Statistical machine translation proposal for Uzbek to English

Alisher Shakirovich Ismailov
Andijan Machine Building Institute
Gulshoda Shamsiyeva
National University of Uzbekistan
Nilufar Abdurakhmonova
Tashkent State University of Uzbek Language and Literature named after Alisher Navoi

Abstract: The machine translation means is a translating one natural language to another natural language automatically [1]. The machine translation is one of the major and the most active areas in natural language processing. The last decade have seen the rise of the use of statistical approaches to the machine translation. The statistical machine translation approaches learn translation parameters automatically from alignment text rather than relying on rule-based approaches. There has been quite extensive work in statistical machine translation area for some language pairs. However, there are very limited research sources available for the Uzbek to English language pair [2]. In this paper, we propose statistical machine translation algorithm for Uzbek to English. The developing English to Uzbek statistical machine translation algorithm is an interesting obstacle from a number of perspectives. The most important challenge is that English and Uzbek are typologically distant languages. The English language has very limited morphology and Uzbek is an agglutinative language with a very rich and productive derivational and inflectional morphology. The Uzbek word structures that can correspond to complete phrases of several words in English when translated. In this paper, propose that will achieve Uzbek to English statistical machine translation algorithm using phrase-base model. Moreover, in order to achieve statistical machine translation we need to develop English-Uzbek corpora. In this paper, we present briefly about English-Uzbek corpora development.

Keywords: Machine translation, natural language, statistical machine translation, corpora

INSTRUCTION

Statistical machine translation (SMT) from English to Uzbek poses a number of problems. Typologically English and Uzbek are very different languages. The English language has very limited morphology and normal sentence order as follows Subject+Verb+Object. The Uzbek language is an agglutinative language with a very rich derivational and inflectional morphology, and sentence order normally is

Subject+Object+Verb. Another issue of practical significance is the lack of large-scale parallel text resources for Uzbek to English. This paper structured as follows: We first briefly discuss issues in statistical machine translation and Uzbek language, and review statistical machine translation methods. We then continue with proposed Uzbek to English statistical machine translation algorithm, and we will briefly explain English-Uzbek corpora and finally conclude our discussion.

ISSUES IN BUILDING A STATISTICAL MACHINE TRANSLATION ALGORITHM FOR UZBEK LANGUAGE

The initial step to build a statistical machine translation algorithm is the compilation of parallel texts, which turns out to be a significant issue for the Uzbek and English pair. There are not many sources of such texts. There is also a limited amount data parallel news corpus available from certain news sources. The main aspect that would have to be seriously considered first for Uzbek language in statistical machine translation is the productive inflectional and derivational morphology. The Uzbek word forms consist of morphemes concatenated to a root morpheme or to other morphemes [3]. Except for a very few exceptional cases, the surface realizations of the morphemes are conditioned by various local regular morphophonemic processes such as vowel harmony, consonant assimilation and elisions [9]. Further, most morphemes have phrasal scopes: although they attach to a particular stem, their syntactic roles extend beyond the stems. The morphotactics of word forms can be quite complicated when multiple derivations are involved [10]. For example, the derived modifier mustahkamlashtiramiz would be broken into surface morphemes as follows:

mustahkam+lashtira+mız

Starting from an adjectival root mustahkam, this word form first derives a verbal stem mustahkamlashtirmoq, meaning, “to make it strong”. A second suffix, the causative surface morpheme +lashtıra which we treat as a verbal derivation, forms yet another verbal stem meaning “to cause” or “to make”. The final suffix, +miz, meaning “we”, “us”. If we translate the word “mustahkamlashtiramiz” into English, would be a”we will make it strong”.

The Uzbek language alphabet has 29 letters. There are 6 vowels: a, e, i, o, u, o`

And 23 consonants: b, d, f, g, h, j, k, l, m, n, p, q, r, s, t, v, x, y, z, g`, sh, ch, ng

The table below illustrates some Uzbek words and their meaning in English language. You can see that some words translated into multiple English words.

Uzbek	English
Go`zal	Beautiful
Men	I, me
Sen, siz	You
U	He, she
Biz	We

Ular	They
Ishdaman	I am at work
O`qimoqchiman	I am planning to read/study
Charchadim	I am tired

STATICTICAL MACHINE TRANSLATION METHODS

Word-based model

In word-based translation method, the basic unit of translation is a word in natural languages [4]. Normally, the translated sentences will be different than original sentence, because of compound words, morphology and idioms. For example, the English word “happy” can be translated in Uzbek language by either “xursand” or “kayfiyati chog`”, depending on context of sentence. Simple word-based translation has difficulties to translate between languages with different fertility. The word-based translation systems work in such that they could map a single word to multiple words, but not the other way around. For example, if we were translating from Uzbek to English language, each word in Uzbek language can be produce any number of English words. However, there is no way to group two English words producing a single Uzbek word. There are some word-based translation systems are the freely available such as GIZA++ package (GPLed), which contains the training program for IBM models and HMM model and Model 6. [5]. Today the word-based translation model is not widely used. The phrase-based systems are more commonly used nowadays. Many phrase-based systems are still using GIZA++ to align the corpus. The alignments are applied to extract phrases or gather syntax rules. [6].

Phrase-based model

The phrase-based translation method’s aim is to reduce the restrictions of word-based translation by translating sequences of words, the translation lengths may differ [4]. These sequences of words are called phrases. The translation phrases found using statistical methods from corpora. The translation chosen phrases will be mapped one-to-one based on a phrase translation table, and then may be reordered for better language structure. This translation table can be learnt based on word-alignment, or directly from a parallel corpus. For morphological rich languages, the phrase-based model will produce better result.

Language model

A language model is a necessary component of statistical machine translation [4]. The language model aids in making the translation as fluent as possible. The language model is a function that takes a translated sentence and returns the probability of its most fluent version. A good language model will for example assign a higher probability to the sentence "the boy is coming from school" than "the school boy coming is". Another function of language model is that it may also help with

word choice. If a foreign word has multiple probable translations, these functions will give better probabilities translations in specific contexts in the target language [7].

PROPOSED METHOD

In order achieve statistical machine translation algorithm for Uzbek to English we apply phrase-based model. When we compare Uzbek-English languages that some words in an Uzbek language translates into multiple English words, or vice versa. The word-based models will have inefficacy in these cases. The figure 1 below illustrates it. The Uzbek input sentence is first segmented into so-called phrases, and then, each phrase is translated into an English phrase. Finally, phrases may be reordered. In Figure 1, the five Uzbek words and five English words are mapped as three phrase pairs.

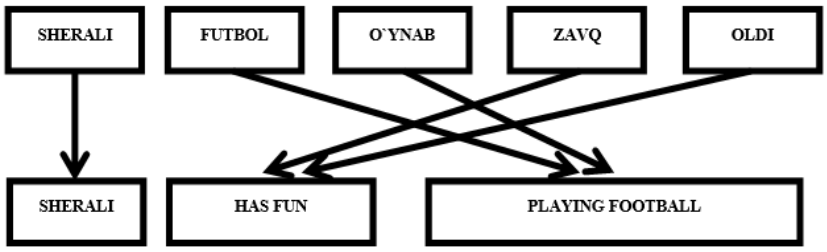


Figure 1. Phrase-based machine translation: The input is segmented into phrases, translated one-to-one into phrases in English and reordered.

The English phrases have to be reordered, so that the verb follows the subject. The Uzbek word Sherali is the subject (name of a person) so it does not translate. The verb in Uzbek “zavq oldi” can be translated in several ways, so we would like to have a translation table that maps. A phrase translation table of English translations for the phrase translation table of Uzbek phrase “zavq oldi” may look like as following:

Uzbek	Translation in English	Probability p(e|f)
Zavq oldi	Has fun	0.5
	Enjoyed	0.3
	Took pleasure	0.15

One of the phrases in Figure 1 is “has fun”. This is an unusual grouping. If we translate word-by-word “zavq”-->enjoyment, “oldi”-->took. Therefore, meaning of the sentence would change dramatically if we translated word by word. In figure 1 example, the phrase changed words depending on context of a sentence. From the example, we have learnt benefits of translation based on phrases instead of words. First, words may not be the best atomic units for translation, due to frequent one-to-many mappings. Secondly, translating group of words instead of single words helps to resolve translation ambiguities. In addition, another advantage is if we have large training corpora, we can learn longer useful phrases. Lastly, the phrase-based model is conceptually much simpler.

Phrase-based model mathematical definition

In this section, we will illustrate the phrase-based statistical machine translation model mathematically. First, we apply the Bayes rule to invert the translation direction and integrate a language model. Therefore, the best English translation e best for an Uzbek input sentence f is defined as

$$\begin{aligned} \mathbf{e}_{\text{best}} &= \text{argmax}_{\mathbf{e}}\, p(\mathbf{e}|\mathbf{f}) \\ &= \text{argmax}_{\mathbf{e}}\, p(\mathbf{f}|\mathbf{e})\, p_{\text{LM}}(\mathbf{e}) \end{aligned}$$

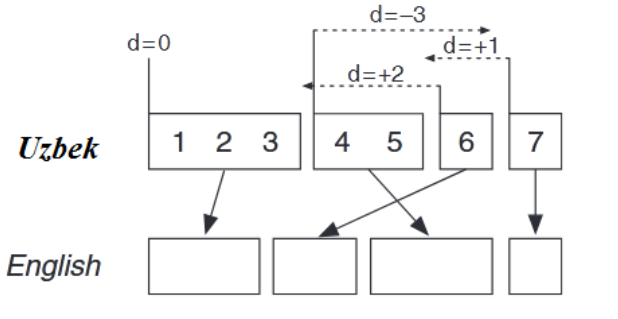
The phrase-based model, we decompose p(f|e) further into

$$p(\bar{f}_1^I|\bar{e}_1^I) = \prod_{i=1}^{I} \phi(\bar{f}_i|\bar{e}_i)\, d(\text{start}_i - \text{end}_{i-1} - 1)$$

The Uzbek sentence f is broken up into I phrases f̄i. Note that this process of segmentation is not modeled explicitly. This means that any segmentation is equally likely. Each Uzbek phrase f̄i is translated into an English phrase ēi. Since we mathematically inverted the translation direction in the noisy channel, the phrase translation probability φ(f̄i|ēi) is modeled as a translation from English to Uzbek language.

For the reordering process, we apply a distance-based reordering model. We consider reordering relative to the previous phrase. We define start i as the position of the first word of the Uzbek input phrase that translates to the ith English phrase, and end i as the position of the last word of Uzbek phrase. The reordering distance is computed as start i − endi−1. In this case, a reordering cost of d(0) is applied. Figure 2 illustrates example.

Instead of estimating reordering probabilities from data, we apply an exponentially decaying cost function d(x) = α|x| with an appropriate value for the parameter α ∈ [0, 1] so that d is a proper probability distribution. This formula simply means



Phrase	Translates	Movement	Distance
1	1–3	start at beginning	0
2	6	skip over 4–5	+2
3	4–5	move back over 4–6	–3
4	7	skip over 6	+1

those movements of phrases over large distances are more expensive than shorter movements or no movement at all.

One thing we have to note that in phrase-based statistical machine translation model the phrase translation table is learnt from data, a predefined model does reordering.

Parallel corpora (English-Uzbek, Uzbek-English)

The statistical machine translation is a probabilistic model learns to generate outputs based on previous observations of translation examples in the given language direction. It is also known as parallel corpora.

Statistical machine translation has advantage of translating without having knowledge of specific language. However in order to get reliable translation statistical machine translation requires sufficient amount of parallel corpora where it provides explanation of words and their translation. Which mean we need to have dictionary that explains the meaning of the words as well. Hence, the main obstacle to build statistical machine translation is to develop parallel corpora of given languages. In this paper, we propose to develop Uzbek-English and English-Uzbek corpora.

COLLECTIONG THE CORPUS

In order to develop parallel corpora we need data. Data is Uzbek-English and English-Uzbek texts of various contexts. Data can be collected few different ways. We collect data from scientific paper that translated by professional translators in electronic .txt format. These papers covers more than a hundred scientific texts in Uzbek translated into English by professional translators. These scientific papers cover a wide range of fields, medicine, history, translation and politics.

EXPIREMENTS & EVALUATION

In order to illustrate the contribution of the Uzbek corpus, we will conduct statistical MT experiments in the English-Uzbek language directions. We first evaluate the quality of the translations in an English-Uzbek translation model. Then professional translators give their opinion accuracy of the translation. Once we conduct our experiment, we will evaluate Uzbek-English statistical machine translation by BLEU score. The BLEU score is an evaluation method to measure the difference between machine and human translations [8]. The BLEU measures by counting and matching n-grams in result translation to n-grams in the reference text, where unigram would be each token and a bigram comparison would be each word pair. The comparison is made regardless of word order. The BLEU method is a modification of a simple precision method.

CONCLUSION

In this paper, we have discussed statistical machine learning algorithm for Uzbek to English language. In order to achieve statistical machine translation we

have collect data as a parallel corpus. Hence, we have proposed to develop English-Uzbek corpora. Moreover, we have discussed methods of statistical machine learning. There are few methods to achieve statistical machine translation we propose phrase-based method for Uzbek to English translation. The phrase-based methods shows had better result when translating agglutinative language.

Reference

[1] Aripov, M., Sharipbay, A., Abdurakhmonova, N., Razakhova B.: Ontology of grammar rules as example of noun of Uzbek and Kazakh languages. In: Abstract of the VI International Conference “Modern Problems of Applied Mathematics and Information Technology - Al-Khorezmiy 2018”, pp. 37–38, Tashkent, Uzbekistan (2018)

[2] Abduraxmonova, N. Z. "Linguistic support of the program for translating English texts into Uzbek (on the example of simple sentences): Doctor of Philosophy (PhD) il dis. aftoref." (2018).

[3] Abdurakhmonova N. The bases of automatic morphological analysis for machine translation. Izvestiya Kyrgyzskogo gosudarstvennogo tekhnicheskogo universiteta. 2016;2 (38):12-7.

[4] Abdurakhmonova N, Tuliyev U. Morphological analysis by finite state transducer for Uzbek-English machine translation/Foreign Philology: Language. Literature, Education. 2018(3):68.

[5] Abdurakhmonova N, Urdishev K. Corpus based teaching Uzbek as a foreign language. Journal of Foreign Language Teaching and Applied Linguistics (J-FLTAL). 2019;6(1-2019):131-7.

[6] Abdurakhmonov N. Modeling Analytic Forms of Verb in Uzbek as Stage of Morphological Analysis in Machine Translation. Journal of Social Sciences and Humanities Research. 2017;5(03):89-100.

[7] Kubedinova L. Khusainov A., Suleymanov D., Gilmullin R., Abdurakhmonova N. First Results of the TurkLang-7 Project: Creating Russian-Turkic Parallel Corpora and MT Systems. Proceedings of the Computational Models in Language and Speech Workshop (CMLS 2020) co-located with 16th International Conference on Computational and Cognitive Linguistics (TEL 2020) .2020/11: 90-101

[8] Abdurakhmonova N. Dependency parsing based on Uzbek Corpus. InProceedings of the International Conference on Language Technologies for All (LT4All) 2019.

[9] A. Ismailov, M. M. A. Jalil, Z. Abdullah and N. H. A. Rahim, "A comparative study of stemming algorithms for use with the Uzbek language," 2016

3rd International Conference on Computer and Information Sciences (ICCOINS), 2016, pp. 7-12, doi: 10.1109/ICCOINS.2016.7783180.

[10] Jalil, Masita & Ismailov, Alisher & Abd Rahim, Noor Hafhizah & Abdullah, Zailani. (2017). The Development of the Uzbek Stemming Algorithm. Advanced Science Letters. 23. 4171-4174. 10.1166/asl.2017.8332.