Augmented attribute grammar in meaning of natural languages sentences

Ualsher Tukeyev
Information systems department
Al-Farabi Kazakh National University
Almaty, Kazakhstan
ualsher.tukeyev@kaznu.kz

Diana Rakhimova
Information systems department
Al-Farabi Kazakh National University
Almaty, Kazakhstan
di.diva@mail.ru

***Abstract*- In work the approach to semantic representation in machine translation of a natural language with use of the augmented attribute grammar offered by the example of the Kazakh language as initial. The essence of the offered approach consists in creation for each sentence of initial (Kazakh) language a ontology, which is used for formation of the text of the sentence of target (Russian, English) language. For representation of ontology offers the formalism of the augmented attribute grammar which feature is inclusion of special semantic rules in view of features of initial (Kazakh) language at a level of representation of words, phrases and the sentence is offered.**

Keywords *- **grammars; natural; language; machine; translation***

INTRODUCTION

Approaches to representation of sense in processing natural languages have various directions: formalisms, based on model - theoretical semantics, in particular, it is semantics of the first order logic[6], formalisms of descriptive logics and frames [6]. In machine translation of natural languages the problem of automatic creation and representation of semantic representation is solved at a stage of the semantic analysis of texts of a natural language. At a stage of the semantic analysis the semantic structure of the source sentence is formed. At present there are some methods and tools used in the semantic analysis. It can be, for example, a tree constructed on the basis of Filmor semantic cases [1], the method of the componential analysis [2], model of conceptual dependence (1968-75 R. Schank) [3], Model of semantics of preference (1964-1972 Y.Wilks) [4]. Each of these methods has the merits and demerits. In the given work the augmented attribute grammar, which based on attribute grammar offered by Donald Knuth (1968) is submitted [5]. The given class of models can be related to a class of models of semantic expansions of syntactic rules: semantic expansions on the basis of a formalism of lamda-calculations, semantic expansions on the basis of a formalism of constraints, semantic expansions on the basis of a formalism of unifications [6]. Difference of an offered formalism of the augmented attribute grammar is in opinion of author's wider scope of grammatical objects of the language, including both a level of lexicon, a level of phrases and sentences.

KNUTH ATTRUBUTE GRAMMAR

Donald Knut has entered attribute grammar (AG), having specified static and dynamic semantics of the programming language.

Attribute grammar AG is the triplet:

$$AG = <G, A, R> \quad (1)$$

Where G the context-free grammar for language,

$$G = <N, T, P, S> \quad (2)$$

Where the N-set of non-terminals, T - set of terminals, P - set of rules, S - is initial symbol.

A-finite set of a attributes for grammar, it is possible to break into two groups still: synthesized attributes and inherited attributes (Paakki, 1995) [7].

R-set of the semantic rules, calculated on the basis of attributes.

The attribute grammar provides a basis for formalization of semantics of the language, based on its context-free grammar (or in practice of the expanded representation of Backus-Naur (Louden, 2003) [8]. For example, detonation semantics, axiomatic semantics and operational semantics of language can be specified with the help of a formalism of attribute grammar by a corresponding choice of attributes. In practice, various programming languages interpret the same syntax, and the attribute grammar forms a basis for definition and the analysis of semantic distinction.

Traditionally attribute grammar has been created and used for work with compilers. On the expiration of time and need of technical, scientific interests the attribute grammar has been applied in various works.

Attribute grammar can be used also to define kinds of activity of compilers of the generator. Certain the grammar, using introductions of methods of the attribute grammar, satisfying some restrictions, it is necessary to consider as executed specifications (Bratko, 2001, chapter 21) [9]. An example of it is system SWI-Prolog. The essence of attribute grammar can be seen in a basis of realization of algorithms of databases which is the engine for optimization of processing

978-1-4673-2743-5/12/$31.00 ©2012 IEEE

bottom-up inquiries (Ramakrishnan, R., Sudarshan, S. 1991) [10]. As have been used for development and definition of algorithms of management by the information (Thirunarayan, 2005) [11].

In the given work represent the approach in understanding and semantic processing of natural languages in machine translation, in particular for group of agglutinative languages, with use of offered augmented attribute grammar (AAG) by the example of machine translation from the Kazakh language on English (Russian). As is known, for agglutinative languages grammatical unambiguity of affixes is characteristic: one affix serves for formation of one grammatical form. The same grammatical forms are reached by the same affixes. As examples of agglutinative languages Turkic, Finno-Ugric languages can serve. Augmented attribute grammar below is considered by the example of grammar and semantics of the Kazakh language.

AUGMENTED ATTRIBUTE GRAMMAR (AAG)

The augmented attribute grammar for the description of sentences of the natural language is represented in the following kind:

$$AAG = <G, A, R^W(A), R^F(A), R^S(A)> \quad (3)$$

Where G - context-free grammar of sentences of a natural language, A - final set of semantic attributes; $R^W(A)$ - set of semantic rules at a level of words, $R^F(A)$ - set of semantic rules at a level of phrases of the sentence, $R^S(A)$ - set of semantic rules at a level of the sentence. Let's consider AAG components more in detail with their binding to stages of machine translation.

THE LEXICAL-SEMANTIC ANALYSES

For definition of semantic rules at a level of words operator ***Value*** for the characteristic of a semantic word meaning is used**.** At a level of the lexical analysis for the account of features of the Kazakh language we shall enter two kinds of objects: **objects of the real world**, we shall designate them through $\mathbf{O^r}$ and **linguistic objects**, we shall designate them through $\mathbf{O^l}$.

Let's consider an example
Оқу *(learning)*
W.Value(***оқу***)= O^r.Value (***оқу)***

Оқу*шы (learner)*
W.Value(***оқу***)= O^r.Value(***оқу***)•O^l.Value(*шы*)

Оқу*лық (textbook)*
W.Value(***оқу***)= O^r.Value(***оқу***)•O^l.Value(*лық*)

Where • the operator of semantic connection. In the first word action is described (to read), in the second example the person (pupil) is described, and in the third example the subject through which action (textbook) is carried out is described. Thus at all words one grammatical root "оқу", but depending on the attached suffix to a root of a word value of word can be various.

Thus, the semantics of the structural form of words is determined by

$$W.Value = O^R.Value \bullet O^l_i.Value \ , \ i=0\ldots n; \quad (4)$$

It allows to have rules of definition of a semantic basis of a word and its characterizing part that allows to find in the subsequent adequate components of this word in target language and to form an adequate word in target language.

In the lexical analysis of the text of the offer break into words, and words on semantic bases (roots) and linguistic units (prefixes, suffixes, etc.). As is defined to what part of speech the pledged word with the corresponding lexical characteristics belongs. For each words in compliance lexical attributes decide on a part of speech, for example for nouns it will be a sort, number, an animateness; for verbs - time, an action's perfection, the person; for adjectives - definition of kind (usual, comparative, excellent), a sort, quantity, etc. At this stage by AAG we define semantics of words with use of semantic attributes. **The semantic attributes** $\boldsymbol{A_{sem}}$ – attributes describing semantic purpose of words to semantic group.

The following main semantic groups have been determined of semantic attributes:

The name of semantic attributes	**Assignment**
A_{sem} (sub)	Attribute describing the person or object carrying out action.
A_{sem} (obj)	Attribute describing object
A_{sem} (act)	Attribute describing action.
A_{sem} (tm)	Attribute describing time
A_{sem} (pl)	Attribute describing a place
A_{sem} (ch.par)	Attribute describing characteristic parameters of object, subject, action, time or a place

Table 1 - Groups of semantic attributes

Respectively, each word should be carried (is classified) to one of these groups.

In above given example the first word *(оқу-learning)* will have lexical attributes of a verb (i.e. an initial form and the presents time) and semantic attribute of action i.e. A_{sem} (act). In the second word *(оқушы-learner)* shows that on the lexical analysis a word root *(оқу)* is the verb with a derivative suffix *(шы)* which by rules is transformed to a noun of singular and the animated form that is determined semantic attribute of the

subject of A_{sem} (sub). In the last word *(оқулық-textbook)* as an identical root and a derivative suffix *(лық)* which by rules is transformed to a noun of singular and an inanimate form that is determined the semantic attribute of object A_{sem} (obj).

THE SYNTAX AND SEMANTIC ANALYSES

The basic semantic elements of the sentence which in aggregate bear sense of the sentence are defined. For definition of semantic values of phrases and sentence semantic rules for group of a noun, a verb, circumstance, and also structure of sentence are entered in view of features of grammar of the Kazakh language.

Source sentence is submitted in the following kind:

$$S=<w_1,w_2,w_i,\ldots,w_n> \quad (5)$$

where S–the input text (sentence), w_i - elements of the text (word), i=1,..,n.

Let's enter designations of words which is the basic semantic elements of the sentence - w_j^*, j=1,..,m; m≤n.

Being based on the basic semantic elements, there are admissible communications of elements in phrase. For simplification of search of elements we consider characteristic syntactic properties of grammar of language (in this case Kazakh). In the basic grammatical rules of the Kazakh language verbs are located in a final part of the sentence, and a nouns at the beginning or on the middle (an exception can be of art text and some quoted statements). Communication is checked on the right on is left from the basic semantic elements, i.e. from a word at which semantic attribute of action (A_{sem} (act))), i.e. at first we find a word of the sentences which has attribute of action. It usually is in the Kazakh language in the sentences end.

$$W_1, W_2, \ldots, W_{i-1}, W_i, \ldots, W_n^*$$

Fig. 1

For determination of semantic values of phrases and sentences semantic rules for group of a noun, a verb, circumstance, and also a sentence structure are entered.

$$R^F(A):S => F \quad (6)$$

$$where \quad F=<f_1,f_2,..,f_k,..,f_m> \quad (7)$$

$$f_k=\{ A_{sem}w_i , A_{sem}w_j\} \quad (8)$$

On the basis of research of semantic formation of phrases the following basic kinds of semantic phrases which are described in table 2 have been developed

№	*Structural forms of phrases*	The description / the Kind of semantic phrases	The basic semantic (dominating) components in a phrase
1	A_{sem} (ch.par) A_{sem} (sub)	The characteristic description of the subject	The subject (sub)
2	A_{sem} (ch.par) A_{sem} (obj)	The characteristic description of the subject	Object (obj)
3	A_{sem} (ch.par) A_{sem} (act)	The characteristic description of the subject	Action (act)
4	A_{sem} (ch.par) A_{sem} (tm)	The characteristic description of the subject	Time (tm)
5	A_{sem} (ch.par) A_{sem} (pl)	The characteristic description of the subject	Place (pl)
6	A_{sem} (ch.par) A_{sem} (ch.par)	Complex (compound) kind characteristic description	Characteristic Parameter (ch.par)
7	A_{sem} (obj) A_{sem} (act)	Object carrying out action	Object (obj) and action (act)
8	A_{sem} (sub)A_{sem} (act)	The subject carrying out action	The subject (sub) and action (act)
9	A_{sem} (tm) A_{sem} (act)	Time of performance of action	Action (act)
10	A_{sem} (pl) A_{sem} (act)	Place of performance of action	Action (act)
11	A_{sem} (act)A_{sem} (act)	Complex (compound) kind of action	Action (act)
12	A_{sem} (sub)A_{sem} (sub)	Complex (compound) kind of the subject	The subject (sub)
13	A_{sem} (sub)A_{sem} (obj)	The subject + object	The subject (sub)
14	A_{sem} (sub)A_{sem} (tm)	The subject + time	The subject (sub)
15	A_{sem} (sub)A_{sem} (pl)	The subject + a place	The subject (sub)
16	A_{sem} (obj) A_{sem}	Object + the	Object (obj) and

	(sub)	subject	the subject (sub)
17	A_{sem} (obj) A_{sem} (obj)	Complex (compound) kind of object	Object (obj)
18	A_{sem} (obj) A_{sem} (tm)	Object + time	Object (obj)
19	A_{sem} (tm) A_{sem} (sub)	Time + the subject	The subject (sub)
20	A_{sem} (tm) A_{sem} (obj)	Time + object	Object (obj)
21	A_{sem} (tm) A_{sem} (tm)	Complex (compound) kind of time	Time (tm)
22	A_{sem} (tm) A_{sem} (pl)	Time + a place	Time (tm) and-or a place (pl)
23	A_{sem} (pl) A_{sem} (sub)	Place + the subject	The subject (sub)
24	A_{sem} (pl)A_{sem} (obj)	Place +объект	Object (obj)
25	A_{sem} (pl) A_{sem} (tm)	Place + time	Place (pl) and-or time (tm)
26	A_{sem} (pl)A_{sem} (pl)	Complex (compound) kind of a place	Place (pl)

Table 2 - The Base phrase structures constructed on semantic attributes and rules

The words having semantic attribute of characteristic parameter (A_{sem} (ch.par)), always precede to described object (the subject, to a place, action, time). The words having semantic attribute of action (A_{sem} (act)), always follows after words with other semantic attributes.

Structure and communications of base phrases were constructed on the basis of grammar rules of the Kazakh language and interrelation of semantic attributes. And in each phrase is determined key word (dominating components in a phrase definition by the table-2) which promotes further using at creation of ontology of the input sentences.

For example:

« қызыл алма » (red apple) f = {A_{sem} (ch.par) A_{sem} (obj)},

« тез келді » (quickly came) f = {A_{sem} (ch.par) A_{sem} (act)},

«үлкен бөлмеде » (in a large room)f = {A_{sem} (ch.par) A_{sem} (pl)},

«мен оқимын» (I am learning) f = {A_{sem} (sub) A_{sem} (act)},

«ауылға бардым» (went to the village) f = {A_{sem} (pl) A_{sem} (act)},

«кешке шақырылды » (invited to the evening) f = {A_{sem} (tm) A_{sem} (act)}.

At concurrences of elements of phrases, the complex (compound) form which carries only one semantic unit comes to light. When $A_{sem}w_i = A_{sem}w_{i+1}$ from the equation 8 follows

$$f=\{ A_{sem}w_i , A_{sem}w_{i+1}\} \rightarrow A_{sem}w_i \quad (9)$$

For example a phrase *«өте әдемі» (very beautiful)*

f = {A_{sem} (ch.par), A_{sem} (ch.par)};

«бара жатыр- is going» f = {A_{sem} (act) A_{sem} (act)}.

That in such cases the phrase can considered as one semantic element of the sentence.

Taking into account grammatical rules of the Kazakh language and semantic interrelations of semantic attributes it is possible to tell, that completion base phrase structures is sufficient.

The set of word-combinations can will consist of compound connections, i.e. the derivative (enclosed) phrases (word-combination) are formed

$$f_m=\{ f_k, A_{sem}w_j\} \quad \text{or} \quad f_m=\{A_{sem}w_i, f_k \} \quad (10)$$

Let's consider an example: *«ақ орамал құлады-* white scarf has fallen» the given word-combination will consist of two separate phrases.

The first *«ақ орамал-* white scarf» $f_1=\{ A_{sem}$ *(ch.par),* A_{sem} *(obj)}*, and second *«орамал құлады-* scarf has fallen» $f_2=\{ A_{sem}$ *(obj),*A_{sem} *(act)}*. In the first phrase dominating component is semantic attribute of object i.e. a semantic nucleus given phrases is the word *«орамал-scarf»*. On the basis of semantic rules and structure of base phrases

f={ A_{sem} *(obj),* A_{sem} *(act)}* the derivative form of a phrase follows, and the full semantic phrase will be submitted in next kind $f=\{ f_1, A_{sem}$ *(act)}*.

On the syntactic analyses by means of semantic rules at level of phrases R^F (A) the set of all possible combinations of semantic phrases F of the sentences that gives the main communications between text elements is calculated. To identification of the general of understanding of the text and creation of ontology to the received semantic phrases it is applied semantic rules at level of sentences R^S (A).

$$R^s(A) \ :F => O(S) \quad (11)$$

where *O(S)*- ontology of sentences *S*

Let's consider an 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 is below given. Syntax of the designs actually describing attribute calculations is resulted only. We shall consider some (complex) components: *батырлар-heroes* major importance is the subject (батыр-hero) and affix of the plural order (лар) at the analysis of semantic rules is defined semantic attribute of the subject A_{sem} (sub).

W. Value (батырлар) =O^R. Value (батыр) •O^L. Value (лар) → A_{sem} (sub): = Asem (sub)

марихында (*in a history*) a semantic basis of a word is the object *(марих-history)* and set of derivative affixes and by semantic rules gives out semantic value of place Asem (pl)

W. Value(тарих)= O^R. Value (тарих)•O^l.$Value_1$ (ын) •O^l.$Value_2$ (да) := A_{sem}(pl)

Such a way we find semantic attributes to each element of the sentence.

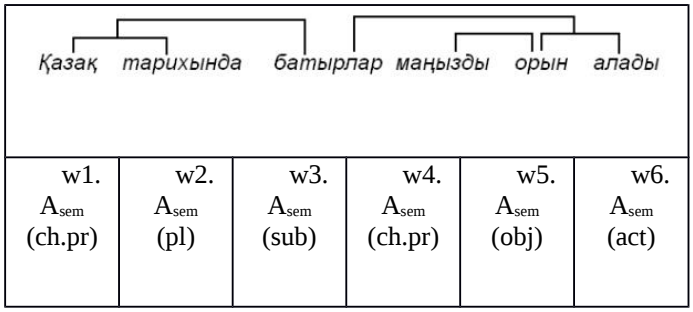

Fig. 2

For definition of semantic values of phrases and sentences semantic rules for group of a noun, a verb, circumstance, and also structure of offers are entered in view of features of grammar of the Kazakh language.

The basic semantic elements of the given offer is w_3* and w_6*. by semantic rules at a level of phrases and the sentences (***$R^F(A)$, $R^S(A)$***), we shall receive satisfying parameters of connection the following base phrases (table 2):

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

Of the received base phrases is under construction semantic combinations (derivative phrases)

$f_8 = \{f_4, w_3\}$, $f_9 = \{w_4, f_2\}$.

By association of semantic phrases in view of semantic actions ***$R^S(A)$*** (rules) the semantic structure of the sentence which will consist of two derivative semantic phrases turns out.

$O(S)=\{f_8,f_9\}$.

Thus, process of formation of ontology of sentence of a natural language will be made with the help of semantic values of elements of the sentence and interrelations (structural and semantic) between them.

THE CONCLUSION

Historically attribute grammar have been advanced in a context of construction of the compiler. In given work expansion and application of attribute grammar for the semantic analysis of the text of a natural language are illustrated. Are developed a number of the algorithms basing not on classical models of representation of knowledge of the text, and on construction ontology of sentences and the syntax-semantic description of words and the phrases, generated by using of semantic attributes and connections between them. As a whole the offered augmented attribute grammar of sentences of the Kazakh language allows to form ontology the sentences, describing all knowledge of the given sentence of the Kazakh language that allows algorithms of machine translation to synthesize on it the text of sentence in target language in view of requirements of grammar of target language.

REFERENCES

Fillmore, Charles J. (1968) "The Case for Case". In Bach and Harms (Ed.): *Universals in Linguistic Theory*. New York: Holt, Rinehart, and Winston, 1-88.

Katz J.J., Fodor J.A. (1963), The structure of the semantic theory, «Language», vol. 39, № 2, 1963, стр. 170—210.

Schank, R. 1969, *A conceptual dependency parser for natural language* Proceedings of the 1969 conference on Computational linguistics, Sång-Säby, Sweden p.1-3.

Wilks, Y. Preference semantics. In Keenan, E.L.(Ed.), The formal semantics of natural language, pp. 329-350, Cambridge Univ. Press.

Knuth, D. E. (1968), *Semantics of context-free languages*. Mathematical Systems Theory 2, 2, pp.95-96, 127-145.

Jurafsky D., Martin J.H. Speech and language processing.Pearson, Prentice Hall, 2nd ed. 2009.

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

Louden, K. C. (2003). *Programming Languages : Principles and Practice*, Second Edition, Thomson – Course Technology.

Bratko, I. (2001) *Prolog Programming for Artificial Intelligence*, Third Edition, Addison Wesley.

Ramakrishnan, R., and Sudarshan, S. (1991) Top-Down versus Bottom-Up Revisited, *Proceedings of the 1991 International Symposium on Logic Programming*, 321-336.

Thirunarayan, K. (2005) On Embedding Machine-Processable Semantics into Documents, *IEEE Transactions on Knowledge and Data Engineering*, *17*(7), 1014-1018.