Journal of Computational Information Systems 6:10 (2010) 3137-3145
Available at http://www.Jofcis.com

# Uyghur-Chinese Statistical Machine Translation by Incorporating Morphological Information

Batuer AISHA[1,2,†], Maosong SUN[3]

[1,3] *Department of Computer Sci. & Technology. State Key Lab on Intelligent Tech. & System. Tsinghua University,*

*Beijing, 100084, China*

[2] *Research in the Key Laboratory of Multilingual Information Technology, School of Information Science and Engineering,*

*Xinjiang University,Urumqi,Xinjiang,830046,China*

**Abstract**

This paper presents a method of machine translation from Uyghur, an agglutinative language with very productive inflectional and derivational morphology, to Chinese, 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 Uyghur language. Experimental results showed that morphological decomposition of Uyghur source is beneficial, specially for smaller-size training corpora. The BLEU score is improved to 25.26 from 13.61 when the input data is tokenized compared to the case without tokenization.

*Keywords:* Uyghur; Chinese; Morphological Information; Machine Translation

## 1. Introduction

Uyghur is spoken by 8.5 million (2004) in China, mostly in the far western Xinjiang Autonomous Region. Uyghur is also spoken by 300,000 in Kazakhstan, and there are Uyghur-speaking communities in Afghanistan, Albania, Australia, Belgium, Canada, Germany, Indonesia, Kyrgyzstan, Mongolia, Pakistan, Saudi Arabia, Sweden, Taiwan, Tajikistan, Turkey, United Kingdom, USA, and Uzbekistan. There are two main languages in Xinjiang Uyghur Autonomous Region: Uyghur and Mandarin Chinese. Mandarin Chinese is not used widely in southern Xinjiang.

About 80 newspapers and magazines are available in Uyghur; 22 TV channels and ten publishers serve as the Uyghur media. The necessity of machine translation system is growing rapidly due to the increase of government documents and translation request.

Although many Uyghur possess skills and sufficient knowledge to work in China or other developing countries, Their lack of knowledge in the Chinese or other languages limit their expression of ideas. Difficulties have often been faced by many at job interviews, appointments of personnel for various

[†] Corresponding author.
*Email addresses*: batur61@163.com (Batuer AISHA)

October, 2010

position and facing examinations in Chinese etc. In addition, Chinese people also face difficulties when trying to understand and communicate in Uyghur when they work or visit Xinjiang.

This work is aimed at addressing these challenges by building Uyghur to Chinese language translator. However, we still face many challenges. 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][17], 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 Uyghur. They are implemented between English to Uyghur MT system[12] and Japanese to Uyghur MT system[11][14][15][16] 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 Uyghur include lots of suffixes. The harmonization about Uyghur 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 Uyghur to Chinese machine translation system, which comes with features such as an inbuilt morpheme or dictionary features. However, there is not a previous work related to Uyghur-Chinese machine translation that we could use as reference for our research. It is a far more difficult task to perform a translation of non-similar languages than similar languages.

Uyghur to Chinese MT is a complex task due to fundamental structural differences of the two languages. Therefore, the development of Uyghur to Chinese MT system must started from the scratch.

In this work we present a Uyghur-Chinese 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 concentrate morpheme and dictionary (61,827 words, Person name( women 4120, men 4093) features on target languages to improve translation models. Lastly, we have to present a scheme to add to the decoder output by correcting words which have out-of-vocabulary(OOV) with respect to the training data and bilingual dictionary feature (solve the data-sparseness problem) and language model to further improve the translations.

The rest of this paper is organized as follows. Section two will discuss some Uyghur language features and its similarities/differences with Chinese which are important points in implementing a successful Phrase based Uyghur to Chinese 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. Uyghur Grammar

The Uyghur language belongs to the Qarluq group of the Turkic language family, which is among the Altaic languages. In grammatical aspects, Uyghur is an agglutinative language with rich and complex morphology. It is possible to produce a very large number of words from the same root with suffixes and the same root word may take many different suffixes in a linear sequence and form different words, increasing the number of out-of-vocabulary words. In Uyghur, a noun or verb could have hundreds of word forms by sequentially adding different suffixes to the word stem.

SMT for Uyghur language gives poor result, if we provide parallel corpus directly, because of the following reasons; First is that Chinese follows SVO(Subject–Verb–Object) word order just like English but Uyghur follows SOV(Subject-Object-Verb) word order, for example, In Uyghur we can say “ ئۇ ئىشىكنى ئاچتى” as well as “ئىشىكنى ئۇ ئاچتى”. Both sentences mean “他/她开了门(he/she opened the door)” in Chinese; secondly Uyghur language is morphologically quite rich for example,

US: سەن قەيەرگە ماڭدىڭ؟  
UST: سەن قەيەر گە ماڭ دىڭ ؟  
CS: 你去哪儿？  
ES: Where are you going?

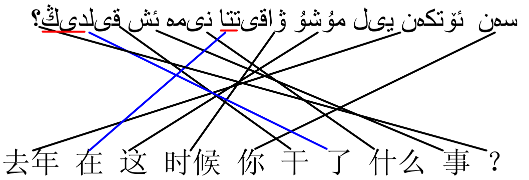
US: بۇ سىزنىڭ قول سومكىڭىزمۇ؟  
UST: بۇ سىز نىڭ قول سومكا ڭىز مۇ ؟  
CS: 这是您的手提包吗？  
ES: Is this your handbag?

US, UST, CS and ES show Uyghur, Uyghur tokenization[5], Chinese and English sentences respectively. It is clear that linguistic information has the potential of improving performance of Statistical Machine Translation systems, especially when limited amounts of parallel training data sets are available (data-sparseness problem). All Uyghur documents are segmented into lemma(stems) or tokens that are inputs for next steps for example, ئۈرۈمچىگە(to Urumqi) is tokenized into two morpheme: ئۈرۈمچى(Urumqi), گە(to),other example ئۈرۈمچىدىن(from Urumqi) is tokenized into two morpheme: ئۈرۈمچى(Urumqi), دىن(from) etc. Appendix A provides more details case affixation type include all suffixes.

So the technique of including morpheme based reordering and morphological processing for SMT for Uyghur gives more accuracy.

All Turkic languages exhibit SOV constituent order but depending on discourse requirements, constituents can be in any order without any substantial formal constraints. The dependency relation of a noun to other words, that is the role it plays in a sentence, is indicated by suffixes for example,

What were you doing this time last year?


سەن ئۆتكەن يىل مۇشۇ ۋاقىتتا نىمە ئىش قىلدىڭ؟
去年 在 这 时候 你 干 了 什么 事 ?

This observation means that the suffixes play the essential roles in Uyghur, and should be treated carefully in Uyghur-Chinese machine translation.

Typologically Uyghur and Chinese are rather distant languages for which rather modest parallel text data exists. Most importantly, Uyghur has complex agglutinative morphology with word structures that can correspond to complete phrases of several words in Chinese when translated.

In our experiments we investigate how different representations of morphology on both the Uyghur and the Chinese, Can help with word and phrase alignment.

We experiment with local word ordering on the Chinese side to bring the word order of specific Uyghur prepositional phrases and verb complexes in line with the morpheme order of the corresponding case marked noun forms and verbs on the Uyghur side.

We are combining phrase-based models with a group of reconstructing rules on Uyghur sentence for producing reordering Uyghur sentence as per Chinese word order. Incorporating morphological information in Uyghur by using a morpheme analyzer.

### 3. Phrase-based MT from Uyghur to Chinese

Supposing we want to translate a source language sentence $S_1^N = S_1 \ldots S_N$ into a target language sentence $T_1^M = T_1 \ldots T_M$, 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

$$\hat{t} = \operatorname{argmax} \Pr(T_1^M / S_1^N) \qquad (1)$$

Typically, Bayes rule is applied, obtaining the following expression

$$\hat{t} = \operatorname{argmax} \Pr(T_1^M)\Pr(S_1^N / T_1^M) \qquad (2)$$

This way, translating $S_1^N$ becomes the problem of detecting which $T_1^M$ among all possible target sentences scores best given the product of two models: $\Pr(T_1^M)$ forms the target language model (The $\Pr(T_1^M)$ is typically the standard n-gram language model), and $\Pr(S_1^N / T_1^M)$ 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 [9][18] based on IBM Model 1-5 and hidden Markov model.

The example of word and phrase alignment is as shown below（include UST-CS and US-CS）.

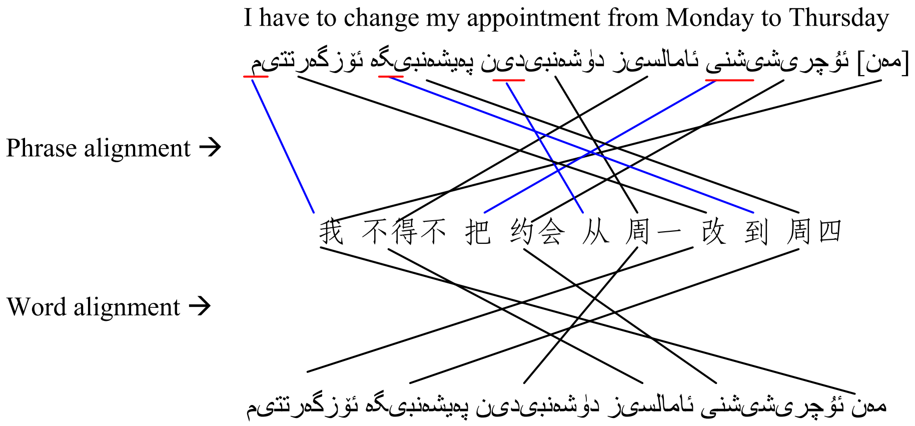

I have to change my appointment from Monday to Thursday
[مەن] ئۇچرىشىشنى ئامالسىز دۈشەنبىدىن پەيشەنبىگە ئۆزگەرتتىم
Phrase alignment →
我 不得不 把 约会 从 周一 改 到 周四
Word alignment →
مەن ئۇچرىشىشنى ئامالسىز دۈشەنبىدىن پەيشەنبىگە ئۆزگەرتتىم

Fig.2 Sample Word and Phrase Alignment

Phrase table created is given to the decoder. The role of the decoder in an SMT system is to globally search a target language so that the value of Pr($S_1^N / T_1^M$) is maximum based on the language model and the translation model trained before. In addition to the difference in the way to construct the language model, the translation model, the SMT systems use different ways to pre-process the input text beforehand using the decoders to make the output text and/or pre-process the output from the decoder to make a better translation text. After this we need a morph generator for combining stems and their corresponding suffixes of Uyghur sentences.

The implementation of decoding process is by segment the source f into a sequence of phrases. A uniform probability is assumed over all segments.

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 Uyghur to Chinese translation based phrase-based translation system. Our baseline is a phrase-based statistical machine translation system.

In the first step, Chinese language model was trained using SRILM (parameter set: n-gram-count -order 5 -interpolate -kndiscount), The language model is a statistical n-gram model estimated using modified Kneser-Ney smoothing. In the second step GIZA++ is used in two translation directions and the grow-diag-final combination method is used to obtain a symmetries alignment matrix.

We employ the phrase-based statistical machine translation framework [2], and use the Moses toolkit, and the SRILM[1] language modeling toolkit [10], and evaluate our decoded translations using the BLEU [13] measure, using a three references translation.

The Uyghur to Chinese MT corpus (UCC) consists of government, religious and law documents translated sentence by sentence into Chinese. Basic statistics about the training part of the UCC are given in Table 1. Table 2 summarizes the BLEU scores on the development and evaluation set for various training corpus and test corpus sizes.

Table 1 Statistics on Uyghur and Chinese Training and Test Data

| Uyghur | | | | |
|---|---|---|---|---|
| | Sentences | Words | Morphemes | Ave. Sent. Length |
| Train | 23,042 | 599,418 | 1,095,607 | 26.01/47.54 |
| Test | 260 | 6,783 | 12,876 | 26.08/49.52 |
| Chinese | | | | |
| Chinese Word Segmenter | Thu-Cws v1.0 | CWS 1.0 | ICTCLAS 1.0 | LDC 1.0 |
| Train Ave. Sent. Length | 23.73 | 24.26 | 24.97 | 25.89 |
| Train words | 546,827 | 559,169 | 575,435 | 596,631 |

Chinese word segment we used ICTCLAS1.0(http://ictclas.org/), Thu-Cwsv1.0 (http://166.111.138.244:8080/wordSegment/FrontPage.jsp) , LDC 1.0 and CWS 1.0 (Our Maximum Entropy Model Toolkit[2] for Chinese Word Segment [6][8]). The problem of Chinese word segmentation can be formally stated as follows. Given a sequence of words $W_i \ldots W_n$, we want to find the corresponding sequence of tags $t_1 \ldots t_n$, drawn from a set of tags $T= t_1 \ldots t_n$ of b, m, s, e tags, where b, m, s, and e stand for begin, middle, single and end of a word, respectively, which satisfies:

$$P(t_1 \ldots t_n | w_1 \ldots w_n)= \prod_{i=1}^{n} P(\mathrm{T}i \mid \mathrm{W}i) \tag{3}$$

The best feature templates used in our Chinese word segmentation experiments:
Unigram features:
$C_{i-2}$: previous two words in sentence
$C_{i-1}$: previous word in sentence
$C_i$: current word in sentence
$C_{i+1}$: next word in sentence
$C_{i+2}$: next two words in sentence
Bigram features:
$C_{i-2} C_{i-1}$, $C_{i-1}C_i$, $C_iC_{i+1}$, $C_{i+1} C_{i+2}$

### *4.2. Experiment Results*

Four experiments have been carried out. One employed phrase-based; one used phrase-based +morpheme, phrase-based + Dict (Bilingual dictionary) and another phrase-based + morpheme + Dict .

[1] http://www-speech.sri.com/projects/srilm/
[2] http://homepages.inf.ed.ac.uk/lzhang10/maxent.html

We evaluate the translation quality with the standard implementation of BLEU, as available for NIST evaluation[3] and with the default setting. An independent implementation of the BLEU metric was used to estimate confidence intervals for all the scores.

The central idea of these two metrics is that the closer a machine translation is to a professional human translation, the better it is. The metrics compute a distance between a candidate machine translation and a human (reference) translation by finding the average n-gram similarity. Each output Chinese translation is compared with three references. The evaluation results of these three experiments are presented in Table 2

Table 2 Comparison of Different Translation Models on the Uyghur-Chinese Task

| Experiments | BLEU | | | |
|---|---|---|---|---|
| | Thu-Cws v1.0 | CWS 1.0 | ICTCLAS 1.0 | LDC 1.0 |
| Word-based | | | | 8.16 |
| Word-based +Morpheme | | | | 11.69 |
| Phrase-based | 10.54 | 13.61 | 11.68 | 11.21 |
| Phrase + Dict | 10.99 | 13.95 | 13.86 | 12.79 |
| Phrase + Morpheme | 19.60 | 24.31 | 24.29 | 22.08 |
| Phrase + Morpheme + Dict | 20.12 | 25.26 | 25.00 | 23.37 |
| | | NIST | | |
| Word-based | | | | 3.22 |
| Word-based +Morpheme | | | | 4.05 |
| Phrase-based | 3.87 | 4.31 | 3.86 | 3.60 |
| Phrase + Dict | 3.92 | 4.32 | 4.35 | 4.16 |
| Phrase + Morpheme | 5.34 | 6.00 | 5.88 | 5.62 |
| Phrase + Morpheme + Dict | 5.48 | 6.05 | 5.90 | 5.72 |

From the results in Table 2, it is clear that the new Phrase-based + Morpheme + Dict SMTS outperforms the phrase-based method greatly.

There are some examples in bellow, which translated from Uyghur to Chinese

US:

<seg id=89> ئىككىنچى باسقۇچ (مايدىن سېنتەبىرگىچە): كەلكۈن مەزگىلى ۋە «بېيجىڭ
ئولىمپىك تەنھەرىكەت يىغىنى» بىخەتەرلىكىنى چۆرىدەپ، خەۋپىنى - تەكشۈرۈش، تۈزەش
خىزمىتىنى ياخشى ئىشلەش لازىم.</seg>

CS:

<seg id=89> 第二 阶段 (مايدىن سېنتەبىرگىچە): 洪水 在 和 «بېيجىڭ 奥运会 期间
چۆرىدەپ، يىغىنى» 安全 隐患 、 تەكشۈرۈش، 整治 工作 لازىم. </seg>

UST:

<seg id=89> ئىككىنچى باسقۇچ ( مايدىن سېنتە بىر گىچە ) : كەل كۈن مەزگىل ى
ۋە « بېيجىڭ ئولىمپىك تەن ھەرىكەت يىغىن ى » بىخەتەرلىك نى ى چۆرىدە پ ،
خەۋپ نى - تەكشۈر ۈش ، تۈز ە ش خىزمىت ى نى ياخشى ئىشلە ش لازىم .</seg>

CS:

<seg id=89> 第二 阶段 ( مايدىن سېنتە 一 你 : 汛期 和 " 北京 奥运会 行为 会议 " 要 以 安全 隐患 检查 、整改 。 </seg>

Reference (manual translation results):

[3] ftp://jaguar.ncsl.nist.gov/mt/resources , we used the version 11b.

<seg id=89> 第二 阶段 ( 5 月 至 9 月 ) : 围绕 汛期 和 北京 " 奥运会 " 安全 做 好 隐患 排查 治理 工作 。 </seg>

In future work, since Uyghur is very much morphologically rich and agglutinative we try to improve the performance of Uyghur tokenize and Chinese word segment. 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.

## 5. Conclusion

In this paper we presented an effective phrase-based SMT system for Uyghur to Chinese machine translation. The translation results can be achieved by applying morpheme-based technique in combination with morphology information from corpus. The results shows that significant improvements are possible by incorporating morphological information to the plain corpus. The results are quite positive but there is quite some room 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. We use morphological analysis to find the base form (token or lemma(stem)) of similar words, thus creating word classes which tend to be translated by the same target word.

**Appendix A:** We give readers a feel for the case affixation include all suffixes, it as shown Table 3.

Table 3 Case Affixation Type

| Type | Suffixes | Example |
|---|---|---|
| Possessive Case | نىڭ | مەكتەپنىڭ(school’s) |
| Dative Case | گە كە قە | مەكتەپكە (to school) |
| Accusative Case | نى | (thschool e) مەكتەپنى |
| Locative Case | تىكى  دا دە تا تە | مەكتەپتە(in school) |
| Ablative Case | دىن تىن | مەكتەپتىن (from school) |
| Similitude Case | دەك تەك | مەكتەپتەك(like school) |
| Limitative Case | قىچە كىچە  غىچە | مەكتەپكىچە (Until school) |

The all suffixes play the essential roles in Uyghur, and should be treated carefully in Uyghur text processing.

## Acknowledgement

The authors would like to thank F. J. Och , P. Koehn , Daniel Marcu and U. Germann , A. Stolcke , K. Papineni for providing the open source Toolkit. This research is a preparation work for the project “Common

Key Technologies and Demonstrating Applications in Information Processing of Minority Languages" of the National Science and Technology Support Program of China.

## References

[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.

[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 Uyghur Tokenization, in Proceedings of IEEE International Conference on Natural Language Processing and Knowledge Engineering. pp.383-387.

[6] Adam 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] Xue. Nianwen, Chinese word segmentation as character tagging, In Journal of Computational Linguistics and Chinese Language Processing, 2003, 8(1): 29-48.

[9] F. J. Och and H. Ney. A Systematic Comparison of Various Statistical Alignment Models. Computational Linguistics, 2003, 29(1): 19-51.

[10] A. Stolcke. Srilm- An Extensible Language Modeling Toolkit. In Proceedings of the International Conference on Spoken language Processing, volume 2, 2002, pp. 901-904.

[11] 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.

[12] Polat KADIR, Koichi YAMADA, Hiroshi KINUKAWA, Comparative Study on Japanese and Uyghur Grammars for An English-Uyghur Machine Translation System, MT Summit X. Conference Proceedings: the tenth Machine Translation Summit; 2005, pp.432-437.

[13] 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

[14] 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.

[15] 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.

[16] Muhtar Mahsut et al., 1995. http://www.kl.i.is.nagoya-u.ac.jp/person/muhtar/snlp95.ps.

[17] 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

[18] F. Och and H. Ney. Improved statistical alignment models, in Proc. ACL, Hong Kong, China, 2000