FORMAL MODELS IN PROCESSING OF JAPANESE LANGUAGE

Petr Horáček
Doctoral Degree Programme (1), FIT BUT
E-mail: ihoracekp@fit.vutbr.cz

Supervised by: Alexander Meduna
E-mail: meduna@fit.vutbr.cz

ABSTRACT

This paper deals with the applications of formal models in describing and processing natural languages, and the Japanese language in particular. The focus is on models that are not currently common in natural language processing – grammars with regulated rewriting (such as matrix grammars) and scattered context grammars. Examples are provided.

1 INTRODUCTION

Natural language processing plays an important role in many practical tasks, such as machine translation or information extraction. In this paper, we propose the usage of regulated grammars and scattered context grammars in the processing of the Japanese language. We present short examples of describing certain elements of the Japanese syntax with a matrix grammar and a scattered context grammar. We also formulate some problems specific for the Japanese language.

2 PRELIMINARIES

In this paper, we assume that the reader is familiar with the basic aspects of the modern formal language theory (see [4]). No prior knowledge of the Japanese language is required, although it can be an advantage (see [1]).

2.1 BASIC DEFINITIONS

Defitinion 2.1 (Context-Free Grammar) *A context-free grammar (CFG) is a quadruple* $G = (N,T,P,S)$*, where* V *is a finite set of nonterminal symbols,* T *is a finite set of terminal symbols* ($N \cap T = \emptyset$)*,* P *is a finite relation from* N *to* $(N \cup T)^*$*, usually represented as a finite set of rules of the form* $A \to x$*, where* $A \in N$ *and* $x \in (N \cup T)^*$*, and* $S \in N$ *is the start symbol.*

Defitinion 2.2 (Matrix Grammar) *A matrix grammar is a pair* $H = (G,M)$*, where* $G = (N,T,P,S)$ *is a context-free grammar and* M *is a finite language over* P ($M \subseteq P^*$)*.*

Defitinion 2.3 (Derivation in Matrix Grammar) *Let $H = (G, M)$ be a matrix grammar, where $G = (N, T, P, S)$. Let $N = A_1, \ldots, A_m$ for some $m \geq 1$. For some $m_i = p_{i_1} \ldots p_{i_j} \ldots p_{i_{k_i}} \in M$, $p_{i_j} : A_{i_j} \rightarrow x_{i_j}$. Then, for $u, v \in (N \cup T)^*$, $m \in M$, $u \Rightarrow v[m]$ in H if there are $x_0, \ldots, x_n$ such that $u = x_0, x_n = v$, and $x_0 \Rightarrow x_1[p_1] \Rightarrow x_2[p_2] \Rightarrow \ldots \Rightarrow x_n[p_n]$ in G, and $m = p_1 \ldots p_n$.*

Defitinion 2.4 (Scattered Context Grammar) *A scattered context grammar (SCG) is a quadruple $G = (N, T, P, S)$, where N is a finite set of nonterminal symbols, T is a finite set of terminal symbols ($N \cap T = \emptyset$), P is a finite set of rules of the form $(A_1, \ldots, A_n) \rightarrow (x_1, \ldots, x_n)$, where $A_1, \ldots, A_n \in N$, $x_1 \ldots, x_n \in (N \cup T)^*$, and $S \in N$ is the start symbol.*

Defitinion 2.5 (Derivation in SCG) *Let $G = (N, T, P, S)$ be a SCG. For $u, v \in (N \cup T)^*$, $p \in P$, $u \Rightarrow v[p]$ in G if $u = u_1 A_1 \ldots u_n A_n u_{n+1}$, $v = u_1 x_1 \ldots u_n x_n u_{n+1}$ and $p = (A_1, \ldots, A_n) \rightarrow (x_1, \ldots, x_n)$, where $u_i \in (N \cup T)^*$ for all $1 \leq i \leq n$.*

Further information regarding regulated rewriting and scattered context grammars may be found in [2] and [5], respectively.

3 NATURAL LANGUAGE PROCESSING AND FORMAL MODELS

Many formal models used in natural language processing (NLP) are based on context-free grammars (or context-free rules). However, while a CFG may be enough to describe a selected subset of a natural language, in general it has insufficient generative power to describe all of the important dependencies present in most natural languages.

We could solve this problem by using context-sensitive (Type-1) or even general (Type-0) grammars, but these grammars are unsuitable for practical implementation (mainly because of the complexity of parsing). Another disadvantage is that long-distance dependencies may be difficult to describe this way. Such dependencies are not uncommon in natural languages, as shown in [5], and partly in this paper.

There is a number of models based on CFG. Examples commonly used in NLP include combinatory categorial grammar (CCG; using combinatory logic) or (lexicalized) tree-adjoining grammar (LTAG; rewriting tree nodes instead of symbols). Statistical approaches are also an important part of modern NLP (see [3]). One of the basic models is a probabilistic CFG (PCFG; also called stochastic, SCFG), which assigns a constant probability to each rule – thus, some derivations become more likely than other.

In this paper, we propose the applications of other well-known formal models, such as models with regulated rewriting (matrix grammar, programmed grammar...) and scattered context grammars, in NLP. Examples of describing elements of the English language using SCG may be found in [5]. This paper focuses on potential applications for the Japanese language.

4 DESCRIBING JAPANESE LANGUAGE

4.1 DESCRIBING JAPANESE LANGUAGE USING CFG

Japanese sentences are in the *subject-object-verb* (SOV) form, with a strict word order (English, for example, uses SVO).

Consider a CFG G with the following rules (nonterminals are in capital letters, S is the start symbol):

1:	S	$\rightarrow$	SP OP VP	5:	NP	$\rightarrow$	N
2:	SP	$\rightarrow$	NP *wa*	6:	N	$\rightarrow$	*kore* | *daigaku*
3:	OP	$\rightarrow$	NP	7:	V	$\rightarrow$	*desu*
4:	VP	$\rightarrow$	V				

One of the possible derivations is:

S $\Rightarrow$ SP OP VP [1] $\Rightarrow$ NP *wa* OP VP [2] $\Rightarrow$ N *wa* OP VP [5] $\Rightarrow$ N *wa* NP VP [3] $\Rightarrow$ N *wa* N VP [5] $\Rightarrow$ N *wa* N V [4] $\Rightarrow$ *kore wa* N V [6] $\Rightarrow$ *kore wa daigaku* V [6] $\Rightarrow$ *kore wa daigaku desu* [7],

which results in a well-formed Japanese sentence, meaning "this is an university" (for the purposes of this paper, we will ignore capitalization and punctuation in example sentences). Likewise, all other successful derivations in G produce a well-formed sentence.

4.2 FORMING QUESTIONS

To change a statement into a question, we can simply append *ka* at the end of the sentence. We can cover this option by adding the following rule to G:

$$8: \quad \text{VP} \quad \rightarrow \quad \text{V } \textit{ka}$$

Taking the previous derivation example, we can now apply the new rule instead of rule 4:

S $\Rightarrow$ SP OP VP [1] $\Rightarrow$ NP *wa* OP VP [2] $\Rightarrow$ N *wa* OP VP [5] $\Rightarrow$ N *wa* NP VP [3] $\Rightarrow$ N *wa* N VP [5] $\Rightarrow$ N *wa* N V *ka* [8] $\Rightarrow$ *kore wa* N V *ka* [6] $\Rightarrow$ *kore wa daigaku* V *ka* [6] $\Rightarrow$ *kore wa daigaku desu ka* [7],

creating a correct Japanese question ("is this an university?").

Adding the following rules to G:

$$9: \quad \text{OP} \quad \rightarrow \quad \text{INT} \quad \Big| \quad 10: \quad \text{INT} \quad \rightarrow \quad \textit{nan}$$

allows us to generate:

S $\Rightarrow$ SP OP VP [1] $\Rightarrow$ NP *wa* OP VP [2] $\Rightarrow$ N *wa* OP VP [5] $\Rightarrow$ N *wa* INT VP [9] $\Rightarrow$ N *wa* INT V *ka* [8] $\Rightarrow$ *kore wa* INT V *ka* [6] $\Rightarrow$ *kore wa nan* V *ka* [10] $\Rightarrow$ *kore wa nan desu ka* [7] ("what is this?")

After adding rules 9 and 10, the following derivation also becomes possible:

S $\Rightarrow$ SP OP VP [1] $\Rightarrow$ NP *wa* OP VP [2] $\Rightarrow$ N *wa* OP VP [5] $\Rightarrow$ N *wa* INT VP [9] $\Rightarrow$ N *wa* INT V [4] $\Rightarrow$ *kore wa* INT V [6] $\Rightarrow$ *kore wa nan* V [10] $\Rightarrow$ *kore wa nan desu* [7]

But *"kore wa nan desu"* ("this is what") is not a well-formed sentence. We need to modify the grammar so that it does not allow this derivation. This may be complicated using CFG only, and we would need to add more rules and nonterminals.

However, to solve this problem, we can easily construct a matrix grammar or a scattered context grammar. We will not need any additional rules or nonterminals, and will be able to preserve the grammatical structure.

4.3 DESCRIBING JAPANESE LANGUAGE USING MATRIX GRAMMAR

Consider a matrix grammar $H = (G, M)$, where G contains the following rules (nonterminals are in capital letters, S is the start symbol):

1:	S	$\rightarrow$	SP OP VP	6:	N	$\rightarrow$	*kore* | *daigaku*
2:	SP	$\rightarrow$	NP *wa*	7:	V	$\rightarrow$	*desu*
3:	OP	$\rightarrow$	NP	8:	VP	$\rightarrow$	V *ka*
4:	VP	$\rightarrow$	V	9:	OP	$\rightarrow$	INT
5:	NP	$\rightarrow$	N	10:	INT	$\rightarrow$	*nan*

and $M = \{1, 2, 3, 4, 5, 6, 7, 8, 9 8, 10\}$.

Note that G is the same CFG we have been discussing so far. But now, according to M, every time we apply rule 9 (rewriting OP to INT), we must apply rule 8 (VP to V *ka*) immediately afterwards – this makes one derivation step in H. Thus, the derivation in question becomes:

S $\Rightarrow$ SP OP VP [1] $\Rightarrow$ NP *wa* OP VP [2] $\Rightarrow$ N *wa* OP VP [5] $\Rightarrow$ N *wa* INT V *ka* [9 8] $\Rightarrow$ *kore wa* INT V *ka* [6] $\Rightarrow$ *kore wa nan* V *ka* [10] $\Rightarrow$ *kore wa nan desu ka* [7] ("what is this?")

It is no longer possible to generate an invalid sentence.

4.4 DESCRIBING JAPANESE LANGUAGE USING SCG

Consider a SCG $G = (N, T, P, S)$ with the following rules (nonterminals are in capital letters, S is the start symbol):

1:	(S)	$\rightarrow$	(SP OP VP)	6:	(N)	$\rightarrow$	(*kore*) | (*daigaku*)
2:	(SP)	$\rightarrow$	(NP *wa*)	7:	(V)	$\rightarrow$	(*desu*)
3:	(OP)	$\rightarrow$	(NP)	8:	(VP)	$\rightarrow$	(V *ka*)
4:	(VP)	$\rightarrow$	(V)	9:	(OP, VP)	$\rightarrow$	(INT, V *ka*)
5:	(NP)	$\rightarrow$	(N)	10:	(INT)	$\rightarrow$	(*nan*)

As with the matrix grammar above, the idea here is that whenever we rewrite OP to INT, we must also rewrite VP to V *ka* (again, this is a single derivation step).

4.5 SPECIFIC PROBLEMS

There are several specific problems we encounter while trying to describe and process the Japanese language. So far we have been using only *romaji* (transcription of Japanese text using Latin alphabet) in the example sentences. However, the Japanese writing system includes three main sets of symbols:

- *kanji* – originally Chinese characters, used to represent word stems (i.e. meaning)
- *hiragana* – a syllabary, mainly used for particles and suffixes; it is also possible to write the whole text in hiragana only (so that the reader does not need to know *kanji* – for example in books for children or Japanese textbooks for beginners)

- *katakana* – another syllabary, mainly used to transcribe words from foreign languages (foreign names, loanwords...)

In modern Japanese, it is common to use the Arabic numerals (although there are *kanji* for numbers), and sometimes Latin letters are also included (for example in abbreviations such as CD, DVD). A typical Japanese sentence might look like this:

日本とチェコスロバキア間の外交関係は1919年に樹立されました。
にほんとちぇこすろばきあかんのがいこうかんけいは1919ねんにじゅりつされました。

The second line shows the same sentence written using *hiragana* only. One of the possible *romaji* transcriptions is *"Nihon to Chekosurobakia kan no gaikou kankei wa 1919 nen ni juritsu saremashita."* The meaning is "The diplomatic relations between Japan and Czechoslovakia started in 1919."

As you may notice in the example, it is not customary to separate words by spaces in Japanese sentences. In practice, we will have to find a way to split the sentence into words (or any suitable syntactic units) before we are able to analyze it using the proposed formal models. We will also need to handle the various ways of writing the same sentence.

5 CONCLUSION

As we tried to illustrate in this paper, both grammars with regulated rewriting and scattered context grammars show great promise in natural language processing. They are powerful formal models, which can describe many dependencies (including long-distance dependencies) found in natural languages in an elegant, intuitive way. At the same time, thanks to their close ties with context-free grammars, they are also relatively simple and straightforward.

ACKNOWLEDGEMENT

This work was partially supported by the BUT FIT grant FIT-S-10-2 and the research plan MSM0021630528.

REFERENCES

[1] E. Banno, Y. Ohno, Y. Sakane and C. Shinagawa. *Genki 1: An Integrated Course in Elementary Japanese*. The Japan Times, 1999

[2] J. Dassow and Gh. Păun. *Regulated Rewriting in Formal Language Theory*. Akademie-Verlag, Berlin, 1989

[3] C. D. Manning and H. Schütze. *Foundations of Statistical Natural Language Processing*. MIT Press, 1999

[4] A. Meduna. *Automata and Languages: Theory and Applications*. Springer, London, GB, 2005

[5] A. Meduna and J. Techet. *Scattered Context Grammars and their Applications*. WIT Press, UK, GB, 2009