Collecting SMT Language Model Training Data for Low Source Language

Mamtily Nighmat[†] and Izumi Yamamoto[††]

[†,††]*Department of Computer Science, Nagoya Institute of Technology, Nagoya, 466-8555 Japan*

Summary
Statistical machine translation (SMT) system basically relies on parallel corpus [1]. Different than Rule based Machine translation (RBMT) approach, capability of SMT system almost depends under the size of corpus. The quality of corpus became a key to build better SMT translation system. In this work, parallel corpus [2] in three languages translated to Uyghur language one by one manually evaluated and applied as train data for Uyghur language model. As a conclusion, comparison between parallel corpus in different grammatical structure language and similar structure language has been discussed.

Key words:
Machine Translation, SMT, Parallel Corpus.

1. Introduction

Building a reliable corpus and language model for SMT system has many difficulty. The time and human cost are always being problems for collecting practical parallel data. Internet provide us with huge amount of random language data. Collected language data mostly include non-parallel translation. SMT system trained by non-parallel corpus doesn't provide decent translation result. To reduce problem above, machine learning technic became popular. The process includes collecting the resource of parallel translated language sentences and correcting them grammatically by applying machine learning. Although, applying machine learning technic reduced large amount of time and human cost, many logical and grammatical mistakes occurs in regrouped or filtered parallel corpus. Retranslation and grammatical correction by human translator always a necessary step in random resources [3, 4, 5, 6].

An open source machine translation project with small amount of resources always restrict by quality of parallel corpus. Voluntarily translated parallel corpus always have qualification problem. Differences of understanding from translators also cause problem about different translation for exact same source sentences. Prepare a basic practical multilingual parallel corpus with better quality control and translation rule helps project with healthy start.

In the research world of Natural Language processing, Uyghur Language is counted as Low source Language. Uyghur language belong to Altaic language, which spoken by 8.5 million people in China. Increase of language resources in Internet, Translation capability from other languages becoming a request from many Uyghur language users. A machine translation program became an interest of Uyghur language researchers. Development of Natural Language processing in Uyghur language had been started by a few researchers by applied similar process in Japanese language. Similarity in word order gained few advantages for Uyghur language. The research about Machine Translation in Uyghur language mainly developed in Rule base machine translation and Statistical machine translation.

Rule based approach has been using resent Uyghur machine translation development. Translation between Japanese, Chinese, English and Turkish language was popular topic for researchers. Likely, Japanese language researcher has solved morphological recognition problem early by building accurate morphological analysis system during building Japanese English translation system as Fig1 (such as CHASEN, Mecab etc).

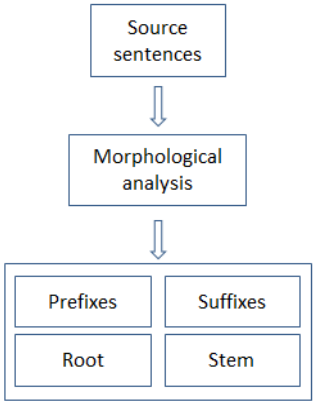


Fig. 1 Morphological Analysis System.

Requirement of morphological analysis system for Uyghur language became a harder task for a few number of researchers. Muhtar, Ogawa, Toyama start with early Japanese Uyghur rule base machine translation system and applied similarity in two languages such as word structure to produce word to word translation system and provide main suffix conversion table for Uyghur language. During

Manuscript received November 5, 2017
Manuscript revised November 20, 2017

the translation from Japanese sentences to Uyghur language, Japanese sentences divided to each words and apply Japanese Uyghur dictionary translate Japanese words one by one to Uyghur words and morphemes. Modification of Uyghur language suffix will combine stem word and suffix to produce modified new word replace stem word, space and suffix. In the final step, the system produce the translation result (Fig 2) [7, 8, 9, 10].

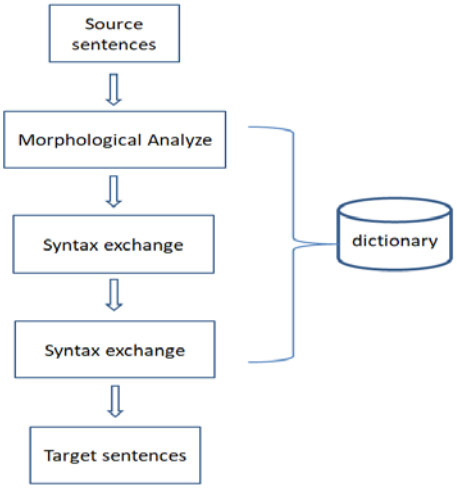


Fig. 2 Rule based Machine Translation System.

Japanese-Uyghur rule based machine translation system mostly developed under one way translation which is from Japanese to Uyghur. Absence of practical Uyghur morphological analysis system causes translation system incapable to provide accurate word and morpheme for translation method. To produce functional morphological analysis system, fully grammatical comparative study and practical solution is necessary.

Even though two languages have similarity in word structure, Japanese sentences structure doesn't provide space between words and Word. There is also no suffix in Japanese language. Abdurihim provide 70,000 morphemes (include word and suffix) and corporate author of Mecab produced an experimental prototype of morphological analysis system.

Statistical Machine Translation research about Uyghur Language is mainly developing in Chinese, English and Japanese language. Although, number of translated documents from government can be assumed as high quality parallel corpus for Chinese-Uyghur SMT system, those unpublished document still is not useful for research propose. A number of Uyghur language users self-constructed a few number of Chinese Uyghur parallel corpus by collecting bilingual book and translating daily used fundamental Chinese sentences from class book and newspaper [11].

The Grammar similarity between English-Chinese also can be apply to English sentences which first translated to Chinese language to build an another reliable English-Uyghur parallel corpus[12].

Although, Japanese language has similar word structure with Uyghur language, a number of qualified basic parallel corpus still need to be increased for in depended research(Fig 3).

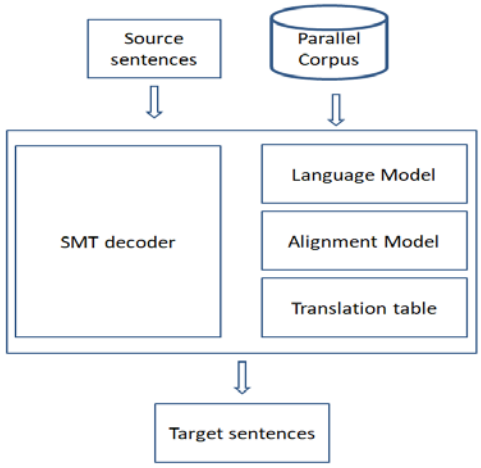


Fig. 3 Statistical based Machine Translation System.

2. Concept

2.1 Building Parallel Corpus

Similar word structure of Japanese and Uyghur language brought much similarity for two languages. However, there are differences too. In Uyghur language, verb always depend on subject that includes pronoun, number, person and tense of object. Japanese language has no conjunction for verb to depend on subject in word form.

Although Uyghur language has officially two alphabet, we applied the Latin based alphabet for this work.

ا	ب	چ	د	ه	ي	ف	گ	غ	ھ	ى
Aa	Bb	Ch, ch	Dd	Ee	Éé	Ff	Gg	Gh, gh	Hh	Ii
ج	ژ	ك	ل	م	ن	ك	و	ۆ	پ	ق
Jj	Jh,j h	Kk	Ll	M m	Nn	Ng, ng	Oo	Öö	Pp	Qq
ر	س	ش	ت	ۇ	ۋ	ۇ	خ	ي	ز	ئ
Rr	Ss	Sh, sh	Tt	Uu	Üü	W w	Xx	Yy	Zz	Prefix or suffix of vowels

Fig. 4 Uyghur Alphabet

Word alignment plays a very important role in SMT system. The idea of this work is to introduce a method to improve Japanese to Uyghur word alignment by increasing language model training data. Most of the translation result failed due to the resources problems. How to overcome the low sources problem by effectively applying existing resources is to increase number of language model. For achieving this porpuse, we have applied 5304 translated parallel corpus used for base line translation model (from Kurohashi & Kawahara lab) which included English, Chinese and Japanese three languages. Each language translated in Uyghur Language independently and result compared in logically and grammatically.

More importantly, the purpose for this test is to build language model and translation model using similar grammatical structured languages to translate different word structured language. There is another result is expected to build more practical and efficient parallel corpus and language model for Japanese-Uyghur SMT system. The result of machine translation result has been evaluated by BLEU and NIST.

We applied phrase based statistical machine translation approach without Morphological analyses. Although the default phrase based statistical translation model will conditioned on movement distance only, few phrases are reordered frequently. Spare data also concern as one of the problem in this test. Some few phrases only occurred in particular time to make statistics less reliable in estimating probability.

Since the 5000 pair sentences translated by voluntarily and correctly, the grammatical difference is still appeared in most of the sentences. Although there is no difficulties for human translators for understanding those different translation in word order, but untrained language model will decrease the accuracy of translation model in Statistical machine translation system.

Example in Tables in chapter 2 shows that grammatical differences when English, Chinese and Japanese language each translated to the Uyghur language.

2.2 Example of Translation Concept

Comparison of translation result in three languages by following steps

I. English to Uyghur (Example in table 1)

5304 English sentences in the corpus will be translated to Uyghur sentences and marked as Uyghur1

II. Chinese to Uyghur (Example in table 2)

5304 Chinese sentences in the corpus will be translated to Uyghur sentences and marked as Uyghur2

III. Japanese to Uyghur (Example in table 3)

5304 Japanese sentences in the corpus will be translated to Uyghur sentences and marked as Uyghur3

IV. Compare each translation of Uyghur sentences (Example in table 4)

Compare three translated Uyghur sentences

Table 1: Example of English to Uyghur

Enlish	Uyghur
The cat took a nap	mushuk chushluk uhlidi
Words can't express my gratitude	rehmitimni ipadilehske sozlirim kemchillik qilidu.

Table 2: Example of Chinese to Uyghur

Enlish	Uyghur
猫睡了午觉。	mushuk chushluk uhlidi.
不太容易说出感谢的话。	Asanliqche rehmet sozi diyelmeslik.

Table 3: Example of Japanese to Uyghur

Enlish	Uyghur
猫が昼寝をした.	mushuk chushluk uhlidi.
なかなか感謝の言葉が出ない	rehmitimni ipadilehske sozlirim kemchillik qilidu.

Table 4: Comparison of each translated Uyghur sentences

Uyghur 1	**Uyghur 2**	***Uyghur 3***
Mushuk chushluk	Mushuk chushluk uhlidi.	mushuk chushluk uhlidi.
Asanliqche rehmet sozi diyelmeslik.	rehmitimni ipadilehske sozlirim kemchillik qilidu.	Asanliqche rehmet sozi diyelmeslik.

From the Example of concept, three languages (English, Chinese and Japanese) each translated to Uyghur language to compare target sentence (Uyghur) word format. We assume the translation from the English and Chinese language has a similar word format, Because of similar grammatical structure between English and Chinese Sentence.

Since Japanese Sentences has very similar word format with Uyghur language, the translation process almost like word to word translation. Although there is verb conjugation differences between two languages.

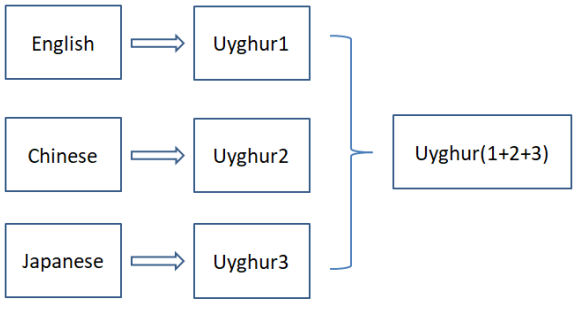


Fig. 5 Individual translation from each language.

2.3 Idea for Training Language Model

In this work, we applied the basic statistical translation model. For language modeling, we used word based ngram model (1~4) since the translation of Uyghur sentences have many misspelling vocabulary cause morphological analysis incorrectly.

Translation result has different word order. We applied all translated Uyghur sentences as language Model training data.

3. Evaluation

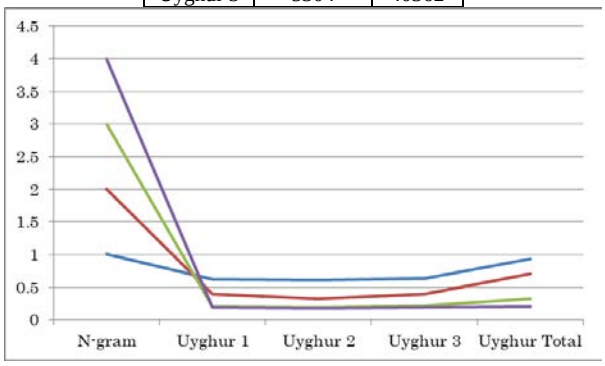
MOSES machine translation toolkit for this test is running under Ubuntu 14.04. Language modeling toolkit SRILM is included in MOSES.

We also applied BLEU Individual, BLEU Cumulative and NIST Individual Scoring system to evaluate the test result.

For evaluating translation score, we apply Japanese sentence. The reason for no testing other two languages is under the low source circumstance, similar word structure languages are easy to perform a better translation result.

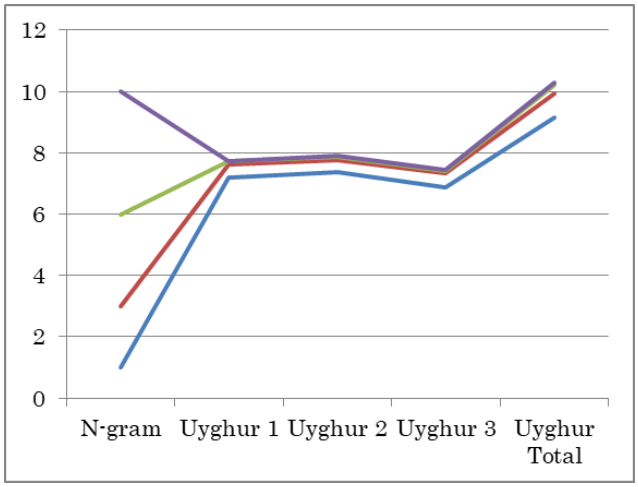
Table 4: Statistics on Translated Uyghur Sentences Training Data

	Sentences	Words
Uyghur 1	5304.	39677
Uyghur 2	5304	40591
Uyghur 3	5304	40362



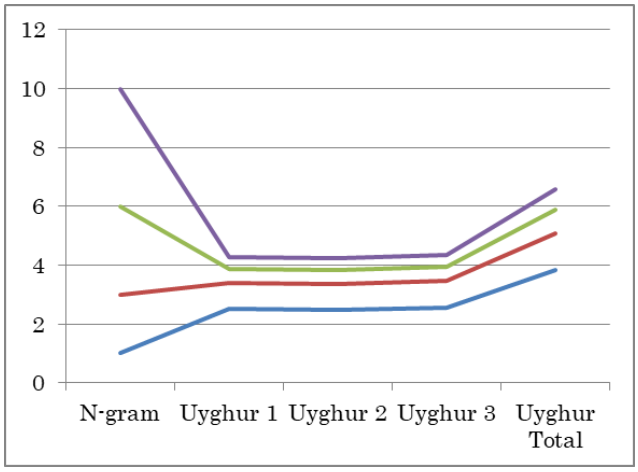
N-gram	1	2	3	4
Uyghur 1	0.617	0.391	0.201	0.192
Uyghur 2	0.615	0.325	0.189	0.179
Uyghur 3	0.632	0.389	0.208	0.183
Uyghur total	0.932	0.711	0.319	0.203

Fig. 6 BLEU Individual Score.



N-gram	1	2	3	4
Uyghur 1	7.19	0.423	0.103	0.016
Uyghur 2	7.37	0.397	0.098	0.028
Uyghur 3	6.89	0.431	0.093	0.031
Uyghur Total	9.14	0.791	0.276	0.089

Fig. 7 NIST Individual Score.



N-gram	1	2	3	4
Uyghur 1	2.51	0.891	0.461	0.408
Uyghur 2	2.48	0.885	0.483	0.398
Uyghur 3	2.56	0.905	0.49	0.391
Uyghur Total	3.83	1.253	0.798	0.707

Fig. 8 BLEU Comulative Score.

As Fig6, Fig7 and Fig8 shows that the SMT train system trained under language model used Uyghur Total model has a few improvement. Although the translation area of language model still limited to the vocabulary of 5304

sentence, the extra language model rate will help the translation to combine better natural sentences.
As we mentioned, the misspelling in translated Uyghur sentences influence part of language modelling incorrectly and evaluation result is limited in word based ngram model. There is still room for update the translation result by applying suffix and stem based language modeling.
This work involves improving the efficiency of the parallel corpus of other cross translated languages, especially for many other low source languages.

4. Conclusion

This paper present an idea of specific way to produce parallel corpus and language model for cross translation SMT system spatially for low source language. Although we present a part of translation result from Japanese to Uyghur for evaluating the increased language model, the translation result is not mainly discussed in this work. We are confident in better language model in targeted area .The very small number of efficient parallel corpus cause Absence of machine evaluation by popular open source Software MOSES. Therefore, manual evaluation covered grading requirement.
In the future work, we plan to insert morphological analyses step before building language model to increase the translation model and translation table probability.

References

[1] Brown, Peter F., John Cocke, (1990) "A statistical approach to machine translation", Computational Linguistics, 16 (2):79–85.
[2] http://nlp.ist.i.kyoto-u.ac.jp
[3] Resnik, Philip, and Noah A. Smith. (2003) "The web as a parallel corpus", Computational Linguistics 29.3 (2003): 349-380.
[4] L. Schwartz, “Monolingual post-editing by a domain expert is highly effective for translation triage,” in Proceedings of the Third Workshop on Post-editing Technology and Practice, 2014, pp. 34–44
[5] N. Aranberri, G. Labaka, “Comparison of post-editing productivity between professional translators and lay users,” in Proceedings of the Third Workshop on Post-editing Technology and Practice, 2014, pp. 20– 33
[6] Taghipour, Kaveh,. (2010) "A discriminative approach to filter out noisy sentence pairs from bilingual corpora", Telecommunications (IST), 5th International Symposium on 2010 : 537-541
[7] Kudo, T. “MeCab: Yet Another Part of Speech and Morphological analyzer”, http://mecab.sourceforge.net/ (Retrieved: 2012.01.25)
[8] 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
[9] 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
[10] 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
[11] Xinghua Dong, Junlin Zhou, Shushen Guo, Turghun Osman, Phrase-based Chinese-Uyghur / Uyghur-Chinese statistical machine translation, Journal of Computer Engineering, vol. 37, no. 9,2011, 16-18
[12] Polat Kadir, Koichi Yamada and Hiroshi Kinukawa. 2004. An English-Uyghur Machine Translation System. In ”Proceedings of The 66th

Mamtily Nighmat 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 Learning 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