Sentence pattern generation for low-resource language pairs
翻訳資源が少ない言語ペアのためのトランスファールール自動構築の開発研究

VARGA István, YOKOYAMA Shoichi, HASHIMOTO Chikara
Yamagata University

Introduction

Highly accurate, fast and economically affordable machine translation systems remain an elusive goal in MT. Although due to the various translation methods and refinements there are a number of success stories, the uncovered territory is still dominant. In the case of low-resource languages or less-common language pairs the question is far more complex: the importance of choosing the most appropriate technique is eclipsed by the resource limitations, whether that's manifested by the lack of personnel or machine translation tools for the languages in question. To overcome these problems, economical and efficient tools and methods are needed.

Research purpose

The goal of our research is an economically viable **Hungarian-Japanese** MT system for gisting purposes. We propose MT techniques that can be reproduced with virtually every languages pair, assuming that there are some monolingual or bilingual tools, however limited. In this current study we are designing a method to automatically generate sentence patterns and grammatical rules, concentrating on low resource languages. Our method assumes the existence of parsers for both languages, as well as the availability of a small bilingual corpus.

Related works

There are already numerous methods for sentence pattern generation. The are a limited number of success stories (Altintas & Güvenir, 2003; Cicekli, 2005), but these methods only work with closely related languages, since they do not need any deep grammatical analysis. Other automated methods make use of large bilingual corpora and a bilingual dictionary, looking for some sort of structural similarity between the counterpart sentences (Watanabe et. al, 2000; Kaji et. al, 1992). However, most of these methods still elude the desired accuracy. We believe that the main reason for this is that these methods work with separate sentence pairs, trying to extract sentence patterns for each pair. Specially with distant languages, whose different grammatical structure offers no help, these methods produce many erroneous, useless or even contradictory results.

Step 1: corpus acquisition

There is no known digital bilingual corpus between Japanese and Hungarian. We built the following corpora:

text type	size (approx. sentence pairs)	acquisition method	comment
translated literature (3rd language)	6500	free download	1-to-1 rare, structurally very inconsistent
translated literature (direct)	15000	scan	structurally inconsistent
help files	50000	free download	noisy translations, probably MT
language books	3600	manual typing	many short sentences, but grammatically rich

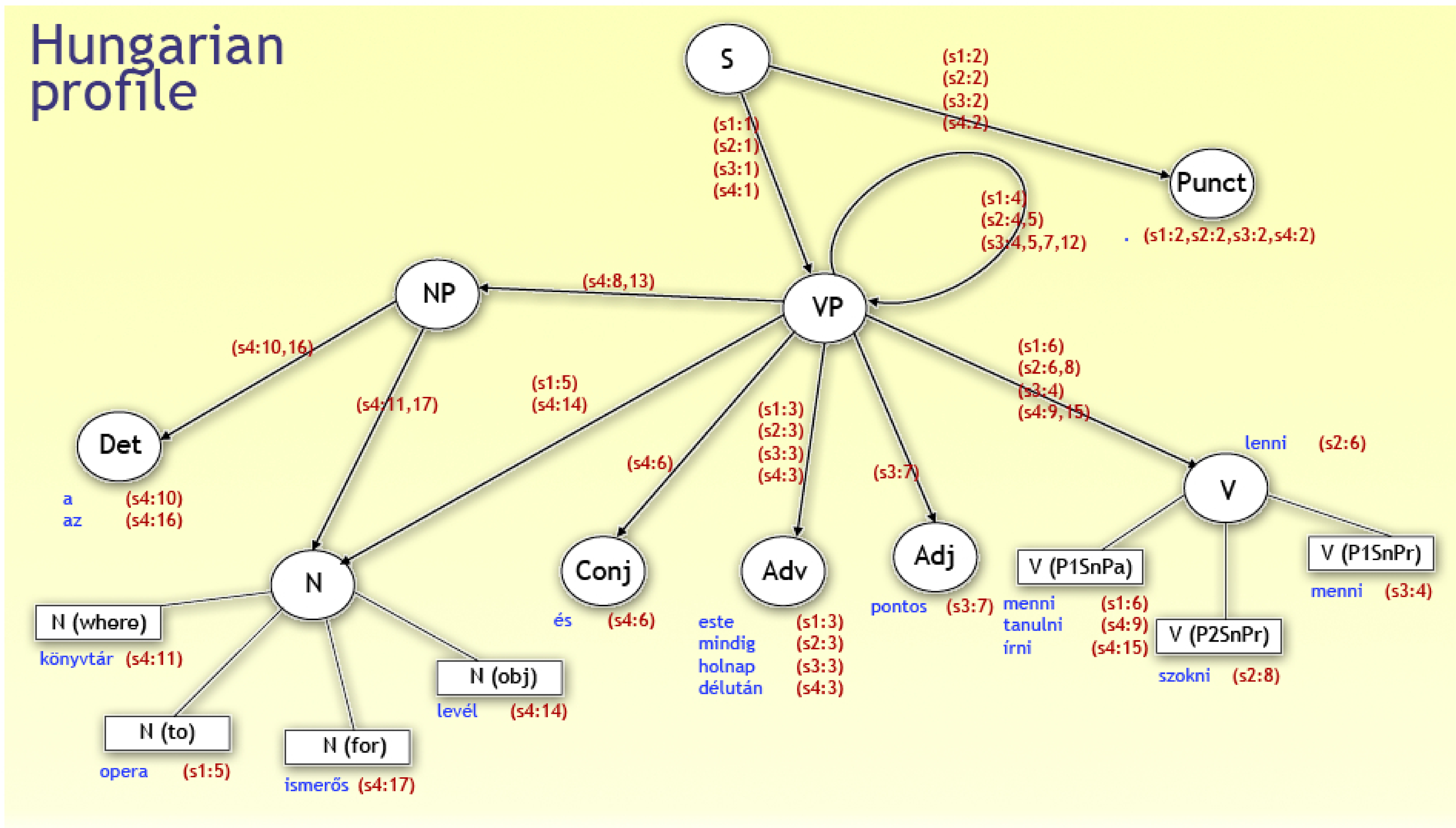
Step 2: grammatical profile generation

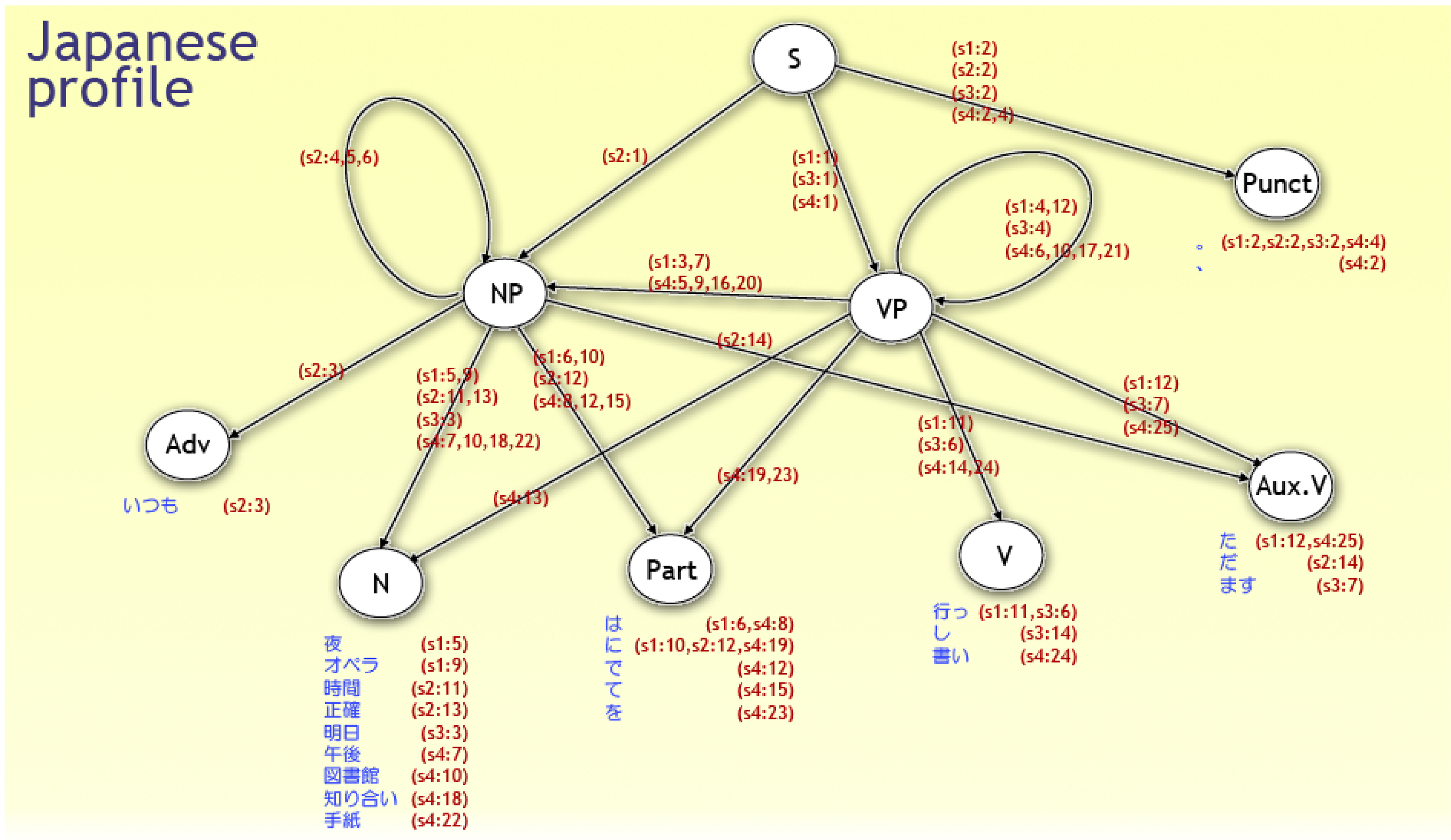
We exemplify our method with a 4-sentence bilingual corpus:

s1: "Este operába mentem."
s2: "Mindig pontos szoktál lenni."
s3: "Holnap megyek."
s4: "Délután a könyvtárban tanultam és leveleket írtam az ismerőseimnek."

s1:「夜はオペラに行った。」
s2:「いつも時間に正確だ。」
s3:「明日行きます。」
s4:「午後は図書館で勉強して、知り合いに手紙を書いた。」

We generate the "grammatical profile" of each language, accumulating all parse trees into one. We are looking for frequent patterns in each language.





Step 3: pattern generation

Method characteristics:

- analyze frequent patterns, instead of each sentence pair
- process from both sides (source and target)
- generate from subtrees, instead of full parsed trees
- buttom-to-top processing: start from the lowest unprocessed level
- generate the most general patterns; subcategorize as needed

```
(1)  while there are "unsolved" source nodes with frequency>threshold
(2)    group the node by its same parent+all child nodes (source pattern(s))
(3)    retrieve the target parse trees
(4)    identify the sub-tree using solved nodes (existing patterns) and/or bilingual dictionary
(5)    if target sub-trees are identified
(6)      consider the most general source pattern(s)
(7)      group the target pattern(s) by the same source pattern(s)
(8)      if all target patterns are the same, save the pattern pair (with frequency nr)
(9)        mark as "solved"
(10)     end if
(11)     else detail source pattern(s) and goto (7)
(12)   end if
(13) end while
```

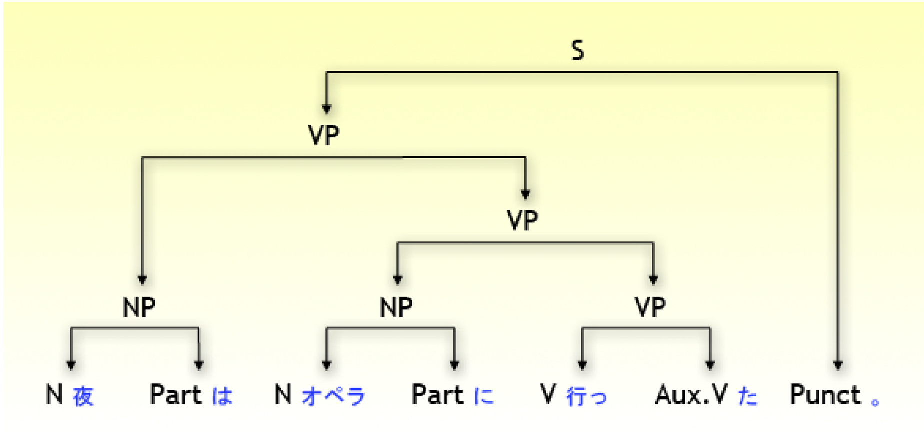
Example:

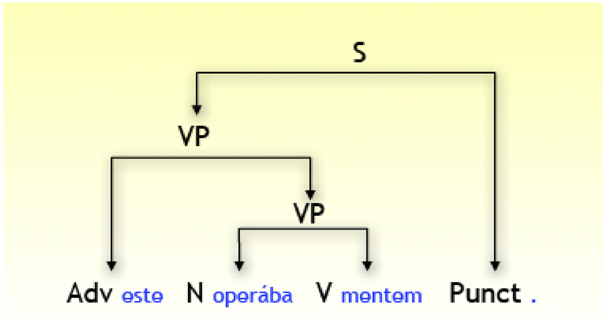
```
"は":     s1,4.   NP (N + "は") → Adv (freq=2) (rule1)
                    N → Adv (from dict)
                    "は" → Ø

"に":     s1.     NP (N + "に") → N "to" (freq=1) (rule2)
                    N → N (from dict)
                    "に" →"to"
          s2.     NP (N + "に") → Ø
          s4.     NP (N + "に") → N "for" (freq=1) (rule3)
                    N → Adv (from dict)
                    "に" →"for"

"た":     s1,4.   VP (V + "た") → V P1SnPa (freq=2) (rule4)
                    V → V (from dict)
                    "た" → P1SnPa (V)

NP + VP: s1.      VP (NP (N + "に") + VP (V + "た")) → VP (N "to" + V P1SnPa) (freq=1) (rule5)
                    NP (N + "に") → N "to" (from rule2)
                    VP (V + "た") → V P1SnPa (from rule4)
```





References:
Altintas, K., Güvenir, H. A (2003): Learning Translation Templates for Closely Related Languages. *KES 2003*, pp. 756-762.
Cicekli, I. (2005): Learning Translation Templates with Type Constraints, *In proceedings of Example-Based Machine Translation Workshop, MT Summit X*, pp. 27-34.
Watanabe, H., Kurohashi, S., Aramaji, E (2000): Finding structural correspondences from bilingual parsed corpus for corpus-based translation, *In proceedings of the 18th conference on Computational linguistics*, pp. 906-912.
Kaji, H., Kida, Y., Morimoto, Y. (1992): Learning translation templates from bilingual text, *In proceedings of the 14th conference on Computational linguistics*, pp. 672-678.