ISSN 1560-7534 • ISSN 1563-0285

ӘЛ-ФАРАБИ атындағы
ҚАЗАҚ ҰЛТТЫҚ УНИВЕРСИТЕТІ

КАЗАХСКИЙ НАЦИОНАЛЬНЫЙ
УНИВЕРСИТЕТ имени АЛЬ-ФАРАБИ

AL-FARABI KAZAKH
NATIONAL UNIVERSITY

СОВМЕСТНЫЙ ВЫПУСК

ВЫЧИСЛИТЕЛЬНЫЕ ТЕХНОЛОГИИ

Том 20

ВЕСТНИК КАЗНУ им. АЛЬ–ФАРАБИ

СЕРИЯ МАТЕМАТИКА, МЕХАНИКА, ИНФОРМАТИКА

№3(86)2015

ЧАСТЬ I

КАЗАХСКИЙ НАЦИОНАЛЬНЫЙ УНИВЕРСИТЕТ имени АЛЬ-ФАРАБИ
ИНСТИТУТ ВЫЧИСЛИТЕЛЬНЫХ ТЕХНОЛОГИЙ
СИБИРСКОГО ОТДЕЛЕНИЯ РАН

ISSN 1560-7534
ISSN 1563-0285

СОВМЕСТНЫЙ ВЫПУСК

по материалам международной научной конференции
"Вычислительные и информационные технологии в науке, технике и образовании"
(CITech-2015)
(24-27 сентября 2015 года)

ВЫЧИСЛИТЕЛЬНЫЕ ТЕХНОЛОГИИ

Том 20

ВЕСТНИК КАЗНУ им. АЛЬ-ФАРАБИ

Серия математика, механика и информатика № 3 (86)

ЧАСТЬ I

АЛМАТЫ – НОВОСИБИРСК, 2015

Table of Contents

Session I. Technological Process Automation and Control

Session II. Information Management, Processing and Security

Choosing The Model for Solving the Problem of Lexical Selection for Kazakh Language on Free/Open-Source Platform Apertium

Aidana Karibayeva, Dina Amirova, and Malika Abakan

Al-Farabi Kazakh National University, Information Systems Chairl,
Al-Farabi av., 71, 050040 Almaty, Kazakhstan
a.s.karibayeva@gmail.com,amirovatdina@gmail.com,mayerabak@gmail.com

a.s.karibayeva@gmail.com,amirovatdina@gmail.com,mayerabak@gmail.com

http://www.kaznu.kz

Abstract. This paper describes process of choosing the model for lexical selection for Kazakh language as a source and target language. We will consider existing models and methods for solving problem of lexical ambiguity. In this paper we will show models which can be applied to Kazakh language. We will consider to rule-based lexical selection and first works which be done to this time for solving this type of problem.

Keywords: machine translation, Apertium, ambiguity, lexiacal selection, HMM, maximum entropy model, MT.

1 Introduction

Today it's important to create machine translation for Kazakh language. When we create machine translation we faced with problem of ambiguity. The ambiguity is an open problem of natural language processing and each machine translation system faces it. Solving the task of ambiguity is a difficult task. Today, there are many algorithms and models of resolving it. Linguists distinguish some kind of ambiguity. There are: lexical, morphological, synsathtic. We will 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 tasks of processing language. Kazakh language has a great number of ambiguity, when translating from English. For example, word 'bet' can translated as 'face' and 'page'. We solved this ambiguity by writing rules, which we will consider at the next section [1].

2 The Apertium platform

Apertium is machine translation platform. The Apertium platform module's work seems like pipeline. The Apertium machine translation system consists following module [2]:

- Deformatter. It separates the text to be translated from the formatting tags. Formatting tags are encapsulated as 'superblanks' that are placed between words in such a way that the remaining modules see them as regular blanks.
- Morphological analyser. For each surface form (that is, for each lexical unit as it appears in the text), the morphological analyser generates one or more lexical forms composed of: lemma (dictionary or citation form), lexical category (or part-of-speech), and inflection information. The morphological analyser executes a finite-state transducer generated by compiling a morphological dictionary for the source language. Lexical units containing more than one word (multiword lexical units) are analyzed as a single lexical unit. Morphological analyser

uses a finite state transducer based on two-level rules (in the case of Kazakh, apertium-kaz.kaz.lexc, apertium-kaz.kaz.twol). This module therefore separates lexemes and processes morphological analysis, and then returns possible lexical forms.

- Part-of-speech (POS) tagger. Apertium's POS tagger is based on a statistical model based on hidden Markov models which processes the result of the application of on constraint-grammar rules (Karlsson 2005), which are used to discard some analyses using simple rules (written in apertium-kaz.kaz.rlx) based on context.
- Lexical transfer. This module uses a bilingual dictionary (apertium-eng-kaz.eng-kaz.dix) which has very simple structure. The module reads each source-language lexical form and finds one or more corresponding target-language lexical forms. Multiword units are translated as a single word.
- Lexical selection. It uses rules that select for those lexical words having many translations, one of the translations in the target language according to context. All rules are written in file apertium-eng-kaz.kaz-eng.lrx . Lexiacal selection is the focus of this paper, and will be described in section 3;
- Structural transfer. This module identifies sequences of lexical forms (phrases or segments), which need syntactical processing (handling of number, prepositions, etc.) to be translated. It uses files with rules, which specify the syntactic transformation as a cascaded process. Transfer rules, which transform lexical-form sequences into a new sequences for the target language, perform the work in this module.
- Morphological generator. From the sequence of target-language lexical forms produced by the structural transfer, it generates a corresponding sequence of target language surface forms. The morphological generator executes a finite-state transducer generated by compiling a morphological dictionary for the target language.
- Post-generator. It takes care of some minor orthographical operations in the target language (for instance, it generates the English form cannot from can and not). This module is generated from file with rules which are very similar in format to dictionary files.

3 The lexical selection

The lexical selection is an open problem of each translation system. One of the main tasks of word processing is the problem of lexical choice, which is associated with the task of word-sense disambiguation. It is the correct choice of the word or term in accordance with the context in which they are used. 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.

3.1 Rule-based lexical selection

In rule-based free/open source platform Apertium [1] this problem is solved by module of lexical selection (F.M. Tyers, M.L. Forcada 2013), where rules are written by hand. As you know personal pronoun 'ol' from Kazakh is translated as 'he', 'she' and 'it' into English. We wrote the rule of lexical selection in which translation is taken by depending located near words. Generally, hand-written rules do not cover the entire context. So, we want to use statistics methods and models to solve this problem, 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, meanwhile in file apertium-eng-kaz.kaz-eng.lrx rules are written to Kazakh into English language pairs.This lexical module in the translation ambiguous word input language to the target language selected one lexical form of all possible with the help of rules depending 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 word's 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 word's 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="zhyl" tags="n.*"/>
    </match>
</rule>
<rule>
    <match lemma="year" tags="n.pl">
    <select lemma="zhas" tags="n.*"/>
    </match>
    <match lemma="old" tags="adj.*"/>
</rule>
```

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

3.2 Statistical-based lexical selection

Statistical-based lexical selection connected with corpora by counting frequency of collocation or words. When we use statistical lexical selection it means that we choose the most likely translation with their probability.

One of the important part of statistical machine translation system is to make corpora of large volumes. One of the difficult task is a collection of parallel corpora, ie, in our case, to gather the corpus of Kazakh and the corpus of the English. Presently, we have been developing a bilingual corpus, which already contains 4255 sentences. We collect this corpora from fairytales, books. Today we are training this corpora, because Apertium works with trained corpora. To receiving corpora-based lexical selection we need aligned corpora, which is not easy to do. As we mention above, all Turkic language have a complex morphlogy.So, some words can be aligned to several words.

We want to use both of type of lexical selection, which is meant above. Because, rule-based did not cover all cases for ambiguity. At the first step of creating statistical-based lexical selection we collect and develop bilingual corpus. Today, corpus-based techniques for lexical selection is widely used. Corpus is main part of any corpus-based lexical selection. Preparing corpora depends on complexity of language. As you know Kazakh has complex morphology. We collect corpus from books, fairytales. At the second step of our work 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).

4 Models

4.1 Hidden Markov model

First suitable model is Hidden Markov model (HMM), which is the main model of statistical modelling in language processing. In HMM model disambiguation is solved by assigning the probability of word. Knowing the most probable tags in context, translation system can decide which translation of word or collocation is adequate. This model requires a big corpus of Kazakh language to receive accurate translation. If a given possible translation appears aligned to a word in a given context more frequently than other possible translations, then we generate a rule which selects the aligned translation in that same context over other translations in that context [3]. We know that the main source of knowledge are dictionaries and encyclopedias. Linguists created thesauri, semantic networks and other specialized structures to establish the relationships between the values. One of the most popular model based on knowledge is a hidden Markov model . Methods based on different variations of hidden Markov model works much better, since it takes into account the context.

The problem of solving the lexical ambiguity can be reformulated as a maximization problem using the formalism of Hidden Markov Models. [16].

Hidden Markov model (Hidden Markov Model) - a statistical model that can be used to solve the classification problem of hidden variables based on observable. A Markov model is a finite state machine, the state transitions performed with a given probability. The process starts from a special starting state, then through discrete points in time it can go into new States.

In a HMM model each state with a given probability corresponds to the observed state. In accordance with the Markov assumption, the current state of the automaton depends only on a finite number of previous, with the change of the law itself states does not change over time. The number of states that need to remember to go to a new state, called the order of the model. The model may be a first order (present condition depends only on the last), n-th order (the current state is dependent on a predetermined number of preceding), and the alternating order (the order is determined according to a certain law).

Hidden Markov models are used to solve three major problems:

- calculating the probability of a observations sequence;
- calculating the most plausible explanation for the observed sequence;
- training model parameters.

For solving the lexical selection problem the words meanings serve as a hidden states, words from the text serve as observations. To use the model, you need to estimate the parameters - the matrix of transition probabilities between states (probability that after a word with the value of the test will met next) and probability matrix of observations (the probability that a given word has a predetermined value). To estimate these parameters using annotated corpus, dictionaries, network documentation. First Order HMM have complexity O (N2T), easy to understand, but for many situations receive insufficient accuracy.

4.2 Maximum entropy model

Second model is Maximum Entropy Markov Model (MEMM) that is used to resolve morphological ambiguity. Morphological ambiguity is the main study object of the problem of determining the word's parts of speech (part of speech tagging). However, modern systems are able to effectively solve the problem using methods of machine learning, such as the support vector machine method or the maximum entropy method, and show the accuracy more than 97

We decided to consider the HMM model as most suitable to Kazakh language and will use this model in machine translation systems, that have Kazakh language as a source and as a target language, namely in MT system from Kazakh into English (and vice versa) and from Russian into Kazakh language.

Maximum entropy lexical selection model includes a set of binary functions and appropriate weights for each function. The feature is defined as combination of 't' and 'c', where 't' is a translation, and 'c' – is a source language context.

5 Results

Current version of system, namely English-Kazakh(and vica versa) can translate simple phrases with ambiguity. In English-Kazakh lexical selection rules and n Kazakh-English lexical selection rules contain about 30 rules.

6 Conclusion

Current lexical selection rules translate some cases of phrases and solve problem of ambiguity in some cases. We have presented a lexical module for Kazakh language on free/open-source platform Apertium. A lexical selection problem is solved by writing rules for words. In the future we would like to use statistical model for more effective solving the problem of lexical selection, namely maximum entropy model. This model shows the high accuracy. So, now we are preparing parallel corpora for English and Kazakh language, 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 we would use supervised learning. This method based on word-alignment from corpora.

We hope that the using of this model to the problem of disambiguation for Kazakh language will give us the opportunity to have a more accurate translation of texts.

In future work is planned to generate lexical selection rules automatically from bilingual corpora to improve translation quality.

References

1. The Apertium machine translation platform: `http://apertium.org/`
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. Francis. M. Tyers, Felipe Sanshez-Martinez, Mikel L. Forcada. Flexible finite-state lexical selection for rule-based machine translation (2012)
4. Daniel Ramage. Hidden Markov Models Fundamentals. CS229 Section Notes. 2007