Ngoc Thanh Nguyen · George A. Papadopoulos
Piotr Jędrzejowicz · Bogdan Trawiński
Gottfried Vossen (Eds.)

LNAI 10449

Computational Collective Intelligence

9th International Conference, ICCCI 2017
Nicosia, Cyprus, September 27–29, 2017
Proceedings, Part II

2 Part II

∅ Springer

Lecture Notes in Artificial Intelligence 10449

Subseries of Lecture Notes in Computer Science

More information about this series at http://www.springer.com/series/1244

Ngoc Thanh Nguyen · George A. Papadopoulos
Piotr Jędrzejowicz · Bogdan Trawiński
Gottfried Vossen (Eds.)

Computational Collective Intelligence

9th International Conference, ICCCI 2017
Nicosia, Cyprus, September 27–29, 2017
Proceedings, Part II

🛡 Springer

Editors

Ngoc Thanh Nguyen
Department of Information Systems,
    Faculty of Computer Science
    and Management
Wrocław University of Science
    and Technology
Wrocław
Poland

George A. Papadopoulos
Department of Computer Science
University of Cyprus
Nicosia
Cyprus

Piotr Jędrzejowicz
Department of Information Systems
Gdynia Maritime University
Gdynia
Poland

Bogdan Trawiński
Department of Information Systems,
    Faculty of Computer Science
    and Management
Wrocław University of Science
    and Technology
Wrocław
Poland

Gottfried Vossen
Department of Information Systems
University of Münster
Münster
Germany

LRLP 2017: Special Session on Low-Resource Languages Processing

Ualsher Tukeyev	al-Farabi Kazakh National University, Kazakhstan
Zhandos Zhumanov	al-Farabi Kazakh National University, Kazakhstan
Madina Mansurova	al-Farabi Kazakh National University, Kazakhstan
Altynbek Sharipbay	L.N. Gumilyov Eurasian National University, Kazakhstan
Rustam Musabayev	Institute of Information and Computational Technologies, Kazakhstan
Zhenisbek Assylbekov	Nazarbayev University, Kazakhstan
Jonathan Washington	Swarthmore College, USA
Djavdet Suleimanov	Institute of Applied Semiotics, Russia
Alimzhanov Yermek	al-Farabi Kazakh National University, Kazakhstan

*CSDMO 2017: Special Session on Cooperative Strategies
for Decision-Making and Optimization*

Dariusz Barbucha	Gdynia Maritime University, Poland
Vincenzo Cutello	University of Catania, Italy
Ireneusz Czarnowski	Gdynia Maritime University, Poland
Joanna Jędrzejowicz	Gdansk University, Poland
Piotr Jędrzejowicz	Gdynia Maritime University, Poland
Edyta Kucharska	AGH University of Science and Technology, Poland
Antonio D. Masegosa	University of Deusto, Spain
Javier Montero	Complutense University, Spain
Ewa Ratajczak-Ropel	Gdynia Maritime University, Poland
Iza Wierzbowska	Gdynia Maritime University, Poland
Mahdi Zargayouna	IFSTTAR, France

IoT-RC 2017: Special Session on Internet of Things – Its Relations and Consequences

Ana Almeida	Porto Superior Institute of Engineering, Portugal
Jorge Bernardino	Polytechnical Institute of Coimbra, Spain
Peter Brida	University of Žilina, Slovakia
Ivan Dolnak	University of Žilina, Slovakia
Josef Horalek	University of Hradec Kralove, Czech Republic
Ondrej Krejcar	University of Hradec Kralove, Czech Republic
Goreti Marreiros	Porto Superior Institute of Engineering, Portugal
Peter Mikulecký	University of Hradec Kralove, Czech Republic
Juraj Machaj	University of Žilina, Slovakia
Marek Penhaker	VSB Technical University of Ostrava, Czech Republic
José Salmeron	Universidad Pablo de Olavide of Seville, Spain
Ali Selamat	Universiti Teknologi Malaysia, Malaysia
Vladimir Sobeslav	University of Hradec Kralove, Czech Republic
Stylianakis Vassilis	University of Patras, Greece
Petr Tucnik	University of Hradec Kralove, Czech Republic

Low Resource Language Processing

Computer Vision Techniques

Intelligent Processing of Multimedia in Web Systems

Combined Technology of Lexical Selection in Rule-Based Machine Translation

Ualsher Tukeyev[✉], Dina Amirova, Aidana Karibayeva,
Aida Sundetova, and Balzhan Abduali

Information Systems Department, Al-Farabi Kazakh National University,
Al-Farabi av., 71, 050040 Almaty, Kazakhstan
ualsher.tukeyev@gmail.com, amirovatdina@gmail.com,
a.s.karibayeva@gmail.com, sun27aida@gmail.com,
balzhanabdualy@gmail.com,
http://www.kaznu.kz

Abstract. This paper describes process of solving the task of lexical selection for English-Kazakh (and vice versa) machine translation system based on combined technology. Proposed combined technology is including the constraint grammar model and maximum entropy model for more effective solution of the problem of lexical selection for English-Kazakh (and vice-versa) language pair. Results are presented by comparing two technologies separately and together in Apertium English-Kazakh (and vice versa) system.

Keywords: Lexical selection · Combined technology · Machine translation · Source language · Target language · Sense · Ambiguity

1 Introduction

Machine translation has two main problems in text processing. First of it is lexical selection, which connected with problem of choosing corresponded translation by context. Second problem of a lexical words is order. Later consider the task of words order in sentence of target language. In word processing the main problem is lexical selection, which leads to disambiguation task.

The work of the translator's program is carried out in several stages. The development of algorithms that allow us to recreate the human ability to understand and choose the right meaning of a word is a difficult task. It consists in choosing the most suitable translation in the context under consideration from all possible translations. The solution of the problem involves finding the probable meanings of words, determining the relationships between these values and the context in which the words were used.

The scientific novelty of paper is in developing combined technology of lexical selection based on the constraint grammar model and maximum entropy model and them sequence applying to Kazakh language as source and target language in translation pairs Kazakh-English and English-Kazakh.

2   Related Works

Scientific research on a word sense disambiguation has long-term history. With a current of years the number of the proposed solutions and their efficiency steadily grew, but the task hasn't received the full decision yet. A large number of methods have been investigated: from the methods based on knowledge, rules, lexicographic sources, training with the teacher at corpus to training methods without teacher, clustering words on the basis of meaning. Among listed, today, training methods with the teacher have shown the best efficiency.

The problem of resolving ambiguity as a separate problem was formulated in the middle of 20th century, almost simultaneously with the advent of machine translation. Since that time, many methods of solving this problem have been developed, but it is still actual.

Tyers et al. in [1] uses the maximum entropy model for performing lexical selection in machine-based translation systems based on rules to English-Basque, English-Catalan. As a learning method, the method of teaching without the teacher (unsupervised methods) is used, which does not require an annotated corpus. The system uses the maximum entropy formalism for lexical selection as in Berger et al. [2] and Mareček et al. [3], but instead of counting the actual lexical selection event in the annotated corpus, they consider fractional occurrences of these events according to the model of the target language. Mareček et al. in [3] showed the model of the maximum entropy of translation in machine translation based on dependencies, which allows to develop a large number of features of functions in order to obtain more accurate translations. Tyers in [4] the general method of teaching rules for the module is described. Monolingual and bilingual corpora can be used for the method. For learning a monolingual corpus the method without the teacher (unsupervised method) is used. Also weighting method is described, based on the principle of maximum entropy. This method allows to take into account all the rules without having to choose between conflicting and overlapping rules.

Rule-based approaches in lexical selection not cover all possible translation of source language. Since it is not always possible to take into account all rules. In the statistical approach, lexical choice also does not guarantee for the correctness of the translation in context. Analyzing both approaches, we came to the conclusion about the development of technology combining the two approaches mentioned above with solutions to lexical selection problems for the Kazakh language in Apertium systems.

3   Combined Technology of Lexical Selection

The combined technology for solving the problem of lexical selection we propose consists of two models: the model of production rules (constraint grammar) and the maximum entropy models (Fig. 1).

In the process of translating a lexically ambiguous word, the system first solves the problem using the model of production rules. The model of production rules gives a good result, but does not completely solve the task. Because rules are not cover all cases of ambiguity of a certain word in any context. A word can have a different senses,
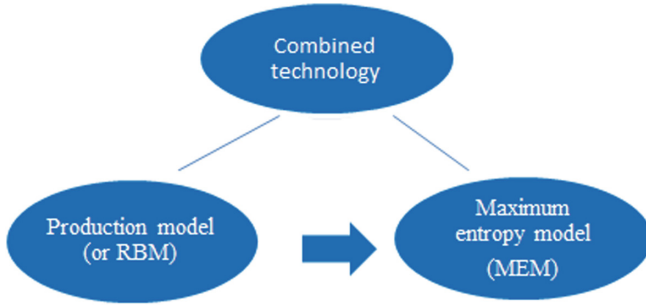
Fig. 1. The models of combined technology

and can be used with different other words in a context. And there can be many such combinations. And writing the rules can take a lot of time. Therefore, for the case when there is no written rule for some case, the selection of translation is solved by using data in semantic cube based on parallel corpora processed by statistical model. So, the ambiguous word is processed using the statistical part of the module.

The models of combined technology performed sequentially. Firstly is used the model of production rules and then maximum entropy model.

The technology of combined technology was introduced in the Apertium [5]. In the aperture, divide into two parts the lexical selection and lexical choice based on the semantic cube (Fig. 2).
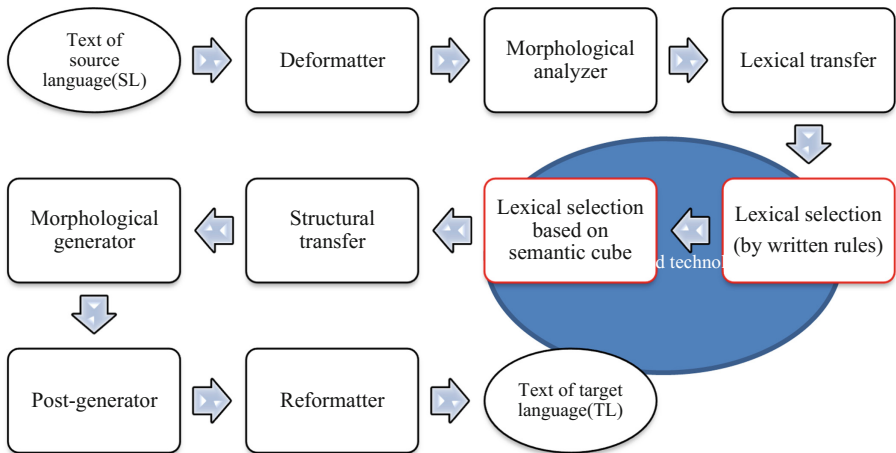
Fig. 2. The combined technology in Apertium work pipeline

3.1 Model of Production Rules

In the Apertium system, the lexical selection problem is solved by writing the rules manually in the lexical selection module using the Constraints Grammar, which has the form of product rules.

The model of product rules is a rule-based model, where knowledge is shown in the form "if (condition), then (action)". The product model can be represented in the following form:

$$i = \langle S, L, A \rightarrow B, T, \rangle$$

here S – case, L – translation, $A \rightarrow B$ – productivity kernel, T – postcondition of the production rule.

The rules of lexical selection are written in the file "apertium-eng-kaz.eng-kaz.lrx" for the English-Kazakh and in the file "apertium-eng-kaz.eng-kaz.lrx" for Kazakh-English pair of languages. All rules are written in XML format. The format of lexical rules is based on the technology of patterns. This module is used to select a translation when the meaning of ambiguous word refers to one part of the speech. For example, the word "year" is more often translated as "жыл"[zhyl], and the phrase "five years old" is translated as "бес жаста"[bes zhasta], that is, it indicates the age of a person. For this case the following rule is written:

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

The meaning of the rule: if after the noun "year" is an adjective "old", then the translation "жас"[zhas] is chosen.

For example word "бас"[bas] has various translation (Fig. 3):
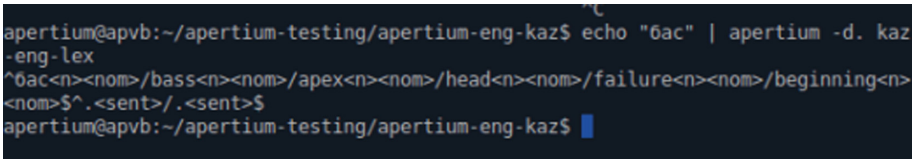
Fig. 3. The translations of word "Бас" in bilingual dictionary of English

The figures below show lexical units to word "бас(bas)". As you see given words have 4 translations.

```
<rule><match lemma= "адам" tags=n.*"/>
  <match lemma= "бас" tags="n.*">
<select lemma="head" tags="n.*"/></match></rule>
<rule>
  <match lemma="фильм" tags= "n.*"/>
  <match lemma= "бас" tags= "n.*"><select lemma= "beginning" tags=
"n.*"/></match></rule>
```

The meaning of the rule: if word "бас"[bas] comes after word "адам"[adam], the translation will be as "head", whereas when it comes after with "фильм" [film] it translated as "beginning"

The model of production rules gives a good result, but does not completely solve the task. Because rules are not cover all cases of ambiguity of a certain word in any context. Kazakh-English and English-Kazakh language pairs in Apertium currently have 98 and 75 lexical selection rules respectively. A word can have a different senses, and can be used with different other words in a context; and there can be many such combinations. And writing the rules also depending of knowledge level of developer. The lexical selection rules are written to the various part of speech, namely to noun, verb, adjective, adverb, preposition and etc. Therefore, for the case when there is no written rule for some case, the selection of translation is solved by using statistics the maximum entropy model.

3.2   Maximum Entropy Model

Lexical selection maximum entropy model includes a set of binary functions and appropriate weights for each function [4]. The feature is defined as $h^s(t, c)$, where t is a translation and c – is a source language context for each source word s:

$$h^s(t, c) = \begin{cases} 1, & c \text{ the value of } t \text{ under the condition } c \\ 0, & other \end{cases} \qquad (1)$$

During the learning process each function is assigned a weight λ^s, and combining the weights as in Eq. (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, n_F- numbers of features for s. Thus, the most probable translation can be found using Eq. (3)

$$\hat{t} = \arg \max_{t \in T_s} p_s(t|c) = \arg \max_{\substack{k=1 \\ t \in T_s}} \sum^{n_F} \lambda_k^s h_k^s (t, c) \qquad (3)$$

where T_s – all possible translations for s source word, λ- weight of known translation.

It is important to note that the rules for the function $h^s(t, c)$ will be different depending on the language pair. Consider an example: 'Mark is a bass player. He fried the bass'. In these sentences the word *bass* is ambiguous. In the first sentence, the sentence is translated into Kazakh as a "*bass guitar*", and in the second as a type of fish "*алабұға"[alabuga*]. Then the function has the form

$$h^s(t, c) = \begin{cases} 1, & if\ t =' \text{бас} - \text{гитара}'\ and\ 'player'\ after\ bass \\ & 0,\ in\ other\ cases \end{cases}$$

In combined technology the maximum entropy model is realized through the construction of a semantic cube.

4 Algorithms of Semantic Cube

The algorithm for implementation consists of the following steps:

Step 1. Create a frequency list of words.

First, to build a cube, we create a frequency list and a list of ambiguous words. The frequency list is a list of the most often met words in the corpus.

Step 2. Create a list of ambiguous words.

After composing the frequency list, it is necessary to find among them ambiguous ones, namely, the lexical ambiguity is taken into account, that is, when possible translations of the required word belong to one part of the speech (Table 1).

Table 1. Example of ambiguous words.

Ambiguous words	Part of speech	Sense 1	Sense 2	Sense 3
String	Noun	жол	жіп	ішек
Order	Noun	рет	жарлық	орден
Part	Noun	бөлік	партия	дене
Small	Adjective	кішкентай	ұсақ	шағын
Thing	Noun	зат	нәрсе	дүние
Discover	Verb	байқау	ашу	табу
Information	Noun	ақпарат	хабар	мәлімет

Step 3. Prepare bilingual parallel corpus.

Next, we prepare a bilingual corpus. Each statistic machine translation consist the greater number of text array. These arrays named as corpora. The word "prepare" means that "unnecessary" words or stop words are deleted, which have a special effect in calculating probabilities, as well as choosing the sense of the word. Such words, for example, articles like *the, a, an*, numbers, punctuation marks, etc. It is also necessary to bring all the words from the context to the basics, that is, to the initial form of the word. This completes the preparation of the corpus.

Step 4. "Training" stage. Semantic cube building.

The result of applying the maximum entropy model to solving the problem of lexical selection is the construction of a "semantic cube". At first the statistical system passes a stage of "training" at which statistical data on the translation of separate words and phrases from source language on target language are taken. The maximum entropy model is trained on a pre-prepared parallel bilingual corpus. On the basis of this stage the cube is formed. The cube represents a three-dimensional set of tables. The tables contain the senses of ambiguous words, the context, the meanings of the probabilities of the senses (translations) of ambiguous words that depend on words from the context.

There two variables can be denote as a linear releationship:

$$amb_{word} : t_i \rightarrow f_i \in C$$

Choosing or finding word's translation basically depends on two main variables - translation and context.

The semantic cube model determines the translation by weight of word's context in parallel corpora. The model gives information about frequency and probability to determined words and his neighbourhood words. The frequency of a translation is determined by the number of occurrences of the features, and when the known features are repeated, the number increases to 1.

The probability of translation calculated by next formula:

$$P_{amb\_word}\left(v_{f_{ij}}|f_n\right) = \frac{v_{f_{ij}}}{\sum_{i=1}^{n} f_n} \tag{4}$$

where, $v_{f_{ij}}$ is the frequence of a word with a certain translation, $\sum_{i=1}^{n} f_n$ – total amount of features in corpora for a particular translation.

The found probability gives the weight of a specific word translation.

The model of semantic cube chooses the translation from the corresponding set of tables of multivalued words maximizing their frequency and probability:

$$\tilde{t} = argmax_{t \in C} \sum_{i=1}^{n} P_{amb_{word}}(t_i|f_n) \tag{5}$$

The advantages of this model it works with parallel corpora.

Step 5. "Testing" stage.

Next is the "testing" stage which is made on another separate text, consisted of sentences, in which there are ambiguous words. During the translation process, this system calculates the most likely translation of the source sentence based on the data obtained during training. By comparing probabilities, as a translation for ambiguous word is selected the sense which has higher probability. It should be noted that the larger the volume of the corpus, the higher the quality of translation.

There are three outputs of the result for this technology. In the first case, when the system recognizes that there is an (lexical) ambiguous word in the sentence, then the word goes into the lexical selection module. First, the ambiguity is solved by using the model of product rules. If there is a written rule in the file for the required ambiguous word, then the translation for the word is selected based on this rule. In the second case, if there is no such rule, then a transition to the next stage occurs, where the problem is solved using the maximum entropy model. In this case, the translation for the required word will be that sense, which probability is higher. In the third case, when the translation cannot be found using the product rules model, or the maximum entropy model, as the default translation selects the meaning that is first specified in the bilingual dictionary.

Compared with other works of Berger et al. [2] and Dell Pietra et al. [6], our model differs in that the whole sentence is used as a context. Tyers et al. [1] use monolingual corpora; we use bilingual parallel corpora and a frequency list of ambiguous words

with certain senses. Tyers uses a monolingual corpus of the source language and a statistical model of the target language. In his model the volume of context for an ambiguous word is four words, two on each side. We have the context of all the words included in the sentence, with the ambiguous word.

In the below figures the results of combined technology are presented (Figs. 4, 5):

Fig. 4. The results to word "оқу"[oku] with combined technology.

At the Fig. 5 Kazakh word "күн"[kun] depending of context is translated as "sun".

Fig. 5. The results to word "күн"[kun] with combined technology

At the Fig. 6 Kazakh word "күн"[kun] depending of context is translated as "day".

Fig. 6. The translation's results to word "күн"[kun] with combined technology

In Fig. 7 is given examples for English-Kazakh language pair. The word "plant" is ambiguous word can be translated into Kazakh as "зауыт"[zauyt] and "өсімдіктер"[osimdikter]. In both situations combined technology model chose right senses, for the first "зауыт"[zauyt] and for the second "өсімдіктер"[osimdikter].

For realization of the combined technology the rules and parallel corpora are used to determine the context. By using corpora for training and testing we distinguish context for ambiguous word. To resolving the task of lexical selection we collect and use parallel corpora taken from different sources such as texts in electronic form from various famous literary novels, fairy tales, open Internet resources, news portals.
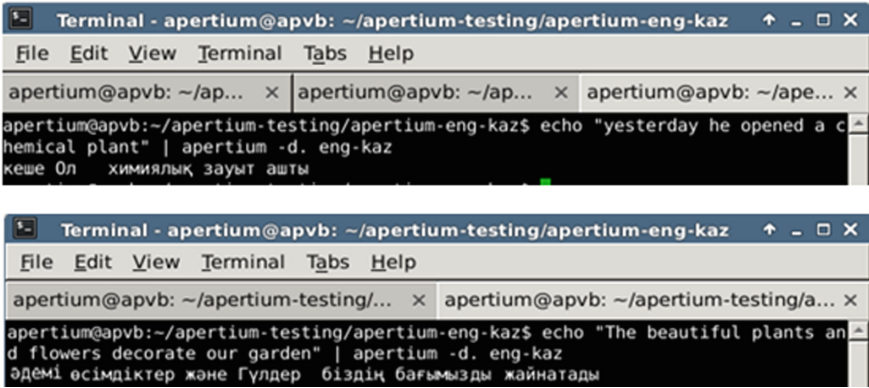
Fig. 7. The translation results for word "plant" with combined technology

To train the part of semantic cube of combined technology we use following corpora:

– corpus_lab_iis, that collected with books, fairy tales.
– Akorda collected from the official website of the President of the Republic of Kazakhstan [7]
– Egov collected from Electronic Government of the Republic of Kazakhstan [8]

We have collected the English-Kazakh parallel corpus of $\sim 30\,000$ sentences.

5 Experiments Results

In experiment for determining the translation were estimated in modes of checking productive rules, maximum entropy model and proposed combined technology in two direction of translation: Kazakh-English and English-Kazakh. Results are given in Table 2.

Table 2. Results of experiments on corpora

Language pair	Total number of sentences	Productive rule model, (%)	Maximum entropy model, (%)	Combined technology, (%)
Kazakh-English	26078	52	72	87
English-Kazakh	26078	48	65	78

By the results we can see that the proposed combined technology works better.

6   Conclusion and Future Work

In this paper was performed a combined technology uniting the model of productive rules and maximum entropy model for solving the problem of lexical selection for English-Kazakh and Kazakh-English language pairs. Also parallel bilingual English-Kazakh corpus has been developed with $\sim 30\,000$ sentences. Experiments of checking productive rules, maximum entropy model and proposed combined technology in two direction of translation: Kazakh-English and English-Kazakh shows better results of proposed combined technology of lexical selection.

In future work we plan to increase the volume of parallel corpora for receiving more exactly results in solving of lexical selection.

References

1. Tyers, F.M., Sánchez-Martınez, F., Forcada, M.L.: Unsupervised training of maximum-entropy models for lexical selection in rule-based machine translation. In: Proceedings of the 18th Annual Conference of the European Association for Machine Translation (EAMT 2015). Antalya, Turkey, pp. 145–153 (2015)
2. Berger, A., Pietra, S.D., Pietra, V.D.: A maximum entropy approach to natural language processing. Comput. Linguist. **22**(1), 39–71 (1996)
3. Mareček, D., Popel, M., Z Žabokrtský, Z.: Maximum entropy translation model in dependency-based MT framework. In: Proceedings of the Joint 5th Workshop on Statistical Machine Translation and MetricsMATR, pp. 201–206, Uppsala, Sweden, 15–16 July 2010
4. Tyers, F.M.: Feasible lexical selection for rule-based machine translation. Ph.D. thesis – Universitat d'Alicante, May 2013, 110 p.
5. https://www.apertium.org/
6. Della Pietra, S., Della Pietra, V., Lafferty, J.: Inducing features of random fields. IEEE Trans. Pattern Anal. Mach. Intell. **19**(4), 1–13 (1997)
7. Akorda. www.akorda.kz
8. Public services and information online. www.egov.kz