Indexed Grammars and Global Index Grammars

Faculty of Information Technology,
Brno University of Technology,
Created as a seminar work for course Modern Theoretical Computer Science,
Adam Husár,
December 18, 2007.

Abstract

This work deals with two types of grammars with controlled derivations, concretely with indexed grammars (IGs) and global index grammars (GIGs). We will provide definitions and some properties of languages generated by these grammars. It turned out that IGs have some properties that are well suited for natural languages processing so a small insight in the natural language processing will be given.

1 Introduction

One of the main interests in grammars with controlled derivations comes from the field of natural language processing. It has shown that context-free languages are not powerful enough to describe some aspects of natural languages. On the other hand, we have context-sensitive languages. Their descriptive power should be big enough to describe almost any language one can think of ([1], page 102). The problem with context-sensitive grammars is that they have some bad properties like exponential complexity of membership algorithm and undecidable emptiness problem.

It is interesting to somehow augment descriptive power of context-free languages so that they will be able to describe a subclass of context-sensitive languages. The aim there is to preserve some nice properties of context-free languages like membership decidability in polynomial time, decidable emptiness problem and closure under union, concatenation and so on. Of course it is not only the domain of natural language processing where such languages could be used, they are useful in also other field of applications like programming languages parsing.

This explains why there was an interest in finding methods to extend grammars with context-free rules by adding some controlling mechanism that controls derivations and augments grammar descriptive power.

There have been introduced some extensions like matrix grammars, random context grammars, forbidding grammars and also indexed grammars with which this paper mainly deals with.

In the following second chapter we will first look at some natural language phenomena and we will show that they cannot be described using classical context-free languages. Further in the third chapter we will introduce indexed grammars and show their properties. In fourth chapter you will be give insight in global index grammars.

2 Motivation - natural languages phenomena

In this chapter, we will take a look at some natural languages phenomena that cannot be described using context-free grammars.

2.1 Reduplication

First feature of natural language with which we will deal is reduplication. Reduplication in linguistics is a morphological process in which a root or stem[1] or part of it is repeated. Is used both in inflections to convey a grammatical function, such as plurality, intensification, etc., and in lexical derivation to create new words [2]. Leads to languages of this form ([1], page 101):

$$\{xx \mid x \in V^*\}\text{, where } V \text{ is a set of grammar symbols.} \qquad (1)$$

In Russian language can be an example of reduplication *чуть-чуть* (very few) which is created from *чуть* (a little, few). Chinese also uses reduplication: 人 *rén* (person), 人人 *rénrén* (everybody). In French language can be reduplication found too: *bon* (good), *bonbon* (bonbon). English uses some kinds on reduplication, mostly for informal expressive vocabulary like rhyming reduplications: *hocus-pokus, pell-mell,* exact reduplication: *fifty-fifty* and ablaut[2] reduplication: *see-saw*, *chit-chat.*

2.2 Multiple and crossed agreements

In languages, agreement is a form of cross-reference between different parts of a sentence or phrase. Agreement happens when one word changes in form depending on which other words it is being related to [4].

Another definition can be found in [5]: The term agreement commonly refers to some systematic covariance between a semantic or formal property of one element and a formal property of another.

For example, one does not say *I is* in English, because is cannot be used when the subject is *I*. The word *is* is said not to agree with the word *I*. This is why the grammatical form is *I am*, even though the verb still has the same function and basic meaning.

Multiple agreements can be modeled by languages of the form

$$\{a^n b^n c^n \mid n \geq 1\}, \{a^n b^n c^n d^n \mid n \geq 1\} \text{ etc. and} \qquad (2)$$

crossed agreements can be modeled by

$$\{a^n b^m c^n d^m \mid n, m \geq 1\} \text{ [1].} \qquad (3)$$

Languages can have no agreement whatsoever, as in Japanese; barely any, as in English; a small amount, as in spoken French; a moderate amount, as in Greek or Latin; or a large amount, as in Czech or other Slavic language.

[1] Root, root word, base, stem, theme, radical - the form of a word after all affixes are removed [3].

[2] Ablaut - a vowel whose quality or length is changed to indicate linguistic distinctions (such as sing sang sung song) [3].

Now, if you will look at language forms (1), (2) and (3), you can see that these languages cannot be modeled using classical context-free grammars. This can be proven using pumping lemma for context-free languages. Another approach must be used. We would still like to have grammar rules in the context-free form because of their properties. In the next chapter we will introduce indexed grammars and their extensions linear indexed grammars and global indexed grammars as a tool that allows us to model reduplication and multiple and crossed agreements.

3 Indexed Grammars

In this chapter the basic definition of an indexed grammar is presented. Also the derivation relation, generated language and some properties of this grammar type are described. Indexed grammars are one of grammars with controlled derivation or regulated grammars. Following definition comes from [1]. In article [6] by Alfred V. Aho were indexed grammars introduced.

Definition 3.1

i) An indexed grammar (IG) is a quintuple

$$G = (N, T, I, P, S),$$

where

- *N*, *T* and *S* are specified as in a context-free grammar,
- *I* is a finite set of productions of the form $A \to w$, with $A \in N$ and $w \in (N \cup T)^*$, and
- *P* is a finite set of productions of the form $A \to \alpha$, with $A \in N$ and $w \in (NI^* \cup T)^*$.

ii) For $x, y \in (NI^* \cup T)^*$, we say that *x* directly derives *y*, written as $x \Rightarrow y$, if either

$x = x_1 A\beta x_2$, for some $x_1, x_2 \in (NI^* \cup T)^*$, $A \in N$, $\beta \in I^*$,

$A \to X_1\beta_1 X_2\beta_2 ... X_k\beta_k \in P$,

$y = x_1 X_1\gamma_1 X_2\gamma_2 ... X_k\gamma_k x_2$, with $\gamma_j = \beta_j\beta$, for $X_j \in N$, and $\gamma_j = \varepsilon$, for $X_j \in T$, $1 \leqq j \leqq k$,

or

$x = x_1 Ai\beta x_2$, for some $x_1, x_2 \in (NI^* \cup T)^*$, $A \in N$, $i \in I$, $\beta \in I^*$,

$A \to X_1 X_2 ... X_k \in i$,

$y = x_1 X_1\gamma_1 X_2\gamma_2 ... X_k\gamma_k x_2$, with $\gamma_j = \beta$, *for* $X_j \in N$, and $\gamma_j = \varepsilon$, for $X_j \in T$, $1 \leqq j \leqq k$.

The first type of derivation rewrites rules from production set *P* in the same way as in derivation in context-free grammars and in addition indices associated with the rule's left-hand side non-terminal are distributed to all right-hand side non-terminals. In the second case, indices are rewritten using rules from subsets of *I*, this allows us to get rid of indices generated by rules from *P* in the sentence form.

iii) The language *L*(*G*) generated by *G* is defined as $L(G) = \{w \in T^* \mid S \Rightarrow^* w\}$.

The power of these grammars remains unchanged if we forbid erasing rules: $\mathcal{L}(I) = \mathcal{L}(\varepsilon I)$. Class of languages generated by the indexed grammar is a proper subset of class generated by the context-sensitive languages and a proper superset of class generated by the context-free languages.

Indexed grammar generated languages can be recognized using nested stack automata where every non-terminal in the actual sentence form has an stack of indices associated with it and also uses one stack as in classical context-free grammar recognition.

4 Global Index Grammars

Indexed grammars served as inspiration for G. Gazdar when he invented the linear index grammars [8]. And this is where the idea of global index grammars came from. They were introduced by José M. Castano [9], [10].

Definition 4.1

i) An global index grammar (GIG) is a 6-tuple

$$G = (N, T, I, S, \#, P),$$

where

- *N*, *T* and *S* are specified as in a context-free grammar,
- *I* is a set of stack indices,
- # is the start stack symbol and
- *P* is a finite set of productions, having the following forms:

a.i	$A \rightarrow_{\varepsilon} \alpha$	*(epsilon),*
a.ii	$A \rightarrow_{[y]} \alpha$	*(epsilon with constraints),*
b.	$A \rightarrow_{x} a\beta$	*(push),*
c.	$A \rightarrow_{\neg x} \alpha a\beta$	*(pop),*

where $x \in I$, $y \in \{I \cup \{\#\}\}$, $A \in N$, $\alpha, \beta \in (N \cup T)^*$ and $a \in T$.

Note the difference between *push* (type b) and *pop* rules (type c): *push* rules require the right-hand side of the rule to contain a terminal in the first position. *Pop* rules do not require a terminal at all. The constraint on *push* rules is a crucial property of GIGs, without it GIGs would be equivalent to a Turing Machine.

ii) Derivation relation ⇒:

a. If A $\rightarrow_{\mu} X_1 ... X_n$ is a production of type (a.), then:

i: $\delta\#\beta A\gamma \Rightarrow_{\varepsilon} \delta\#\beta X_1 ... X_n\gamma$ or

ii: $z\delta\#\beta A\gamma \Rightarrow_{[z]} z\delta\#\beta X_1 ... X_n\gamma$.

b. If A $\rightarrow_{\mu} X_1 ... X_n$ is a production of type (b.), then:

$\delta\#wA\gamma \Rightarrow_{z} z\delta\#wX_1 ... X_n\gamma$. *(leftmost derivation)*

c. If A $\rightarrow_{\mu} X_1 ... X_n$ is a production of type (c.), then:

$z\delta\#wA\gamma \Rightarrow_{\neg z} \delta\#wX_1 ... X_n\gamma$. *(leftmost derivation)*

Where $\beta, \gamma \in (N \cup T)^*, \delta \in I^*, z \in I \cup \{\varepsilon\}, w \in T^*$ and $X_i \in (N \cup T)$.

iii) The language $L(G)$ generated by G is defined as $L(G) = \{w \in T^* \mid S \Rightarrow^* w\}$.

José Castaňo describes many properties of GIG in his PhD. thesis [10]. Similarly as for indexed grammars, class of languages generated by the global index grammars is a proper subset of class generated by the context-sensitive languages and a proper superset of class generated by the context-free languages. Equivalence of IG's and GIG's power was not proven yet.

For GIG generated languages recognition can be used the two-stack pushdown automata. If we impose the same restrictions as are defined for the GIG rules and derivation relation (Push rules must have a terminal on the left of the right-hand side and for rules manipulating stack we must perform a leftmost derivation.), we get an equivalent model.

5 Conclusions

As seen in the second chapter, there are some phenomena of natural languages like multiple agreements and reduplication. They cannot be described by context-free grammars and two types of directed grammars: indexed grammars and global index grammars, which could be applied to solve these problems, were described.

Natural languages processing is a problem that, I am afraid, can be hardly solved by any syntactic approach. Natural languages are still evolving and far as I know, for the English language, there is not even a book describing rules of English grammar as there is for example for the Czech language. Especially for the English as a globally used language, in all corners of the world, the language is always a little different.

Approach called linguistic pragmatism or neopragmatism begins to be used instead. It does not concern with some syntactic rules and is based rather on some statistical description. This approach was used for example by Google in their quite well-working translator [11] and it seems that with today's computing power and enough data on the internet for creation of needed statistics, the natural language processing problem can be approached in this non-systematic way.

References

[1] Rozenberg, G., Salomaa, A. et al.: *Handbook of Formal Languages*, vol. 2. Springer Publishing, 2004.

[2] *Reduplication*. Wikipedia, 2007. Document available on the WWW: <http://en.wikipedia.org/wiki/Reduplication>.

[3] *WordNet Search - 3.0*. Princeton Universtiy, 2007. Document available on the WWW:<http://wordnet.princeton.edu/perl/webwn>

[4] *Agrement*. Wikipedia, 2007. Document available on the WWW: <http://en.wikipedia.org/wiki/Agreement_(linguistics)>

[5] Corbett, G. G.: Agreement: terms and boundaries. *Proceedings of the 2001 Texas Linguiostic society conference, 2001*. Document available on the WWW: http://www.surrey.ac.uk/LIS/SMG/projects/agreement/Papers/texas.pdf

[6] A. V. Aho, Indexed grammars - an extension of context-free grammars, *Journal of the Association for Computing Machinery*, vol. 15(4), 1968.

[7] Xiao, H.: *On NLP Formalism: from Indexed Grammars to Global Index Gramma*rs, Research Survey. Queen's University, 2005. Document available on the WWW: <http://www.cs.queensu.ca/home/xiao/gig/NLP.html>.

[8] G. Gazdar, Applicability of indexed grammars to natural languages, Natural Language Parsing and Linguistic Theories (Dordrecht), D. Reidel, 1988.

[9] Castano, J. M.: Gigs: Restricted context-sensitive descriptive power in bounded polynomial-time. Proc. of Cicling, 2003.

[10] Castano, J. M.: Global Index Languages. PhD. Thesis, The Faculty of the Graduate School of Arts and Sciences, Brandeis University, 2004. Document available on the WWW: <http://www.cs.brandeis.edu/~jcastano/thesis3.pdf>

[11] Application Google Translate. Google, 2007. Avaiable on the WWW:<http://www.google.com/translate_t>.