mainly discussed about statistical based machine translation system and connectivity between morphological information and dictionary. Standard SMT system does well translation beyond tremendous parallel corpora and still can`t solve the disconnection problem between translation and source sentence. We implement a system which includes a learning capacity for low source languages to produce better translation without huge parallel corpus. There is still having more manual step in translation system. In the future work we will do more experiment to add more learning approach on the system.

Acknowledgments

This work was supported by the Hori Sciences & Arts Foundation in Japan.

References

[1] Polat Kadir, Koichi Yamada and Hiroshi Kinukawa. 2004. An English-Uyghur Machine Translation System. In ”Proceedings of The 66th National Convention of IPSJ”, pages 51-52, Information Processing Society of Japan, Tokyo, Japan
[2] Yasuhiro Ogawa, Muhtar Mahsut, Katsuhiko Toyama and Yasuyoshi Inagaki. 1997. Japanese- Uighur Machine Translation based on Derivational Grammar: A Translation of Verbal Suffixes, IPSJ SIG-Notes, NL-120-1
[3] Yasuhiro Ogawa, Muhtar Mahsut, Kazue Sugino, Katsuhiko Toyama and Yasuyoshi Inagaki. 2000. Verbal Phrase Generation based on Derivational Grammar in Japanese-Uighur Machine Translation, Journal of Natural Language Processing, 7(3): 57-77
[4] Muhtar Mahsut, Yasuhiro Ogawa and Yasuyoshi Inagaki. 2001. Translation of Case Suffixes on Japanese-Uighur Machine Translation, Journal of Natural Language Processing, 8(3):123-142
[5] Hamit Tomur and Anne Lee. 2003. Modern Uyghur Grammar. Yildiz, Istanbul, Turkiye
[6] Yoshiyuki Watanabe, Shigeki Matsubara, Katsuhiko Toyama, Yasuyoshi Inagaki. 2000.Einichi Douji Tsuuyaku-no Tame-no Zenshinteki Nihongo Seisei, Proceedings of The Sixth Annual Meeting of The Association for Natural Language Processing , pages 272-275
[7] Kiyotaka Uchimoto, Satoshi Sekine, Hitoshi Isahara. 1999. Japanese Dependency Structure Analysis Based on Maximum Entropy Models. In ”Proceedings of the 9th conference on European chapter of the Association for Computational Linguistics”, Bergen, Norway, pages 196 . 203
[8] The Phrase-based translation system is based on the software released at the 2009 NAACL Workshop on Statistical Machine Translation (http://www.statmt.org/wmt09/)
[9] Maimitili Nimaiti, Paerhati Abudukadeer, Izumi Yamamoto.2011. An Experiment on Japanese-Uighur Machine Translation with MOSES. In” IPSJ SIG 203 Technical Report”, Tokushima, Japan,
[10] Schmid, H.: Probabilistic part-of-speech tagging using decision trees. In: Proceedings of International Conference on New Methods in Language Processing (1994)
[11] Koehn, P., et al.: Moses: Open source toolkit for statistical machine translation. In: Proceedings of the 45th Annual Meeting of the Association for Computational
[12] Stolcke, A.: Srilm – an extensible language modeling toolkit. In: Proceedings of the Intl. Conf. on Spoken Language Processing (2002)
[13] Papineni, K., et al.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, University of Pennsylvania, pp. 311–318 (2002)
[14] http://chasen-legacy.sourceforge.jp/

Maimitili Nimaiti is currently a PhD student at Nagoya Institute of Technology, Japan. He received his B.S. degrees in computer science from Xinjiang University in 2007 and his MS degree from Nagoya Institute of Tecnology in 2012. M.S. His focus is on Machine Lwarning and Natural Language Processing.

Izumi Yamamoto received her MS and PhD in Japanese linguistics from Nagoya University in 1992 and 1996, respectively. She is currently (2014) a Professor of Computer Science at Nagoya Institute of Technology of Nagoya City, Japan. Her current research interests include Japanese literature, intelligent informatics Japanese language education