Lexical Selection in Machine Translation of Russian to Kazakh

D.Rakhimova, M.Abakan
Laboratory of Intelligent Information Systems, Institute of Mathematics and Mechanics, al-Farabi Kazakh National University, Almaty, Kazakhstan
di.diva@mail.ru; mayerabak@gmail.com;

Abstract

This article presents a method of resolving lexical ambiguity of words in an automatic text processing for different groups of natural languages that have not marked corpus. The proposed method is based on creating context vectors by which held the semantic analysis of the text. This method has been successfully applied in the machine translation from Russian into Kazakh. The practical results presented.

1 Introduction

Resolution of lexical ambiguity - is the establishment of the word meaning in some context [1]. For a human the process of eliminating ambiguity is largely subconscious and does not present any difficulties. Despite this, as a computational problem, the task of resolving lexical ambiguity is a difficult task. The resolution of lexical ambiguity is used to improve the accuracy of classification methods and clustering of texts, increasing the quality of machine translation, information retrieval and other applications.

The task of resolving lexical ambiguities (word sense disambiguation) occurred in 50-ies of the last century as a subtask of machine translation. Since then, researchers have proposed a great number of methods to solve this problem, but it remains more relevant today. The resolution of lexical ambiguity is one of the central tasks of text processing. To solve the problem it is necessary to identify possible meanings of words and the relationships between these meanings and the context in which words were used. At the moment, the main source of meanings are dictionaries and encyclopedias. Thesauri, semantic networks and other specialized structures are created by linguists to establish the relationships between the meanings of the words. However, creating such resources requires an enormous effort.

2 Overview of scientific works and approaches

The importance of task resolution of lexical ambiguity is difficult to overestimate. The electronic library ACL (The Association for Computational Linguistics) contains more than 700 articles on this topic [1]. Obviously, the solution of this problem is a prerequisite for a full understanding of natural language. As there is no recognized ways to determine, where the meaning of one word ends and another begins, it is problematic to formalize the task of eliminating the ambiguity.

Next, we will consider existing approaches to the definition of values, context, and comparison methods of different approaches to the resolution of lexical ambiguity.

In 1993-94 [3] David Yarowsky made the observation and determined that the length of microcontext may vary depending on the type of ambiguity.

He suggested that to resolve local ambiguities 3-4 words of context are enough, while for the semantic ambiguities a

Gönderme ve kabul tarihi: 17.09.2014-25.10.2014

larger box, consisting of 20-50 words is needed. Thus, researchers still have not come to a consensus regarding the optimal length of microcontext. Additionally, for the resolution of lexical ambiguity in some works phrases and syntactic relations are used.

So D. Yarowsky [3] found that for the same combinations of two words, the likelihood that the relevant words in the same values ranges from 90-99%.

This observation is used in many modern heuristics works. So, one value for the phrase (one sense per collocation), i.e. the appropriate words in the same phrase must have the same meaning.

Thematic context researches appeared somewhat later than microcontext and for several years was actively discussed in the field of information retrieval [4]. Modern works mainly combine thematic and microcontext approaches.

William Gale and others [5] have improved the accuracy of their method from 86% to 90%, expanding the context of the 12 words in the target environment up to 100 words. In addition, they showed that the importance of words in context decreases with distance from the target word. In their works[6] they showed that in the same thematic contexts the meanings of the corresponding ambiguity words are the same (one sense per discourse).

There is also an approach based on learning on marked blocks. The success of this approach depends on the availability of large annotated collections of texts. Rapid progress in the automatic determination of the parts of speech and syntactic analysis has been made, in particular, due to the large markup enclosures, such as Penn Treebank [7]. Models that derived from the annotated corpus methods of machine learning show good performance in many problems in natural language processing.

It is possible to distinguish two dominant approaches from the set of all existing algorithms for solving problems of lexical ambiguity.

The first approach of lexical ambiguity resolution is based on external sources of knowledge (knowledge-based methods). This approach can be easily adapted to the documents obtained from any source and not tied to a specific language. **The second approach** is based on machine learning. Algorithms based on this approach show good results in comparison with the algorithms presented in the recent literature, however, they require the training on documents similar to the processed further. This is due to the problem of sparseness of language.

3 Description of the context vector method.

Methods based on external knowledge sources have several advantages, so they attract researchers` interests. These methods can be easily adapted to the documents obtained from any source, in contrast to methods based on learning, which is applied only to the words that are available in the marked case. Another important advantage of these methods is that they do not depend on the availability of tagged corpus and can be easily applied to any other languages.

In the current work, the solution of lexical ambiguity of words based on Bag of Words (BoW) model will be proposed. The Bag of words (BoW) model is one of the two methods of representing context feature vector [7] for supervised learning technology.

Another method is a method of vector of colocational features which represents the words left and right of the target word to determine its meaning. **Method context feature vectors** (CV) is an unordered set of a certain length, the most frequent context words generated by processing a certain body of text for the target words. Then for each sense of the target word forms a binary

vector CV. In this model, the text (e.g. a sentence or a document) is represented as a set (multiset) with his words, disregarding grammar and even word order, but keeping many in the form of vectors.

The task of disambiguation in text can be easily represented as a task multivalued mappings:
Let X *and* Y — an arbitrary set. A multivalued mapping from the set X into Y is called every display :

$$F:X\rightarrow\Omega(Y),$$

which we will call this mapping from X in Y. Where each input word $x_i \in T$ of text T should be attributed to one of the output values of the classes $m_j, i \in M_i$, where M_i —the set of meanings of the word x_i. F- representation function of multivalued mappings.
To obtain knowledge about the external sources we must have information about the elements of the text (grammar , syntax properties) and relationships between them. However, a full analysis of the text, you can replace the partial. To optimize the analysis of the text we will consider the word context that used only to describe and highlight a specific group of values. As result, we will build a set of meaning vectors of allocating a noun, verb and adjectives groups, for efficiency building a complete semantic mapping for each unit complex word

Lets consider the multivalued mapping for the case when the source language is Russian and the target language is Kazakh. Consider the class of ambiguous words, which are called **lexical homonyms**, i.e. sound and grammatical match different linguistic units, which are not semantically related to each other.
For example, the word "*коса*" - braid, braiding hair, in kazakh "*бұрым*" (hairstyles), and "*коса*" spit - subject to mowing grass, kazakh "*орақ"* (agricultural tools); the word "*лук",* onion as plant, in kazakh *"пияз"*or *"лук"* like weapon for throwing arrows , in kazakh "*садақ"*.
Unlike ambiguous words, lexical homonyms do not have subject-semantic relationship, i.e. they have no common semantic features by which you could judge the polisemantism words. In this work, will be considered this kind of multiple meanings of words and will be the method of resolving this issue . Below is the segment tables of multivalued mappings (m -mappings) for ambiguous words (in this case homonyms)

$$X^m \rightarrow Y^m$$

where $X^m = \{a_k\}$, a_k- initial form of ambiguous words that have the k-th value. Y^m- represented as a matrix consisting of elements CV , that are corresponding words in context for each a_k values.

$$Y_{ij}^m = b_{ij}\mu_{ij}$$
$$Y_{ij}^m = \{(b_{1j}\mu_{1j}), (b_{2j}\mu_{2j}), (b_{3j}\mu_{3j}),\}$$

where b_{ij} - elements of a particular group of CV, $i=1,3$ (where b_1- verb group , b_2-noun group , b_3- adjective group), and for each element is given by the ratio of preference (relativity) μ_{ij} of given element in text, in the following range $0 \leq \mu_{ij} \leq 1$.
If after a full lexical and syntactic analysis of the sentence ambiguous words show up in the text, then on the basis of the approach is determined by the availability of appropriate CV words the context of a set of vectors of the multivalued mapping a_k. If such b_{ij} words of was found, then in accordance with its relativity to one or another meaning a_k meanings was selected.
Suppose that the word a_k (where $k=1,2$) have two different meaning values : a_1| a_2. If defined one or more elements from a_1 , the system output will give the required value of a_1.
In some situations, the preferred analysis and selection on the basis of the coefficient $\mu_{ij.}$, it happens when the text includes several

items CV different value.

Lets introduce the new notation $p(a_k)$- that is the number of b_{ij} in sentence for values a_k. If in the sentence will be determined :

$p(a_1) > p(a_2)$ then will be determined value a_1.; If will be determined the same number of words that are elements of CV $p(a_1)=p(a_2)$, then the decision will be made on the basis of the analysis of the coefficients μ_{ij} , with the help of which will be determined by the weight of the preferred meaning of the word in the context of proposals .

Depending on the grammatical characteristics, communication and relationship between the words for each element b_{ij} we enter a specific value as s coefficient.

For example, the basic steps performed to a particular subject will reveal its essence, therefore, the verb group was assigned the largest value from relatively nominal and prepositional groups. Preference setting to be made on the basis of comparison of the sums of the coefficients of one or another value .

$$S_k=\sum p_i(b_j)*\mu_{ij}$$

where $p_i(b_j)$ -number of elements KV i-th group for values k , i=1,2,3 and j=1,..,n Using the proposed method, the view function F multivalued mappings can be represented as a set of rules applied to the matrix elements of the context groups, and the coefficients of preference.

4 Practical results

In the implementation of the system of machine translation from Russian into Kazakh language found many difficulties in the description of grammatical rules and organization of data on different levels of the analyzer and generator. Taking into account the representation of the data and grammatical and semantic properties of different languages multivalued mapping was to present tabular data and their attributes [8]. The data for the ambiguous words were presented in m-mappings table, which is represented in Figure-1.

Database | SQL | Data | Design | DDL

Refresh

RecNo	id_omon	id_verb	koef_verb	id_noun	koef_noun	id_adj	koef_adj
				Click here to define a fi			
1	7000	7	0,4	4177	0,3	21	0,2
2	7000	23	0,4	2747	0,3	51	0,2
3	7000	57	0,4	421	0,3	52	0,2
4	7000	65	0,4	422	0,3	123	0,2
5	7000	70	0,4	3494	0,3	951	0,2
6	7000	71	0,4	2209	0,3	<null>	0
7	7000	129	0,4	2082	0,3	<null>	0
8	7000	268	0,4		0	<null>	0
9	7000	293	0,4	<null>	0	<null>	0
10	7000	396	0,4	<null>	0	<null>	0
11	7000	408	0,4	<null>	0	<null>	0
12	7002	290	0,4	<null>	0	<null>	0
13	7002	329	0,4	<null>	0	<null>	0
14	7002	2393	0,4	<null>	0	<null>	0
15	7004	3	0,4	<null>	0	<null>	0
16	7004	74	0,4	<null>	0	<null>	0
17	7004	92	0,4	<null>	0	<null>	0
18	7004	161	0,4	<null>	0	<null>	0
19	7004	301	0,4	<null>	0	<null>	0

Figure-1. The segment of m-mappings table ambiguous words of the Russian language.

For quick search and filling facilities in m-mappings table presents only the id numbers of the meanings in database dictionary not whole words.

In id_omon column fills the numbers of fundamentals of polysemantic words from the main table of the word; in columns id_verb, id_noun, id_adj respectively filled with the id number of context-related words of certain groups (verb-verbs, noun -nouns, adj- adjectives and adverbs); in columns koef_verb, koef_noun and koef_adj respectively fills the coefficients of preference μ_{ij} for each group.

For example : multivalued word "*лук*" — onion "пияз" or weapon, bow "*садақ*".

$X(a_k)::=Y(a_1)|Y(a_2)$

X(*лук*)::=Y(*пияз*)|Y(*садақ*)

where

$$Y(a_1)::=\begin{cases} b_1\ (садить, чистить, готовить, ...) \\ \quad (to\ plant, to\ clean,\ to\ cook, ...) \\ b_2\ (овощ, сад, огород, ...) \\ \quad (vegetable, garden, kitchen\ garden, ...) \\ b_3\ (большой, зеленый, сочный...) \\ \quad (big, green, juicy...) \end{cases}$$

$$Y(a_2)::=\begin{cases} b_1\ (стрелять, воевать, биться, ...) \\ \quad (to\ shoot, be\ at\ war, to\ fight, ...) \\ b_2\ (стрела, войн, стрельба, ...) \\ \quad (arrow, wars, firing, ...) \\ b_3\ (тугой, деревянный, удобный, ...) \\ \quad (taut, wooden, comfortable, ...) \end{cases}$$

Figure-2. Elements of CV for ambiguous words "лук"

Here given some examples :
*я купил **лук**.- I bought **onions/ bow**.*
*я купил зеленый **лук**.- I bought green **onions**.*
*я стрелял из **лука**.- I shot a **bow**.*
*я в саду стрелял из **лука**.- I in the garden shot a **bow**.*

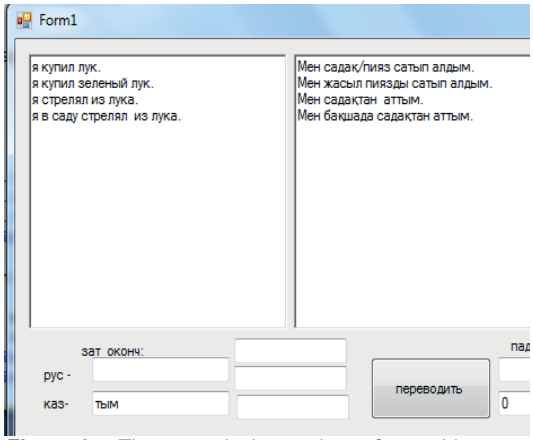

Figure-3. The practical results of machine translation of simple sentence with ambiguity words from Russian into Kazakh language.

In the example, the first sentence is limited in the context of information and in this case it is not clear which the meaning of the word *"лук"* the author means. Undefined b_{ij} – elements of CV. Therefore, the output was thrown wide range of values for post-editing by the user. In the following examples the various forms of relative context of the ambiguous words in the text was shown. In the last sentence defines two different value items CV. By the context of the action *"стрелял"* (*shot*) the proposal relates to the value of a_2, and place *"в саду" (in the garden)* the execution relates to the value of a_1.

In such conflicting situations the number and ratios of preference μ_{ij} of multiple meanings of words are considered.

For deciding the choice will be made the greatest value S_k.
$p_2(b_{ij})=1$ b_{1j} *(стрелял(shot))* $\mu_{1j}=0,4$;
$p_1(b_{ij})=1$ b_{2j} *(в саду(in the garden))* $\mu_{2j}=0,3$;

The proposed method of multivalued mappings and solving problems with multi-tasking words were applied to a simple sentence in the system of machine translation from Russian into Kazakh language. Practical implementation is done in the programming language C#, MS Visual Studio with DB SQLite Expert for 10,000 words units. The comparative analysis was done to test the resolution of the problems of ambiguity in the modern online translators (Sanasoft http://www.sanasoft.kz , Pragma6 http://translate.ua/ru/pragma-6x, Audaru http://audaru.kz) from Russian into Kazakh language. The results of the test show that considered machine translation systems do not determine the ambiguity of words and give one of the options of values to the output language, which often does not conform to the desired sense.

5 Conclusion

For the solution of problems regarding the resolution of lexical ambiguities there were tasked and solved the following tasks:

1. To determine the values for each word, related to the text;
2. To choose the most suitable value of meaning based on the context in which the word exists.

Most of the modern works are based on predefined values: lists of words found in dictionaries, translations into foreign languages, etc. The advantage of this method is an improvement of the good qualities of the classical approach based on external sources of knowledge through the application of the method of CV and multivalued mappings. In contrast to the method of the neighboring words and phrase structures, the method of CV handles all components of the sentence, and not just standing around ambiguous words. Due to this, semantically more complete analysis of the text comes out. Taking into account not only the number of elements of context, but the introduction of the preference factor for each individual item types of context vectors improved the quality of machine translation. This method can be successfully applied in various systems of automatic text processing and semantic search for a variety of natural languages.

Gratitude. This work is carried out under the grant of the Ministry of education and science of the Republic of Kazakhstan.

6 References

[1] Word Sense Disambiguation: Algorithms and Applications (Text, Speech and Language Technology), Ed. by **AGIRRE E., EDMONDS P. G.**.— 1 edition.— Springer, 2007.—November

[2] **YAROWSKY D**., One sense per collocation // HLT '93: Proceedings of the workshop on Human Language Technology.— Morristown, NJ, USA: Association for Computational Linguistics, 1993.— Pp. 266–271.

[3] **TURDAKOV D**., Recommender System Based on User-generated Content // Proceedings of the SYRCODIS 2007 Colloquium on Databases and Information Systems.— 2007.

[4] **GALE W. A., CHURCH K. W., YAROWSKY D.**, A method for disambiguating word senses in a large corpus. // Computers and the Humanitzes.— Vol. 26.— 1993.— Pp. 415–439

[5] **GALE W. A., CHURCH K. W., YAROWSKY D**. One sense per discourse // HLT '91: Proceedings of the workshop on Speech and Natural Language.— Morristown, NJ, USA: Association for Computational Linguistics, 1992.— Pp. 233–237

[6] **RICHENS R. H.**. Interlingual machine translation // Computer Journal.— Vol. 3.— 1958.— Pp. 144–147.

[7] **JURAFSKY D., MARTIN J. M**., Speech and Language Processing // second edition, Pearson Prentice Hall, New Jersey pp.640-644.

[8] **TUKEYEV U.A. , RAKHIMOVA D.R. et al.,** Development of morphological analysis and synthesis for machine translation from Russian into Kazakh using multivalued mapping tables. Computer processing of Turkic languages. First International Conference: Proceedings / Astana L.N.Gumilev ENU Publishing House, 2013, 182-191.(in Kazakh)