Proceedings of the International Workshop on Intelligent Information Systems
IIS'2011, September 13-14, 2011, Chisinau, Republic of Moldova

Morpho-syntactic Disambiguation Using Attribute Grammars

Constantin Ciubotaru, Veronica Gîsca

Abstract: Syntactic structure of language can be defined by context-free grammars. Context-free grammars do not facilitate generation of context sensitive aspects, for example, agreement between different parts of a sentence. This allows generation of ambiguous sentences. To solve morpho-syntactic disambiguation we have proposed attribute grammars (AG). AG are extension of context-free grammars, where attributes are associated with grammar symbols, and semantic rules define values of the attributes.

Keywords: context-free grammars, attribute grammars, parser, morpho-syntactic disambiguation, bottom-up analysis, semantic rules.

1 Introduction

Natural language modeling, natural language processing is rather a lengthy process that involves detailed analysis of basic rules of communication.

A problem often encountered is the one of ambiguity. While people easily solve the problem of disambiguation, computational techniques are not sophisticated enough.

Computer operates strictly embodied elements, with algorithms and mathematical models well determined. For this reason, attempts to represent natural language by formalisms understood by the computer are made. To solve the disambiguation at morpho-syntactic level the formalism of attribute grammars (AG) is proposed.

2 Attribute Grammars

AG are extension of context-free grammars, where attributes are associated with grammar symbols, and semantic rules define values of the attributes.

Thus, certain aspects of natural language such as agreements between words, subcategories, etc. can be easily shaped.

In an attribute grammar, a set of attributes is attached to each symbol. The attribute values are calculated according to the rules attached to

grammar productions, called *semantic rules*. A semantic rule defines computation of an attribute in the left side of production – and then the attribute is called *synthesized* – or an attribute of a symbol from the right side of production – and then the attribute is called *inherited* [1].

So in formal terms the attribute grammar is defined as follows:

Definition 1 [1]. $GA = (V_T, V_N, V_S, A, P, S)$,

where V_N *– nonterminal alphabet symbols,*

V_T *– terminal alphabet symbols,*

A *– set of attributes,*

V_S *– set of semantic rules,*

P *– set of productions of type* $A \to \alpha$, *where* $\alpha \in (V_T \cup V_N)^*$;

S *– axiom.*

To demonstrate the proposed method, a simple grammar was constructed with V_T*={ v (verb), n (noun), adj (adjectiv), pron (pronoun), num (numeral), adv (adverb), art (article), pp (preposition), interj (interjection), conj (conjunction)};* V_N*={NP (noun phrase), VP (verb phrase), ADJP (adjectival phrase), PP (propositional phrase), ADVP (adverb phrase)* }[2]. The set of attributes is defined as: *A={ number, gender, case, person}.* For the rule $NP \to n\ adj$, for example, one of the semantic functions is: ***if*** *n. number = adj. number* ***&*** *n.gender = adj.gender* ***then*** *NP. number = n. number, NP.gender = n.gender; NP.case = n.case.*

Using attribute grammar more information can be formalized, which then can be used to solve problems encountered in natural language processing. One of the most difficult problems encountered in natural language processing is the ambiguity that is possibility to give two or more interpretations for a construction or its component. Often, these multiple interpretations are completely different, and in a particular context the speaker needs to choose the appropriate meaning of a word. This process is called *disambiguation*.

3 Morpho-syntactic Disambiguation

Morpho-syntactic ambiguity is characterized by a word belonging to the same or different parts of speech [2].

One word, however, can have multiple entries for different parts of speech, as having a different semantics, for example, the Romanian verb *a*

acorda can be translated in the legal field – to make an agreement and *a acorda* – in the music industry – to adjust the tone settings [4].

Therefore, the first step in achieving the morpho-syntactic disambiguation method is the annotation of each word from the sentence with lexical morphological attributes. To define the set of attributes we have used the lexicon RRRLT[1] *(Reusable Resources for the Romanian Language Technology*) developed at the Institute of Mathematics and Computer Science of the Academy of Sciences of Moldova. Computational resources include a database of words with their linguistic information.

The lexicon consists of words and their information about the morphological categories and possible syntactic functions. After annotating each word with attributes, the attribute grammar is defined which will be used for morpho-syntactic analysis of the sentence.

Semantic rules represent attribute values, calculated according to the rules attached to grammar productions. Systems based on rules solve the problem of ambiguity quiet well, but their creation is a difficult task and requires a high linguistic qualification.

Syntactic analysis techniques are used to automate the analysis of sentences. Syntactic analysis techniques used in natural language processing differ from those used for instruction parsing of programming languages. This difference comes from the fact that programming languages have a deterministically pronounced character, while in natural language the ambiguity is an obvious feature.

Syntax description of simple sentences of the Romanian language using attribute grammars allows the use of formal methods in expanding the parser.

Syntactic analysis, which cannot be a stand-alone application in natural language analysis, is used in combination with a method of semantic analysis represented by semantic rules. These rules are created to solve some problems related to the agreement between the different parts of speech. The analysis process is automated using ascending left to right (LR) analysis techniques.

There are several types of LR parsers differentiated by the structure of parsing tables and used grammars. We will use the LALR (1) parser

[1] http://imi201.math.md/elrr/

which consists of: input tape, stack, output tape and parsing tables. Parsing tables constructing is an important step that determines the efficiency of parser, because these tables take an important part of the analysis management [3].

In order to evaluate attributes during syntactic analysis, LALR(1) parser is modified by adding a parallel stack in which attribute values are stored for each terminal and nonterminal symbol. Integration of attributes evaluation with syntactic analysis has led to the use of semantic elements, thus making morpho-syntactic disambiguation.

4 Conclusion

In this article the method of morpho-syntactic disambiguation of simple sentence from Romanian using attribute grammars is described. To define the attributes set the computational linguistic resources developed at the Institute of Mathematics and Computer Science of the Academy of Sciences of Moldova was used.

The process of semantic rules evaluation was integrated with a syntactic ascending LR parser.

References

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

[2] D. Tufiş. *Automated disambiguation of words from parallel corporas using translation equivalents.* Iaşi, 2002, http://www.racai.ro/~tufis/papers/tufis-sisc12002.pdf (in Romanian).

[3] F. Hristea. *Introduction to natural language processing with apllications in Prolog.* Publishing house of the University of Bucharest, 2000, 309 p., (in Romanian).

[4] D. Irimia. *Romanian language grammar*. Polirom, 2004, 543 p., (in Romanian).

Constantin Ciubotaru[1] ,Veronica Gîsca[2] ,

[1] Institute of Mathematics and Computer Science of the Academy of Sciences of Moldova
E-mail: chebotar@math.md

[2] Institute of Mathematics and Computer Science of the Academy of Sciences of Moldova
E-mail: veronica.gisca@gmail.com