Vol.2011-NL-203 No.5
2011/9/16

MOSES を用いた日本語ウイグル語機械翻訳実験

マミテリ・ニマテ†　パルハト・アブドカディル††
山本いずみ†††

日本語・ウイグル語機械翻訳システムの開発にあたり、Moses を用いて翻訳モデルを作成し、実験を行った。結果として、単純な文では正確な翻訳文を得ることができたが、複雑になるにつれ、精度が下がった。その原因は、接尾辞の処理にあると考えられる。日本語もウイグル語も膠着語であり、多彩な語尾変化を有する。正確な翻訳を行うためには、特にウイグル語形態素解析システムの精度を高める必要がある。

An Experiment on Japanese-Uighur Machine Translation with MOSES

Maimitili NIMAITI† and Paerhati ABUDUKADEER††
and Izumi YAMAMOTO†††

This paper presents an experiment of machine translation between Japanese and Uighur, by incorporating morphological information into a statistical machine translation model. The basic idea is the agglutinated suffixes should be treated carefully so as to make correct translation, because they play important roles in the Uighur language. Experimental results showed that morphological decomposition of Uighur source is necessary, specially for smaller-size training corpora.

1. Introduction

Uighur is spoken by 8.5 million (2004) in China, mostly in the far western Xinjiang Autonomous Region. The necessity of machine translation system is growing rapidly due to the increase of government documents and translation request.

One of them is the lack of training data in both automatic speech recognition and machine translation(MT), especially for low-resource languages. 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[1][14] , phrases based[2] , and syntax based[3] Statistical Machine Translation(SMT). Other researchers consider the development of better decoding algorithms[4]. There are very few researches on MT in Uighur. They are implemented between English to Uighur MT system[10] and Japanese to Uighur MT system[9][12][13] uses the rule based approach, whose all knowledge from linguists is externalized as a set of inference rules. 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 approach has several drawbacks related to time consumption and rule conflict. In this work we pay much attention to linguistic information than translating approaches. This is the first Japanese to Uighur machine translation system in SMT, which comes with features such as an inbuilt morpheme or dictionary features. However, there is not a previous work related to Japanese-Uighur machine translation in SMT that we could use as reference for our research. Japanese to Uighur MT is a complex task due to fundamental structural differences of the two languages. Therefore, the development of Japanese to Uighur MT system must started from the scratch. In this work we present a Japanese-Uighur SMT 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

† 情報処理学会
Information Processing Society of Japan
†† 情報処理学会
Information Processing Society of Japan

concentrate morpheme and dictionary features on target languages to improve translation models. The rest of this paper is organized as follows. Section two will discuss some Uighur language features and its similarities/differences with Japanese which are important points in implementing a successful Phrase based Japanese to Uighur translation system. Section three mainly explains statistical phrase based machine translation. Section four illustrates the corpus and shows the experimental results. Finally, Concludes the paper and recommends future works will be discussed.

2. Grammatical Comparison of Japanese and Uighur

Uighur, like all the other Turkic languages, has a word order of subject + object +verb (SOV), and is considered to be an agglutinative language with very productive inflectional and derivational suffixation process in which a sequence of inflectional and derivational morphemes get affixed to a word stem. In Uighur, a verb could have hundreds of word forms by sequentially adding different affixes to the word stem. Japanese, which is also considered to be an agglutinative language, also has the same word order and morphological features as Uighur. Some researches show that this morphological and syntactic closeness is sufficient to obtain a relatively good translation result from Japanese into Uighur on a transfer approach. In the following sections, we will make a comparison between Japanese and Uighur in two different levels: morphology and syntax with a close attention focused on their differences. We could find that in both Japanese and Uighur, word forms are generated by attaching many suffixes denoting case, mood, person, tense, etc. to one word stem as seen in Table 1.Generally, Japanese and Uighur share a significant amount of morphological and syntactic features in common. However, there are also some differences in word formation of nouns, verbs,

Table 1

Kur+al+mi+ghan+liktin("as it was not seen")	
見ら＋れ＋な＋かった＋ので	
Kur/見ら	:stem
Al /れ	:passive voice
Mi/な	:negation
Ghan/かった	:past tense
Liktin/ので	:causal form

3. Phrase-based MT from Japanese to Uighur

Supposing we want to translate a source language sentence $S_1^N = S_1 \ldots S_N$ into a target language sentence $E_1^M = E_1 \ldots E_M$, 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

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

Typically, Bayes rule is applied, obtaining the following expression

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

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.

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.

The standard implementation of a decoder is essentially an beam search algorithm. The current state of the art decoder is the factored decoder implemented in the Moses toolkit [7]. 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 [8][15] 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.

4. Experiment

4.1 Tools and Data preparation

Our experiments are on Uighur to Japanese translation based phrase-based translation system. Our baseline is a phrase-based statistical machine translation system. Uighur to Japanese MT corpus (UJC) consists of 1582 Japanese sentence from the story book and text book translated sentence by sentence into Uighur sentence in Uighur Latin Alphabet(ULA). Basic statistics about the training part of the UJC are given in Table 1.

Table 2 Statistics on Uighur and Japanese Training and Test Data

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

4.2 Morphological Decomposition

In the first step, Japanese corpus was morphological decomposed by Mecab. Mecab is one of the morphological analysis system for Japanese Language. Uighur corpus already has spaces between each word.

4.3 Language Model

In the second step Japanese language model was trained using SRILM (parameter set: ngram-count-order 4 -interpolate -kndiscount), The language model is a statistical n-gram model estimated using modified Kneser-Ney smoothing. Result is showed in Figure-1.

Figure1, Language model

```
食べ物 について 昔 から 言わ れ て いる 常識 も あり ます 。
NULL ({ }) yimekliklerghe ({ 1 }) nispiten ({ 2 }) burundin ({ 3 4 5 6 }) dep ({
 7 }) kilwatqan ({ 8 9 }) kuzqarashlarmu ({ 10 }) bar ({ 11 12 }) . ({ 13 })
# Sentence pair (1570) source length 15 target length 22 alignment score : 3.559
22e-28
医学 的 に 正しい もの は ほとんど ない の です が 、 多く の 人 が まだ 信じ て
 い ます 。
NULL ({ 6 9 }) dohturluq ({ 1 }) ilmi ({ 2 }) buyiche ({ 3 }) toghra ({ 4 }) ner
siler ({ 5 }) asasen ({ 7 }) yoq ({ 8 }) bolsimu ({ 10 11 }) , ({ 12 }) nurghun
({ 13 14 }) kishler ({ 15 16 }) yene ({ }) ishinip ({ 18 19 }) yuruydu ({ 17 20
21 }) . ({ 22 })
# Sentence pair (1571) source length 13 target length 19 alignment score : 8.634
15e-23
生活 が 変わっ て も 、 昔 の 常識 を そのまま 信じ て いる の は 変 です 。
NULL ({ }) turmush ({ 1 }) uzgersimu ({ 2 3 4 5 }) , ({ 6 }) burunqi ({ 7 8 }) k
uzqarashlargha ({ 9 10 17 18 }) shu ({ }) petti ({ 11 }) ishinip ({ 12 13 }) yur
uydighan ({ 14 }) kishiler ({ 15 16 }) hazirmu ({ }) bar ({ }) . ({ 19 })
# Sentence pair (1572) source length 18 target length 27 alignment score : 2.909
22e-38
時々 その 常識 が 正しい か どう か 、 どうして その 常識 が でき た か 、 考え
て み た ほう が いい でしょ う 。
NULL ({ }) bezi ({ 1 }) chaghlarda ({ }) shu ({ 2 11 }) kuzqarash ({ 3 12 }) tog
hrimu ({ 4 5 13 14 }) qandaq ({ 6 7 8 16 }) , ({ 9 }) nimishqa ({ 10 }) shundaq
({ }) kuz ({ 23 }) qarashlar ({ 25 26 }) bolghan ({ 15 }) , ({ 17 }) oylinip ({
18 }) baqqan ({ 20 21 22 }) yahshi ({ 24 }) he ({ 19 }) . ({ 27 })
```

4.4 Translation Model

In the third step GIZA++ is used in two translation directions and the grow-diag-final-and combination method is used to obtain a symmetries alignment matrix. We employ the phrase-based statistical machine translation framework ,

4.5 Phrase Tablel

In this step use the Moses toolkit, and the SRILM language modeling toolkit, Uighur to Japanese MT corpus (UJC) consists of 1582 Japanese sentence from the story book and text book translated sentence by sentence into Uighur sentence in Uighur Latin Alphabet(ULA). Result showed in Figure-2

Figure-3. Phrase Table

```
～ と 聞い た ||| -- dep soridi ||| 1 0.0637613 1 0.00388214 ||| ||| 1 1
～ と 聞い た 。 ||| -- dep soridi . ||| 1 0.0629504 1 0.00377466 ||| ||| 1 1
～ と 褒め られ まし た ||| mundaq dep maxtidi ||| 1 0.0004457 1 0.00640884 |||
||| 1 1
～ と 褒め られ まし た 。 ||| mundaq dep maxtidi . ||| 1 0.000440031 1 0.006231
4 ||| ||| 1 1
～ と 言い たかっ ||| -- digusi bar ||| 1 0.0261438 0.5 0.00117079 ||| ||| 1 2
～ と 言い たかっ ||| -- digusi ||| 1 0.0261438 0.5 0.0511245 ||| ||| 1 2
～ と 言い たかっ た ||| -- digusi bar idi ||| 0.333333 0.0146405 1 6.85819e-05
||| ||| 3 1
～ と 言い たかっ た の ||| -- digusi bar idi ||| 0.333333 0.0015745 1 6.85819e-
05 ||| ||| 3 1
～ と 言い たかっ た の です ||| -- digusi bar idi ||| 0.333333 1.26365e-05 1 6.
85819e-05 ||| ||| 3 1
～ と 言い まし た ||| -- didi ||| 1 0.00112496 0.333333 0.0289073 ||| ||| 2 6
～ と 言い まし た ||| ~ dep jawap berdi ||| 1 0.00125268 0.166667 0.000139507 |
|| ||| 1 6
～ と 言い まし た ||| ~ didi ||| 1 0.00100655 0.333333 0.0289073 ||| ||| 2 6
～ と 言い まし た ||| ~ digen ||| 1 2.29404e-05 0.166667 0.00561567 ||| ||| 1 6
～ と 言い まし た 。 ||| -- didi . ||| 1 0.00111066 0.333333 0.0281069 ||| |||
2 6
～ と 言い まし た 。 ||| ~ dep jawap berdi . ||| 1 0.00123675 0.166667 0.000135
645 ||| ||| 1 6
～ と 言い まし た 。 ||| ~ didi . ||| 1 0.000993746 0.333333 0.0281069 ||| |||
2 6
～ と 言い まし た 。 ||| ~ digen . ||| 1 2.26487e-05 0.166667 0.00546019 ||| ||
| 1 6
～ と 言っ た よ ||| , ~ degenghu ||| 1 0.00492459 0.5 0.00103467 ||| ||| 1 2
～ と 言っ た よ ||| ~ degenghu ||| 1 0.00492459 0.5 0.00492877 ||| ||| 1 2
～ と 言っ た よ 。 」 ||| , ~ degenghu " . ||| 1 0.00296461 0.5 0.00021314 |||
||| 1 2
～ と 言っ た よ 。 」 ||| ~ degenghu " . ||| 1 0.00296461 0.5 0.00101532 ||| ||
| 1 2
～ と 言っ て ||| ~ dep bolghandin ||| 1 0.00698566 1 0.00144923 ||| ||| 1 1
～ と 言っ て から ||| ~ dep bolghandin kiyin ||| 1 0.00182235 1 9.45152e-05 |||
 ||| 1 1
～ と 言っ て から 食べ ます ||| ~ dep bolghandin kiyin yeydu ||| 1 0.000202483
1 3.24989e-06 ||| ||| 1 1
～ など も あり まし た ||| -- qatarliqlarmu bar ||| 1 0.00023395 1 0.00153952 |
|| ||| 1 1
～ など も あり まし た 。 ||| -- qatarliqlarmu bar . ||| 1 0.000230974 1 0.0014
9689 ||| ||| 1 1
```

5. Experiment Result

There was only the employed phrase-based Experiment had been carried out. We evaluate the translation quality with human evaluation. In this experiment 100 sentences(include 30 simple sentence and 35 complex sentences and 35 phrases) have been tested. Result is presented in Table 2

Table2. Evaluation of the Result

	Simple sentences	Complex sentences	phrase
Total	30	35	35
Translated	16	0	22
Untranslated	14	35	13

Because of the grammatical similarity, if there is well dictionary translation, except suffix, those translated vocabulary can be right in the position of each sentence. The experiment result is presented in Table 3

Table 2 Translation Result of Japanese to Uighur

明日 は 富士山 へ 行き ます。	ete bolsa 富士山 barsa barimen .
私 は 歌 が 上手 です 。	men naxshigha usta
私 の 名前 は 田中 です 、あなた は 誰 です か ？	men ismi 田中 iken 、あなた bolsa 誰 iken ?
私 は 日本語 を 話せ ませ ん 。	men yapon tilini 話せ bolmaydu .
明日 家 へ 帰ら なけれ ば なり ませ ん 。	ete oyi barsa 帰ら qilmisa bolmaydu .
彼 は どこ から 来 た 。	u yaqqa yerlerde yaghuz .
私 の 名前 は 山本 です 、 日本人 です 。	men ismi yamamoto iken , yapunluq .
明日 雨 が 降る そう だ 。	ete yamghur kup 降る shundaq .
子供 の 時 から 、 忘れ て は いけ ない 、忘れ て は いけ ない 、と 教え られ 、 忘れ た と 言っ て は 叱ら れ て き た 。	baliliq chaglardin bashlap , untup qalma , untup qalma dap ughutup , untup qaldimdim dep berse bolsa 叱ら bolup keldi .
昔 の 人 は 、 自然 に 従っ た 生活 を し て い た から 、 神 の 与え 給う た 忘却 作用 で ある 睡眠 だけ で 、 充分 、 頭 の 掃除 が でき た 。	burunqi ademler bolsa , tebietke 従っ turmush ni hokum bolghachqa , huda ata qilghen untush hizmitini qildighan oyqu bilenla , asasen mengning taziliqini qilalighan .
北京 を 訪れ 帰郷 し た とき の こと で ある 。	beijinggha qilinghan seperdin qaytip keliwatqanda bolghan ish ibaret .
10 月 から 3月 まで 雨 が 多い です 。	10-aydin 3-ayghiche yamghur kup iken .
広島 で おいしい お 土産 を 買い まし た から 、 来週 持っ て 行き ます 。	hiroshima din setiwalghan sowghatni keler hapte elip barimen .

6. Conclusion

In this paper we presented an phrase-based SMT system for Uighur to Japanese machine translation. The results shows that morphological decomposition, increasing of dictionary size and parallel corpus are necessary for both languages. The results aren`t positive and there are quite some rooms for improvement. Our current work involves improving the quality of our current system as well as expanding this approach to other Turkic languages. This approach is efficient when translating from highly inflected languages like other Turkic languages.

7. Future Work

In future work, we try to improve the dictionary and morphological analysis system for Uighur. since Uighur is very much morphologically rich and agglutinative we try to improve the performance of Uighur tokenize Since there is much deficiency in the parallel corpora we try to get more corpora from different domains in such a way that it will cover all the wordings. As a new member of SMT Machine Translation family. Japanese-Uighur Machine Translation system has been developed by few ways. There are also a few works and resource in the past which can rely on it. In the future. We propose to collect the resource and make evaluations on the near future.

Reference

1) F. J. Och, Statistical Machine Translation:From Single-Word Models to Alignment Templates, 2002
2) P. Koehn., F. J. Och, Daniel Marcu. Statistical Phrase-Based Translation. In Proceedings of the Human Language Technology Conference. 2003,pp. 127-133. Microsoft Office,
3) Kenji Yamada and Kevin Knight. A Decoder for Syntax-based Statistical MT. Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics . 2002, pp. 303-310.
4) P. Koehn. A Beam Search Decoder for Phrase Based Statistical Machine Translation Models. A Technical Manual of the Pharaoh decoder.2003, pp.13-148.
5) Batuer Aisha and Maosong Sun. 2009. A Statistical Method for Uighur Tokenization, in Proceedings of IEEE International Conference on Natural Language Processing and Knowledge Engineering. pp.383-387.
6) 3137-3145 6Adam L. Berger, Stephen A. Della Pietra, and Vincent J. Della Pietra., A Maximum Entropy Approach to Natural Language Processing, Association for Computational Linguistics, 1996, pp. 39-71.
7) P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst. Moses: Open source toolkit for statistical machine translation. in Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL'07) – Companion Volume, June 2007.
8) F. J. Och and H. Ney. A Systematic Comparison of Various Statistical Alignment Models. Computational Linguistics, 2003, 29(1): 19-51
9) Muhtar Mahsut, Yasuhiro Ogawa, Kazue Sugino Yasuyoshi Inagaki. Utilizing Agglutinative Features in Japanese-Uighur Machine Translation. MT SummitVIII: Machine Translation in the Information Age, Proceedings, 2001, pp.217-222.
10) Polat KADIR, Koichi YAMADA, Hiroshi KINUKAWA, Comparative Study on Japanese and Uighur Grammars for An English-Uighur Machine Translation System, MT Summit X. Conference Proceedings: the tenth Machine Translation Summit; 2005, pp.432-437.
11) K. Papineni, S. Roukos, T.Ward, W.J. Zhu:BLEU: a Method for Automatic Evaluation of Machine Translation. In: ACL Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Philadelphia, Pennsyl-vania , 2002, pp. 311-318
12) Muhtar Mahsut, Yasuhiro Ogawa, Kazeu Sugino, Katsuhiko Tuyama and Yasuyoshi Inagaki. An Experiment on Japanese-Uighur Machine Translation and Its Evaluation. AMTA 2004,LNAI 3265, 2004, pp.208-216.
13) Muhtar Mahsut, Fabio CASABLANCA, Katsuhiko TOYAMA and Yasuyoshi INAGAKI. Particle Based Machine Translation for Altaic Languages The Japanese - Uighur Case. In Proceeding of the 3rd Pacific Rim International Conference on Artificial Intelligence, Vol.2. Beijing, 1994, pp.725-731.
14) F. J. Och and H. Ney. Discriminative training and maximum entropy models for statistical machine translation. In Proc. of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), 2002, pp. 295-302
15) F. Och and H. Ney. Improved statistical alignment models, in Proc. ACL, Hong Kong, China, 2000