著者（英）
Mamtily Nighmat
学位名
博士 工学
学位授与番号
00006442
学位授与年月日
2019-03-27
URL
http://id.nii.ac.jp/1476/00006442/
The Development of Japanese-Uyghur Statistical Machine Translation System with Morphological Information
語形情報に基づく日本語・ウイグル語の統計的機械翻訳システムの開発

2019

Mamtily Nighmat
As explained in this thesis, we applied an open-source phrase-based statistical machine translation (SMT) toolkit (MOSES) to build a Japanese to Uyghur machine translation system. We further attempted to explore the results of a rich sourced language, in this case Japanese, translated to a low sourced grammatically similar language: Uyghur. Phrase-based SMT is a complex structure that uses several other components at least including the translation model, language model, and decoder model. The main emphasis of this thesis is to compare and analysis base phrase-based SMT model with the implemented model of adding other features from preprocessing and post-processing to the phrase-based SMT model and a parallel corpus. Preprocessing includes evaluating parallel corpora, adding morphological derivation into the Uyghur language corpus, and assessing effects of translation results under a phrase-based statistical approach. Post-processing includes morphological combinations of translation results.

More efficient statistical approaches are available for analyzing bilingual text corpora in preprocessing than phrase-based SMT approaches such as the factory-based SMT model and the Tree to Tree SMT model, but no earlier reported work has addressed Japanese–Uyghur SMT. The size of language sources and requirements of large amounts of human work have strictly restricted further progress. Compared to other SMT methods, phrase-based SMT can generate translation directly between surface forms and especially translate phrases rather than single words. Phrase-based SMT is assumed to produce fair translation results in languages that have similar grammatical structures. Compare to very few sourced languages such as Uygur language, Japanese language is a rich sourced language and includes not only include huge bilingual datasets parallel to the English but also has tremendous size of morphological dictionaries in many different approaches. This richness gives Japanese language great benefits of high probability for a phrase translation table and a more naturally trained language model in SMT.

For experimentation, we prepared about 5000 parallel sentences. For training a phrase-based machine translation system, we applied manual work to evaluate a parallel corpus, aa pruning phrase translation table, and a parsed small part of training of a parallel dataset. Differences in evaluation results arise between machine evaluation and human evaluation for the machine translation results. Therefore, we submit both BLEU and human evaluation scores and compare the evaluated results. From the test score, it is difficult to distinguish the progress of automatic evaluation than giving human evaluation under the low number of training data and test data for translation.

Finally, we implement the first Japanese to Uyghur SMT system. Although the structure is simple, phrase-based machine translation, receiving a correct translation result by application of calculated probability in small parallel corpora is challenging. Nevertheless, we
obtained positive results by evaluating and morphologically decomposing the Uyghur language in parallel corpora. We also obtained better results by application of preprocessing to the language model, phrase table, and post-processing of translation results. Some understandable translated results are produced by this system. Many structures exist in the SMT system. We only applied phrase-based machine translation. There is much work to do for simply adjusting language features to many other frameworks for future compatibility to rich source languages in the SMT field. Furthermore, for a whole SMT system, the parallel corpus is the most important resource. Therefore, we must collect additional resources not only in Japanese but also in other main research languages such as English and Chinese.
Acknowledgement

I would like to express my gratitude to my supervisor Izumi YAMAMOTO for her knowledgeable guidance, support and great patience through on this research.

I am very thankful for graduated master student RAYILE ELI for her valuable assistant work and help.

I would like to express my sincere gratitude to my supervisor IZUMI Yamamoto for her great patience, support and invaluable feedback during the work of this thesis.

I also thank Professors in intelligence field, espacially Ichi TAKUMI for his valuable feedback and help.

A special thanks goes to my wife who has supported me through this year.
目 次

第1章 INTRODUCTION 1
  1.1 Brief History of Machine Translation .......................... 1
  1.2 Approaches for Computational Machine Translation Systems .......... 2
  1.3 Statistical Machine Translation ...................................... 3
  1.4 Phrase-based Machine Translation System .......................... 4
  1.5 Thesis Overview .................................................. 5

第2章 CONCEPTS OF JAPANESE GRAMMAR AND UYGHUR GRAMMAR 9
  2.1 Comparison in Morphology .......................................... 9
  2.2 Noun ............................................................... 10
  2.3 Verb ............................................................... 10
  2.4 Sentence Style .................................................... 12
  2.5 Overall ......................................................... 14

第3章 JAPANESE–UYGHUR MACHINE TRANSLATION SYSTEM 15
  3.1 Machine Translation for Uyghur Language ......................... 16
    3.1.1 Statistical machine translation system .......................... 19
  3.2 Workflow of SMT .................................................. 21
    3.2.1 Language modeling ............................................ 22
    3.2.2 Discounting method ......................................... 25
    3.2.3 Linear interpolation ......................................... 26
    3.2.4 Translation model ............................................ 26
  3.3 Phrase based approach ............................................. 30
  3.4 Word alignment .................................................... 35
  3.5 Corpus Evaluation ................................................ 36
    3.5.1 Evaluation techniques ....................................... 38
    3.5.2 Effect of parallel corpora ................................... 41
  3.6 The JU Parallel Corpus ........................................... 41
    3.6.1 Evaluation of SMT systems ................................... 42

第4章 EVALUATION OF PARALLEL CORPUS 45
  4.1 Experiment Objective .............................................. 45
  4.2 Data Preparation ................................................ 46
  4.3 Proposed Model .................................................. 47
  4.4 Evaluation Method ................................................. 48
  4.5 Evaluation Results and Issues .................................... 51
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.6</td>
<td>Error analysis</td>
<td>54</td>
</tr>
<tr>
<td>5.1</td>
<td>Proposed Method</td>
<td>55</td>
</tr>
<tr>
<td>5.1.1</td>
<td>“Organize the corpus” method</td>
<td>55</td>
</tr>
<tr>
<td>5.1.2</td>
<td>Formula derivation</td>
<td>57</td>
</tr>
<tr>
<td>5.2</td>
<td>Tools Employed</td>
<td>58</td>
</tr>
<tr>
<td>5.3</td>
<td>Experiment</td>
<td>59</td>
</tr>
<tr>
<td>5.3.1</td>
<td>Translation model calculation</td>
<td>59</td>
</tr>
<tr>
<td>5.3.2</td>
<td>Phrase-based translation model</td>
<td>61</td>
</tr>
<tr>
<td>5.4</td>
<td>Language Model Estimation</td>
<td>63</td>
</tr>
<tr>
<td>5.4.1</td>
<td>“be verb” organized corpus</td>
<td>63</td>
</tr>
<tr>
<td>5.4.2</td>
<td>“do verb” organized corpus</td>
<td>65</td>
</tr>
<tr>
<td>5.4.3</td>
<td>“Mixed verb” corpus</td>
<td>67</td>
</tr>
<tr>
<td>5.5</td>
<td>Evaluation Result</td>
<td>68</td>
</tr>
<tr>
<td>5.5.1</td>
<td>Translation model</td>
<td>68</td>
</tr>
<tr>
<td>5.5.2</td>
<td>Phrase-based translation</td>
<td>69</td>
</tr>
<tr>
<td>5.5.3</td>
<td>Language model</td>
<td>69</td>
</tr>
<tr>
<td>6.1</td>
<td>Applied Models in Experiments</td>
<td>75</td>
</tr>
<tr>
<td>6.2</td>
<td>Original Model</td>
<td>75</td>
</tr>
<tr>
<td>6.3</td>
<td>Proposed Model</td>
<td>76</td>
</tr>
<tr>
<td>6.4</td>
<td>Result</td>
<td>77</td>
</tr>
<tr>
<td>6.5</td>
<td>Conclusion</td>
<td>79</td>
</tr>
<tr>
<td>7.1</td>
<td>Conclusion</td>
<td>83</td>
</tr>
<tr>
<td>7.2</td>
<td>Future Work</td>
<td>84</td>
</tr>
<tr>
<td>7.2.1</td>
<td>Test sentences</td>
<td>89</td>
</tr>
<tr>
<td>Acknowledgement</td>
<td></td>
<td>3</td>
</tr>
<tr>
<td>Reference</td>
<td></td>
<td>85</td>
</tr>
<tr>
<td>Appendix</td>
<td></td>
<td>89</td>
</tr>
</tbody>
</table>
2.1 Translation Method for Japanese Language and Uyghur Language 13
3.1 Rule-Based Machine Translation Approach. 17
3.2 Rule-Based Machine Translation Example. 18
3.3 Rule-Based Machine Translation Approaches. 19
3.4 Rule-Based Machine Translation Process Example. 20
3.5 Architecture of Statistical Translation Systems. 20
3.6 Statistical Translation Systems and Corpora 37
4.1 UAY and ULY. 46
4.2 Evaluation Model of Statistical Translation Systems. 47
5.1 "Be verb" Organized Corpus. 71
5.2 "Do verb" Organized Corpus. 72
6.1 SMT original model. 76
6.2 SMT proposed model. 77
6.3 NIST score 78
6.4 BLEU Cumulative Score 79
6.5 BLEU Individual score 80
表 目 次

2.1 Parsing the Sentence ("couldn’t go") ........................................... 10
2.1 Persons and Numbers of Nouns in Uyghur .................................. 10
2.1 Verb Conjugation in Japanese–Uyghur ...................................... 11
2.2 Subject and Verb Relations in Uyghur Sentences ......................... 11
2.3 Verbs in Japanese Sentences and Uyghur Sentences ...................... 12
2.1 Case Particle of Japanese and Case Suffix of Uyghur ...................... 13
2.2 Verb ‘to be’ ............................................................................. 13
2.3 Verb ‘to do’ ............................................................................. 14
2.4 Verb ‘to be + to do’ ................................................................... 14
3.1 Japanese–Uyghur Parallel Sentence Word Alignment .................... 36
4.1 English to Uyghur ..................................................................... 49
4.2 Chinese to Uyghur ..................................................................... 49
4.3 Comparison of Two Translated Uyghur Sentences ......................... 50
4.4 English to Japanese ................................................................... 50
4.5 Chinese to Japanese ................................................................... 50
4.6 Comparison of Two Japanese Sentences ...................................... 50
4.7 Translated Japanese to Uyghur .................................................. 50
4.8 Translated Japanese to Uyghur .................................................. 50
4.9 Comparison of Two Translated Uyghur Sentences ......................... 51
4.1 #Example Sentence No. 58 ....................................................... 51
4.2 #Example Sentence No. 148 ..................................................... 51
4.3 #Example Sentence No. 103 ..................................................... 52
4.4 #Example Sentence No. 153 ..................................................... 52
4.5 #Example Sentence No. 959 ..................................................... 52
4.6 #Example Sentence No. 4301 .................................................. 52
4.7 #Example Sentence No. 77 ....................................................... 53
4.8 #Example Sentence No. 39 ....................................................... 53
4.9 #Example Sentence No. 719 ..................................................... 53
4.10 #Example Sentence No. 722 .................................................... 53
5.1 "be verb" Sentences ................................................................. 57
5.1 Japanese-to-Uyghur Word Alignment .......................................... 60
5.2 Japanese-to-Uyghur Word Alignment .......................................... 61
5.1 100 "be verb" Organized Corpus ............................................... 64
5.2 100 Unorganized Corpus ........................................... 64
5.3 200 "be verb" Organized Corpus .................................... 64
5.4 200 Unorganized Corpus ............................................. 64
5.5 400 "be verb" Organized Corpus ..................................... 64
5.6 400 Unorganized Corpus ............................................. 64
5.7 600 "be verb" Organized Corpus ..................................... 64
5.8 600 Unorganized Corpus ............................................. 64
5.9 800 "be verb" Organized Corpus ..................................... 65
5.10 800 Unorganized Corpus ............................................ 65
5.11 1178 "be verb" Organized Corpus .................................. 65
5.12 1178 Unorganized Corpus .......................................... 65
5.13 100 "do verb" Organized Corpus .................................... 65
5.14 100 Unorganized Corpus ............................................ 65
5.15 500 "do verb" Organized Corpus .................................... 66
5.16 500 Unorganized Corpus ............................................ 66
5.17 1000 "do verb" Organized Corpus ................................... 66
5.18 1000 Unorganized Corpus ........................................... 66
5.19 1500 "do verb" Organized Corpus ................................... 66
5.20 1500 Unorganized Corpus ........................................... 66
5.21 2000 "do verb" Organized Corpus ................................... 67
5.22 2000 Unorganized Corpus .......................................... 67
5.23 2723 "do verb" Organized Corpus ................................... 67
5.24 2723 Unorganized Corpus .......................................... 67
5.25 500 "mixed verb" Organized Corpus ................................. 67
5.26 500 Unorganized Corpus .......................................... 67
5.27 1200 Unorganized Corpus .......................................... 68
5.28 1200 Unorganized Corpus .......................................... 68
5.29 100 "Organized Mixed Verbs" ...................................... 68
5.30 100 "Unorganized Mixed Verbs" ................................... 68
5.1 Results Obtained Using the Proposed Method ......................... 72
6.1 Verb ‘To be’ Sample 1 .............................................. 80
6.2 Verb ‘To be’ Sample 2 .............................................. 80
6.3 Verb ‘To be’ Sample 3 .............................................. 80
6.4 Verb ‘To do’ Sample 1 .............................................. 81
6.5 Verb ‘To do’ Sample 2 .............................................. 81
6.6 Verb ‘To do’ Sample 3 .............................................. 81
6.7 Verb ‘To be + To do’ Sample 1 ..................................... 81
6.8 Verb ‘To be + To do’ Sample 2 ..................................... 81
6.9 Verb ‘To be + To do’ Sample 3 ..................................... 81
Machine translation (MT) systems are old computer applications: they have been used since 1949 to produce a translation from one human source language to another language automatically. Many ways exist to produce such translations.

1.1 Brief History of Machine Translation

The history of machine translation can be divided into five main periods\[1\]\[2\]\[3\].

- Warren Weaver presented the first idea of machine translation in 1949. Yehoshua Bar Hillel led the first machine translation symposium at the Massachusetts Institute of Technology (MIT) in July 1952. A group of researchers from Georgetown University worked with International Business Machines Inc. (IBM) and developed first basic automatic translator. The translator produced results of 60 Russian sentences translated into English. The developers claimed then that machine translation would not be a difficulty within five years. The first journal article describing machine translation was published with the title of "Mechanical translation devoted to the translation of languages utilizing machines" by Victor Yngve.

- The early 1960s marked the second period. Grammatical parsing has been emphasized, with many parsers developed with grammar of different types. In September 1961, many computer scientists and linguists including Lamb, Harper, Garvin, Hockett, Kay, and Vauquois participated in the "First International Conference on Machine Translation of Languages and Applied Language Analysis of Teddington." In 1964, Automatic Language Processing (ALPAC) created as a committee with the US government for additional research into machine translation.
• The third period started from 1966. Toma started to Develop SYSTRAN1 as a translation system between Russian and English. In 1976, Colmerance led a group of researchers and created the WEATHER system under a project by the University of Montreal. In 1978, Fujitsu Ltd. created a rule-based translation system called AT-LAS2.

• Many Japanese companies started to build their machine translation systems between English and Japanese during the fourth period, which began in the 1980s. In 1982, Sharp Corp. introduced their rule-based machine translator: DUET. In 1983, NEC Inc. marketed their interlingua-based machine translation system: Honyaku Adaptor II based on PIVOT algorithm. In 1986, another Japanese company, Oki Ltd., introduced a rule-based translation machine, PENSEE, for translation between English and Japanese. Researchers at Hitachi Ltd. created their own rule-based machine translation system, the Hitachi Computer Aided Translation System (HICATS), based on a transfer approach.

• The fifth period started in the 1990s. In 1993, the Consortium for Speech Translation Advanced Research (C-STAR) was created. It became a base for C-STAR I, which is trilingual translation system between English, German, and Japanese languages. In 1998, Softissimo introduced the REVERSO translator. In the 2000s, the Japanese laboratory, ATR, developed the ALPH system based on example approaches between English, Japanese, and Chinese. In 2005, Google (Alphabet) introduced a web translation system, METIS-II, which was developed using a hybrid translation approach that includes rule-based, example-based, and statistic-based machine translation approaches.

1.2 Approaches for Computational Machine Translation Systems

Different architectures have been implemented at different times by different work throughout machine translation history. Computational machine translation systems are generally divided into rule-based, corpus-based, and hybrid approaches.
• Rule-based machine translation systems are a main subject in the machine translation field to the present day. Transfer and Interlingua are two main approaches for rule-based machine translation[4]. Different levels of linguistic rules are applied in rule-based machine translation systems. The rules are applied by morphological and syntactic analysis, lexical transfer and syntactic generation. In standard rule-based machine translation system include translation rules, morphological dictionaries, and dictionaries of bilingual or more set languages. Building those parts of rule-based machine translation systems requires computer scientists to be knowledgeable about the linguistic field and to implement rules in all linguistic aspects. These processes are very time consuming because they require collecting those linguistic resources and accumulating linguistic knowledge[5].

• Corpus-based approaches are comparable to rule-based approaches and are also designated as data-driven machine translation. They provide clear solutions for knowledge acquisition problems of rule-based machine translation systems. Corpus-based machine translation (CBMT) can be divided to example-based machine translation (EBMT) and statistical machine translation (SMT). CBMT generates a translation model automatically by learning and analyzes bilingual text corpora[6][7][8][9].

• Hybrid approaches combine the rule-based approach and corpus-based approaches. They were proposed to work under fewer translation rules, language knowledge resources, and translation models trained from parallel text corpora. Learning the syntactic generation rules from corpus-based models is a main purpose of these systems. Complete systems include steps of morphological parsing, lexical or transfer, and syntactic generation[10][11].

1.3 Statistical Machine Translation

Since IBM introduced a machine translation model based on machine learning, progress in statistical machine translation systems has increased considerably [12][13]. Applied machine learning methods generates translation and language models from parallel data. They have overcome difficulties of rule-based approaches, which manually code tremendous sets of translation rules in linguistic knowledge. Groups of large training parallel datasets provide
the possibility for supervised machine learning approaches such as the Canadian Hansards corpus and the European Parliament corpus[14].

For translation, the IBM model is limited in terms of word-to-word translation. Phrase-based machine translation systems were introduced in the late 1990s by researchers[15][16][17][18]. To produce translations, phrase-based models apply mapping to construct translations between phrases from the source language and target language. Google developed web translation tools based on phrase-based SMT approach, which considerably increased the translation quality of CBMT approaches in the 15 years. Such approaches are widely applied today.

Many efficient approaches have been developed based on SMT approaches such as syntax-based model, factored-based model, Tree-to-String, and the Tree-to-Tree model. They all have performed well by applying and adding morphological feature to a bilingual training corpus to which phrase-based models are not applied[19][20][21][22].

1.4 Phrase-based Machine Translation System

Phrase-based SMT approach is representative of a recent machine learning framework in machine translation. As described in this thesis, we applied a translation method based on a phrase-based statistical approach trained in non-edited bilingual text corpora as a baseline system for comparison with the phrase-based SMT trained in edited corpora. Further information about how phrase-based approach works will be discussed in Chapter 2.

Recent studies have examined statistical machine translation (SMT) between the Uyghur language and other languages such as Chinese language. Although Japanese and Chinese languages share some character similarity, because of word order differences, we were unable to apply the existing structure of language modeling, translation table, and morphological features to models we would like to implement. In this work, we compare grammars of both languages and analyze the results of the phrase-based machine translation results in the Japanese–Uyghur SMT system. Then we compare a standard phrase-based SMT model with the implemented model of adding other features to preprocessing and post processing to the base phrase-based SMT model. Preprocessing includes adding morphological derivation to the Uyghur language corpus and assessment of how it affects translation results under
the phrase-based statistical approach. Post-processing includes morphological combination of translation results.

1.5 Thesis Overview

In machine translation systems, resources for many languages are limiting factors. Especially for low source languages such as the Uyghur corpus, low-quality machine translation systems are still available. Two important key steps can be used for implementing better performance of SMT for Uyghur language with another target language. The first step is to increase the parallel corpus efficiency to organize a corpus to improve the alignment probability and achieve better translation performance without a huge parallel corpus. The second step is to apply morphological features to a parallel corpus and translation results. Training a phrase-based SMT system by morphologically analyzing the corpus and morphologically composing divided suffix in translation results will produce understandable translation results. Particularly, this thesis emphasizes the following points.

- Improving the alignment
  Japanese and Uyghur languages have many word structure similarities, but they also have differences. In the Uyghur language, nouns and verbs have different forms according to the person, number, and tense of the subject. By contrast, Japanese has no such conjugation: verbs are independent of the person, number, and subject. In SMT systems, word alignment plays an important role, but because of the quantity of the resources of Uyghur language and Japanese, it is difficult to find correct alignment. The idea proposed in this thesis is to introduce a method to improve Japanese-to-Uyghur word alignment.

- Organization method
  Most errors in Japanese-to-Uyghur SMT result from lack of resources and translation result problems. An efficient means of overcoming the low sources issue is to use existing data in an effective way by evaluating and morphologically analyzing parallel corpora, increasing the Language model and morphologically processing the translation result. This can be achieved by organizing existing and available data from JEC from Kurohashi & Kawahara lab. The data for English, Japanese, and Chinese, then adding the Uyghur language, constitute four languages corpora called JECU 5304 × 4 sentence alignment.
This study specifically examines adoption of an efficient algorithms model: a language model to learn how to enhance the alignment probability.

• Experiment and Evaluation
For all aspects of parallel corpus evaluation, morphological analysis, and result processing, we conducted machine evaluation, human evaluation, and a model calculation evaluation method. For deep examination of the proposed idea of "resource organize," the parallel corpus JECU 5304 × 4 sentence alignment was organized. Results of the language model in the same and different structure sentences were compared. We also analyzed the translation model and phrase-based model probability with a language model and compared the standard phrase-based SMT system with a modified phrase-based SMT (translation results modified) to implement our phrase-based SMT model.

The structure of this thesis is the following.
• Chapter 1 presents an introduction of machine translation systems and a brief introduction of this thesis.

• Chapter 2 explains details about grammatical study and translation approach about how Uyghur language applies translation from Japanese language.

• Chapter 3 explains the theoretical background of SMT and research in Japanese language and Uyghur language. Also, we describe a models formula baseline for the second experiment, and describe corpus evaluation based on techniques and effects.

• Chapter 4 shares our result in how a morphologically analyzed corpus affects the SMT result. The language model trained by morphological analyzed corpus will help a translation decoder to generate sentences that are more understandable.

• Chapters 5 explains our language modeling used for the current research. Because we are developing a Japanese–Uyghur SMT system, we even more strongly realized that the machine learning step is extremely important for collection of resources. We must
also add manual resources to supervised machine learning method for base comparative study. These steps are expected to be a base to help us to build our unsupervised model for our future Japanese–Uyghur machine translation system.

- Chapter 6 presents our experiment method and comparison results of standard phrase-based SMT approach with modified phrase-based SMT approach. Results show that the morphological analysis process to parallel corpus and translation result will improve the translation accuracy for low source languages. We listed three word format translated example sentences: ’to be’, ’to do’, and ’to be + to do’.

- Chapter 7 This chapter explains our future step for current research and summarizes chapters above.
第2章 CONCEPTS OF JAPANESE GRAMMAR AND UYGHUR GRAMMAR

As a member of an agglutinative language family, a verb in the Uyghur language can have hundreds of word forms by sequential addition of different suffixes to the word stem. Japanese, which is also regarded as an agglutinative language, has the same word order and morphological features as the Uyghur language [subject + object + verb]. Actually, an earlier study (Ogawa 1997, 2000) [23] [24] showed that morphological and syntactic similarity is sufficient to prove good translation results from Japanese into Uyghur on a transfer approach. One report (Kadir 2004, 2005) [25] [26] explained the English–Uyghur machine translation system as efficient based on existing Japanese–English rule-based machine translation. However, the two languages have a few differences. We will make a brief comparison between Japanese language and Uyghur language on two different levels, morphology and syntax, with close attention devoted to their differences.

2.1 Comparison in Morphology

Word information from Table 1 shows that in both Japanese sentences and Uyghur sentences, word forms are generated by combining suffixes to one word stem.

Examples show that Japanese language and Uyghur language share common morphological and syntactic features. However, differences in word formation of nouns, verbs, etc. exist. The following section presents some differences in Japanese grammar and Uyghur grammar.
2.2 Noun

Uyghur nouns have ownership to object and generate sentences by companioning some person, number, etc. When attaching a different suffix, a noun in Uyghur language expresses different ownership to an object. Furthermore, this very same suffix will simultaneously show different person and number categories (Tomur, 2003) [27]. Table 2 shows the word ”glass” with two categories of person and number.

<table>
<thead>
<tr>
<th>Person</th>
<th>Singular</th>
</tr>
</thead>
<tbody>
<tr>
<td>1st person</td>
<td>Istakan. im (my glass)</td>
</tr>
<tr>
<td>2nd person</td>
<td>Istakan. ing (your glass)</td>
</tr>
<tr>
<td>3rd person</td>
<td>Istakan. i (his/her glass)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Person</th>
<th>Plural</th>
</tr>
</thead>
<tbody>
<tr>
<td>1st person</td>
<td>Istakan. imiz (our glass)</td>
</tr>
<tr>
<td>2nd person</td>
<td>Istakan. inglar (your glass)</td>
</tr>
<tr>
<td>3rd person</td>
<td>Istakan. i (their glass)</td>
</tr>
</tbody>
</table>

Table 2.1: Persons and Numbers of Nouns in Uyghur

Japanese sentences require no grammatical category or change for a word form.

2.3 Verb

A Japanese verb makes ”katsuyo, 活用” (changing word forms of the verb stem) before they conjugate to show different tenses and moods, etc. In fact, ”katsuyo” is a difference in Japanese language from the Uyghur language. There is no such inflection of verbs before
conjugation in Uyghur. However, in most situations, many verbal suffixes in Japanese correspond to Uyghur. Table 3 presents similarities of verb conjugation in Japanese and Uyghur (stem verb ”go”).

<table>
<thead>
<tr>
<th>Japanese verb stem</th>
<th>行く</th>
<th>行</th>
<th>行</th>
</tr>
</thead>
<tbody>
<tr>
<td>Katsuyogobi</td>
<td>く</td>
<td>か</td>
<td>か</td>
</tr>
<tr>
<td>suffix</td>
<td>none</td>
<td>せる</td>
<td>させる</td>
</tr>
<tr>
<td>Uyghur verb stem</td>
<td>bar</td>
<td>bar</td>
<td>bar</td>
</tr>
<tr>
<td>Suffix</td>
<td>none</td>
<td>ghuz</td>
<td>ghuzul</td>
</tr>
</tbody>
</table>

Table 2.1: Verb Conjugation in Japanese–Uyghur

There are still differences according to grammatical categories of nouns between Japanese language and Uyghur language (Tomur and Lee, 2003). Singular and plural are expressed on subjects and verbs in Uyghur sentences. Verb would be affected by different inflectional forms of the subject such as person, pronoun, and number. The number and tense are also expressed by the suffixes of verbs. However, Japanese sentences have no subject–verb agreement as Uyghur sentences do. The table below shows the subject and verb relation in Uyghur sentences.

<table>
<thead>
<tr>
<th></th>
<th>Number</th>
<th>Present tense</th>
<th>Past tense</th>
</tr>
</thead>
<tbody>
<tr>
<td>1st</td>
<td>Singular</td>
<td>Bar-iman</td>
<td>Bar-dim</td>
</tr>
<tr>
<td>Person</td>
<td>Plural</td>
<td>Bar-imiz</td>
<td>Bar-duk</td>
</tr>
<tr>
<td>2nd (formal)</td>
<td>Singular</td>
<td>Bar-isiz</td>
<td>Bar-tingiz</td>
</tr>
<tr>
<td>Person</td>
<td>Plural</td>
<td>Bar-isilar</td>
<td>Bar-tinglar</td>
</tr>
<tr>
<td>2nd (play)</td>
<td>Singular</td>
<td>Bar-isan</td>
<td>Bar-ting</td>
</tr>
<tr>
<td>Person</td>
<td>Plural</td>
<td>Bar-isilar</td>
<td>Bar-tinglar</td>
</tr>
<tr>
<td>3rd</td>
<td>Singular</td>
<td>Bar-idu</td>
<td>Bar-di</td>
</tr>
<tr>
<td>Person</td>
<td>Plural</td>
<td>Bar-idu</td>
<td>Bar-di</td>
</tr>
</tbody>
</table>

Table 2.2: Subject and Verb Relations in Uyghur Sentences

Table 2.2 shows second person plural (formal form) and second person plural (play form) as having the same word format. The same table also shows third person singular form and plural form as having the same word format.

From the explanation above, Japanese verbs and Uyghur verbs have many similarities. However, the two following points must be mentioned.
Table 2.3: Verbs in Japanese Sentences and Uyghur Sentences

<table>
<thead>
<tr>
<th>Stem(verb)</th>
<th>行く</th>
<th>bar</th>
</tr>
</thead>
<tbody>
<tr>
<td>Causitive</td>
<td>せる</td>
<td>ghuz</td>
</tr>
<tr>
<td>Passive</td>
<td>（ら）れる</td>
<td>il</td>
</tr>
<tr>
<td>Aspect</td>
<td>ている</td>
<td>iwat</td>
</tr>
<tr>
<td>Negation</td>
<td>ない</td>
<td>ma</td>
</tr>
<tr>
<td>Tense</td>
<td>た</td>
<td>idu</td>
</tr>
<tr>
<td>Person</td>
<td>(-)</td>
<td>idu</td>
</tr>
<tr>
<td>Modal</td>
<td>だろう</td>
<td>ghandur</td>
</tr>
<tr>
<td>Mood</td>
<td>ね</td>
<td>he</td>
</tr>
</tbody>
</table>

- Katsuyogobi only require verbs in Japanese sentences; verbs in Uyghur sentences do not.

- Verbs in Uyghur sentences are heavily dependent on the subject format; Japanese sentences are not.

2.4 Sentence Style

Japanese sentences and Uyghur sentences both belong to the Altaic language system, mostly with subject + object + verb (SOV) word order. Therefore, Japanese sentences can be translated directly to the target Uyghur sentences by sequentially translating word by word. As we explained in the previous section, because Uyghur verbs follow ‘subject–verb agreement’, it is important to identify the case category of nouns. Japanese language and Uyghur language both express the sentence meaning by case forms. Adding nominal case suffixes to the subject (such as noun) can generate case forms. There is always a correspondence case suffix to a case particle for translating Japanese to Uyghur. Comparison of the case category in both languages produces the results presented in the table below.

Relations between phrasal units in both languages are very similar. Comparing the rule ‘Bunsetsu’ (Watanabe et al., 2000; Uchimoto et al., 1999) with dependency structure in the Uyghur language, there are three similar points as described below.

- Dependence of a word on another word is from left to right.
### Table 2.1: Case Particle of Japanese and Case Suffix of Uyghur

<table>
<thead>
<tr>
<th>Case name</th>
<th>Case particle in Japanese</th>
<th>Case Suffix in Uyghur</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nominative Case</td>
<td>は / が¹</td>
<td>-</td>
</tr>
<tr>
<td>Possessive Case</td>
<td>の</td>
<td>-ning</td>
</tr>
<tr>
<td>Dative Case</td>
<td>へ</td>
<td>iwat</td>
</tr>
<tr>
<td>Objective Case</td>
<td>を</td>
<td>-ni</td>
</tr>
<tr>
<td>Locative Case</td>
<td>に</td>
<td>-da</td>
</tr>
<tr>
<td>Ablative Case</td>
<td>から</td>
<td>-din</td>
</tr>
</tbody>
</table>

- Dependency connects two independent sentences.
- A head word can only be linked by one dependent word.

From those three similarities, any change in Japanese sentences can be mapped by Uyghur word order. As in other languages, Uyghur language sentences also have three common styles.

![Fig. 2.1: Translation Method for Japanese Language and Uyghur Language](image)

<table>
<thead>
<tr>
<th>Japanese</th>
<th>私は留学生です。</th>
</tr>
</thead>
<tbody>
<tr>
<td>Uyghur</td>
<td>Man chatallik oqughuchi</td>
</tr>
</tbody>
</table>

**Table 2.2: Verb ‘to be’**
### Table 2.3: Verb ‘to do’

<table>
<thead>
<tr>
<th>Japanese</th>
<th>私は外国語を勉強します。</th>
</tr>
</thead>
<tbody>
<tr>
<td>Uyghur</td>
<td>曼沙特特利呢欧罕曼</td>
</tr>
</tbody>
</table>

### Table 2.4: Verb ‘to be + to do’

<table>
<thead>
<tr>
<th>Japanese</th>
<th>私は外国人ですが、日本語で会話できます。</th>
</tr>
</thead>
<tbody>
<tr>
<td>Uyghur</td>
<td>曼沙特利克包思满胡，亚盼特里达格齐勒亚曼</td>
</tr>
</tbody>
</table>

#### 2.5 Overall

As an agglutinative language, Uyghur language has rich and complex morphology. Very large number of words will generated from the same root word with different suffixes. Newly generated suffixes will take many different suffixes in a sequence and generate different words, which will produce the same root related vocabulary and produce difficulties in translating to other language if the root words are not clear.
Machine translation has proposed many approaches. Rule-based machine translation is one of the most applied translation approaches. Rule-based machine translation systems produce translation results from source language sentences to target language sentences by application of hand-coded rules which are not very capable of covering the translated source material acceptably. A statistical post-editing approach introduced by Dugast[32], makes changes automatically to the output results of RBMT to reduce some manual work before the outputs are completed by human translators.

Although the RBMT system is still very active in many languages, a considerable number of researchers have joined SMT research in the last 15 years. Their number is growing. By those researchers’ efforts, SMT systems have been developed in many efficient ways, showing decent performance.

The Language Weaver SMT system and post-editing machine translation reported by Flournoy and Duran[33] produced results of product documents compared to translation results from scratch and achieved four-fold acceleration in a pilot study.

The reason for SMT’s greater popularity than RBMT systems are inferred to be those explained below.

- Linguistic knowledge obviation is an attractive point of SMT systems. Although, there are several linguistic approaches implemented based on SMT approaches, the standard framework only requires a bilingual parallel date, which only requires translation of source sentences to localize SMT.

- Low cost is another attractive point. An SMT translation system can be built from scratch using bilingual data that can be collected from story books, news websites from
the internet, open government documents, or any other existing source. Some efficient open-source toolkits such as MOSES are also available.

Many positive results underscore the potential of localization in SMT. However, it is difficult to apply to complete translation without a human reviewer. Although syntactic generation has produced good results compared to natural styles of translation by application of language models learned from numerous parallel data, it is unlikely to have all output sentences parallel to human translations semantically. It must be reviewed.

Japanese and Uyghur are similar languages in terms of word format. Most importantly, both languages have agglutinative morphology with word structures that can correspond to complete phrases of several words in the source language when translated.

In terms of morphological aspects, many recent research results have explained Finnish (an agglutinative language similar to Uyghur) with large corpora in the European Parliament (Europarl) corpus.

### 3.1 Machine Translation for Uyghur Language

Research related to machine translation in Uyghur language mainly developed in rule-based machine translation and SMT.

- Rule-based machine translation system

Rule-based approaches have been used in recent Uyghur machine translation development. Translation between Japanese, Chinese, English, and Turkish languages has been a popular spot for researchers. Japanese language researchers have solved morphological recognition problems early by building accurate morphological analysis systems during building of Japanese–English translation systems such as CHASEN and Mecab. The requirements of a morphological analysis system for Uyghur language became a harder task for a few researchers. Muhtar, Ogawa, and Toyama started with early Japanese–Uyghur rule base machine translation system and applied similarity in two languages such as word structure to produce a word-to-word translation system and provided the main suffix conversion table
for Uyghur language. During the translation of Japanese sentences to Uyghur language, Japanese sentences divided to each word and apply preloaded Japanese–Uyghur dictionary table translated Japanese words one-by-one to Uyghur words and morphemes. Modification of Uyghur language suffixes will combine stem words and suffixes to produce modified new words to replace stem words, spaces, and suffixes. As the final step, the system produces the translation result.

The Japanese–Uyghur rule-based machine translation system was mostly developed under one-way translation from Japanese to Uyghur. The absence of a practical Uyghur morphological analysis system leads to a translation system that is incapable of providing accurate words and morphemes for translation methods. To produce a functional morphological analysis system, a fully grammatical comparative study and a practical solution are necessary.

Even though the two languages have similarity in terms of their word structure, the sentence structure does not provide space between words. Moreover, word suffixes does do not apply in the Japanese language. Abdurihim provides 70,000 morphemes (including words and suffixes). A corporate author of Mecab produced an experimental prototype of the morphological analysis system.

Rule-based machine translation systems for input sentence translation into output sentences require three phases: morphological analysis; syntax transfer, which translates from the dictionary from the source into the target language; and morphological generation as shown in Fig. 4.1.

![Fig. 3.1: Rule-Based Machine Translation Approach.](image)

Rule-based machine translation systems depend strongly on dictionary and grammar that
can not generate a translation of unseen sentences without new resources. This is an extremely laborious task to collect such extremely large amounts of data. Translation approaches are always based on word-to-word translation. An example is shown in Fig. 4.2.

Rule-based machine translation systems output target text from the source text by application of three main processes as shown in Figs. 4.3 and 4.4.

- Preprocessing decomposes the source sentence to a stem word and morpheme, as we call morphological analysis. Decomposition differs among languages. Because Japanese language has no space in sentences, adding spaces between words is also a step of morphological analysis. By contrast, Uyghur language has spaces between words. Morphological analysis of Uyghur language is mainly a process of dividing words into a stem word and a morpheme.

- Intermediate processing uses a dictionary to translate decomposed stem words and morphemes to a target stem word and a morpheme.

- Post-processing combines the translated target stem word and morpheme and generates the translation result.
3.1. Machine Translation for Uyghur Language

3.1.1 Statistical machine translation system

Recent research investigating Uyghur SMT machine translation is mainly developing in Chinese, English, and Japanese languages. Although the number of translated documents from the government can be assumed to be possible to produce a high-quality parallel corpus for Chinese–Uyghur machine translation, such unpublished documents are still not available for research purposes. Many Uyghur language users have self-constructed a few Chinese–Uyghur parallel corpora by translating from daily used fundamental Chinese sentences[?][?]. Grammar similarity between English-Chinese also is applicable to English sentences which first translated to the Chinese language to build another reliable English–Uyghur parallel corpus.

The SMT system model architecture has a structure of the target language depending on a language model and translation model between words of the source and target sentences. The language model represents fluency of generated sentences; the translation model presents a translation between two languages to match input and output determined by the lexicon.
Fig. 3.4: Rule-Based Machine Translation Process Example.

Fig. 3.5: Architecture of Statistical Translation Systems.
model and alignment model. The lexicon model represents the translation of words in the source language. The alignment model forms a mapping between two languages. This algorithm provides the highest probability of the target sentences, as shown in Fig. 4.3.

A statistical machine translation system presents benefits over other approaches through a statistical method that can generate seen or unseen target sentences. Although the Japanese language has similar word structures with those of Uyghur language, several qualified basic parallel corpora still must be increased for independent research.

Consequently, we specifically examine the corpus evaluation find out how to generate a qualified translation with collected data and also the efficient way to extend corpus.

### 3.2 Workflow of SMT

As we introduced in chapter 1, SMT research started in the early 1950s. The IBM model of the SMT system applied word-based translation with no linguistic knowledge from either the source or target language[12][13]. Researchers applied morphological information[28][29] and other linguistic features such as syntactic and phrasal approaches[15][16][17][18][30][31].

Statistical machine translation[12][13] applies the Bayes theorem. Given the task of finding target string u (assumed as Uyghur), the SMT model maximizes the probability of string u given string j by a given source string j (we assume it as Japanese)

\[
P(u|j) \propto P(j|u)P(u),
\]

where translation model \(P(j|u)\) represents the probability that the source language (e.g., Japanese) is the translation of the target language (e.g., Uyghur), and the language model \(P(u)\) denotes the probability of seeing that Uyghur language language. This decomposition is attractive because it splits the problem into two subproblems. Finding the best translation \(\tilde{u}\) is done by selecting the alternative that yields the highest probability:

\[
\tilde{u} = \arg \max_{u \in u^*} P(u|j) = \arg \max_{u \in u^*} P(j|u)P(u). \tag{3.1}
\]
For a rigorous implementation of this point, one would have to conduct an exhaustive search by going through all strings $u^*$ in the native language. To explain the structure of the phrase-based machine translation system, we will describe the basic framework and main parts of SMT in the next few subsections.

Statistical machine translation is a data driven method an entirely different approach to the machine translation, this has a significant result in the translation field. The translation process generates a sentence based on statistical models to give the most probable word sequences, seen or unseen. The algorithm calculates from the given language pair from parallel corpora to generate translation and language models. The advantage is that, because there is a sufficient corpus for a given language pair, machine translation systems are theoretically developed for the new language pair. Some outstanding training parallel corpora for research in SMT include the Europarl corpus reported by Koehn.

The SMT formulated by Bayes decision using target language model $p(T)$ and translation model $p(S | T)$:

$$p(S | T) = \arg\max_t p(T) p(S | T)$$  \hspace{1cm} (3.2)

Consequently, the best translation is given by language model and translation model maximization of the probability.

The IBM model based on a machine learning method introduces the translation model and phrase-based translation model, which present very exciting developments for representing natural language. The phrase-based translation model [Och, 2002; Koehn, 2003] is used in the machine translation application industry for Google Translate and for Omniscient Technology. Since this approach was developed, translation quality has increased considerably during recent decades.

### 3.2.1 Language modeling

In this chapter, using an example, one can explain how to build a language model. Language models were proposed originally to solve speech recognition problems and are still a main subject of recent speech recognition systems and also in many fields of natural language processing. The reason for building a language model for machine translation is to learn from language corpus datasets to estimate language model probability $P_{LM}$, where LM represents
a language model, to have more natural language style and making the decision in candidate translation in the target language. As shown below, the Markov assumption is

\[ P(w_1, w_2, \ldots, w_n) = p(w_1)p(w_2|w_1) \cdots p(w_n|w_1, w_2, \ldots, w_{n-1}), \tag{3.3} \]

where \( w_1, \ldots, w_n \) denotes a sequence of \( n \) words. Also, \( p(w_n|w_1, w_2, \ldots, w_{n-1}) \) are conditional probabilities estimated using relative frequency. As Kneser and Ney introduced\[34\] it, often with a sort of smoothing model called Kneser–Ney smoothing. In an example of a very useful trigram model, in the given sentence "they are gone STOP", we always have "**" at the beginning of sentences and \( w_n \) for "STOP" to prove this is the finite set of a string. Then we would have the following.

\[ p(they\text{are}gone\text{STOP}) = q(they|\ast\ast)q(are|\ast, they)q(gone|they, are)q(STOP|are, gone) \tag{3.4} \]

As the trigram model, the assumption for each word depends on the last two words.

The language model is widely used in natural language processing; for instance, SMT, speech recognition, and spelling checkers. Phrase-based translation also makes extensive use of it.

The monolingual corpus for the target language provided a language model. The language model represents fluency of generated sentences; it represents natural language sentences from the highest probability distribution. The best language model probability is conditioned on the prior two or more words, which predict a whole seen or unseen sentence.

As calculated, the probability for how many times sentences are seen in \( N \) data is

\[ p(x_1, \ldots, x_n) = \frac{c(x_1, \ldots, x_n)}{N}, \tag{3.5} \]

where \( p(x_1, \ldots, x_n) \) is a probability distribution over a sentence \((x_1, \ldots, x_n) \in V\).

This is, however, a very poor model. It will assign probability 0 to any sentence that is not seen in the training data.
In NLP, there are many ways to define a language model; a most important practical example is the trigram model, which is a direct application of Markov models. In the trigram Language model probability distribution, the P model is

\[ p(x_1 \ldots x_n) = \prod_{i=1}^{m} q(x_i | x_{i-2}, x_{i-1}), \]  

(3.6)

where a word \( x_i \) is conditioned on the preceding two words \( x_{i-2}, x_{i-1} \). For instance, for a sentence

\[ \text{He is a STOP} \]

it would be

\[ p(\text{He is a STOP}) = q(\text{He} | *,*) \times q(\text{is} | *, \text{He}) \times q(\text{a} | \text{He, is}) \times q(\text{Stop} | \text{is, a}). \]

When \( x_i \) can be any member of the training data, it will give a very large number. In (Collins, 2013) modeling started from maximum-likelihood estimates as shown below.

\[ q(w | u, v) = \frac{c(u, v, w)}{c(u, v)} \]  

(3.7)

With probability distribution \( q \), a word conditioned on the preceding two words can be calculated by counting: the number of times that the trigram \( (u,v,w) \) is observed from the training data divided by number of times that the bigram \( (u,v) \) is observed from the training data. This model engenders two problems.

· The estimates will be zero when the numerator \( c(u,v,w) \) counts typically a very large or count in numerator being zero. This will engender systematic underestimation of many probability distributions.

· The estimate will not well define where denominator \( c(u,v) \) is equal to zero. Recall that an unreasonable number of parameters exists in the model.

Smoothed estimates of the trigram model can engender strengths and weakness of trigram language models that are extremely useful in practice. The idea will be based on counting
bigrams and unigrams to smooth the estimates depending on trigrams. Two smoothing methods are commonly used for estimation.

### 3.2.2 Discounting method

A discounting method trigram model is a very usable estimation method in the experiment. This is the easiest way to calculate the language model. It reflects aspects of the training corpus, it will systematically estimate the probability of grams seen in the corpus. The discounting method that considers discounted counts is the following.

\[
c^*(v, w) = c(v, w) - \beta \tag{3.8}
\]

The typical value \( \beta = 0.5 \) is a value between 0 and 1. The discounting method is shown below.

\[
q(w | u, v) = \begin{cases} 
\frac{c^*(u,v,w)}{\alpha(v) \times \sum_{c(u,v,w)} q(w)} & (c(u, v, w) > 0) \\
\frac{c(u, v)}{q(W)} & (c(u, v, w) = 0)
\end{cases} \tag{3.9}
\]

Define \( c(u,v,w) \) as the number of the times that trigram \((u,v,w)\) is observed in the training corpus. Also, \( c(u,v) \) is the number of times that the bigram \((u,v)\) is observed in the corpus.\(^?\)

From the evaluation of language model perplexity, one obtains the following.

\[
\prod_{i=1}^{m} p(x^i) \tag{3.10}
\]

Collins (2013) described measurements of the quality of language model as a probability for entire sentences in the training corpus. Actually, the greater the quantity \( i \) becomes the better the quality becomes.

This discounting method trigram model is used in the experiment chapter to calculate the language model probability distribution and evaluate the corpus performance for SMT translation.
3.2.3 Linear interpolation

The linear interpolation trigram model uses three estimates to define the trigram model:

\[ q(w \mid u, v) = \gamma_1 \times q(w \mid u, v) + \gamma_2 \times q(w \mid v) + \gamma_3 \times q(w). \] (3.11)

For three additional parameters of the models \( \gamma_1 \geq 0, \gamma_2 \geq 0, \gamma_3 \geq 0 \), the following are used.

\[ \gamma_1 + \gamma_2 + \gamma_3 = 1 \] (3.12)

This model will estimate the entire situation with no difficulty with the numerator or denominator.

3.2.4 Translation model

Development of the translation model in SMT system is based mostly on IBM models. The source language corresponds to the target language at word level translation by application of word alignment. The mathematical function is presented below.

\[ n = a(m) \] (3.13)

Variables \( m \) and \( n \) are the respective positions of the target language and source language that can be mapped by an alignment function. Word alignment can be found by the word alignment process that maximizes \( P(a\mid u, j) \) as presented below (assume \( u \) to Uyghur, \( j \) to Japanese).

\[ a = \arg \max \limits_a P(a\mid u, j) \] (3.14)
By applying the IBM translation model based on words[12], a word can be sought using the Expectation-Maximization (EM algorithm) procedure, which can ascertain the probability of word-based translation[35]. Translation in IBM model 1 also has taken place in our experiment to compare two other translation models. In our further explanation, we mainly follow the idea of Koehn[36], because he explained his phrase-based machine translation[15] based on IBM models. We replace the original variable in examples such as e (English) and f (French) to u (Uyghur) and j (Japanese). Furthermore, we replace the word sequence variables i and j to n and m for ease of further explanation.

As we explained, IBM Model 1 is based on word level translation. The probability of sentences $P(u|a, j)$ is estimated solely by word translation probabilities $t(u_m|j_n)$. Therein, $u_m$ corresponds to $j_n$.

The following formation describes the translated string $u$ and alignment $a$, given the source string $j$ in string length of $l$.

$$P(u, a|j) = \epsilon \sum_{l} P(u|j) \Pi_{l_u=1}^{l} t(u_m|j_{a(m)})$$  \hspace{1cm} (3.15)

To produce the sum of probability to 1, $\frac{\epsilon}{l_j + 1} \Pi_{l_u=1}^{l} t(u_m|j_{a(m)})$ must be normalized based on product $t$.

We have the following by definition.

$$P(u|j) = \Sigma_a P(u, a|j) = \epsilon \sum_{l_j} \Pi_{l_u=1}^{l} t(u_m|j_{a(m)})$$ \hspace{1cm} (3.16)

Applying two formulas above, we can calculate $P(a|u, j)$ as shown below.

$$P(a|e, f) = \frac{P(u, a|j)}{P(u|j)} = \Pi_{l_u=1}^{l} \frac{t(u_m|j_{a(m)})}{\sum_{l_j=0}^{l} t(u_m|j_{n})}$$  \hspace{1cm} (3.17)

Actually, the E-step in the EM procedure will be finished by obtaining the probability of $a$.

We can also re-estimate $t(e—f)$ based on performing M-step by defining the count function as shown below.
\[
c(u|j; u, j) = \sum_a P(u, a|j) \sum_{j=1}^{l} \delta(u, u_m) \delta(j, j_{a(m)})
\]

(3.18)

Therein, \( \delta \) is equal to 1 in conditions of \( a = b \) in \( \delta(a, b) \), or 0, which is called the Kronecker function. By re-estimating \( t(u—j) \), one obtains

\[
t(u|j; u, j) = \frac{\sum_{u,j} c(u|j; u, j)}{\sum_{u,j} \sum_{u,j} c(u|j; u, j)}.
\]

(3.19)

The EM procedure is able to determine \( c(u|j; u, j) \) and \( t(u|j) \).

Actually, IBM model 1 is an inefficient model for the estimation of probabilities for a practical SMT model. However, because the EM algorithm is useful to even higher useful IBM model, IBM model 1 is still a useful choice for applying estimation of translation quality. The IBM model applies in only one to many directions. Researchers have extended a feature and overcome this problem by allowing alignments in two directions[37].

The difference between a translation model and a language model is a corpus by which the translation model uses a parallel corpus to present a translation between two languages to match input and output. The language model uses one target language’s monolingual corpus. However, a translation model should be learned from a bilingual text for which source language must be paired with the target language.

In the late 1980s and early 1990s, IBM started using statistical approaches to translation problems. They formed a central part of modern statistical translation systems. The IBM model invented by (Brown, 1993), with a word alignment model, played an important role in SMT.

For any target sentences \( t = t_1\cdots t_m \) paired with source sentences

\[
p(s_1\cdots s_m | t_1\cdots t_l, m)
\]
and for translation between two languages, alignment variable \(a_i\), which the target word \(t_j\), is aligned to source word \(P\) under the length of source sentences words \((s_1\cdots s_m \mid t_1\cdots t_l, m)\). For IBM Model 2, conditional distribution over the target language and alignment is

\[
p(s_1\cdots s_m \mid t_1\cdots t_l, m) = \prod_{i=1}^{m} q(a_i \mid i, l, m) t(s_i \mid T_{a_i}).
\] (3.20)

\(t(s_i \mid T_{a_i})\) is a translation probability parameter of generating a target language text from the source language.

\(q(a_i \mid i, l, m)\) is an alignment probability parameter that can be interpreted as a probability of a variable.

\(a_i\) takes value \(t\), conditioned on lengths \(m\) and \(l\) of the target sentence and source sentences.

Applying IBM Model 2 with the language model and translation model translation of any target language is done as

\[
A = \arg\max_T P(T)p(S \mid T) = \arg\max_T \{0\ldots l\} \prod_{i=1}^{m} q(a_i \mid i, l, m) t(s_i \mid T_{a_i}).
\] (3.21)

Parameter estimation with fully observed data for translation probability and alignment probability parameters using maximum-likelihood estimates are as shown below.

\[
t_{ML}(s \mid t) = \frac{c(t, s)}{c(t)}
\] (3.22)

\[
q(a \mid i, l, m) = \frac{c(j \mid i, l, m)}{c(i, l, m)}
\] (3.23)

From this method to estimate parameters by simple counting from the training corpus archive, where \(c(j \mid i, l, m)\) is the number of times that a source sentence of \(l\) together with target sentence lengths of \(m\). To date we have been able to obtain a translation model by counting \(c(j \mid i, l, m)\) \(c(i, l, m), c(j, u), c(j)\).

IBM Model 1 gives the most convenient method as

\[
q(a \mid i, l, m) = \frac{1}{l + 1},
\] (3.24)
where \( l \) represents the lengths of source sentences. This engenders IBM Model 1, for any source language translation into the target language as shown below.

\[
p(s_1 \cdots s_m \mid t_1 \cdots t_l, m) = \frac{1}{(l+1)^m} \prod_{i=1}^{m} t(S_i \mid T_{ui})
\]

(3.25)

According to IBM model 1, IBM model 2, and maximum likelihood (ML) estimation for IBM model 2 and under language model \( p(J) \), we can define the translation of any Uyghur sentences \( U \) as

\[
A = \arg\max_{T} p(T) \left( \frac{1}{(l+1)^m} \prod_{i=1}^{m} t(S_i \mid T_{ui}) \right).
\]

(3.26)

For that, we define \( c(J,U) \) as the number of the times word \( J \) is aligned to word \( U \) in the training data, \( c(J) \) as the number of the times that \( J \) is aligned to any target word, \( l \) as the source sentences of length, and \( m \) as the target sentence of length. The higher the alignment probability becomes, the higher the translation probability is: they are proportional.

IBM Model 1 closely resembles the IBM Model 2 algorithm that can be estimated using the EM algorithm. In chapter 4, formula derivation is provided according to this algorithm for the calculation translation model.

### 3.3 Phrase based approach

Phrase-based SMT models translate a small number of words in sequence to produce fairly better translation results than IBM word-based SMT model to become a currently best-performing SMT system. However, phrase-based approaches still depend on alignment from a bilingual corpus.

We will give a brief review about how phrase-based approaches generate a translation rule from parallel data because we mainly applied this approach as an experiment setup.

Different from IBM models, phrase-based machine translation approaches can have many-to-many alignment between the source language and target language. One illustrative example is the famous sample in phrase-based machine translation between English and German.
(This is a small house — das ist ein kleines haus) These are divisible to the word translation table.

\begin{align*}
  (\text{the} & \rightarrow \text{der}) \\
  (\text{is} & \rightarrow \text{ist}) \\
  (\text{a} & \rightarrow \text{ein}) \\
  (\text{small} & \rightarrow \text{kleines}) \\
  (\text{house} & \rightarrow \text{haus})
\end{align*}

can also be divided to a phrase translation table.

\begin{align*}
  (\text{this is} & \rightarrow \text{das ist}) \\
  (\text{a} & \rightarrow \text{ein}) \\
  (\text{small} & \rightarrow \text{kleines ein}) \\
  (\text{house} & \rightarrow \text{haus})
\end{align*}

From the example given above, two translations have the same correct result. The phrase-based approach produced many-to-many alignment. Multi-word expression options represent significant progress over IBM models. Target sentences generated by phrases rather than each word make this a very useful method.

From the definition\cite{36}, a phrase-based lexicon is a set of small numbers of words or strings. Each group of a string or word is a tuple \((u,j,p)\):

- \(u\) is one or more Uyghur words (not restricted) in sequence;
- \(j\) is one or more Japanese words (not restricted) in sequence; and
- \(p\) is a score for the translation probability. It can be any real number.

As an example,

\begin{align*}
  (\text{man bir , 私 は,12}) \\
  (\text{sizchu , あなた は,0.9}) \\
  (\text{har halda , どうやる,-6})
\end{align*}
Phrase-based models are used directly in IBM models. Phrase-based SMT approaches allow string entries for which more than one word might be a source or target both sides. This method has much-improved translation quality. Significantly, it generates a translation for many other language pairs. The lexical entries on both the source and target side are more advanced factors than IBM models (word-to-word statistical translation models).

Phrase-based models obtained using a bilingual corpus are used to calculate the probability of relative frequency to produce good translation results that became the best-performing SMT system applied to the translation industry today. Phrase-based machine translation performs multiple alignments between the source language and target language.

Phrase-based models are the idea of a phrase-based lexicon. A lexicon might have millions of entries. These large numbers of lexicon translation examples are extracted by first IBM translation models to drive alignment. They are thereby based on those alignments. They apply these phrase-based entries.

Phrase-based models consist of three comments:

- phrase-based lexicon A phrase-based lexicon $L$ is a set of lexical entries, each $p_k$ should be a member of the set of phrases $P$; each word in $x$ is translated exactly once, where each lexical entry is a tuple $(s, t, g)$.

  - $s$ is a sequence of one or more source words.

  - $t$ is a sequence of one or more target words.

  - $g$ is a score for lexical entry. It could be any real number.

As an example, one can consider a sample trained from a phrase-based machine translation model, with German translated into English:

$$S = \frac{\text{wir müssen auch diese kritik ernst ehmen}}{1 \quad 2 \quad 3 \quad 4 \quad 5 \quad 6 \quad 7}$$
\[ T = \text{wemustalsotakethiscriticismseriously} \]

The phrase-based lexicon entry:

\((\text{wirauch, wemustalso, } -1.5)\)

\((\text{nehmen, take, } -1.8)\)

\((\text{diesekiritik, thiscriticism, } -0.2)\)

\((\text{Ernst, seriously, } -1.2)\)

same phrase-based lexicon entry with the word number and g score:

\((1, 3, \text{wemustalso, } -1.5)\)

\((7, 7, \text{take, } -1.8)\)

\((4, 5, \text{thiscriticism, } -0.2)\)

\((6, 6, \text{seriously, } -1.2)\)

Derivation is a sequence of phrases. The phrase stores can indicate how well target phrases match the underlying source sentence.

\[ y = (1, 3, \text{wemustalso}), (7, 7, \text{take}), (4, 5, \text{thiscriticism}), (6, 6, \text{seriously}) \]

so,

\[ T(y) = \text{wemustalsotakethiscriticismseriously} \]
• A trigram language model is used for the language translated: the target language, such as q(also| we, must). The trigram model is introduced in chapter 2.4. This parameter can be trained on very large quantities of target text. The trigram model is going to be invaluable for providing a prior distribution over which sentences are unlikely in the target language.

• A distortion, typically negative, will penalize things from moving too far in the translation process. Phrases should satisfy the distortion limits presented below.

\[
|t(p_k) + 1 - s(p_{k+1})| \leq d
\]  

(3.27)

\[
|1 - s(p_1)| \leq d
\]  

(3.28)

Actually, \(d \geq 0\) is a model parameter. The typical value is four. The reason is that \(d=4\) can reduce the search space of possible translation, and will make translation more efficient. However, this formed distortion value of four can improve translation performance, capable of limitation that phrases are moving too far, which translations tend to be poor. The distortion value controls the language reordering.

The phrase-based models include the following (Collins, xxxx):

\[
f(y) = h(t(y)) + \sum_{k=1}^{L} g(p_k) + \sum_{k=1}^{L-1} \gamma \times |t(p_k) + 1 - s(p_{k+1})|
\]  

(3.29)

- \(f(y)\) maximizes the sum of these three scores for the particular derivation.
- \(t(y)\) is the target-language string for derivation \(y\).
- \(h(t(y))\) is the log probability for the string \(j(y)\) under the trigram language model.
- \(g(p_k)\) is the score for the phrase \(p_k\).
– $\gamma$ is a distortion of the model, typically chosen to optimize translation performance and quality.

Alignment model with phrase-based translation model is

$$A = \arg\max_{\gamma} p(y) p(x \mid y) = \arg\max f(y), \quad (3.30)$$

wherein $\max f(y)$ represents the highest score for any state in the set. Also, $y(x)$ is exponential in size: the number of elements or the number of valid derivations grows at a high geometric rate with respect to the sentence length. $y(x)$ implies a large set of possible derivations.

This model functions according to the calculation making the translation model more efficient. Empirically, it is often shown to improve the translation performance. For many language pairs, it is preferable to disallow consecutive phrases that are a long distance from one other because this will engender poor translation.

### 3.4 Word alignment

Word alignment is the heart of machine translation. From word alignment, a translation model can be trained. In the experiment chapter, we also specifically examine word alignment during corpus evaluation. Word alignment as a defined function is

$$A = \arg\max_{\gamma} p(S, a \mid T, m). \quad (3.31)$$

For any target string $T$, source string $S$, and source sentence length $m$, this model gives conditional probability $P(S, a \mid T, m)$ [4,5].

The word alignment is proportional to the language model and translation model. The higher alignment probability is the higher translation probability. To improve word alignment, researchers have applied different approaches. Direct modeling of the target language constituents (Yamada and Knight, 2001; Zollmann, 2008) approach can lead to marked improvement in translation performance.
Another approach is reordering of the source language corpus. The advantage of this method is making the source language word structure more similar with the target language word order. This method is applicable to many language pairs such as French to English, German to English, English to Chinese, English to Hindi, and English to Urdu.

In this research, through the corpus, all aspects of evaluation will provide a method to enhance the word alignment probability. For instance, Japanese–Uyghur parallel sentence word alignment is shown in Table 1, we assume Uyghur is the target language, as follows.

Japanese :  世界を旅するのが好きですか。

Uyghur language : dunyani sayahet qilishni yahshi koremsiz?

English : do you like to travel around the world

<table>
<thead>
<tr>
<th></th>
<th>Dunya</th>
<th>ni</th>
<th>sayahet</th>
<th>qilish</th>
<th>ni</th>
<th>yahshi</th>
<th>koremsiz</th>
<th>?</th>
</tr>
</thead>
<tbody>
<tr>
<td>世界</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>を</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>旅</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>する</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>のが</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>好き</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>です</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>か</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 3.1: Japanese–Uyghur Parallel Sentence Word Alignment

### 3.5 Corpus Evaluation

There is always improvement in the development of techniques from evaluation. Through evaluation, the problems of approaches can be identified. Solutions can be devised for improved methods.

Since the invention of the first electronic computer, researchers have been talking about machine translation projects on computers. It has been twenty years now since the field awoke to the importance of evaluation (Gaizauskas, 1998). "What corpus should be used?" "Is it necessary to build a new one?" is often explored for NLP research or application-building
fields. These require a framework for corpus evaluation.

Statistical machine translation models rely on a translation model and language models able to represent natural language sentences from the highest probability distribution. Translation models fit with parallel corpora that can be learned from bilingual corpora that source language must be paired with the target language. The language model fits with the fundamental importance of monolingual corpora which for the target language provided a language model.

As the figure shows, these translation and language models directly rely on monolingual corpora and bilingual corpora. Consequently, machine translation research is highly dependent on robust evaluation.

Corpus evaluation for machine translation not only for SMT but also rule-based machine translation approaches, like any other task in the natural language processing, depend strongly on strict assessment. As introduced before, to produce good translation performance, training corpus data are always required. Aligned sentences can make corpora that are translations of one another.
Evaluating the corpus to ascertain methods to improve word alignment and obtain better results is necessary for "good" translation. However, "good" requires some context for comparison. The input of the sentences might have several translation models, all of which might be correct. By evaluating the JECU-basic-sentence parallel corpora, we investigated that a Japanese sentence by a human translator has two translation models, both of which are correct in the Uyghur language.

The same data applied for the whole system makes corpus evaluation developed slowly because of the merits of approaches based on the introduced data method. The progress in the corpus evaluation yielded the same result in two decades, showing which better input will produce good output. Corpus validation was not evaluated. The field has no answer for how to evaluate a corpus. From the 1990s, this problem was regarded as caused by inadequate corpora. However, now there are groups of training corpora we can get from the web site.

The demand for corpora has increased recently. The web provides huge boundless supplies of data for many languages. Therefore, building corpora has become easier. Still, corpora quality might vary and even differ according to the task. Therefore, a corpus evaluation method is necessary. Designing an evaluation method is relevant to a broad range. Some criteria for evaluation are expected to be useful.

Many approaches can be used to evaluate a corpus. However, setting a standard for evaluation is a difficult task because of the flexibility of natural language, for which different systems should be compared and estimated. The corpus evaluation techniques are the following.

3.5.1 Evaluation techniques

Evaluation parameters are calculated by humans with a computer standard.
1. Human evaluation

There are various machine translation systems. They need different parallel data to train the system. The most usual and accurate way to evaluate a corpus is human evaluation. Through manual evaluation, we can set a "standard" to a translation system; which can learn from the human evaluators what is "correct". Although this approach presents way too much time and difficulty for human evaluators to estimate a large parallel translated
corpus, collecting accurate data is necessary.

When evaluating the performance of a processing resource, usually humans provide a “gold standard” against which comparison can be made. Because different people have different ideas of what is correct, it is not easy to determine a gold standard. Therefore, to solve this difficulty, more than two evaluators must assess their resources. Metrics for evaluation in information extraction such as IAA are based on human arguments. When two people cannot agree on one, IAA can give same annotation “correctly”.

Automatic evaluation methods are extensive implementations of human labor. Research has demonstrated that automatic evaluation methods produce reasonable results against human evaluators. Therefore, this automatic evaluation score is applicable for the evaluation of large datasets.

A way to increase human evaluation reliability is computing inter-annotator agreement measures on a manually annotated corpus. Applications have been directly affected by the quality of manual annotations. Because there are mistakes in the manual annotation, machine learning tools will follow human mistakes [Reidsma and Carletta, 2008; Schluter, 2011]. Therefore, the “gold standard” of corpora is observed by the indication of errors in a manual evaluation.

For this research, we evaluate a corpus manually by translation of four cross languages using the proposed evaluation model in chapter 3. There is no inter-annotator agreement.

For assessing system or application performance automatically against a trusted “gold standard”, MUC evaluation metrics of precision, recall, and F-measure are used. Next, a brief introduction of each will be given. The corpus is evaluated by the researchers using these measurement candidates [gate].

2. Precision

The number of documents retrieved correctly divided by the total number of documents retrieved gives the precision. It is the percentage of the number of items correctly identified. It measures the actual correct identified number in the system. Precision is the number of words from the candidate that are found in the correct translation divided by total number
of words in the candidate. To compute the precision, there is at least one manual translation for each training sentence.

\[
\text{Precision} = \frac{\text{correct} + \frac{1}{2}\text{partial}}{\text{correct} + \text{spurious} + \text{partial}}
\]  

(3.32)

3. Recall

Recall measures the missing correct number of items in the system. Recall is also consistent with human evaluators: at least one human translation example is needed for each testing sentence. The recall learned from relevant items is corrected, with true positives divided by the total number of relevant documents the system should have produced. Recall candidates provide a higher rate. The system is not missing correct items.

\[
\text{Recall} = \frac{\text{correct} + \frac{1}{2}\text{partial}}{\text{correct} + \text{missing} + \text{partial}}
\]  

(3.33)

4. F-measure

This is a combination of recall and precision. The F-measure is a variant of accuracy that is unaffected by negative results. It is a harmonic mean of precision and recall.

\[
F - \text{measure} = \frac{(\beta^2 + 1) p * R}{R}
\]  

(3.34)

In a popular setting \( \beta = 1 \). It is the relative importance recall and precision. The advantage of the F-measure is that it heavily penalizes small values of precision and recall. A geometric interpretation that does not include any false positive or false negative.

· Judging: under the MUC metrics calculation to compare one corpus to another, human judgment is included. As in comparable precision, recall and F-measure exercises, humans will judge all candidates. For creation of a publication-quality collocations dictionary in (52), one can ask evaluators such as native speakers, linguistic students, and professional lexicographers, all of whom had worked on the second edition of the Oxford Collocations Dictionary to judge and discussed, after the evaluation metric calculation.
3.5.2 Effect of parallel corpora

- Quantity: One earlier study (Seljan, 2010) used the CorAl sentence alignment tool to evaluate parallel corpora. The result showed that larger and annotated corpora are always reliable and better [sanja,3483]. Most researchers agree that more is better.

- Quality: To classify parallel corpora as being of high quality or poor quality, an earlier report (Yildiz, 2014) describes development of a machine-learning-based classifier, showing that obtaining better translation performance requires a high-quality corpus. There is improvement in the blue score using only some training data. The use of more data leads to higher SMT accuracy.

Name entity is an effect factor in the corpus. The evaluations in (52) names are excluded. Names in Japanese and Uyghur or all kinds of languages are difficult to classify. The name is a classification task. The name in the corpus should be for male or female. Without a doubt for a task at sub-sentence level, such as word alignment, online dictionary, semantic networks better and more annotated corpora are needed [Sanja].

3.6 The JU Parallel Corpus

Machine translation (MT) perform translation of one natural language (source language) into another language (target language) automatically, from the Middle of the twentieth century; it became the most interesting topic among those of the natural language processing (NLP) field. Recently, language pairs are playing an important role in SMT.

In any machine translation approaches to generate a target language regarded as a statistical problem. MT relies on large parallel corpora, with algorithms to learn word level dictionaries and phrase-based patterns and to translate unseen sentences. The resulting machine translation system for better performance of the translation quality theoretically shows that the most parallel corpora is the better generate. However, large parallel corpora are available for few languages independently. The remainder of the languages are regarded as low source languages, including the Uyghur language.

In the research world of natural language processing, Uyghur Language is counted as a
low source language. Uyghur language is an Altaic language spoken by 11 million people in China. An increase of language resources on the internet, translation capabilities from other languages are becoming requested from many Uyghur language users. A machine translation program became an interest of Uyghur language researchers. Development of Natural Language processing in Uyghur language started with a few researchers by application of the process of the Japanese language. Similarity in word order gained few advantages for Uyghur language. Research about machine translation in Uyghur language mainly developed in rule-based machine translation and SMT.

Building a reliable corpus for SMT systems presents many difficulties. The time and human costs always present difficulties for collecting practical and accurate parallel data. The internet provides huge amounts of language data. Collected language data mostly include non-parallel translations. To reduce the difficulties described above, machine learning techniques became popular. The process includes collecting resources of parallel translated language sentences and correcting them grammatically by application of machine learning. Applying machine learning technique has reduced large amounts of human and time costs, the quality of regrouped or filtered parallel corpora still presents logical errors in distinguishing lexical aspects between sentences. Retranslation has always been a necessary step with random resources.

An open-source machine translation project with few resources is always restricted in the problem of quality in a parallel corpus. Voluntarily translated parallel corpora always have qualification problems. Differences of understanding from translators also cause problems related to different translations for the exact same source sentences. Preparing a basic practical multilingual parallel corpus with better quality control and translation rules to produce a target language helps a project to have a good start.

Japanese language has similar word structure to that used by Uyghur language, but several qualified basic parallel corpora still must be increased for independent research.

3.6.1 Evaluation of SMT systems

Many approaches can be used for evaluation. Papineni introduced the first automatic measurement in 2001 and widely applied it for evaluation of machine translation[40]. Bilingual
Evaluation Understudy (BLEU) calculates the level of similarity between candidate translation from MT, and also presents several reference translations based on precision of a specific n-gram model. The following formula define the BLEU score:

\[
\text{BLEU}_n = \text{brevity} - \text{penalty} \exp \sum_{i=1}^{n} \lambda_i \log \text{precision}_i \tag{3.35}
\]

\[
\text{brevity} - \text{penalty} = \min 1, \frac{\text{output} - \text{length}}{\text{reference} - \text{length}}. \tag{3.36}
\]
第4章 EVALUATION OF PARALLEL CORPUS

Used in this chapter, 5304 × 4 translated parallel corpora (from Kurohashi & Kawahara lab) included English, Chinese, and Japanese language. Each language was translated independently into Uyghur Language, with results compared to the lexical base.

More importantly, the purpose is building more practical and efficient parallel corpora for use with a Japanese–Uyghur machine translation system. The process has been translated and evaluated fully manually.

4.1 Experiment Objective

For corpus evaluation by humans, the main goal is to obtain a high quality parallel corpus. The corpus quality is a key to building a better SMT translation system. In this work, a corpus evaluation model proposed and parallel corpora in three languages were translated to Uyghur language one-by-one. Then they were manually evaluated. Results of comparison between parallel corpora in different grammatical languages and similar structure languages are discussed.

Translated parallel corpora always have qualification difficulty. Differences of understanding from translators also cause difficulties related to different translation for the exact same source sentences. Preparing a basic practical multilingual parallel corpus with better quality control and translation rules to produce the target language helps the project get a good start.

Japanese and Uyghur languages have many word structure similarities. Nevertheless, they also have differences of nouns and verbs. Uyghur language has different forms according to the person, number, and tense of the subject. Japanese has no such conjugation: verbs are independent of the person, number, and subject.
Through human evaluation of parallel corpora to find errors by provided corpus evaluation, one can provide a standard corpus for Japanese–Uyghur and another cross language to achieve better translation performance using a qualified, trained corpus.

### 4.2 Data Preparation

To study Uyghur language with cross languages, JEC-basic-sentence-v1-2 (5304 × 4) was translated manually into Uyghur language. The completed corpus was designated as JECU-basic-sentence (5304 × 4). The translation was done by a single person.

We performed corpus evaluation using the JECU (5304 × 4) basic sentences dataset and based on Japanese–Uyghur, English–Uyghur, and Chinese–Uyghur sentence alignment. Inter-annotator agreement calculation was not required. This provided us with a gold standard for evaluating the sentence alignment procedure.

Uyghur language has 32 letters consisting of 8 vowels and 24 consonants. The letters in Uyghur language have letters of two types: Uyghur Arabic Yeziq (UAY) and Uyghur Latin Yeziq (ULY), as shown in Figure 4.1.

![UAY and ULY](image_url)

**Fig. 4.1: UAY and ULY.**
In the corpus, preparation steps for the Uyghur language are written in Uyghur Arabic Yeziq (ULY). Our Linux system was not capable of accommodating fonts of two types, Uyghur Arabic Yeziq (UAY) is currently unavailable for Linux systems.

This corpus and evaluation method used cross translation resources for low source languages. In addition, the JECU-basic-sentence can be extended to other languages or corpus sizes. Therefore, Uyghur Latin Yeziq (ULY) was chosen for the data input.

4.3 Proposed Model

We chose a random sentence from JECU to run manually in the Japanese–Uyghur, English–Uyghur, and Chinese–Uyghur translation systems; then we compared target languages to ascertain whether the same results can be given as shown in Table Figure 4.2.1. In this model, we evaluate the same STM system with a different number of a parallel corpora.

![Evaluation Model of Statistical Translation Systems](image)

Fig. 4.2: Evaluation Model of Statistical Translation Systems.

Modeling details above are summarized as explained below:
• English to Uyghur.
  Two base English sentences were selected as the base source language and were translated to Uyghur sentences.

• English to Chinese.
  Two Chinese sentences were translated from base English sentences, with one well translated and one translated mistakenly both grammatically and logically.

• English to Japanese & Chinese to Japanese.
  Four Japanese sentences were selected from two base English sentences. Two Chinese sentences were translated from the two base English sentences. Two were well translated into four languages. The other two had logical and grammatical translation problems. These four Japanese sentences were compared. Problems of correct and incorrect translations are discussed later.

• Japanese to Uyghur
  Four Japanese sentences were translated to four other Uyghur sentences and were compared to the first two Uyghur sentence from method one. The results are discussed later.

4.4 Evaluation Method

Total 5304 × 4 sentences were selected for this experiment. The evaluation will be divided into three methods.

• Method 1

  Using method 1, we evaluate Uyghur translation from English and Chinese sentences. Two translated sentences were compared.
4.4. Evaluation Method

For method 2, we evaluate Uyghur translation from English and Japanese with translation from Chinese and Japanese individually. Then we compare the two evaluation results for assimilation translation.
### Table 4.3: Comparison of Two Translated Uyghur Sentences

<table>
<thead>
<tr>
<th>Uyghur Sentence E1</th>
<th>Uyghur Sentence C1</th>
</tr>
</thead>
<tbody>
<tr>
<td>mushuk chushluk uhlidi.</td>
<td>mushuk chushluk uhlidi.</td>
</tr>
<tr>
<td>rehmitimni ipadilehske sozlirim kemchillik qilidu.</td>
<td>Asanhqche rehmet sozi diyelmeslik.</td>
</tr>
</tbody>
</table>

### Table 4.4: English to Japanese

<table>
<thead>
<tr>
<th>English Sentence</th>
<th>Japanese Sentence</th>
</tr>
</thead>
<tbody>
<tr>
<td>The cat took a nap.</td>
<td>猫が昼寝をした.</td>
</tr>
<tr>
<td>Words can not express my gratitude.</td>
<td>なかなか感謝の言葉が出ない</td>
</tr>
</tbody>
</table>

### Table 4.5: Chinese to Japanese

<table>
<thead>
<tr>
<th>Chinese Sentence</th>
<th>Japanese Sentence</th>
</tr>
</thead>
<tbody>
<tr>
<td>mao shui le wu jiao.</td>
<td>猫が昼寝をした.</td>
</tr>
<tr>
<td>bu tai rong yi shuo chu gan xie de hua.</td>
<td>なかなか感謝の言葉が出ない</td>
</tr>
</tbody>
</table>

### Table 4.6: Comparison of Two Japanese Sentences

<table>
<thead>
<tr>
<th>Japaneses Sentence E</th>
<th>Japanese Sentence C</th>
</tr>
</thead>
<tbody>
<tr>
<td>猫が昼寝をした.</td>
<td>猫が昼寝をした.</td>
</tr>
<tr>
<td>なかなか感謝の言葉が出ない.</td>
<td>なかなか感謝の言葉が出ない</td>
</tr>
</tbody>
</table>

### Table 4.7: Translated Japanese to Uyghur

<table>
<thead>
<tr>
<th>Japaneses Sentence E</th>
<th>Uyghur Sentence 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>猫が昼寝をした.</td>
<td>mushuk chushluk uhlidi.</td>
</tr>
<tr>
<td>なかなか感謝の言葉が出ない.</td>
<td>rehmitimni ipadilehske sozlirim kemchillik qilidu.</td>
</tr>
</tbody>
</table>

### Table 4.8: Translated Japanese to Uyghur

<table>
<thead>
<tr>
<th>Japanese Sentence C</th>
<th>Uyghur Sentence 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>猫が昼寝をした.</td>
<td>mushuk chushluk uhlidi.</td>
</tr>
<tr>
<td>なかなか感謝の言葉が出ない.</td>
<td>rehmitimni ipadilehske sozlirim kemchillik qilidu.</td>
</tr>
</tbody>
</table>
4.5 Evaluation Results and Issues

Through comparison by evaluation method and in translation process among four languages using the proposed evaluation model, there are results of five types in JECU-basic-sentence alignment corpus evaluation.

- The quality of both corpora includes misaligned pairs or poor translation quality pairs.

<table>
<thead>
<tr>
<th>J</th>
<th>三年生が卒業証書づくりをした</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>E</td>
<td>The third year students receive diplomas.</td>
<td>x</td>
</tr>
<tr>
<td>C</td>
<td>san nian ji xue sheng de dao le bi ye zheng shu.</td>
<td>x</td>
</tr>
<tr>
<td>U</td>
<td>3ji yillik okuhuchiliri okush putturush kinish kisini aldi.</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 4.1: #Example Sentence No. 58

<table>
<thead>
<tr>
<th>J</th>
<th>彼がそれを key word にする</th>
<th>x</th>
</tr>
</thead>
<tbody>
<tr>
<td>E</td>
<td>He made that a key word.</td>
<td>0</td>
</tr>
<tr>
<td>C</td>
<td>ta na zhe ge zuo guan jian ci.</td>
<td>0</td>
</tr>
<tr>
<td>U</td>
<td>u buni asasliq soz kildi.</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 4.2: #Example Sentence No. 148

- Sentences include a disambiguation component, which is a word with more than two meanings.
It was investigated by the press.

Use that as the base.

It’s all in one ear and out the other

He’s getting ready for his university entrance exams.

Different translators have different expressions.

Many grammatical similarities were observed between Japanese and Uyghur in the evaluation. The ”be verb” produced positive results for generating Uyghur from a Japanese text. However, negative results were obtained for conjunctions of Japanese translations, which have several meanings in the Uyghur language.
• "Do verb" has several translations in the Uyghur dictionary. For that reason, translation results became incorrect. Comparison between the target language with two other languages (English and Chinese) demonstrates that the method does not always yield the same result. Comparison of texts, even if using approximate matching, is bound to yield some incorrect results.

| J | その言葉を誰もが一度は耳にした | 0 |
| E | Everyone’s heard that once. | 0 |
| C | na ju hua shui dou zhi shao ting guo yi ci | 0 |
| U | gepni kopchilik az digende bir qetim bolsa quliquqgha qilghan. | x |

Table 4.8: #Example Sentence No. 39

| J | 私はその質問を多くの方から受けます。 | 0 |
| E | Many people ask me the same question. | 0 |
| C | wo shou dao le hen duo ren de wen ti. | 0 |
| U | Man bu sualni kopligen ademledin anglidim. | x |

Table 4.9: #Example Sentence No. 719

| J | Xが評価を受けた。 | 0 |
| E | X was judged. | 0 |
| C | X shou dao le ping jia. | 0 |
| U | X bolsa bahalandi./ X bolsa bahalashni qobul qildi. | x |

Table 4.10: #Example Sentence No. 722
4.6 Error analysis

We estimate that translation mistakes should be corrected by more than one native speaker or linguistic professional. For sentences, disambiguation should be annotated. More effort by more professionals can increase the sentence translation consistency.

We estimate that the performance of the SMT system can be improved by clarification of all of these mistakes.

For providing corpus evaluation results in all aspects with increased performance of SMT method, we will use model calculation and mathematical analyses.
5.1 Proposed Method

5.1.1 ”Organize the corpus” method

The very small number of available parallel corpora obviates the use of machine translation systems. To resolve this difficulty for the low source language, we propose the ”organize the corpus” method by reordering the parallel corpus according to ”be verb” and ”do verb” items. This method produces an efficient parallel corpus to achieve higher language model probability than an unorganized parallel corpus and also produces an efficient way to extend corpora.

Better translation performance is achieved with the good probability under a trigram language model. Under a good translation model, the word alignment is proportional to those of the language model and translation model according to the formula. The higher alignment probability is the higher translation probability.

\[
A = \arg \max f(y) = \arg \max p(y) p(x \mid y)
\]

To improve the word alignment, we provide the ”organize the corpus” method to enhance the word alignment probability by increasing the language model. The higher the language probability becomes, the more fluently the target language can be used.
The collected parallel corpus is evaluated by calculating the translation model, phrase-based model, and language model. Comparison between an organized corpus and an unorganized corpus is described in the experiment section. The specific way to produce efficient parallel corpus for cross translation SMT system is explained below.

- The parallel corpus organized by English "be" and "do" verbs. The Japanese and Uyghur grammars have two differences of verb stems in a specific order. First, a Japanese verb makes "katsuyo"; Uyghur verbs do not. Secondly, Japanese verb forms are independent of the categories of person, number, and tense; the Uyghur verb forms have different word structures according to the person, number, and tense just as English verbs have. Therefore, the JECU-basic sentence alignment is divided according to English "be" and "do" verbs.

- The "organize the corpus" method is applied. The JECU-basic-sentences (5304 × 4) collected a training parallel corpus generally organized into three types. In this work, with Uyghur as a target language, the language model is estimated. We organize the corpus JECU-basic sentences by division into "be verb" and "do verb" and "mixed verb" parallel corpora, based on Japanese–Uyghur, English–Uyghur, and Chinese–Uyghur sentence alignment. There are 1178 × 4 "be" verb sentences, 2723 × 4 "do" verb sentences, and 1404 × 4 verb mix sentences in the training parallel corpora.

- A "mixed verb" corpus is organized designated as "organized mixed verb" corpus. These two corpora are compared.

The "mixed verb" sentences’ language model estimation was very low, with almost identical probability to that of the unorganized corpus. Therefore, we apply the organized method to the "mixed verb" sentences, with one complex sentence divided into sample sentences. A comparison experiment is conducted with "organized mixed verb" sentences and "unorganized mixed verb" sentences.

As the method used for organization, we put all "be" verb sentences together based on the JECU-basic sentence alignment.

Building a corpus with organization by "be verb" or "do verb" not only makes the collected resources more efficient; it is also an efficient means of extending the corpus as proposed.
Training the language model using the "organize the corpus" method, in contrast to the derivate translation model calculation formula and phrase-based translation model, is a direct method to obtain better target language results. For the phrase-based translation model, the higher the language score that is obtained, the more fluent the target language performance becomes.

### 5.1.2 Formula derivation

Formula derivation is used for calculation of the translation model. In chapter 2, the translation model was introduced. According to IBM definition models, we propose a formula for the calculation.

1. Formula derivation

First, based on IBM Model 2, we have the following.

\[
p(j_1 \cdots j_m \mid u_1 \cdots u_l, m) = \prod_{i=1}^{m} q(a_i \mid i, l, m) t(j_i \mid u_{a_i})
\]

\[
A = \arg \max_a p(U) p(J \mid U) = \arg \max_j \in \{0, \ldots, l\} \prod_{i=1}^{m} q(a_i \mid i, l, m) t(j_i \mid u_{a_i})
\]

The meaning of the formula is described in chapter two.

Secondly, the following is obtained from IBM Model 1.

\[
p(j_1 \cdots j_m \mid u_1 \cdots u_l, m) = \prod_{i=1}^{m} \frac{1}{(l + 1)^m} \times t(J_i \mid U_{a_i})
\]

\[
A = \arg \max_a p(U) p(J \mid U) = \arg \max_j \in \{0, \ldots, l\} \frac{1}{(l + 1)^m} \prod_{i=1}^{m} t(J_i \mid U_{a_i})
\]

The maximum-likelihood estimate is shown below.
$$t_{ML}(J \mid UU) = \frac{c(u,j)}{c(u)}$$

Derivation:

$$p(j_1 \cdots j_m \mid u_1 \cdots u_l, m) = \prod_{i=1}^{m} \frac{1}{l+1} \times t(J_i \mid U_{ai})$$

$$= \frac{1}{(l+1)^m} \prod_{i=1}^{m} t(J_i \mid U_{ai})$$

$$= \frac{1}{(l+1)^m} \prod_{i=1}^{m} \frac{c(u,j)}{c(u)}$$

$$A = \arg\max_u p(U) p(J \mid U) = \arg\max_{j \in \{0, \ldots, l\}} \frac{1}{(l+1)^m} \prod_{i=1}^{m} t(J_i \mid U_{ai})$$

$$= \arg\max_{j \in \{0, \ldots, l\}} \frac{1}{(l+1)^m} \prod_{i=1}^{m} \frac{c(u,j)}{c(u)}$$

Therein, the source language is Japanese; the target language is the Uyghur language.

\(c(u,j)\) represents the number of times that the Uyghur word is aligned to the Japanese word.

\(c(u)\) represents the number of times that the Uyghur word is aligned to any Japanese word in the training data.

\(m\) is a Japanese sentence of length, where word i in the Japanese sentence is aligned to word j in the Uyghur sentence.

### 5.2 Tools Employed

For translation model calculation, two pair alignment sentences are selected for Japanese and Uyghur. The translation model is calculated using one-to-one word alignment and many-to-one alignment situation.
5.3. Experiment

For the phrase-based machine translation model two parallel sentences were selected in Japanese and Uyghur.

In this experiment, a 5101 × 2 parallel corpus is chosen for language model estimation for the target language.

For the language model, translation model, and phrase-based translation model calculation, parallel corpus translation was done by a single person.

We assume the Uyghur language translated from Japanese, for which the source language is Japanese and the target language is the Uyghur language.

5.3 Experiment

The experiment was conducted with three model calculation, including translation model, phrase-based translation model, and language model estimation.

5.3.1 Translation model calculation

• one-to-one alignment

For all translation model calculations, we assume that the word alignment is annotated by humans. The Uyghur text is generated from the Japanese text, where l=6 and m=6. The following alignment is specified in the table. The training parallel sentence is from #5279.
Table 5.1: Japanese-to-Uyghur Word Alignment

<table>
<thead>
<tr>
<th>X</th>
<th>X</th>
</tr>
</thead>
<tbody>
<tr>
<td>が</td>
<td>ni</td>
</tr>
<tr>
<td>想像 1</td>
<td>oylimaq</td>
</tr>
<tr>
<td>に</td>
<td>bolsa</td>
</tr>
<tr>
<td>難</td>
<td>qiyin</td>
</tr>
<tr>
<td>くない 1</td>
<td>emes</td>
</tr>
</tbody>
</table>

\[ S = X \text{が想像に難くない} \]

\[ T = X \text{ni oylimaq qiyin emes}. \]

\[
p(j_1 \cdots j_m | u_1 \cdots u_l, m) = \frac{1}{(l + 1)^m} \prod_{i=1}^{m} \frac{c(u, j)}{c(u)}
\]

\[
= t(X|X)t(が | ni)t(想像 | oylimaq)t(に | bolsa)t(難 | qiyin)t(くない | emes)
\]

\[
= \frac{1}{(6 + 1)^6} \left[ \frac{1}{6} \times \frac{1}{6} \times \frac{1}{6} \times \frac{1}{6} \times \frac{1}{6} \times \frac{1}{6} \right]
\]

\[
= \frac{1}{7^6} \times \frac{1}{6^6}
\]

\[
\cdot \text{many-to-one alignment.}
\]

\[ S = \text{彼が自分で椅子の高さを調整します。} \]

\[ T = u \text{ozi orunduqning igizlikini tengshidi.} \]

The training parallel sentence from #5167, with Japanese translated into Uyghur language, where \( m = 10 \) and \( l = 9 \). The following alignment is specified in the table below.

\[
p(j_1 \cdots j_m | u_1 \cdots u_l, m) = \frac{1}{(l + 1)^m} \prod_{i=1}^{m} \frac{c(u, j)}{c(u)}
\]

\[
= t(彼 | U)t(が | bolsa)t(自分 | ozi)t(で | ozi)t(椅子 | orunduq)t(の | ning)
\]
5.3. Experiment

<table>
<thead>
<tr>
<th>彼</th>
<th>U</th>
</tr>
</thead>
<tbody>
<tr>
<td>が</td>
<td>bolsa</td>
</tr>
<tr>
<td>自分</td>
<td>ozi</td>
</tr>
<tr>
<td>で</td>
<td>ozi</td>
</tr>
<tr>
<td>椅子</td>
<td>orunduq</td>
</tr>
<tr>
<td>の</td>
<td>ning</td>
</tr>
<tr>
<td>高さ</td>
<td>igizliki</td>
</tr>
<tr>
<td>を</td>
<td>ni</td>
</tr>
<tr>
<td>調整</td>
<td>tengshi</td>
</tr>
<tr>
<td>します</td>
<td>di</td>
</tr>
</tbody>
</table>

Table 5.2: Japanese-to-Uyghur Word Alignment

\[
t(\text{高さ}) | \text{igizliki})t(\text{を}) | \text{ni})t(\text{調整}) | \text{tengshi})t(\text{します}) | \text{di})
\]

\[
= \frac{1}{(9 + 1)^{10}} \left( \frac{1}{10} \times \frac{1}{10} \times \frac{1}{10} \times \frac{1}{10} \times \frac{1}{10} \times \frac{1}{10} \times \frac{1}{10} \times \frac{1}{10} \times \frac{1}{10} \right)
\]

\[
= \frac{1}{10^{10}} \times \frac{1}{10^{10}}
\]

5.3.2 Phrase-based translation model

The phrase-based translation model is introduced in chapter 2. In this section, we use Japanese as the source language, and the Uyghur language as the target language: \(d= 4, \gamma = -1\). The calculations are presented below.

1. \# 5123 sentence selected.

\[
S = \frac{\text{魚} が 元気に 水の中を 泳ぐ}{1 \ 2 \ 3 \ 4 \ 5 \ 6}
\]

\[
T = \text{beliq bolsa suda kogulluk uzup oynawatidu}.
\]

The condition of the phrase-based translation model is satisfied, as shown below.

\cdot \(p_4\) is a member of the set of phrases \(P\).

\cdot Each word is translated exactly once.

\[
|t(p_k) + 1 - s(p_{k+1})| \leq d \& |1 - s(p_i)| \leq d
\]
\[ |2 + 1 - 4| = 1 \leq d \]
\[ |5 + 1 - 3| = 3 \leq d \]
\[ |3 + 1 - 6| = 2 \leq d \]
\[ |1 - 1| = 0 \leq d \]

It is expressed completely as the following.

\[
f(y) = h(t(y)) + \sum_{k=1}^{L} g(p_k) + \sum_{k=1}^{L-1} \gamma \times |t(p_k) + 1 - s(p_{k+1})|
\]

\[
= \log q(\text{beli}q|\ast, \ast) + \log q(\text{bolsa}|\text{beli}q, \ast) + \log q(\text{suda}|\text{beli}q, \text{bolsa}) + \log q(\text{kogulluk}|\text{bolsa}, \text{suda}) + \log q(\text{uzup}|\text{suda}, \text{kogulluk}) + \log q(\text{oynawatidu}|\text{kogulluk}, \text{uzup}) + g(1, 2, \text{beli}q\text{bol}sa) + g(4, 5, \text{suda}) + g(3, 3, \text{kogulluk}) + g(5, 6, \text{uzup}oynawatidu)
\]

\[
+ |1 - 1| \times \gamma + |2 + 1 - 4| \times \gamma + |5 + 1 - 3| \times \gamma + |3 + 1 - 6| \times \gamma
\]

\[
= 6 \times \gamma
\]

\[
= -6
\]

2. Same training sentence with different phrase order compared to the preceding sentence.

\[ Y = (1, 2, \text{beli}q\text{bol}sa), (3, 3, \text{kogulluk}), (4, 5, \text{suda}), (6, 6, \text{uzup}oynawatidu) \]

Condition of phrase-based translation model is satisfied, as shown below.

\cdot \text{p}_4 \text{ is a member of the set of phrases P.}
\cdot \text{Each word is translated exactly once.}

\[
|t(p_k) + 1 - s(p_{k+1})| \leq d \text{ & } |1 - s(p_1)| \leq d
\]

\[
|2 + 1 - 3| = 0 \leq d
\]
\[ |3 + 1 - 4| = 0 \leq d \]
\[ |5 + 1 - 6| = 0 \leq d \]
\[ |1 - 1| = 0 \leq d \]

It is expressed completely as the following.

\[
f(y) = h(t(y)) + \sum_{k=1}^{L} g(p_k) + \sum_{k=1}^{L-1} \gamma \times |t(p_k) + 1 - s(p_{k+1})
\]

\[
= \log q(\text{be}\text{-}\text{verb}|*,*) + \log q(\text{bolsa}\text{-}\text{be}\text{liq}, \text{bolsa}) + \log q(\text{kogulluk}\text{-}\text{bolsa}, \text{suda})
+ \log q(\text{uzup}\text{-}\text{suda}, \text{kogulluk}) + \log q(\text{oynawatidu}\text{-}\text{kogulluk}, \text{uzup})
\]

\[
+ g(1,2, \text{be}\text{liq}\text{-}\text{bolsa}) + g(3,3, \text{kogulluk})) + g(4,5, \text{suda}) + g(5,6, \text{uzup}\text{-}\text{oynawatidu})
\]

\[
+ |1 - 1| \times \gamma + |2 + 1 - 3| \times \gamma + |3 + 1 - 4| \times \gamma + |5 + 1 - 6| \times \gamma
\]

\[
= 0 \times \gamma
\]

\[
= 0
\]

### 5.4 Language Model Estimation

For the organized JECU-basic sentences, 1178 \( \times \) 4 "be verb" corpus, 2723 \( \times \) 4 "do verb" corpus, and 1404 \( \times \) 4 "mixed verb" corpus, the language model is estimated and compared with same amounts of unorganized corpora. The estimation applies by the IBM trigram model \( q(w \mid u, v) = \frac{c^t(u,v)}{c(u)} \) \( c(u,v) \) greater than zero.

#### 5.4.1 “be verb” organized corpus

First, 1178 "be verb" and random 1178 unorganized sentences were selected. For all the bigrams listed \( (u,v), u=X, c(u,v) \) is greater than zero, with probability estimation based on the discount count. For the unigram "bolsa, is", the bigrams are listed by six groups in the table.
<table>
<thead>
<tr>
<th>$x$</th>
<th>$c(x)$</th>
<th>$c^*$</th>
<th>$P$</th>
</tr>
</thead>
<tbody>
<tr>
<td>bolsa</td>
<td>49</td>
<td></td>
<td></td>
</tr>
<tr>
<td>u bolsa</td>
<td>12</td>
<td>11.5</td>
<td>0.235</td>
</tr>
<tr>
<td>awu bolsa</td>
<td>3</td>
<td>2.5</td>
<td>0.051</td>
</tr>
<tr>
<td>kopchilik bolsa</td>
<td>1</td>
<td>0.5</td>
<td>0.010</td>
</tr>
<tr>
<td>X bolsa</td>
<td>4</td>
<td>3.5</td>
<td>0.071</td>
</tr>
<tr>
<td>bar bolsa</td>
<td>10</td>
<td>9.5</td>
<td>0.193</td>
</tr>
</tbody>
</table>

Table 5.1: 100 "be verb" Organized Corpus

<table>
<thead>
<tr>
<th>$x$</th>
<th>$c(x)$</th>
<th>$c^*$</th>
<th>$P$</th>
</tr>
</thead>
<tbody>
<tr>
<td>bolsa</td>
<td>27</td>
<td></td>
<td></td>
</tr>
<tr>
<td>u bolsa</td>
<td>2</td>
<td>1.5</td>
<td>0.056</td>
</tr>
<tr>
<td>awu bolsa</td>
<td>1</td>
<td>0.5</td>
<td>0.018</td>
</tr>
<tr>
<td>kopchilik bolsa</td>
<td>1</td>
<td>0.5</td>
<td>0.018</td>
</tr>
<tr>
<td>X bolsa</td>
<td>1</td>
<td>1.5</td>
<td>0.018</td>
</tr>
<tr>
<td>bar bolsa</td>
<td>1</td>
<td>0.5</td>
<td>0.018</td>
</tr>
</tbody>
</table>

Table 5.2: 100 Unorganized Corpus

<table>
<thead>
<tr>
<th>$x$</th>
<th>$c(x)$</th>
<th>$c^*$</th>
<th>$P$</th>
</tr>
</thead>
<tbody>
<tr>
<td>bolsa</td>
<td>87</td>
<td></td>
<td></td>
</tr>
<tr>
<td>u bolsa</td>
<td>18</td>
<td>17.5</td>
<td>0.201</td>
</tr>
<tr>
<td>awu bolsa</td>
<td>3</td>
<td>2.5</td>
<td>0.028</td>
</tr>
<tr>
<td>kopchilik bolsa</td>
<td>1</td>
<td>0.5</td>
<td>0.005</td>
</tr>
<tr>
<td>X bolsa</td>
<td>9</td>
<td>8.5</td>
<td>0.098</td>
</tr>
<tr>
<td>bar bolsa</td>
<td>14</td>
<td>13.5</td>
<td>0.155</td>
</tr>
</tbody>
</table>

Table 5.3: 200 "be verb" Organized Corpus

<table>
<thead>
<tr>
<th>$x$</th>
<th>$c(x)$</th>
<th>$c^*$</th>
<th>$P$</th>
</tr>
</thead>
<tbody>
<tr>
<td>bolsa</td>
<td>195</td>
<td></td>
<td></td>
</tr>
<tr>
<td>u bolsa</td>
<td>37</td>
<td>36.5</td>
<td>0.187</td>
</tr>
<tr>
<td>awu bolsa</td>
<td>5</td>
<td>4.5</td>
<td>0.023</td>
</tr>
<tr>
<td>kopchilik bolsa</td>
<td>2</td>
<td>1.5</td>
<td>0.007</td>
</tr>
<tr>
<td>X bolsa</td>
<td>18</td>
<td>17.5</td>
<td>0.089</td>
</tr>
<tr>
<td>bar bolsa</td>
<td>25</td>
<td>24.5</td>
<td>0.125</td>
</tr>
</tbody>
</table>

Table 5.4: 200 Unorganized Corpus

<table>
<thead>
<tr>
<th>$x$</th>
<th>$c(x)$</th>
<th>$c^*$</th>
<th>$P$</th>
</tr>
</thead>
<tbody>
<tr>
<td>bolsa</td>
<td>54</td>
<td></td>
<td></td>
</tr>
<tr>
<td>u bolsa</td>
<td>9</td>
<td>8.5</td>
<td>0.157</td>
</tr>
<tr>
<td>awu bolsa</td>
<td>4</td>
<td>3.5</td>
<td>0.065</td>
</tr>
<tr>
<td>kopchilik bolsa</td>
<td>2</td>
<td>1.5</td>
<td>0.028</td>
</tr>
<tr>
<td>X bolsa</td>
<td>8</td>
<td>7.5</td>
<td>0.139</td>
</tr>
<tr>
<td>bar bolsa</td>
<td>10</td>
<td>9.5</td>
<td>0.175</td>
</tr>
</tbody>
</table>

Table 5.5: 400 “be verb” Organized Corpus

<table>
<thead>
<tr>
<th>$x$</th>
<th>$c(x)$</th>
<th>$c^*$</th>
<th>$P$</th>
</tr>
</thead>
<tbody>
<tr>
<td>bolsa</td>
<td>301</td>
<td></td>
<td></td>
</tr>
<tr>
<td>u bolsa</td>
<td>56</td>
<td>55.5</td>
<td>0.184</td>
</tr>
<tr>
<td>awu bolsa</td>
<td>5</td>
<td>4.5</td>
<td>0.015</td>
</tr>
<tr>
<td>kopchilik bolsa</td>
<td>2</td>
<td>1.5</td>
<td>0.005</td>
</tr>
<tr>
<td>X bolsa</td>
<td>31</td>
<td>30.5</td>
<td>0.101</td>
</tr>
<tr>
<td>bar bolsa</td>
<td>30</td>
<td>29.5</td>
<td>0.098</td>
</tr>
</tbody>
</table>

Table 5.6: 400 Unorganized Corpus

<table>
<thead>
<tr>
<th>$x$</th>
<th>$c(x)$</th>
<th>$c^*$</th>
<th>$P$</th>
</tr>
</thead>
<tbody>
<tr>
<td>bolsa</td>
<td>76</td>
<td></td>
<td></td>
</tr>
<tr>
<td>u bolsa</td>
<td>10</td>
<td>9.5</td>
<td>0.125</td>
</tr>
<tr>
<td>awu bolsa</td>
<td>4</td>
<td>3.5</td>
<td>0.046</td>
</tr>
<tr>
<td>kopchilik bolsa</td>
<td>2</td>
<td>1.5</td>
<td>0.0197</td>
</tr>
<tr>
<td>X bolsa</td>
<td>15</td>
<td>14.5</td>
<td>0.190</td>
</tr>
<tr>
<td>bar bolsa</td>
<td>11</td>
<td>10.5</td>
<td>0.138</td>
</tr>
</tbody>
</table>

Table 5.7: 600 “be verb” Organized Corpus

<table>
<thead>
<tr>
<th>$x$</th>
<th>$c(x)$</th>
<th>$c^*$</th>
<th>$P$</th>
</tr>
</thead>
<tbody>
<tr>
<td>bolsa</td>
<td>27</td>
<td></td>
<td></td>
</tr>
<tr>
<td>u bolsa</td>
<td>2</td>
<td>1.5</td>
<td>0.056</td>
</tr>
<tr>
<td>awu bolsa</td>
<td>1</td>
<td>0.5</td>
<td>0.018</td>
</tr>
<tr>
<td>kopchilik bolsa</td>
<td>1</td>
<td>0.5</td>
<td>0.018</td>
</tr>
<tr>
<td>X bolsa</td>
<td>1</td>
<td>1.5</td>
<td>0.018</td>
</tr>
<tr>
<td>bar bolsa</td>
<td>1</td>
<td>0.5</td>
<td>0.018</td>
</tr>
</tbody>
</table>

Table 5.8: 600 Unorganized Corpus
5.4. Language Model Estimation

### Table 5.9: 800 "be verb" Organized Corpus

<table>
<thead>
<tr>
<th>x</th>
<th>c(x)</th>
<th>c*</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>bolsa</td>
<td>396</td>
<td></td>
<td></td>
</tr>
<tr>
<td>u bolsa</td>
<td>81</td>
<td>80.5</td>
<td>0.203</td>
</tr>
<tr>
<td>awu bolsa</td>
<td>6</td>
<td>5.5</td>
<td>0.013</td>
</tr>
<tr>
<td>kopchilik bolsa</td>
<td>2</td>
<td>1.5</td>
<td>0.003</td>
</tr>
<tr>
<td>X bolsa</td>
<td>42</td>
<td>41.5</td>
<td>0.105</td>
</tr>
<tr>
<td>bar bolsa</td>
<td>34</td>
<td>33.5</td>
<td>0.084</td>
</tr>
</tbody>
</table>

### Table 5.10: 800 Unorganized Corpus

<table>
<thead>
<tr>
<th>x</th>
<th>c(x)</th>
<th>c*</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>bolsa</td>
<td>222</td>
<td></td>
<td></td>
</tr>
<tr>
<td>u bolsa</td>
<td>14</td>
<td>13.5</td>
<td>0.061</td>
</tr>
<tr>
<td>awu bolsa</td>
<td>5</td>
<td>4.5</td>
<td>0.020</td>
</tr>
<tr>
<td>kopchilik bolsa</td>
<td>2</td>
<td>1.5</td>
<td>0.006</td>
</tr>
<tr>
<td>X bolsa</td>
<td>18</td>
<td>17.5</td>
<td>0.078</td>
</tr>
<tr>
<td>bar bolsa</td>
<td>11</td>
<td>10.5</td>
<td>0.047</td>
</tr>
</tbody>
</table>

### Table 5.11: 1178 "be verb" Organized Corpus

<table>
<thead>
<tr>
<th>x</th>
<th>c(x)</th>
<th>c*</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>bolsa</td>
<td>567</td>
<td></td>
<td></td>
</tr>
<tr>
<td>u bolsa</td>
<td>105</td>
<td>104.5</td>
<td>0.184</td>
</tr>
<tr>
<td>awu bolsa</td>
<td>6</td>
<td>5.5</td>
<td>0.009</td>
</tr>
<tr>
<td>kopchilik bolsa</td>
<td>2</td>
<td>1.5</td>
<td>0.003</td>
</tr>
<tr>
<td>X bolsa</td>
<td>74</td>
<td>73.5</td>
<td>0.129</td>
</tr>
<tr>
<td>bar bolsa</td>
<td>34</td>
<td>33.5</td>
<td>0.059</td>
</tr>
</tbody>
</table>

### Table 5.12: 1178 Unorganized Corpus

<table>
<thead>
<tr>
<th>x</th>
<th>c(x)</th>
<th>c*</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>bolsa</td>
<td>322</td>
<td></td>
<td></td>
</tr>
<tr>
<td>u bolsa</td>
<td>20</td>
<td>19.5</td>
<td>0.061</td>
</tr>
<tr>
<td>awu bolsa</td>
<td>5</td>
<td>4.5</td>
<td>0.014</td>
</tr>
<tr>
<td>kopchilik bolsa</td>
<td>2</td>
<td>1.5</td>
<td>0.004</td>
</tr>
<tr>
<td>X bolsa</td>
<td>26</td>
<td>25.5</td>
<td>0.079</td>
</tr>
<tr>
<td>bar bolsa</td>
<td>11</td>
<td>10.5</td>
<td>0.032</td>
</tr>
</tbody>
</table>

5.4.2 "do verb" organized corpus

The 2723 "be verb" and random 2723 unorganized sentences were selected. For all the listed bigrams (u,v), u = X, c(u,v) is greater than zero 0. Probability estimation is based on the discount count. For the unigram "qilimen, do", the bigrams are listed in six groups in the table.

### Table 5.13: 100 "do verb" Organized Corpus

<table>
<thead>
<tr>
<th>x</th>
<th>c(x)</th>
<th>c*</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>qilish</td>
<td>7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>yardem qilish</td>
<td>2</td>
<td>1.5</td>
<td>0.214</td>
</tr>
<tr>
<td>hel qilish</td>
<td>2</td>
<td>1.5</td>
<td>0.214</td>
</tr>
<tr>
<td>tejirbe qilish</td>
<td>1</td>
<td>0.5</td>
<td>0.071</td>
</tr>
<tr>
<td>doklat qilish</td>
<td>2</td>
<td>1.5</td>
<td>0.214</td>
</tr>
<tr>
<td>serip qilish</td>
<td>1</td>
<td>0.5</td>
<td>0.071</td>
</tr>
</tbody>
</table>

### Table 5.14: 100 Unorganized Corpus

<table>
<thead>
<tr>
<th>x</th>
<th>c(x)</th>
<th>c*</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>qilish</td>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>yardem qilish</td>
<td>1</td>
<td>0.5</td>
<td>0.125</td>
</tr>
<tr>
<td>hel qilish</td>
<td>1</td>
<td>0.5</td>
<td>0.125</td>
</tr>
<tr>
<td>tejirbe qilish</td>
<td>1</td>
<td>0.5</td>
<td>0.125</td>
</tr>
<tr>
<td>doklat qilish</td>
<td>1</td>
<td>1.5</td>
<td>0.125</td>
</tr>
<tr>
<td>serip qilish</td>
<td>1</td>
<td>0.5</td>
<td>0.125</td>
</tr>
</tbody>
</table>
### EXPERIMENT OF LANGUAGE MODELING

<table>
<thead>
<tr>
<th>x</th>
<th>c(x)</th>
<th>c*</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>qilish</td>
<td>42</td>
<td></td>
<td></td>
</tr>
<tr>
<td>yardem qilish</td>
<td>11</td>
<td>10.5</td>
<td>0.251</td>
</tr>
<tr>
<td>hel qilish</td>
<td>10</td>
<td>9.5</td>
<td>0.226</td>
</tr>
<tr>
<td>tejirbe qilish</td>
<td>1</td>
<td>0.5</td>
<td>0.011</td>
</tr>
<tr>
<td>doklat qilish</td>
<td>3</td>
<td>2.5</td>
<td>0.059</td>
</tr>
<tr>
<td>serip qilish</td>
<td>2</td>
<td>1.5</td>
<td>0.035</td>
</tr>
</tbody>
</table>

Table 5.15: 500 "do verb" Organized Corpus

<table>
<thead>
<tr>
<th>x</th>
<th>c(x)</th>
<th>c*</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>qilish</td>
<td>27</td>
<td></td>
<td></td>
</tr>
<tr>
<td>yardem qilish</td>
<td>3</td>
<td>2.5</td>
<td>0.092</td>
</tr>
<tr>
<td>hel qilish</td>
<td>6</td>
<td>5.5</td>
<td>0.203</td>
</tr>
<tr>
<td>tejirbe qilish</td>
<td>1</td>
<td>0.5</td>
<td>0.018</td>
</tr>
<tr>
<td>doklat qilish</td>
<td>1</td>
<td>1.5</td>
<td>0.056</td>
</tr>
<tr>
<td>serip qilish</td>
<td>1</td>
<td>0.5</td>
<td>0.018</td>
</tr>
</tbody>
</table>

Table 5.16: 500 Unorganized Corpus

<table>
<thead>
<tr>
<th>x</th>
<th>c(x)</th>
<th>c*</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>qilish</td>
<td>70</td>
<td></td>
<td></td>
</tr>
<tr>
<td>yardem qilish</td>
<td>19</td>
<td>18.5</td>
<td>0.264</td>
</tr>
<tr>
<td>hel qilish</td>
<td>15</td>
<td>14.5</td>
<td>0.207</td>
</tr>
<tr>
<td>tejirbe qilish</td>
<td>3</td>
<td>2.5</td>
<td>0.036</td>
</tr>
<tr>
<td>doklat qilish</td>
<td>3</td>
<td>2.5</td>
<td>0.036</td>
</tr>
<tr>
<td>serip qilish</td>
<td>4</td>
<td>3.5</td>
<td>0.050</td>
</tr>
</tbody>
</table>

Table 5.17: 1000 "do verb" Organized Corpus

<table>
<thead>
<tr>
<th>x</th>
<th>c(x)</th>
<th>c*</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>qilish</td>
<td>11</td>
<td></td>
<td></td>
</tr>
<tr>
<td>yardem qilish</td>
<td>27</td>
<td></td>
<td></td>
</tr>
<tr>
<td>hel qilish</td>
<td>10</td>
<td></td>
<td></td>
</tr>
<tr>
<td>tejirbe qilish</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>doklat qilish</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>serip qilish</td>
<td>1</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 5.18: 1000 Unorganized Corpus

<table>
<thead>
<tr>
<th>x</th>
<th>c(x)</th>
<th>c*</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>qilish</td>
<td>94</td>
<td></td>
<td></td>
</tr>
<tr>
<td>yardem qilish</td>
<td>21</td>
<td>20.5</td>
<td>0.218</td>
</tr>
<tr>
<td>hel qilish</td>
<td>18</td>
<td>17.5</td>
<td>0.186</td>
</tr>
<tr>
<td>tejirbe qilish</td>
<td>5</td>
<td>4.5</td>
<td>0.047</td>
</tr>
<tr>
<td>doklat qilish</td>
<td>4</td>
<td>3.5</td>
<td>0.037</td>
</tr>
<tr>
<td>serip qilish</td>
<td>4</td>
<td>3.5</td>
<td>0.037</td>
</tr>
</tbody>
</table>

Table 5.19: 1500 "do verb" Organized Corpus

<table>
<thead>
<tr>
<th>x</th>
<th>c(x)</th>
<th>c*</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>qilish</td>
<td>89</td>
<td></td>
<td></td>
</tr>
<tr>
<td>yardem qilish</td>
<td>19</td>
<td>18.5</td>
<td>0.207</td>
</tr>
<tr>
<td>hel qilish</td>
<td>16</td>
<td>15.5</td>
<td>0.174</td>
</tr>
<tr>
<td>tejirbe qilish</td>
<td>3</td>
<td>2.5</td>
<td>0.028</td>
</tr>
<tr>
<td>doklat qilish</td>
<td>1</td>
<td>0.5</td>
<td>0.056</td>
</tr>
<tr>
<td>serip qilish</td>
<td>2</td>
<td>1.5</td>
<td>0.0116</td>
</tr>
</tbody>
</table>

Table 5.20: 1500 Unorganized Corpus
5.4. Language Model Estimation

5.4.3 "Mixed verb" corpus

We used 1200 "mixed verb" sentences and 1200 randomly selected unorganized sentences. For the "mixed verb" parallel corpus diversification, the same bigram showed no more frequently than the "be verb" and "do verb" organized sentences shown in the table.

Table 5.21: 2000 "do verb" Organized Corpus

<table>
<thead>
<tr>
<th>x</th>
<th>c(x)</th>
<th>c*</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>qilish</td>
<td>116</td>
<td></td>
<td></td>
</tr>
<tr>
<td>yardem qilish</td>
<td>19</td>
<td>18.5</td>
<td>0.159</td>
</tr>
<tr>
<td>hel qilish</td>
<td>22</td>
<td>21.5</td>
<td>0.185</td>
</tr>
<tr>
<td>tejirbe qilish</td>
<td>5</td>
<td>4.5</td>
<td>0.038</td>
</tr>
<tr>
<td>doklat qilish</td>
<td>7</td>
<td>6.5</td>
<td>0.056</td>
</tr>
<tr>
<td>serip qilish</td>
<td>4</td>
<td>3.5</td>
<td>0.030</td>
</tr>
</tbody>
</table>

Table 5.22: 2000 Unorganized Corpus

<table>
<thead>
<tr>
<th>x</th>
<th>c(x)</th>
<th>c*</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>qilish</td>
<td>110</td>
<td></td>
<td></td>
</tr>
<tr>
<td>yardem qilish</td>
<td>16</td>
<td>15.5</td>
<td>0.140</td>
</tr>
<tr>
<td>hel qilish</td>
<td>17</td>
<td>16.5</td>
<td>0.150</td>
</tr>
<tr>
<td>tejirbe qilish</td>
<td>4</td>
<td>3.5</td>
<td>0.031</td>
</tr>
<tr>
<td>doklat qilish</td>
<td>2</td>
<td>1.5</td>
<td>0.013</td>
</tr>
<tr>
<td>serip qilish</td>
<td>2</td>
<td>1.5</td>
<td>0.013</td>
</tr>
</tbody>
</table>

5.4.3 "Mixed verb" corpus

We used 1200 "mixed verb" sentences and 1200 randomly selected unorganized sentences. For the "mixed verb" parallel corpus diversification, the same bigram showed no more frequently than the "be verb" and "do verb" organized sentences shown in the table.

Table 5.23: 2723 "do verb" Organized Corpus

<table>
<thead>
<tr>
<th>x</th>
<th>c(x)</th>
<th>c*</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>qilish</td>
<td>165</td>
<td></td>
<td></td>
</tr>
<tr>
<td>yardem qilish</td>
<td>25</td>
<td>24.5</td>
<td>0.148</td>
</tr>
<tr>
<td>hel qilish</td>
<td>22</td>
<td>21.5</td>
<td>0.130</td>
</tr>
<tr>
<td>tejirbe qilish</td>
<td>9</td>
<td>8.5</td>
<td>0.051</td>
</tr>
<tr>
<td>doklat qilish</td>
<td>9</td>
<td>8.5</td>
<td>0.051</td>
</tr>
<tr>
<td>serip qilish</td>
<td>6</td>
<td>5.5</td>
<td>0.033</td>
</tr>
</tbody>
</table>

Table 5.24: 2723 Unorganized Corpus

<table>
<thead>
<tr>
<th>x</th>
<th>c(x)</th>
<th>c*</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>qilish</td>
<td>142</td>
<td></td>
<td></td>
</tr>
<tr>
<td>yardem qilish</td>
<td>25</td>
<td>24.5</td>
<td>0.172</td>
</tr>
<tr>
<td>hel qilish</td>
<td>19</td>
<td>18.5</td>
<td>0.130</td>
</tr>
<tr>
<td>tejirbe qilish</td>
<td>4</td>
<td>3.5</td>
<td>0.025</td>
</tr>
<tr>
<td>doklat qilish</td>
<td>5</td>
<td>4.5</td>
<td>0.031</td>
</tr>
<tr>
<td>serip qilish</td>
<td>3</td>
<td>2.5</td>
<td>0.017</td>
</tr>
</tbody>
</table>

Table 5.25: 500 "mixed verb" Organized Corpus

<table>
<thead>
<tr>
<th>x</th>
<th>c(x)</th>
<th>c*</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>ning</td>
<td>248</td>
<td></td>
<td></td>
</tr>
<tr>
<td>waqit ning</td>
<td>2</td>
<td>1.5</td>
<td>0.006</td>
</tr>
<tr>
<td>nuqti ning</td>
<td>1</td>
<td>0.5</td>
<td>0.002</td>
</tr>
<tr>
<td>mushuk ning</td>
<td>2</td>
<td>1.5</td>
<td>0.006</td>
</tr>
</tbody>
</table>

Table 5.26: 500 Unorganized Corpus

<table>
<thead>
<tr>
<th>x</th>
<th>c(x)</th>
<th>c*</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>ning</td>
<td>224</td>
<td></td>
<td></td>
</tr>
<tr>
<td>waqit ning</td>
<td>2</td>
<td>1.5</td>
<td>0.0066</td>
</tr>
<tr>
<td>nuqti ning</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>mushuk ning</td>
<td>1</td>
<td>0.5</td>
<td>0.002</td>
</tr>
</tbody>
</table>
According to the tables, "mixed verb" sentence language model estimation is very low, and is almost equal to or smaller than the unorganized corpus. Consequently, the "mixed verb" sentences were organized, with one complex sentence divided into sample sentences. Comparison between "organized mixed verb" sentences and "unorganized mixed verb" 100 sentences is shown below.

Table 5.27: 1200 Unorganized Corpus

<table>
<thead>
<tr>
<th>x</th>
<th>c(x)</th>
<th>c*</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>ning</td>
<td>534</td>
<td></td>
<td></td>
</tr>
<tr>
<td>waqit ning</td>
<td>4</td>
<td>3.5</td>
<td>0.006</td>
</tr>
<tr>
<td>nuqti ning</td>
<td>2</td>
<td>1.5</td>
<td>0.003</td>
</tr>
<tr>
<td>mushuk ning</td>
<td>3</td>
<td>2.5</td>
<td>0.005</td>
</tr>
</tbody>
</table>

Table 5.28: 1200 Unorganized Corpus

<table>
<thead>
<tr>
<th>x</th>
<th>c(x)</th>
<th>c*</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>ning</td>
<td>542</td>
<td></td>
<td></td>
</tr>
<tr>
<td>waqit ning</td>
<td>3</td>
<td>2.5</td>
<td>0.005</td>
</tr>
<tr>
<td>nuqti ning</td>
<td>1</td>
<td>0.5</td>
<td>0.001</td>
</tr>
<tr>
<td>mushuk ning</td>
<td>2</td>
<td>1.5</td>
<td>0.002</td>
</tr>
</tbody>
</table>

Table 5.29: 100 "Organized Mixed Verbs"

<table>
<thead>
<tr>
<th>x</th>
<th>c(x)</th>
<th>c*</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>men</td>
<td>17</td>
<td></td>
<td></td>
</tr>
<tr>
<td>men oylaymen</td>
<td>9</td>
<td>8.5</td>
<td>0.5</td>
</tr>
<tr>
<td>men diqqet</td>
<td>1</td>
<td>0.5</td>
<td>0.029</td>
</tr>
<tr>
<td>men tirishtim</td>
<td>1</td>
<td>0.5</td>
<td>0.029</td>
</tr>
</tbody>
</table>

Table 5.30: 100 "Unorganized Mixed Verbs"

<table>
<thead>
<tr>
<th>x</th>
<th>c(x)</th>
<th>c*</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>ning</td>
<td>43</td>
<td></td>
<td></td>
</tr>
<tr>
<td>waqit ning</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>nuqti ning</td>
<td>1</td>
<td>0.5</td>
<td>0.011</td>
</tr>
<tr>
<td>mushuk ning</td>
<td>1</td>
<td>0.5</td>
<td>0.011</td>
</tr>
</tbody>
</table>

5.5 Evaluation Result

In this section, the results are given respectively for the three models.

5.5.1 Translation model

The translation model calculation result shows that the translation probability is extremely low. The issue in this experiment is that word alignments are annotated by a human. This is the ideal situation. Thereby, these alignments would be present in this one pair parallel corpus training data. Annotating word alignments is an extremely laborious task.

The training data used for this study were 5304 × 4 sentences. This method is far more time consuming than the other method. More importantly, from this perspective, it is difficult to make the parallel corpus more efficient.
5.5.2 Phrase-based translation

The phrase-based model provides the highest score from the number of states in which all of them have a score. Therefore, in the experiment, the second phrase lexicon is applied as a translation result for which \( f(y) = 0 \).

For training sentences \( S \) and \( T \), the translation goal is going to be to seek the derivation in the set which has the highest score on sub-function \( f \) of \( y \).

\[
A = \arg \max f(y) = \arg \max p(y) p(x \mid y)
\]

A good translation should be a derivation with good probability under a trigram language model and good probability under a translation model. The language model scores in this phrase-based model indicate that the result of translation is a sentence of the target language (Uyghur language).

Therefore, according to this phrase-based translation model calculation, to increase the translation performance, a key step is training the language model. The efficient corpus will give a better language model probability.

5.5.3 Language model

In the trigram language model estimation \( q(w \mid u, v) = \frac{c^*(u,v)}{c(w)} \) \( c(u,v) \) greater than zero.

· The 2723 "do verb" organized sentences compared to the same number of unorganized sentences show 73% more positive results for the language model of the organized parallel corpus than for an unorganized corpus.

· The 1178 "be verb" organized sentences show a 100% more positive result than for the same amount of unorganized sentences.

1. The unigram (denominator) of organized sentence counting is invariably more than unorganized sentences in the experiment.
2. The bigram (numerator)

- When collected training data in the organized parallel corpus include more bigrams \( c(u,v) \), the positive result indicates which language model of the organized corpus probability estimation is greater than the unorganized parallel corpus, as shown below.

\[ P_{organized} > P_{unorganized} \]

- When collected training data in the organized parallel corpus include fewer bigrams \( c(u,v) \), the negative result indicates which language model of organized corpus probability estimation is less than the unorganized parallel corpus, as shown below.

\[ P_{organized} < P_{unorganized} \]

3. The 1200 "mixed verb" sentences show a difference of only 0.001: an almost identical probability result in contrast to the unorganized same amount of sentences. For the increase of the language model probability, the "mixed verb" sentences are organized. Comparison between "organized mixed verb" sentences and "unorganized mixed verb" in 100 sentences show positive results. The probability of "organized mixed verb" improved considerably after sentence reordering.

The language model estimation results of organized and unorganized parallel sentences are presented in the graph below.
Fig. 5.1: "Be verb" Organized Corpus.
In the graph, blue color higher than red color represents the "organize the corpus" method gave a more positive result. Results obtained using the proposed method are presented in the table below.

<table>
<thead>
<tr>
<th>Sentences</th>
<th>Training corpus</th>
<th>Positive (%)</th>
<th>Negative (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>be verb</td>
<td>1178</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>do verb</td>
<td>2723</td>
<td>73.5</td>
<td>26.5</td>
</tr>
<tr>
<td>be and do</td>
<td>3901</td>
<td>86.5</td>
<td>13.5</td>
</tr>
</tbody>
</table>

Table 5.1: Results Obtained Using the Proposed Method

The proposed "organize corpus" method: P (blue) greater than P (red) shows an 86.5 % positive result. For the 13.5% negative result, complex sentences were tested, demonstrating that the "organized mixed verb" probability can be increased considerably.

The collected training sentences, as implemented, play a key role in the calculation, when narrowing down the range of data, the language model probability can be improved. Consequently, the more organized a corpus is, the higher the quality of the SMT translation experiment can become.
The collected training sentences played a key role in the calculation, when narrowing down the range of data, the language model probability can be improved. Consequently, a more organized corpus will yield higher quality results from the SMT translation experiment.
第6章 TRANSLATION
EXPERIMENT AND
COMPARISON RESULTS

In this chapter, we explain our experiment setup, which includes an open-source Phrase-based Statistical Machine Translation system: MOSES[39]. Japanese–Uyghur language pair is interesting for SMT for an important reason: both agglutinative languages have the same word order and have different ”Subject–Verb Agreement” vocabularies. Japanese cases are few in this respect.

Standard phrase-based SMT examples and modified phrase-based SMT systems for Japanese language and Uyghur language exist. Therefore, we will compare the results obtained using two experiment models trained using the Japanese–Uyghur training corpus.

6.1 Applied Models in Experiments

MOSES can be divided into two main components, which are the training pipeline and the decoder. The training pipeline is a collection of many tools[39]. MOSES has also extended new features for producing translations based on linguistic knowledge such as factor-based and tree-based knowledge. To our experiment, we mainly applied a phrase-based model for small corpora.

6.2 Original Model

The first model is applied purely in standard phrase-based SMT model. Because the Uyghur lexicon is naturally divided in corpora, there will be no further change for these translation methods, except to parse the Japanese corpus to lexical units using morphological analytical tools.
6.3 Proposed Model

The second model applies the original model with pre-editing of the training corpus, intermediate processing of the phrase table, and the translation table (refers to dictionary of SMT) and post-editing of translation results. We assume that pre-editing of the training corpus and post editing of the translation results will generally produce understandable translation results. We also assume the intermediate processing will improve rule-based translation systems to generate better translations for a specific area.

Experiment steps are explained below.

- We divided sentences into phrases from the Japanese–Uyghur corpus. Because both languages have similar word order, the respective vocabularies can be almost entirely paired.
- As the next step, we divided Uyghur words into stem words and morphemes in the parallel phrase group. This processed Uyghur phrase group is applied to train the proposed model.
6.4 Result

Standard Japanese–Uyghur SMT experiment by application of the MOSES toolkit, 5304 sentences were used for training. The language model has been created from one-gram model to four-gram model; each language model has attended translation and is scored using BLEU. With the very low and informal translated parallel corpus, the system performance was very
low. The score of the very useful trigram model was less than 0.1.

The probability of sparse data also strongly influenced the translation results such as the example presented in the Appendix. The translation probability of ”,” has many quite high probabilities in the phrase table. The missing of other real possible translations is unpredictable if only applying by filtering the specific number of probability, which includes the ”、” translation probability.

From the result of evaluation using the NIST individual score, BLEU cumulative score, and BLEU individual score, we receive a few improvements by application of morphological processing to the standard SMT system. However, translation results are influenced strongly by the N-gram number. Figures 6.3, 6.4, and 6.5 show that by increasing the N-gram number, the translation score will decrease. The morphological process for the SMT system by modifying the parallel corpus and translation result, is suitable for a limited vocabulary area. Key processes for wide application of the SMT system are increasing the high-quality training corpus and for more work on the morphological-process-specific part of the SMT system.

<table>
<thead>
<tr>
<th>Ngram</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Without MA</td>
<td>4.912</td>
<td>0.677</td>
<td>0.038</td>
<td>0.002</td>
</tr>
<tr>
<td>With MA</td>
<td>4</td>
<td>0.652</td>
<td>0.03</td>
<td>0.012</td>
</tr>
</tbody>
</table>

Fig. 6.3: NIST score
6.5 Conclusion

Major results of our work can be summarized as explained below.

- Some number of phrase sentences will be translated to the correct form of Uyghur sentences by the standard phrase-based translation system. Because of word order similarity, the following are true.

- The suffix decomposition of the parallel corpus and suffix composition of the translation result will increase the translation accuracy.

- Pre-processing (as suffix decomposition of the parallel corpus) shows increased translation understandability.

- Post-processing does not increase the translation accuracy without preprocessing of parallel corpus.

A translated example is listed below.

Fig. 6.4: BLEU Cumulative Score
Fig. 6.5: BLEU Individual score

<table>
<thead>
<tr>
<th>N-gram</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Without MA</td>
<td>0.745</td>
<td>0.398</td>
<td>0.211</td>
<td>0.102</td>
</tr>
<tr>
<td>With MA</td>
<td>0.907</td>
<td>0.743</td>
<td>0.381</td>
<td>0.096</td>
</tr>
</tbody>
</table>

Table 6.1: Verb ‘To be’ Sample 1

<table>
<thead>
<tr>
<th>Japanese</th>
<th>私 は 歌 が 上 手 で す</th>
</tr>
</thead>
<tbody>
<tr>
<td>U (Human)</td>
<td>Men nahshigha usta</td>
</tr>
<tr>
<td>U (Standard)</td>
<td>men ?? usta</td>
</tr>
<tr>
<td>U (modified)</td>
<td>men naxsha gha usta</td>
</tr>
</tbody>
</table>

Table 6.2: Verb ‘To be’ Sample 2

<table>
<thead>
<tr>
<th>Japanese</th>
<th>私 は 学 生 で す</th>
</tr>
</thead>
<tbody>
<tr>
<td>U (Human)</td>
<td>Men oqughuchi</td>
</tr>
<tr>
<td>U (Standard)</td>
<td>men oqughuchi</td>
</tr>
<tr>
<td>U (modified)</td>
<td>men oqughuchi</td>
</tr>
</tbody>
</table>

Table 6.3: Verb ‘To be’ Sample 3

<table>
<thead>
<tr>
<th>Japanese</th>
<th>私 は 先 生 で す</th>
</tr>
</thead>
<tbody>
<tr>
<td>U (Human)</td>
<td>Men oqutquchi</td>
</tr>
<tr>
<td>U (Standard)</td>
<td>men ustaz</td>
</tr>
<tr>
<td>U (modified)</td>
<td>men ustaz</td>
</tr>
</tbody>
</table>
Table 6.4: Verb ‘To do’ Sample 1

<table>
<thead>
<tr>
<th>Japanese</th>
<th>Ete fuji teghigha bariman</th>
<th>○</th>
</tr>
</thead>
<tbody>
<tr>
<td>U (Human)</td>
<td>Ete ?? ?? ?? ?? barimen</td>
<td>✗</td>
</tr>
<tr>
<td>U (modified)</td>
<td>Ete bolsa ?? gha barimen</td>
<td>✗</td>
</tr>
</tbody>
</table>

Table 6.5: Verb ‘To do’ Sample 2

<table>
<thead>
<tr>
<th>Japanese</th>
<th>Man Englizcha uginiman</th>
<th>○</th>
</tr>
</thead>
<tbody>
<tr>
<td>U (Human)</td>
<td>Man Engliz ?? ?? uginiman</td>
<td>△</td>
</tr>
<tr>
<td>U (modified)</td>
<td>Man Engliz ni ?? uginish qiliman</td>
<td>△</td>
</tr>
</tbody>
</table>

Table 6.6: Verb ‘To do’ Sample 3

<table>
<thead>
<tr>
<th>Japanese</th>
<th>Hiroshimada yarlik mahsulat setiwalghachqa, keler hapte elip barimen</th>
<th>○</th>
</tr>
</thead>
<tbody>
<tr>
<td>U (Human)</td>
<td>hiroshima ?? ?? setiwaldi ?? keler hapte elip barimen</td>
<td>✗</td>
</tr>
<tr>
<td>U (modified)</td>
<td>hiroshima din ?? setiwel ?? keler hapte elip barimen</td>
<td>✗</td>
</tr>
</tbody>
</table>

Table 6.7: Verb ‘To be + To do’ Sample 1

<table>
<thead>
<tr>
<th>Japanese</th>
<th>Kachkicha hizmat qilghan bolsammu, wazipini tugitalmidim</th>
<th>○</th>
</tr>
</thead>
<tbody>
<tr>
<td>U (Human)</td>
<td>kach ghicha hizmat ni qilidi ??, ?? tugimidi</td>
<td>✗</td>
</tr>
<tr>
<td>U (modified)</td>
<td>kach ghicha hizmat ni qilidim ghu, hizmat tugimidi</td>
<td>△</td>
</tr>
</tbody>
</table>

Table 6.8: Verb ‘To be + To do’ Sample 2

<table>
<thead>
<tr>
<th>Japanese</th>
<th>Qolayliq dukandin narsa setwalghan idim, bahasi qimmat bulup katti</th>
<th>○</th>
</tr>
</thead>
<tbody>
<tr>
<td>U (Human)</td>
<td>konbini da setwal narsa ?? qildim, baha qimmat boldi</td>
<td>✗</td>
</tr>
<tr>
<td>U (modified)</td>
<td>konbini da setwal narsa ni qildim, baha qimmat boldi</td>
<td>✗</td>
</tr>
</tbody>
</table>

Table 6.9: Verb ‘To be + To do’ Sample 3
第7章 Conclusion and Future work

7.1 Conclusion

In this thesis, corpus evaluation is presented as a specific means of producing an efficient parallel corpus for use with a cross-translation SMT system intended for application to a low source language. The very small number of efficient parallel corpora currently obviates the use of machine translation systems. Our research goal is to enhance SMT performance by the effective use of a corpus that produces better translation without a huge parallel corpus. Corpus evaluation is implemented by two means.

First, human evaluation is applied using the proposed evaluation model. The JEC-basic sentence alignment is extended into four languages by full translation into Uyghur language of the 5304×4 sentences of the (Kurohashi& Kawahara) corpus. We investigate issues of the bilingual corpus and compare three pairs of corpora to ascertain whether they give the same result in SMT. The evaluation result demonstrates that this testing method requires a larger and higher-quality parallel corpus to train the model.

Secondly, the parallel corpus was evaluated by comparison using calculated results obtained using a language model, translation model, and phrase-based translation model. For increased word alignment probability for SMT, the "organize the corpus" method is proposed. With the proposed translation model calculation derivation formula, the probability results were found to be very low, rendering it difficult to train an efficient corpus for the low source language. To obtain a better phrase-based translation model, the language model score plays an important role. The model’s calculation result shows that "organize the corpus" method is applicable for increased parallel corpus efficiency, and for extending parallel corpora.

This work involves improvement of the efficiency of the respective parallel corpora of other cross-translated languages, especially for many other low source languages. We will apply this method to train a low source language decoder that teaches the SMT system how to
separate complex sentences to obtain a more efficient parallel corpus and thereby achieve better-quality translation performance.

Major results of our work can be summarized as presented below.

- Parallel Corpora: Low accurate corpus increase the misalignment in the model.
- Tagged corpus must be increased because Uyghur language has no open-source OCR software

### 7.2 Future Work

In this work, we have discussed corpus evaluation for a Japanese–Uyghur SMT system. The corpus was evaluated in terms of two aspects: human evaluation and model calculation. A cross-language evaluation model, "organize the corpus" method, and a formula derived for a translation model are proposed. These methods still must be trained for SMT system performance.

- We plan to increase the JECU-basic-sentence alignment quality for better evaluation results. The collected parallel corpus will be checked and cleared of poor translation sentences and alignment mistakes.

- Collect more organized parallel corpora for Japanese–Uyghur.

In this work, we presented some results obtained through this Japanese-to-Uyghur phrase-based SMT study. This language pair is interesting for SMT for several reasons: Both agglutinative languages have the same word order and have differences of "Subject–Verb Agreement" vocabulary, although Japanese examples are few in this respect.
Reference


Conclusion and Future work


7.2. Future Work


.1 Test sentences

- **Japanese:** 考慮の整理学
- **Uyghur:** pikerning retlesh ilimi

- **Japanese:** 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。
- **Uyghur:** balliq chaglardin bashlap, untup qalma, untup qalma dap ughutup, untup qaldimdim disem til ishitten idim.

- **Japanese:** そのせいもあって、忘れることができた。忘れることに恐怖心を抱き続けている。
- **Uyghur:** shuning sewebimu bolup, untup qaldighan ishtin qorqush hisyatini uzemge hemra qilip keldim.

- **Japanese:** 悪いと決めてしまう。
- **Uyghur:** yaman dep bikitewitish.

- **Japanese:** 学校が忘れず、良く覚え、と命じるのは、それなりの理由がある。
- **Uyghur:** mektep untup qalma, isingde chingtut, dep buyrq chushurishing uzighe chushluk sewebi bar.

- **Japanese:** 教室は知識を与える。
- **Uyghur:** sinip bilim beridu.

- **Japanese:** 知識を増やすのを目標にする。
- **Uyghur:** bilim ni kupeytishni nishan qilghan bolidu.

- **Japanese:** せっかく与えたものを片端から、捨ててしまっている。
- **Uyghur:** ming teslikte berghen nersini ishletmestinla tashiwetsek yahshi emes.

- **Japanese:** 良く覚えておく。
- **Uyghur:** isingde ching saqla.

- **Japanese:** 覚えているかどうか、ときどき試験をして調べる。
- **Uyghur:** este tutqan tutmighanliqi, daim imtahan elip sinap turidu.

- **Japanese:** 覚えていなければ減点して警告する。
- **Uyghur:** este saqlimisaq numur tartip agahlanduridu.

- **Japanese:** 点はいいほうがいいに決まっているから、みんな知らず知らずのうちに、忘れのを怖がるようになる。
- **Uyghur:** numur yahshi bolsa yahshi ish dep bikitiliwetkenliktin, kupchilik bilmey bilmeyla untup qalishtin qorqudighan bolup qalidu.

- **Japanese:** 教育程度が高くなればなるほど、そして、頭が高いと言われれば、言われるほど、知識をたくさんもっている。
- **Uyghur:** maarip sewiyisi qanche yuquri bolsa, hemde kallisi qanche kallsi utkur digenche, bilimni shunche kup qubul qilidu.

- **Japanese:** つまり、忘れないでいるものが多い。
Uyghur: Dimek, untup qalmaydighan nerse kup.

Japanese: 頭の優秀さは、記憶力の優秀さとしばしば同じ意味を持っている。

Uyghur: Kallining utkuriliki bolsa, este tutush qabiliyiting yuqurliqi bilen asasi jeheppin ohhash menini uz ichge alidu.

Japanese: それで、生き字引と言うような人間ができる。

Uyghur: Shundaq qilip, ulge bolginudek kishge aylanghan bolidu.

Japanese: これまでの教育では、人間の頭脳を、倉庫のようなものだと見てきた。

Uyghur: Hazirghe qeder maarip bolsa, kishlerning mengisini, sang gha uhshap ketidu dep qarighan.

Japanese: 知識をどんどん蓄積する。

Uyghur: Bilim tadirji kupiyip maghidu.

Japanese: 倉庫は大きければ大きいほどよいらしい。

Uyghur: Sang qanche chong bolghansiri shunche yahshi.

Japanese: 中にたくさんのものが詰まっていればいるほど結構だとなる。

Uyghur: Ichqhe qanche kup nerse qachilansa shunche ihham kurinidu.

Japanese: せっかく蓄積しようとしている一方から、どんどんものがなくなってしまったりしてはことだから、忘れるな、が合言葉になる。

Uyghur: Aran saqlay dep turush bilen birge, tadirji untup qalidighan ish bolghachqa, untup qalma, bolsa daimliq suzge aylinip qalidu.

Japanese: ときどき在庫検査をして、なくなっていないかどうかをチェックする。

Uyghur: Gaxida sangni tekshurup, tughidimu tughimidimu dep tekshurup turidu.

Japanese: それがテストである。

Uyghur: U del intamahan bolidu.

Japanese: 倉庫としての頭にとっては、忘却は敵である。

Uyghur: Meghe dighen sanggha nispiten, untush bolsa dushmen hisaplinidu.

Japanese: 博識は学問のある証拠であった。

Uyghur: Bilimlik dighen bolsa bilimi kup ekenlikining ispati bolidu.

Japanese: ところが、こういう人間頭脳にとって恐るべき敵が現れた。

Uyghur: Emma, mushunda kishlening mengisige nispiten qorqumushluq dushmen peyda boldi.

Japanese: コンピューターである。

Uyghur: U bolsimu computerdin ibaret.

Japanese: これが倉庫としては素晴らしい機能を持っている。

Uyghur: Bu sanggha nispiten isil rolni hazrilghan.

Japanese: いったん入れたものは決して失わない。

Uyghur: Bir qetim kirghen nerse zadila uchup ketmeydu.
・Japanese:必要なときには、きっと、引き出すことができる。
Uyghur :lazim waqitta ,derhal elip ishletkili bolidu.

・Japanese:整理の完全である。
Uyghur :retlesh mukemmel bolidu.

・Japanese:コンピューターの出現、普及に伴って、人間の頭を倉庫として使うことに、疑問がわいてきた。
Uyghur :computernig utturgha chiqishi,kengiyishigha eghiship ,kishlernign mengisini sang dep ishlitidighan ishlargha sual bilen qarash kupiyip qaldi.

・Japanese:コンピューター人間をこしらえていたのでは、本物のコンピューターにかなうわけがない。
Uyghur :computer ademni yasaydighan bolsaq ,heqiqi computerni yengiwalghili bolmaydighan ish yoq.

・Japanese:そこでようやく創造的人間ということが問題になってきた。
Uyghur :shundaq qilip,aran teslikte ijatchan adem dighen nerse mesiligne aylendi.

・Japanese:コンピューターのできないことをしなくては、というのである。
Uyghur :computer qilalmaydighan ishlarni qiliaydighan bolsa ,dighen ish din ibaret.

・Japanese:人間の頭はこれからも、一部は倉庫の役をはたし続けなくてはならないだろうが、それだけではいけない。
Uyghur :ademlarning mengisi bundin kiyinmu,birqisimi sangnig hizmitini utesh bilenla qalmay,bashqa hizmetlernimu utushi zurur.

・Japanese:新しいことを考え出す工場でなくてはならない。
Uyghur :yengni bersini oylap tapidighan zawut bolmisa bolmaydu.

・Japanese:倉庫なら、入れたものを紛失しないようにしておくべきだが、物を作り出すにはそういう保存保管の能力だけでは仕方がない。
Uyghur :sang bolsa ,qachilap qoyghan nersileri buzulp ketmeydighan qlip saqlap qoysila bolidu,nersini yasap chiqishini shundaq saqlap bashqurush iqdidari bilen amal qilghili bolmaydu.

・Japanese:第一、工場にやたらなものが入っていては作業能率が悪い。
Uyghur :birinchi,zawutni qalaymiqan bersiler kirip qalsa hizmet unumni tuwanlep ketidu.

・Japanese:余計なものは処分して広々としたスペースをとる必要がある。
Uyghur :artuq bersilerini birterap qilip ,keng bolghan boshluqni ichish bekmu muhim.

・Japanese:それがと言って、すべてのものを捨ててしまうのは仕事にならない。
Uyghur :shundaq dep hemmee bersini tashliwetsek hizmet bolmuydu.

・Japanese:整理が大事になる。
Uyghur :retlesh bolsa muhi bolidu.

・Japanese:倉庫だって整理は欠かせないが、それはあるものを取り除く整理である。
Uyghur: sanggha nispiten retlesh bolsa lazim, u bolsimu bar nersini elip tashlashtin ibaret.

Japanese: この工場の整理に当たることをするのが、忘却だ、工場として能率をよくしようと思えば、どんどん忘れてやってなくてはいけない。

Uyghur: bu zawutni retlesh hizmiti bolsa, untush, zawutning ish unumini yahshilay dep oylisaq, peydin-pey untup ketmisek bolmaydu.

Japanese: そのことが、今の人間にはよくわかっていない。

Uyghur: bu zawatni retlesh hizmiti bolsa, untush, zawutning ish unumini yahshilay dep oylisaq, peydin-pey untup ketmisek bolmaydu.

Japanese: それで工場の中を倉庫のようにして喜んでいる人が現れる。

Uyghur: bu zawatni retlesh hizmiti bolsa, untush, zawutning ish unumini yahshilay dep oylisaq, peydin-pey untup ketmisek bolmaydu.

Japanese: それでも、倉庫としてもうまく機能しない頭を育ててしまいかねない。

Uyghur: bu zawut disekmu, sang disekmu yahshi ishlimeydighan mengini terbiylep qaldighan isthin saqlanqili bolmaydu boludu.

Japanese: コンピューターには、こういう忘却ができないのである。

Uyghur: computer ni bolsa, mushundaq untush ni qilaymaydu.

Japanese: コンピューターには倉庫に専念させ、人間の頭は知的工場に重点を置くようにするのが、これからの方向でなくてはならない。

Uyghur: computer ni bolsa sanguila oylap, ademning mengisini bilim zawutini oylayshi muhim orungha qoyudighan qilish bolsa, bundin kiyinki yunilishtin ibaret.

Japanese: それには、忘れることに対する偏見を改めなくてはならない。

Uyghur: shuning uchun, untush dighen kuzshe ilmasliqi uzgertmisak bolmaydu.

Japanese: そして、そのつもりになってみると、忘れるのは案外、難しい。

Uyghur: shuning bilen, shundaq qilmaqchi bolup baqsaq, untush bilekchila tes bolup ketidu.

Japanese: 例えば、なんか突発の事件が起こったとする。

Uyghur: mesilen, qandaqtur alahide weqe yuzberghen bolsun.

Japanese: その渦中の人は、あまりのことに、あそれもこれもいろいろなことが一時に殺到する。

Uyghur: shuning ichide qalghan hademghe hedidin ziyade, undaq ish bundaq ish bolup herhil ishlar birla waqitta busup kelidu.

Japanese: 頭の中へどんどんいろいろなことが入ってきて、混乱状態に落ちている。

Uyghur: mengighe tedirji nurghunlighan ishlar kirip, qalaymiqan haletke chushup qalidu.

Japanese: 茫然失失、どうしていいか分からない。

Uyghur: uzini yoqutup qoyup, zadi nime ish qilish kireklklin bilmeq qalidu.

Japanese: これが「忙しい」のである。

Uyghur: bu bolsa aldirashchiliqtiq ibaret.

Japanese: 「忙」の字は、心（りっしんぺん）を亡くしていると書く。

Uyghur: [aldirash] digen suz bolsa, kungulni yoqutup qoyush dighen haltte yeziidu.
1. Test sentences

- Japanese: 忙しいと頭が働かなくなってしまう。
  Uyghur: aldirash bolsa menge ishlimey qalidu.

- Japanese: 頭を忙しくしてはいけない。
  Uyghur: mengni aldirash qilwetsek bolmaydu.

- Japanese: ガラクタのいっぱいの倉庫は困る。
  Uyghur: ishletmeydihgan herhil nersiler liq qachilanghan sang bolsa qiylinip qalidu.

- Japanese: 平常の生活で、頭が忙しくてはいけない。
  Uyghur: adettiki turmushta bolsa, meghe aldirash bolsa bolmaydu.

- Japanese: 人間は、自然に、頭の中を整理して、忙しくならないようになっている。
  Uyghur: adem tebila mengni retlep, aldirash bolmaydighan halatte qilidu.

- Japanese: 睡眠である。
  Uyghur: uhlash tin ibaret.

- Japanese: 眠ってからしばらくすると、レム睡眠というものが始まる。
  Uyghur: ohlap azraq waqit utkendin kiyin , rim uyqu dighen nerse bashlinidu.

- Japanese: このレムの間に、頭はその日のうちにあったことを整理している。
  Uyghur: bu rim turghan waqitta , meghe shu kunde bolghan ishlarni retlep chiqidu.

- Japanese: 記憶しておくべきこと、すなわち、倉庫に入れるべきものと、処分してしまってよいもの、忘れるものとの区分けが行われる。
  Uyghur: iside saqlaydighan ishlarni , mundaq disek ,sanggha salidighan nersiler, birterep qilwetse boldighen nersiler,untuydighan nersilerni ayrish eilp berilidu.

- Japanese: 自然忘却である。
  Uyghur: tebi untush din ibaret.

- Japanese: 朝目を覚まして、気分爽快であるのは、夜の間に、頭の中がきれいに整理されて、広々としているからである。
  Uyghur: ettighen urundin turup,kunul rahet boldighan ish bolsa, kechide , mengning ichi chirayliq retlinip,kengiyip qalghanliqi uchun.

- Japanese: 何かの事情で、それが妨げられると、寝覚めが悪く、頭が重い。
  Uyghur: qandaqtur mesilide, u tusqunluqqa uchrup,uyqudin oyghanmaq tesliship,bash eghir.

- Japanese: 朝の時間が、思考にとって黄金の時間であるのも、頭の工場の中がよく整頓されて、動きやすくなっているからにおかなければならない。
  Uyghur: ettighendi waqit,oylashqa nispiten altun waqit bolup,mening zawuti taza retlinip,herket asan bolghanliqtin bashqa nerse emes.

- Japanese: 昔の人は、自然に従った生活をしていたから、神の与えられた忘却作用である睡眠だけでも、充分、頭の掃除ができていた。
  Uyghur: burunqi ademler bolsa,tebietke masliship turmuch kechurghen bolghachqa,huda ata qil-
ghen untush hizmitini qildighan oyqu bilenla, asasen mengning taziliqini qilalighan.

- Japanese: とそこが、今の人間は、情報通多といわれる社会に生きている。
  - Uyghur: emma, hazirqi ademlar bolsa, uchur kup bolghan jemiyette yashawatidu.

- Japanese: どうしても不可欠なものがあり、頭にたまりやすい。
  - Uyghur: qandaqla qilsimu lazim bolmighan nersiler mengighe qapliship qelish asan.

- Japanese: 夜のレム睡眠くらいでは、処理できないものが残る。
  - Uyghur: kechitiki rim oyqu bilen birterp qilaymaydighan nersiler qilip qalidu.

- Japanese: これをそのままにしておけば、だんだん頭の中が混乱し、常時、「忙しい」状態になる。
  - Uyghur: bu nersileri shu halette qoyup qoyasaq, peydin-ney mengining ichi qalaymiqanliship , adette qapliship qelish asan.

- Japanese: ノイローゼなども、そういった原因から起こる。
  - Uyghur: noyruze mu shuning sewebidin peyda boloidu.

- Japanese: かつては、忘れてはいけない、忘れてはいけない、と言っていられた。
  - Uyghur: burun, untup qalsang bolmaydu, untup qalsang bolmaydu dep kelghen idurq.

- Japanese: 倉庫として頭を使った。
  - Uyghur: sang dep qarap mengni ishlitep kelghen.

- Japanese: 中が広々していたからである。
  - Uyghur: ichi kengyip ketkenlikitin bolghan.

- Japanese: このごろは入れるものが多かったのに、スペースには限りがある。
  - Uyghur: hazir kirghuzighan nersiler kuplep ketkenlikitin, suret ning chikige yitip qalidu.

- Japanese: そん上、倉庫だけではなく工場としてものを割り出さなくてはいけない。
  - Uyghur: uning ustighe, sang deplae emes, zwut dep nersilerini yasap chiqmisaq bolmaydu.

- Japanese: 場ふさぎがごろごろしているのは不都合である。
  - Uyghur: urun etkuch ching bolmisa yahshi bolmaydu.

- Japanese: 忘れる能力が求められるようになる。
  - Uyghur: ntush qabiliyite telep qilinidighan boloidu.

- Japanese: これまで、多くの人はこんなことは考えたこともないから、さあ、忘れてみよ、と言われても、さっさと忘れられるわけではない。
  - Uyghur: hazirghiche, kupleghan kishler bundaq ishlarni oylapmu qoymighan bolghachqa, aha, untup baqay, depmu, asanla untup ketidighan ishlar yoq.

- Japanese: しかし、入れるものがあれば、出るものがなくてはならない。
  - Uyghur: emma, kirghen nerse bolghandin kiyn, chiqmidaq nerse bolmisa bolmaydu.

- Japanese: 入れるだけで、出さなくては、爆発してしまう。
  - Uyghur: kiguzhush bilenla bolup, chiqarmisaq, partilap ketidighan gep.
1. Test sentences

- Japanese:食べ物を食べると。
  Uyghur :yimekilki neyish.

- Japanese:消化して吸収すべきものを吸収したら、その残りは体外へ排泄する。
  Uyghur :hezim qilib qobil qilidighan nersiler ni qobil qilib, eship qaqlininelse bolsa beden siritigha chiqirwetidighan gep.

- Japanese:食べなければならない、排泄しなければ痛みである。
  Uyghur :yiyish bilenla, chiqirish bolmisa qawziyet bolup qalidu.

- Japanese:これまでの、倉庫式教育は、うっかりしていると、この痛みをつくりかねなかった。
  Uyghur :hezirghiche, sangche maaripta bolsa, bush tursaq, qawziyet ni kelturup chiqarmaydu dighili bolmaydu.

- Japanese:どんなに摂取したら、どんなに排泄しないといけない。
  Uyghur :kupkep sumirsek, kulep chiqarmisaq bolmaydu.

- Japanese:忘却はこの不可欠な排泄に当たる。
  Uyghur :untush bolsa mushu kem bolsa bolmaydighan chiqirishqa teng dimektur.

- Japanese:目のかたちにするのは大きな誤りである。
  Uyghur :kuzimizni qisiwalsaq chong hatalashqan bolimiz.

- Japanese:勉強し、知識を習得する一方で、不要になった物を、処分し、整理する必要がある。
  Uyghur :ughumup, bilimni qobil qilish bilen birghe, lazim bolmighan nersileri birterp qilib, retleshmu lazim bolindu.

- Japanese:何か大切で、何かそうでないか。
  Uyghur :nim muhim, nime muhim emes.

- Japanese:これが分からないと、古新聞一枚だって、整理できないが、いちいちそれを考えているひまはない。
  Uyghur :buni bilmisek, biwaste sezgu bilsela, keyni keynidan lazim bolmighan nersileri ayrip toldurup tizip chiqidighan gep.

- Japanese:自然のうちに、直観的に、あとあと必要なものを、不要らしいものを区別して、新陳代謝をしている。
  Uyghur :tebi halda, biwaste sezgu bilsela, keynidan lazim bolmighan nersileri ayrip toldurup tizip chiqidighan gep.

- Japanese:頭をよく働かせるには、この “ 忘れる ” ことが、きわめて大切である。
  Uyghur :mengni yahshi ishlitish uchun, bu niyish niyish bolsa bekmu muhim.

- Japanese:頭を高能率の工場にするためにも、どうしても絶えず忘れて行く必要がある。
  Uyghur :mengni unumi yahshi zawutqa aylandurush uchun, qandaqyl qilib bolsimu, zadi untup qelish muhim bolindu.

- Japanese:忘れるのは価値観に基づいて忘れる。
・Uyghur :untush bolsa qimmet qarishigha asaslinip untulinidu.

・Japanese:面白いいと思っていますは、些細なことも忘れない。
・Uyghur :qiziqarliq dep oylighan ish bolsa,inchike ishlardimu zadila untup qelinmaydu.

・Japanese:価値観がしっかりしていないと、大切なものを忘れ、つまらないものを覚えていることになる。
・Uyghur :qimmet qarishi eniq bolmighan bolsa,ruhum nersilerni untup qelip, keraksiz nersilerni isida saqlap yuruydighan ishlar bolidu.

・Japanese:これについては、さらに考えなくてはならない。
・Uyghur :buningha nispiten,yenisim oylanmisaq bolmaydu.

・Japanese:食事の文明論
・Uyghur :tamaq mediniyiti

・Japanese:世界のほとんどの地域において一日三回食事をするようになったのは、つい近頃のことである。
・Uyghur :dunyaning herqaysi jaylirigha nisiten birkunde uch waq tamaq yeydighan bolghili ,yeqinqi bir nechche yil bolid.

・Japanese:蛇のように食いだめをすることができない体なので、人間は一日に一度は食事をしなくてはならない。
・Uyghur :yilangha ohshash tamaqni saqlap yeyelmighenliktin,adem bir kunde bir qetim qandaqla bolmusun tamaq yimise bolmaydu.

・Japanese:だが、日に何回食べるべきかということを決めるのは人体生理の問題ではなくて、それぞれの社会の問題である。
・Uyghur :emma,kunghe qanche qetim yise bolidu dighenni adem bedininig ajritip chiqirishining mesilisi bolupla qalmastin, belki jemiyetningmu mesilisidin ibaret.

・Japanese:西ニューギニア高地のモニ族の社会では、正式の食事ということは一日一回しかない。
・Uyghur :gharibi niginia ighizlikidi moni milliti no jemiyetide bolsa,resmi tamaq waqtig birkunde bir qetimla iken.

・Japanese:午後二時ごろ一日の畑仕事が終わって家に帰ってつくる食事を家族全員がそろって食べる。
・Uyghur :chushtin kiyin 2 lerde bir kunluk etiz ishlorini tughitup uyghe qaytip kilip hemmeylen yighilip tamaq yeydu.