Syntax Driven Japanese-Czech Translation

PhD Thesis

Petr Horáček

Department of Information Systems
Faculty of Information Technology
Brno University of Technology

March 10, 2011

Outline

Outline

Motivation

Natural language processing (NLP)

- Important application area of formal language theory
  - One of the key aspects behind the creation of the theory
- Currently covering a wide area of various tasks
  - Mostly aimed at practical applications

Machine translation

- One of the oldest NLP tasks (but still room for improvement)
- Practical applications

Japanese-Czech translation

- Little research in the area
- Good relations between Japan and the Czech republic
- Growing interest in Japanese culture and language in Europe

Outline

Context-free grammar

Definition

A context-free grammar (CFG) 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 relation from N to $(N \cup T)^*$, usually represented as a finite set of *rules (productions)* of the form $A \to x$, where $A \in N$ and $x \in (N \cup T)^*$
- $S \in N$ is the *start symbol*

Derivation step and generated language

Let $u, v \in (N \cup T)^*$ and $p = A \to x \in P$. Then, *uAv directly derives uxv* according to p in G, written as $uAv \Rightarrow_G uxv\ [p]$ or simply $uAv \Rightarrow uxv$.

$$L(G) = \{w : w \in T^*, S \Rightarrow^* w\}$$

Matrix grammar

Definition

A matrix grammar is a pair $H = (G, M)$, where

- $G = (N, T, P, S)$ is a context-free grammar
- M is a finite language over P ($M \subseteq P^*$)

Derivation step

For $x, y \in (N \cup T)^*$, $m \in M$,

$$x \Rightarrow y[m]$$

in H if there are $x_0, \ldots, x_n$ such that $x = x_0, x_n = y$, and

1. $x_0 \Rightarrow x_1[p_1] \Rightarrow x_2[p_2] \Rightarrow \cdots \Rightarrow x_n[p_n]$ in G, and
2. $m = p_1 \ldots p_n$

Outline

Formal language theory and natural languages

Classic formal models in NLP – problems

- CFG – insufficient generative power
- Context-sensitive and general grammars – unsuitable for practical use (complexity of parsing)

Solution in NLP

1. Find new approaches, create new models
2. Modify known models – often based on CFG

Parallel in formal language theory

- Regulated rewriting (matrix grammar, programmed grammar. . .)
- Scattered context grammar
- . . .

Machine translation

Translation systems

- Bilingual vs. multilingual
- Unidirectional vs. bidirectional

Approaches to translation

1. Direct translation
2. Interlingua
   - Internal abstract representation
3. Transfer
   - Separate abstract representation for source and target language

New trends

- Rule-based vs. corpus-based systems
- Statistical approaches

Outline

Publications

Published

- Horáček, P.: Formal Models in Processing of Japanese Language. In *Proceedings of the 16th Conference and Competition STUDENT EEICT 2010 Volume 5*, Faculty of Information Technology BUT, 2010

Submitted

- Horáček, P., Meduna, A.: Syntax Driven Japanese-Czech Translation (AFL 2011)
- Horáček, P.: Parse Driven Translation (STUDENT EEICT 2011)

Syntax driven translation

Translation grammar (basic idea)

- A grammar that generates two corresponding sentences (input and translation) in one derivation
- Based on CFG (usually)
- Each rule has two right-hand sides – one generates the input sentence, other the corresponding output sentence
- One left-hand side – always rewriting the same nonterminal

Example

- Rule:

$$1 : E \rightarrow E + T \ , E \ T +$$

- Derivation step:

$$(E, E) \Rightarrow (E + T, E \ T +) \ [1]$$

Parse driven translation

Idea

- Based on the the idea of syntax driven translation and translation grammars
- Two grammars (input and output), corresponding rules share labels
- Input sentence and output sentence – same parse (sequence of rules used in derivation, denoted by their labels)
- Example – rules:

Input grammar	Output grammar
$1 : E \rightarrow E + T$	$1 : E \rightarrow E\ T\ +$

- Note: the two corresponding rules do not need to rewrite the same nonterminal

Parse driven translation

Translation in practice (idea)

1. Parse the input sentence using input grammar – we get a sequence of rules (parse)

$$S_I \Rightarrow^* x_I[\alpha]$$

2. Generate the translation using output grammar – apply the rules of output grammar according to the sequence from step 1

$$S_O \Rightarrow^* x_O[\alpha]$$

Parse translation grammar

Definition

A parse translation grammar is a 5-tuple $H = (G_I, G_O, \Psi, \varphi_I, \varphi_O)$, where

- $G_I = (N_I, T_I, P_I, S_I)$ and $G_O = (N_O, T_O, P_O, S_O)$ are CFGs, card P_I = card P_O = card Ψ,
- Ψ is a set of *rule labels*,
- φ_I is a bijection from Ψ to P_I and φ_O a bijection from Ψ to P_O.

Translation

Translation $T(H)$ is a set of pairs of sentences:

$$T(H) = \{(w_I, w_O) : \quad w_I \in T_I^*, w_O \in T_O^*, \\ S_I \Rightarrow_{G_I}^* w_I[\alpha], S_O \Rightarrow_{G_O}^* w_O[\alpha], \\ \alpha \in \Psi^*\}$$

Parse translation matrix grammar

Definition

A parse translation matrix grammar is a 7-tuple $H = (G_I, M_I, G_O, M_O, \Psi, \varphi_I, \varphi_O)$, where

- (G_I, M_I) and (G_O, M_O) are matrix grammars, card $M_I =$ card $M_O =$ card Ψ,
- Ψ is a set of *matrix labels*,
- φ_I is a bijection from Ψ to M_I and φ_O a bijection from Ψ to M_O.

Translation

Translation $T(H)$ is a set of pairs of sentences:

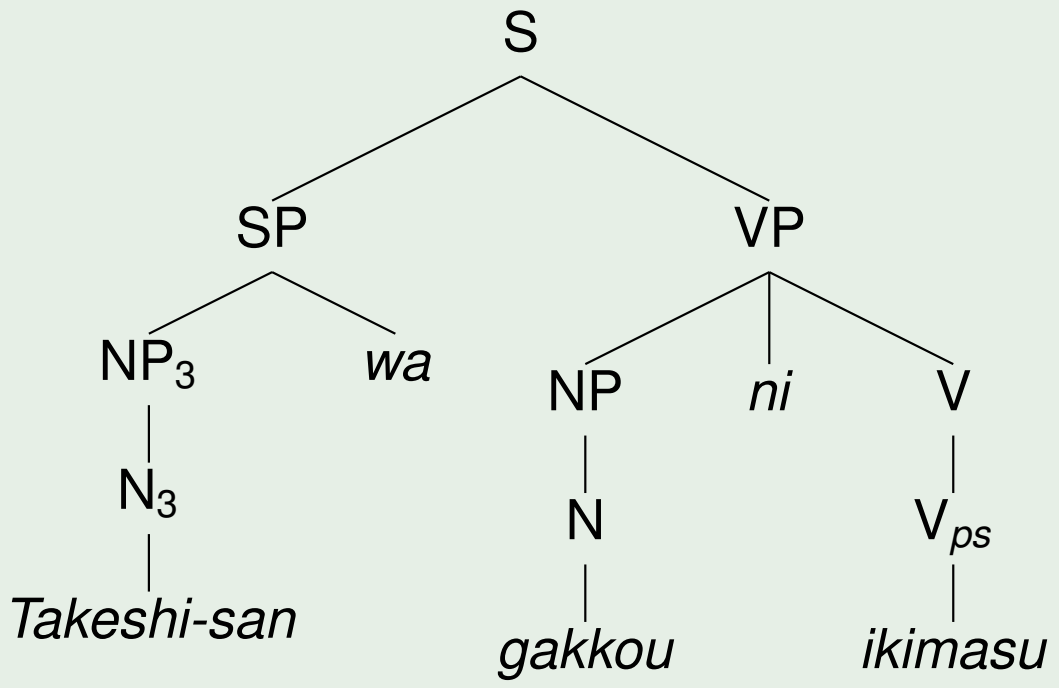
$$T(H) = \{(w_I, w_O) : \quad w_I \in T_I^*, w_O \in T_O^*, \\ S_I \Rightarrow^*_{(G_I,M_I)} w_I[\alpha], S_O \Rightarrow^*_{(G_O,M_O)} w_O[\alpha], \\ \alpha \in \Psi^*\}$$

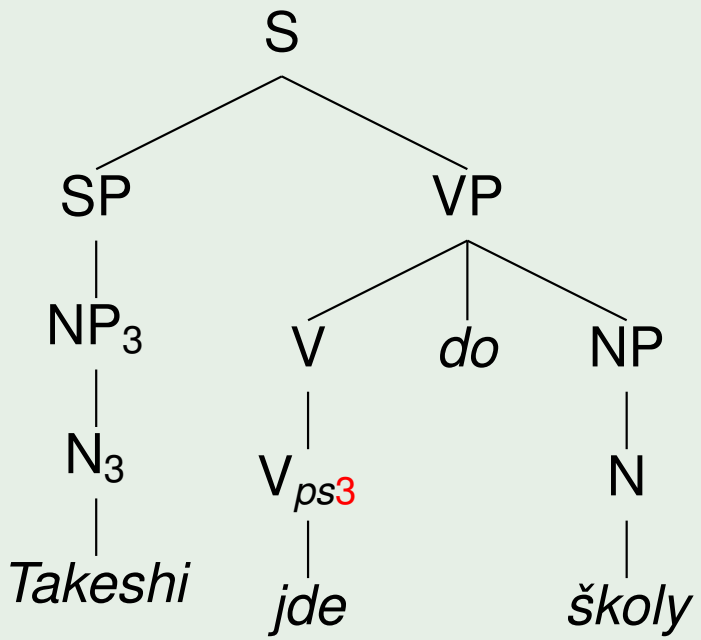
Japanese-Czech translation example

Example

watashi wa gakkou ni ikimasu	*já jdu do školy*
anata wa gakkou ni ikimasu	*ty jdeš do školy*
Takeshi-san wa gakkou ni ikimasu	*Takeshi jde do školy*

Example





Japanese-Czech translation example

Rules

1a:	SP	$\rightarrow$	NP_1 *wa*	1a:	SP	$\rightarrow$	NP_1
1b:	SP	$\rightarrow$	NP_2 *wa*	1b:	SP	$\rightarrow$	NP_2
1c:	SP	$\rightarrow$	NP_3 *wa*	1c:	SP	$\rightarrow$	NP_3
2:	V	$\rightarrow$	V_{ps}	2a:	V	$\rightarrow$	V_{ps1}
				2b:	V	$\rightarrow$	V_{ps2}
				2c:	V	$\rightarrow$	V_{ps3}

Matrices

A:	1a 2	A:	1a 2a
B:	1b 2	B:	1b 2b
C:	1c 2	C:	1c 2c

Outline

Conclusion

Goals

- Study and define formal models suitable for describing natural language syntax, with focus on the Japanese language
- Propose methods and formalisms that can be used in Japanese-Czech translation
- Create translation rules

Further research

- Syntax analysis with matrix grammars
- Theoretical study of the proposed models and their properties
- Practical applications of the translation system

References

- Dassow, J., Păun, Gh.: *Regulated Rewriting in Formal Language Theory*. Berlin: Springer, 1989, ISBN 3-540-51414-7.
- Meduna, A.: *Automata and Languages: Theory and Applications*. Springer Verlag, 2005, ISBN 1-85233-074-0.
- Meduna, A., Techet, J.: *Scattered Context Grammars and their Applications*. WIT Press, UK, WIT Press, 2010, ISBN 978-1-84564-426-0.
- Mitkov, R.: *The Oxford Handbook of Computational Linguistics*. Oxford University Press, 2004, ISBN 978-0-19-927634-9.
- Novotný, M.: *S algebrou od jazyka ke gramatice a zpět*. Academia Praha, 1988.
- Rozenberg, G., Salomaa, A.: *Handbook of Formal Languages: Volume I*. Springer Verlag, 1997.

Thank you for attention