DEVELOPMENT OF CONCEPTUAL MODEL OF MACHINE TRANSLATION OF THE KAZAKH LANGUGE

D.R. Rakhimova

Department of Information Systems, Al-Farabi Kazakh National University,

Almaty, Kazakhstan

Abstract

In article it is presented conceptual model of machine translation. Models and algorithms of realization of the lexical and syntactic module of machine translation are described In this paper proposed an approach to machine translation of the Kazakh language using the proposed work of augmented attribute grammars.

The essence of this offer method is to create a sentences for each source (Kazakh) language of relationships, which is used to form the text of sentences of the target (Russian and English) language.

For representation of relationships of the offer in the device of the augmented attribute grammar which feature is inclusion of special semantic rules taking into account features of the Kazakh language at level of representation of words, phrases and whole sentences.

Key words: machine translation, model, Kazakh language, relationships, attribute grammar, semantics.

Introduction

Machine translation (MT) is the application of computers to the task of translating texts from one natural language to another. One of the very earliest pursuits in computer science, MT has proved to be an elusive goal, but today a reasonable number of systems are available which produce output which, if not perfect, is of sufficient quality to be useful in a number of specific domains.

Machine translation, sometimes referred to by the abbreviation MT (not to be confused with computer-aided translation, machine-aided human translation MAHT and interactive translation) is a sub-field of computational linguistics that investigates the use of software to translate text or speech from one natural language to another.

On a basic level, MT performs simple substitution of words in one natural language for words in another, but that alone usually cannot produce a good translation of a text, because recognition of whole phrases and their closest counterparts in the target language is needed. Solving this problem with corpus and statistical techniques is a rapidly growing field that is leading to better translations, handling differences in linguistic typology, translation of idioms, and the isolation of anomalies.[citation needed]

Current machine translation software often allows for customization by domain or profession (such as weather reports), improving output by limiting the scope of allowable substitutions. This technique is particularly effective in domains where formal or formulaic language is used. It follows that machine translation of government and legal documents more readily produces usable output than conversation or less standardized text.

Improved output quality can also be achieved by human intervention: for example, some systems are able to translate more accurately if the user has unambiguously identified which words in the text are names. With the assistance of these techniques, MT has proven useful as a tool to assist human translators and, in a very limited number of cases, can even produce output that can be used as is (e.g., weather reports).

The progress and potential of machine translation has been debated much through its history. Since the 1950s, a number of scholars have questioned the possibility of achieving fully automatic machine translation of high quality. Some critics claim that there are in-principle obstacles to automatizing the translation process.

History

In the 1950s, The Georgetown experiment (1954) involved fully automatic translation of over sixty Russian sentences into English. The experiment was a great success and ushered in an era of substantial funding for machine-translation research. The authors claimed that within three to five years, machine translation would be a solved problem.

Real progress was much slower, however, and after the ALPAC report (1966), which found that the ten-year-long research had failed to fulfill expectations, funding was greatly reduced. Beginning in the late 1980s, as computational power increased and became less expensive, more interest was shown in statistical models for machine translation.

The idea of using digital computers for translation of natural languages was proposed as early as 1946 by A. D. Booth and possibly others. Warren Weaver wrote an important memorandum "Translation" in 1949. The Georgetown experiment was by no means the first such application, and a demonstration was made in 1954 on the APEXC machine at Birkbeck College (University of London) of a rudimentary translation of English into French. Several papers on the topic were published at the time, and even articles in popular journals (see for example Wireless World, Sept. 1955, Cleave and Zacharov). A similar application, also pioneered at Birkbeck College at the time, was reading and composing Braille texts by computer.

Conceptual model of translation of Kazakh language

Kazakh language is refers Altai family, Turkic branch, Kypchak group, Nogai-Kypchak subgroup. According to its typology and morphological structure, Kazakh language refers to agglutinative languages. That's why new words in it are formed by adding affixes, suffixes and grammatical endings successively to the root or stem of the word. Affixes and functional words are used instead of prepositions and prefixes.

For example: *to the left - солға қарай; came - келді; till tomorrow - ертеңге дейін, for people - адамдар үшін, about people - адамдар туралы , came from Almaty - Алматыдан келді .*

There are a lot of polysemic and homonymous words.

For example: *аға - brother, аға оқытушы – senior teacher. Am -1. Name; 2. Horse; 3. Shoot.*

In linguistics there is a classification of languages on the basis of typology of a word order in the offer. It is based on an order in which a subject (subject), the predicate (verb) and a direct object (object) stand in the offer. According to this classification there are 6 possible types of languages

SVO — Subject Verb Object
SOV — Subject Object Verb
VSO — Verb Subject Object
VOS — Verb Object Subject
OSV — Object Subject Verb
OVS — Object Verb Subject

The Kazakh language belongs to SOV typology. It means that in absolute majority sentences of the Kazakh language the subject and addition will be connected with a predicate on the right, and the predicate will be connected both with a subject, and with addition at the left. The communication distance between pair subject predicate will be more than distance between pair addition predicate.

Taking into account all peculiar characteristics and difficulties of Kazakh language, there appear a lot of obstacles in machine translation development, especially at semantic analysis of a text.

Figure-1 shows a conceptual diagram of the machine translation:

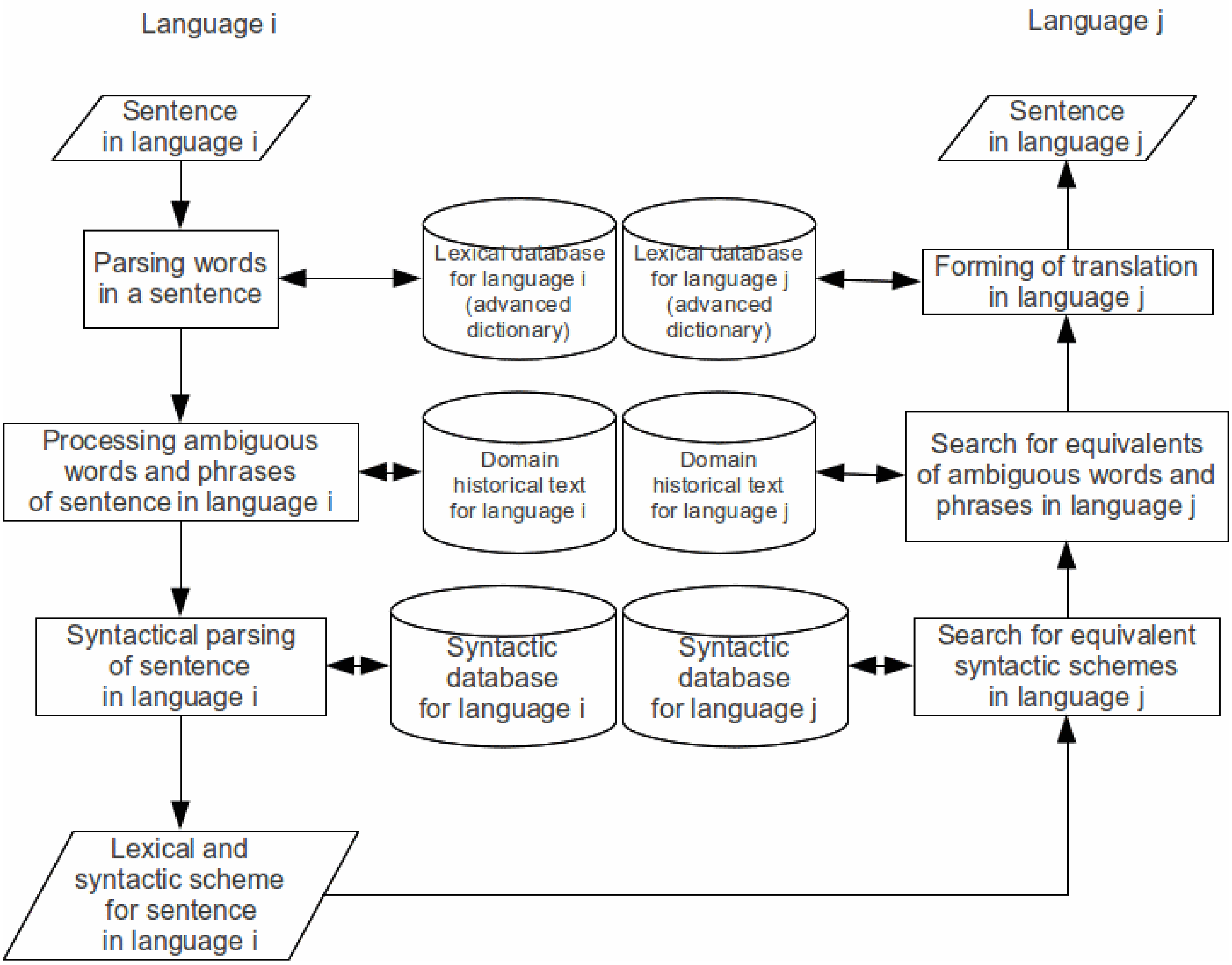

Language i
Language j
Sentence in language i
Sentence in language j
Parsing words in a sentence
Lexical database for language i (advanced dictionary)
Lexical database for language j (advanced dictionary)
Forming of translation in language j
Processing ambiguous words and phrases of sentence in language i
Domain historical text for language i
Domain historical text for language j
Search for equivalents of ambiguous words and phrases in language j
Syntactical parsing of sentence in language i
Syntactic database for language i
Syntactic database for language j
Search for equivalent syntactic schemes in language j
Lexical and syntactic scheme for sentence in language i

Fig.1. Conceptual diagram of machine translation

Analysis of different approaches in the field shows that the highest quality computer translation is obtained using the augmented attribute grammars. When using this approach in computer translation, among other things, the following modules are required:

- Lexical analysis module, carrying out an analysis of words, which make up the sentence to be translated.
- Syntactical parsing module, carrying out the process of sentence analysis for the composition.
- Ambiguous words processing module and phrases, which determines their correct translation depending on the context of use.
- Sentence structures translation module. This module takes into account strict order of words in both languages and allows by description of structure of the input language (Kazakh) to pick up its match in the target language (English, Russian).
- Database of translation dictionary for Kazakh and English/ Russian words used for direct translation from one language to another.

Development the lexical-semantics analysis of words and parse of the Kazakh language on the basis of augmented attribute grammar

In this paper we propose an approach to machine translation of the Kazakh language using the proposed work of augmented attribute grammars.

The essence of this offer method is to create a sentence for each source (Kazakh) language of relationships, which is used to form the text of sentences of the target (Russian and English) language.

For representation of relationships of the offer in the device of the augmented attribute grammar which feature is inclusion of special semantic rules taking into account features of the Kazakh language at level of representation of words, phrases and whole sentences.

The attribute grammar was offered by Donald Knut in which to each rule of context-free grammar of language the semantic rule of determination of value of terminal and non-terminal symbols of grammar is attributed. The augmented attribute grammar for the description sentences of the Kazakh language is represented in the following look is defined as:

$$AAG = <G, A, R^{W}(A), R^{F}(A), R^{S}(A)>, \quad (1)$$

where G — context-free grammar of sentences of the Kazakh language, A – a final set of semantic attributes; $R^{W}(A)$– a set of semantic rules at level of words, $R^{F}(A)$- a set of semantic rules at level of phrases of the sentence, $R^{S}(A)$- a set of semantic rules of sentence level.

To determine the semantic rules to the level of words used to describe the attribute of the *Value* of the semantic meaning of the word. At the level of lexical analysis to account for the peculiarities of the Kazakh language, we introduce two kinds of objects: real-world objects, which we denote by O^{r} and linguistic objects, which we denote by O^{l}.

For example, the word "*оқулық*" (textbook, book) consists of real-world object O^{r} "*оқу*" (to read) and the linguistic object O^{l} "*лық*" (affix of Kazakh language).

Then the word W - "оқулық"will be determined by the semantic formula:

$W.Value= O^{r}.Value \bullet O^{l}.Value$, where the •-operator of semantic connections.

Aforementioned examples show great number of prefixes, affixes and semantic values of the real world and language, all of them are called $\boldsymbol{A^{sin}}$ – syntactic attributes of our grammar.

Semantic attributes $\boldsymbol{A^{sem}}$ are calculated at the syntactic level according to semantic rules (actions) of the definite grammar and analysis of syntactic attributes. Semantic attributes (A^{sem}) are assigned to syntactic attributes' computation and application of semantic roles of these elements. The main semantic attributes have been defined, and they show object (obj), subject (sub), action (act), time (tm), place (pl), characterizing parameter (ch.pr).

In the aforementioned example the first word will have semantic attribute of an action, that is A^{sem}(obj), the second word – A^{sem}(act).

Let's examine one more example: *үстелде* (*on the table*), notional stem of the word is an object (үстел – table) + multiple derivative affixes gives semantic value of place-A^{sem}(pl)

W. $Value^{R}$(үстелде)=Rt. $Value^{R}$(үстел)•Af. $Value^{L}{}_{i}$(де)

{W. $Value^{R}$ (obj) := Rt. $Value^{R}$(obj)•Af. $Value^{L}{}_{i}$
Laws
{R(A) W. $Value^{R}$ (sub)::=…
SEMANTICS
Asem(pl):=R(Asem(sub))}}

After the analysis of elements of the entrance text and assignment of a semantic word meaning on the basis of the augmented attribute grammar we define roles (status) and relationships (links) between elements at syntactic level for the cognitive meaning of the text.

We reveal the basic semantic elements of the offer which bear the main sense of the offer. In the basic case this certain action and the subject/object, is defined as:

$$S=\{w_1,w_2,w_3,\dots,w_n\}, \quad (2)$$

where the S-input text (sentence), w - text elements (words) w_i^*- is a main conceptual elements of the sentence (text), where i = 1, .., m; m ≤ n. Based on the main elements of meaning, we allowed all links elements in the phrase (word combination).

For simplification of search of elements it is considered characteristic syntactic properties of grammar of language (in this case Kazakh). Link checked on the left side of the main elements of meaning. See figure2.

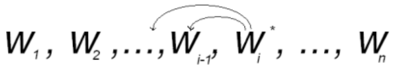


Fig.2.

Thus, we have compiled a set of phrases $\{f_k\}$ -, which are meaningful links.

$$f_k=\{w_i,w_j\} \quad (3)$$

And as the set of phrases can consist of compound relationships, i.e. the difficult (enclosed) word combinations are formed: $f_m=\{ f_k,w_j\}$ or $f_m=\{w_i, f_k \}$;

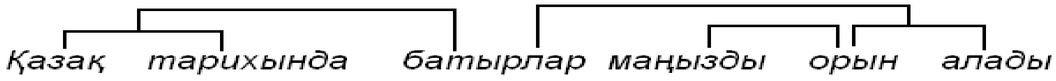
Let’s examine the following example

Қазақ тарихында батырлар маңызды орын алады- In the kazakh history heroes occupy an important place.

Let's consider interrelations between words and semantic attributes

Syntax of language of the description of the expanded attribute grammar

Such the way is found semantic attributes to each element of the sentence and their semantic relations. (See figure 3)



$w_1.A^{sem}$(ch.pr) | $w_2.A^{sem}$(pl) | $w_3.A^{sem}$(sub) | $w_4.A^{sem}$(ch.pr) | $w_5.A^{sem}$(obj) | $w_6.A^{sem}$(act)

Fig.3.

The basic semantic elements of the given sentence is w_3^* and w_6^*. Based on syntactic and semantic rules we will receive following semantic phrases:

$f_1=\{w_3,w_6\}$, $f_2=\{w_5,w_6\}$, $f_3=\{w_2,w_3\}$, $f_4=\{w_1,w_2\}$, $f_5=\{w_4,w_5\}$, $f_6=\{f_4,w_3\}$, $f_7=\{w_3,f_2\}$;

United all semantic links of elements with the account semantic actions (rules) we receive semantics (context) of the sentences.

For determination of semantic values of phrases and sentences semantic rules for group of a noun, a verb, circumstance, and also a sentence structure taking into account features of grammar of the Kazakh language are entered.

Conclusion

In this paper we consider a system of machine translation from the Kazakh language into English/Russian language. Models and algorithms of realization of the lexical and syntactic module of machine translation are described. In general, the proposing augmented attribute grammar of sentences of the Kazakh languages allows to create a private relationships proposal describing more knowledge of the proposal of the Kazakh language, algorithms that allows the machine translation of her sentences is to synthesize the target language to meet the requirements of the target language grammar. In article the method of expanded attribute grammar for improvement of quality of machine translation of the Kazakh language is presented. The developed models and algorithms were implemented in a program of machine translation.

References

[1]Knuth, D. E. (1968), Semantics of context-free languages. Mathematical Systems Theory 2, 2, pp. 127--145.

[2]D. E. Knuth: (1990), The genesis of attribute grammars. Proceedings of the international conference on Attribute grammars and their applications LNCS, vol. 461, pp. 1--12. Some informal, historical information.

[3]Paakki, J. (1995), Attribute Grammar Paradigms - A High-Level Methodology in Language Implementation, ACM Computing Surveys, 27(2), pp. 196--255.

[4]Neven, F. (2005), Attribute grammars for unranked trees as a query language for structured documents, Journal of Computer and System Sciences, 70, pp. 221--257.

[5]Matthew S. Dryer.(2005), Order of Subject, Object, and Verb . TWAC. 28 January 2005

[6]Заболеева-Зотова А.В. (2008), Атрибутная грамматика формального документа "Техническое задание"// Известия ВолгГТУ. Серия "Актуальные проблемы управления, вычислительной техники и информатики в технических системах": Межвузовский сборник научных статей. – Волгоград: ВолгГТУ,– Val.4, № 2. pp..39--43.

[7]Aho A.V.,Ullman J.D. (1978), The theory of parsing,translation and compiling, vol.1, Prentice Hall, Englewood Cliffs.

[8]Ganzinger H. (1980), Transforming denotational semantics into practical attribute grammars. Lecture Notes Сотр.Sci., vol.94, Springer-Verlag, pp. 1--64.

[9]Куликовская Л.К, Мусаева Э. Н. (2006), Грамматика казахского языка в таблицах и схемах в сопоставлении с грамматикой русского языка/Учебное пособие. Алма-ата

[10]М. Оразов (1991), Семантика казахского языка /, Алма-Ата Рауан , pp 23--68

[11]http://www.divms.uiowa.edu/~slonnegr/plf/Book/Chapter3.pdf

[12]http://www.cs.wright.edu/~tkprasad/papers/Attribute-Grammars.pdf

[13]http://paramax.susu.ru/study/TRLP.pdf

[14]http://www.trtu.h12.ru/p_209.htm

[15]http://en.wikipedia.org/wiki/Machine_translation

[16]http://ru.wikipedia.org/wiki/SOV