3rd International Conference on Computer Processing
in Turkic Languages (TURKLANG 2015)

Choosing the model for solving the problem of lexical selection for English-Kazakh language pair in the free/open-source platform Apertium.

Dina Amirova[1]

Al-Farabi Kazakh National University, Information Systems Department, Al-Farabi av., 71, 050040, Almaty, Kazakhstan

Abstract

This paper describes rule-based lexical selection for English-Kazakh pair in the free/open-source platform Apertium and describes the model for lexical selection that can be applied to Kazakh language as a target language. The problem of lexical selection is one of the main tasks of word processing, which is associated with the task of word-sense disambiguation. It will be not difficult to choose the correct meaning of words to people, but for machines it isn't simple. Despite the long history of its existence, a word sense disambiguation still is the developing branch of the knowledge. The machine translation system Apertium consists of several modules. One of them is the lexical module which is considered there. This lexical module in the translation ambiguous word of input language to the target language selected one lexical form of all possible with the help of rules depending on the context. Rules are hand-written. There given examples of rules that are used to selection of right sense of ambiguous words. Also to solve this problem can be used statistical models. In the paper would be considered a statistical model, maximum entropy model, which is used for solving a problem of lexical selection. Maximum entropy model shows high accuracy in different systems. The use of two systems, rule-based selection and statistical-based selection, for solving the problem of lexical selection can give a more accurate translation of texts. In the paper will be considered the works which are done to this time for solving the problem of lexical selection.

1

* *E-mail address:* amirovatdina@gmail.com

Keywords: machine translation; lexical selection; ambiguity; Apertium; Maximum Entropy Model.

1. Introduction

Recently, the role of the Republic of Kazakhstan in the international arena increases, which leads to a significant increase of interest in our country from the world community. Today English language is recognized as international language. The official language of the Republic of Kazakhstan is Kazakh. The scope of work of translators is increasing every year. Accordingly, the creation and developing of automated translation from English into Kazakh is very important and useful for people who want to translate to Kazakh.

One of the main tasks of processing of texts is the task of a lexical selection which is connected to the task of a word sense disambiguation. It is the correct choice of the word or term in accordance with the context in which they are used. Solving the task of ambiguity is one of the central word processing task. Word-sense disambiguation is used in different areas: to improve the quality of machine translation, improve the accuracy of methods of classification and clustering texts, information retrieval and other applications.

Today, there are many algorithms and models of resolving it. Linguists distinguish next kind of ambiguity: lexical, morphological, syntactic, Let's consider lexical ambiguity. Lexical selection is a choosing one translation of the word in target language by context of source language. Lexical selection is a main task of processing language.

2. Rule-based lexical selection in Apertium platform

Apertium is a platform of machine translation which development started with financing from the governments of Spain and Catalonia at Alicante University (Universitat d'Alacant). It is a free software which is published by developers according to GNU GPL conditions. To create the new system of machine translation one needs develop linguistic data (dictionaries, rules) in accurately specified XML formats [1]. The Apertium machine translation system consists following modules: deformatter, morphological analyser, part-of-speech (POS) tagger, lexical transfer, lexical selection, structural transfer, morphological generator, post-generator [2].

In rule-based free/open source platform Apertium [3] the problem of lexical selection is solved by module of lexical selection (F.M. Tyers, M.L. Forcada 2013). The rules are written by hand. The rules of lexical selection are written a way in which translation is taken by depending located near words. Hand-written rules do

not always cover the entire context. So to solve this problem we use statistics methods and models, which connected with training corpora to generate rules automatically.

Rule-based lexical selection is written in file apertium-eng-kaz.eng-kaz.lrx for language pairs from English into Kazakh. This lexical module in the translation ambiguous word input language to the target language selects one lexical form of all possible using rules which depend on the context. All the rules are written in the XML-format.

The content of the lexical rules:

<rule> - start of rule;
<match lemma="the word in english/kazakh" defining word;
tags="part of speech" tag of the words part of speech,
for example, noun - "n", adjective - "adj", and etc.;
<select lemma="selected word" selection of a particular ambiguous word translation;
tags="part of speech" tag of the words part of speech;
</match>,
</rule> - closing of the relevant tags.

Example of lexical selection rule for ’zhas’:

```
<rule>
<match lemma="year" tags="n.pl">
<select lemma="" tags="n.*"/>
</match>
</rule>
<rule>
<match lemma="year" tags="n.pl">
<select lemma="" tags="n.*"/>
</match>
<match lemma="old" tags="adj.*"/>
</rule>
```

(Example from apertium-eng-kaz.eng-kaz.lrx)

3. Statistical-based lexical selection

Statistical-based lexical selection chooses the most likely translation with their

probability. Statistical-based lexical selection based on counting frequency of collocation or words in corpora. One of the main part of statistical machine translation system is to make corpora especially corpora of large volume. One of the difficult task is a collection of parallel corpora, in our case, to gather the corpus of Kazakh and the corpus of the English. Now we have been developing a bilingual corpus, which contains 4255 sentences. Corpora are collected from fairytales, books. Today we are training these corpora. To receive corpora-based lexical selection we need aligned corpora, which is not easy to do. Kazakh language has a complex morphology. So, some words can be aligned to several words.

Because of rule-based lexical selection do not cover all cases of ambiguity, we want to use both of type of lexical selection, which were meant above. For creating statistical-based lexical selection we collect and develop bilingual corpus. Then we are training system by adding words to monolingual dictionary of Kazakh (apertium-kaz.kaz.lexc) and English (apertium-eng-kaz.eng.dix) language and adding to bilingual dictionary (apertium-eng-kaz.kaz-eng.dix).

Maximum Entropy Model

Today there are different types of models and methods of solving the problem of lexical ambiguity, which are used to solve it. One of this model's Maximum Entropy model [4].

Model maximum entropy lexical selection includes a set of binary functions and appropriate weights for each function. The feature is defined as $h^s(t,c)^2$, where t is a translation, and c – is a source language context.

$$h^s(t,c)^2 = \begin{cases} 1, & if\ value\ t\ under\ condition\ c \\ 0, & in\ other\ cases \end{cases}$$

During the learning process each function is assigned a weight λ^s, and combining the weights as in Equation (2) gives the probability of a translation t for word s in context c.

$$p_s(t|c) = \frac{1}{Z} exp \sum_{k=1}^{n_F} \lambda_k^s h_k^s(t,c), \qquad (2)$$

where Z – is a normalizing constant. Thus, the most probable translation can be found using equation (3)

$$\hat{t} = \underset{t \in T_s s}{argmax}\ p_s(t|c) = \underset{t \in T_s s}{argmax} \sum_{k=1}^{n_F} \lambda_k^s h_k^s(t,c) \qquad (3).$$

It is important to note that the rules for the feature $h^s(t,c)^2$ will be different

depending on the language pair.

4. Results

Today there are 85 rules in English-Kazakh lexical selection. The system can translate simple phrases and sentences with ambiguity [5].

5. Conclusion

In this paper was described a lexical module for English-Kazakh language pair in the free/open-source platform Apertium, where lexical selection problem is solved by writing rules for words. In the future we would like to use maximum entropy model for more effective solving the problem of lexical selection. Because this model shows the high accuracy. We are preparing parallel corpora for English and Kazakh languages now, which are collected from fairy-tales and books. Then we would train it as statistical machine translation system has a property of «self-learning». As a method would be used supervised learning. This method based on word-alignment from corpora.

References

1. Apertium: http://en.wikipedia.org/wiki/Apertium
2. Sundetova, A., Karibayeva A., Tukeyev Ua.: STRUCTURAL TRANSFER RULES FOR KAZAKH-TO-ENGLISH MACHINE TRANSLATION IN THE FREE/OPEN-SOURCE PLATFORM APERTIUM. Proceedings of the International Conference on Computer processing of Turkic Languages ”TURKLANG’14”, Istanbul(2014)
3. The Apertium machine translation platform: http://apertium.org/
4. Francis Morton Tyers. Feasible lexical selection for rule-based machine translation. Ph.D. thesis. – Universitat d’Alacant. – May, 2013.
5. Сундетова А. М., АПЕРТИУМ ПЛАТФОРМАСЫНДАҒЫ АҒЫЛШЫН-ҚАЗАҚ МАШИНАЛЫҚ АУДАРМА ЛЕКСИКАЛЫҚ МОДУЛІ. Международная научная конференция студентов и молодых ученых «Фараби әлемі». – Алматы: «Қазақ университеті», 2014. – С. 145.