

Head Finalization Reordering for Chinese-to-Japanese Machine Translation

Dan Han *, Katsuhito Sudoh †, Xianchao Wu §,
Kevin Duh ‡, Hajime Tsukada †, Masaaki Nagata †

* National Institute of Informatics

† NTT Communication Science Laboratories

§ Baidu Japan, Inc.

‡ Nara Institute of Science and Technology(NAIST)

Outline

- Introduction
- Motivation & Objective
- Syntactic-Based Reordering Rules
- Other Reordering Issues
- Conclusion

Introduction

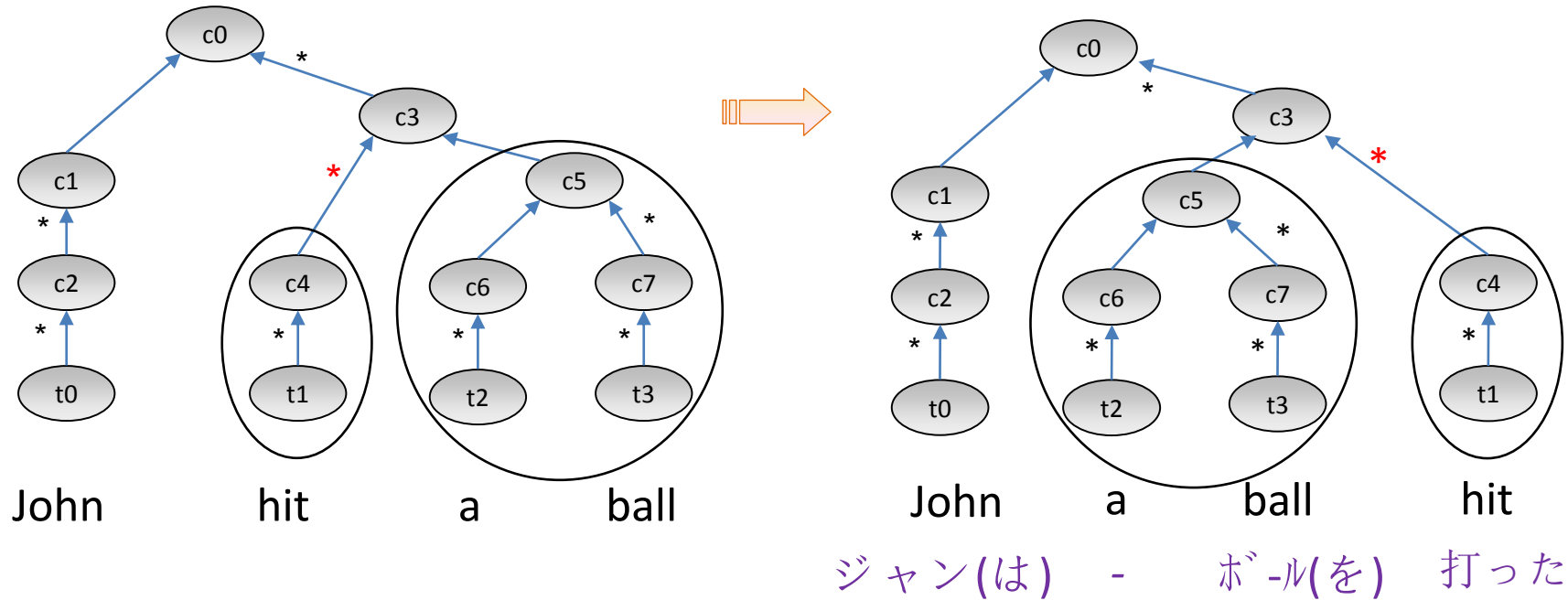
- English and Chinese are head-initial languages, while Japanese is a typical head-final language.
- Head Finalization Reordering Rule for English-Japanese MT (Isozaki et al. 2010).
 - Using the parsed result of Enju, an HPSG parser (Miyao and Tsujii, 2008) that outputs the syntactic heads.



Methodology

Move syntactic heads to the end of the corresponding syntactic constituents

Introduction (cont.)



-- H.Isozaki, K.Sudoh 2010

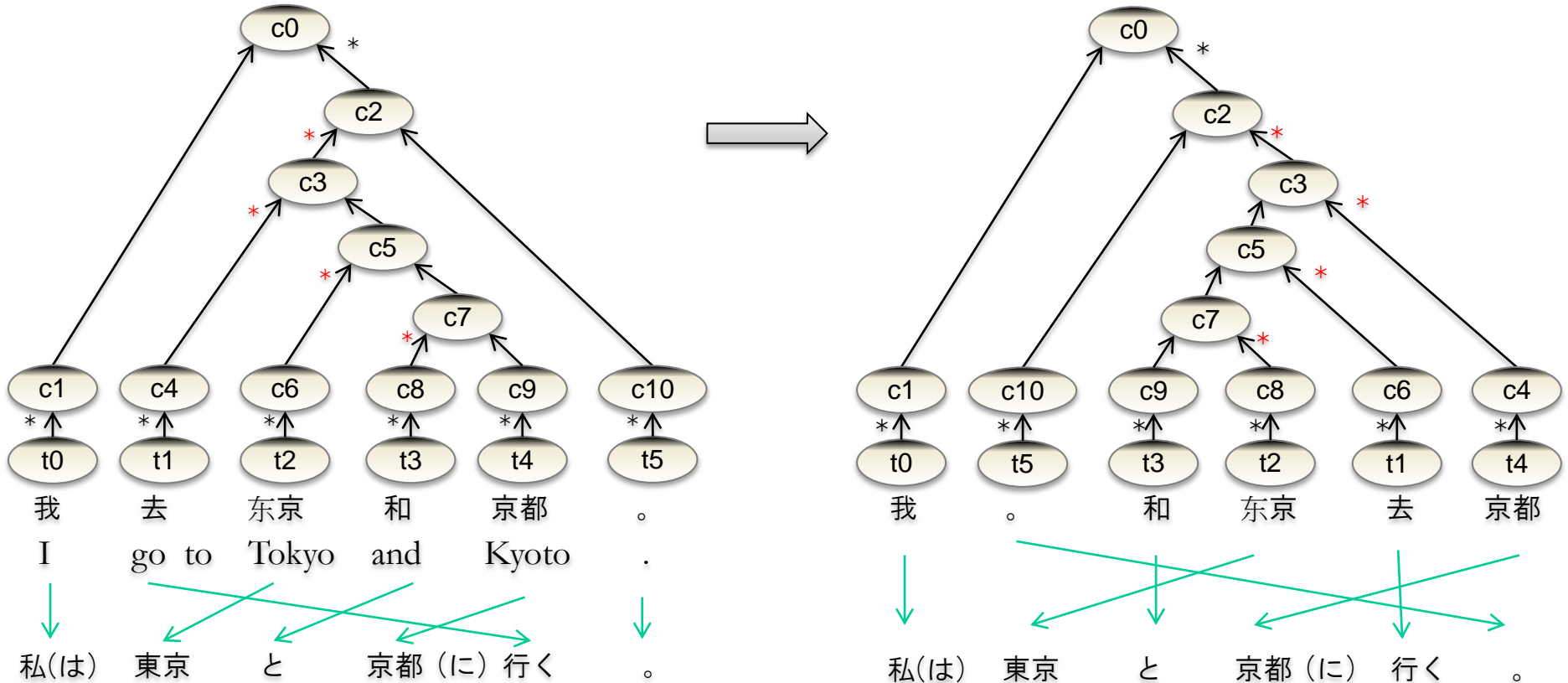
* Indicate the syntactic head

● Head-Final English (HFE)

Motivation

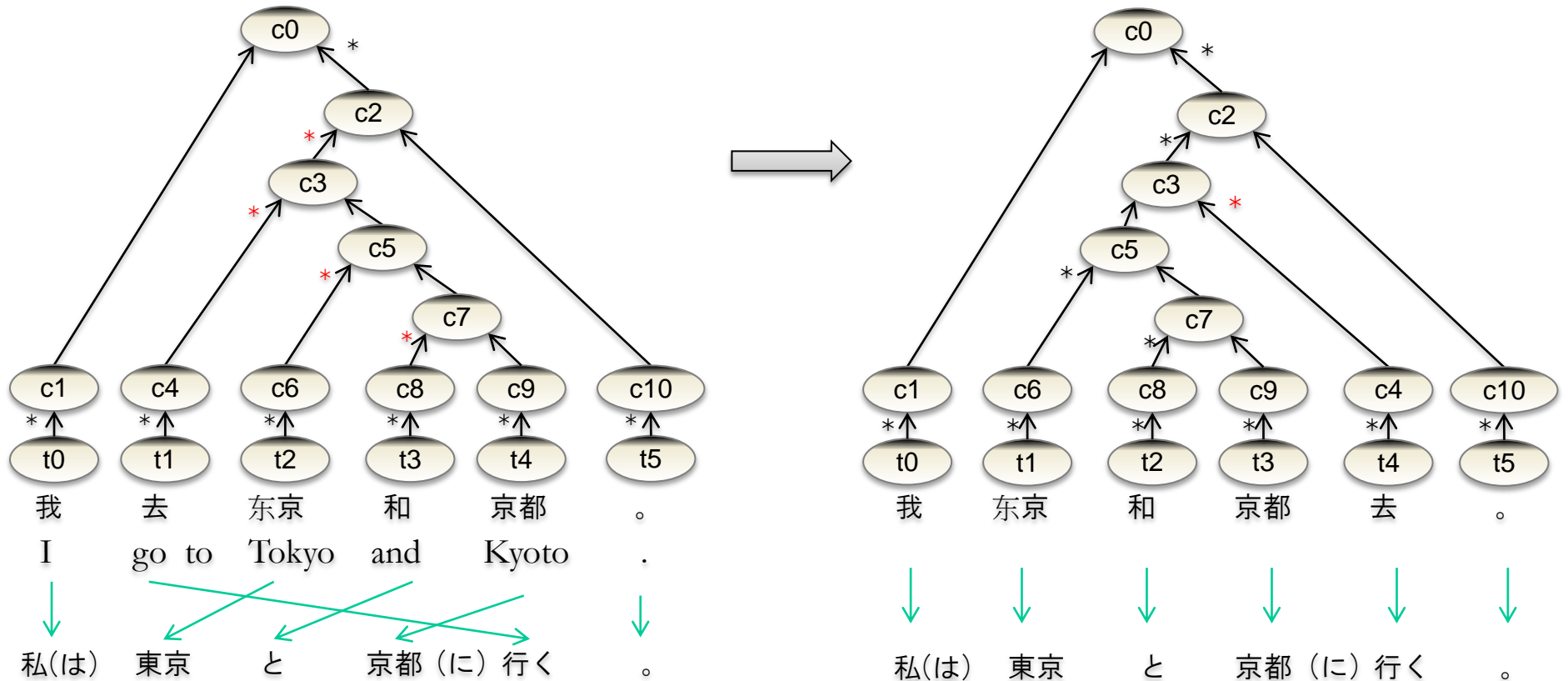
- HF works well for Japanese-English, but not for other language pairs.
- Attempt to remedy it at the level of syntactic-head analysis.
- Discrepancies in Head Definition among Chinese, Japanese, and English cause reordering issues while implementing HF.
- In this work, example of analysis and solution for Chinese-to-Japanese.

Problem!



- Head-Final Chinese (HFC)

Perfect!



- Refined-Head-Final Chinese (Refined-HFC)

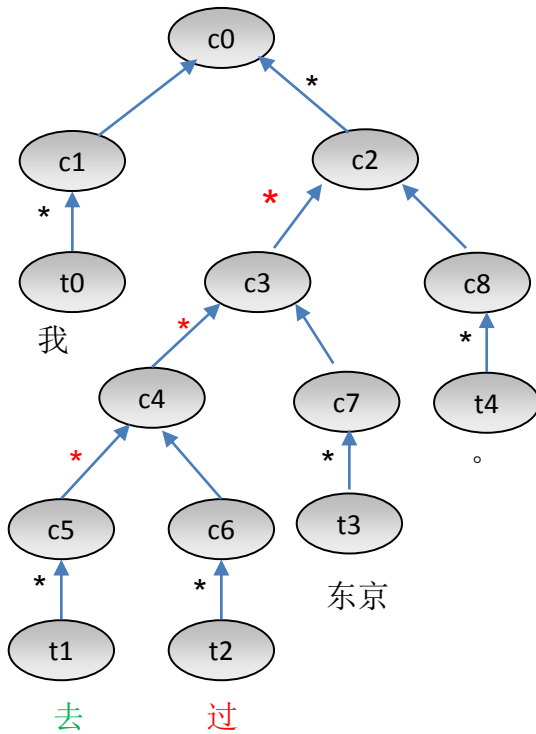
Objective

- Present detailed syntactic analysis of reordering issues.
- Define novel reordering rules based on HF and linguistically inspired refinements.

Syntactic-based Reordering Rules

- Aspect Particle
- Adverbial Modifier 'bu4'
- Sentence-final Particle
- Et cetera
- Punctuation
- Coordination

Aspect Particle



Example for Aspect Particle

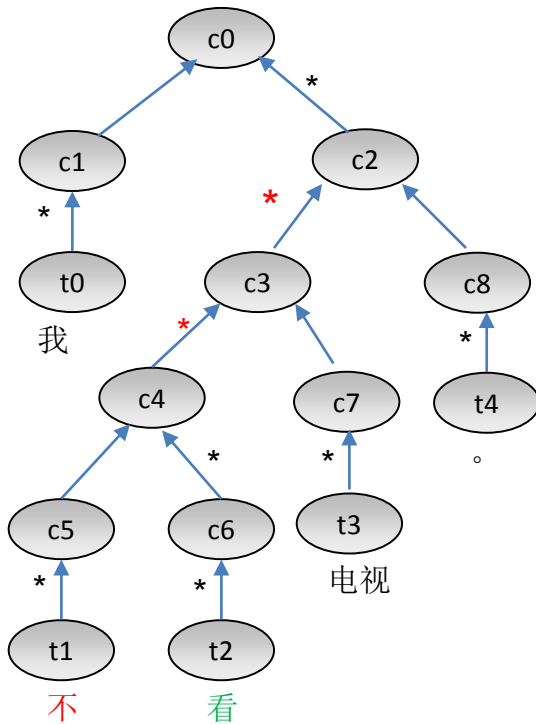
Ch 我去过东京。

En I have been to Tokyo.

HFC	我 (I)	东京(Tokyo)	过(have)	去(been to)
Ja	私(は)	東京(に)	い	った
R-HFC	我	东京	去	过

Arrows indicate the mapping: HFC to Ja (downward), Ja to R-HFC (upward), and HFC to Ja (downward). A crossed arrow indicates that the English '去' (green) maps to the Japanese 'い' (green) and the English '过' (red) maps to the Japanese 'った' (red).

Adverbial Modifier 'bu4'



Example for Adverbial Modifier bu4

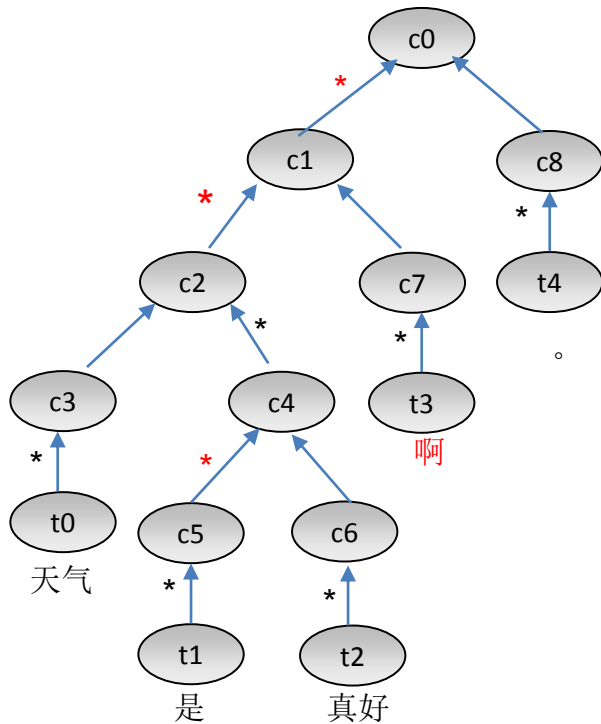
Ch 我不看电视。

En I do not watch TV.

HFC	我(I)	电视(TV)	不(do not)	看(watch)
Ja	私(は)	テレビ(を)	見	ない
R-HFC	我	电视	看	不

Arrows indicate the mapping: HFC to Ja (I to 私, TV to テレビ), and HFC to R-HFC (I to 我, TV to 电视, 不 to 不, 看 to 看). A crossed arrow indicates that the HFC '不' and '看' are swapped in the Ja translation to '見' and 'ない'.

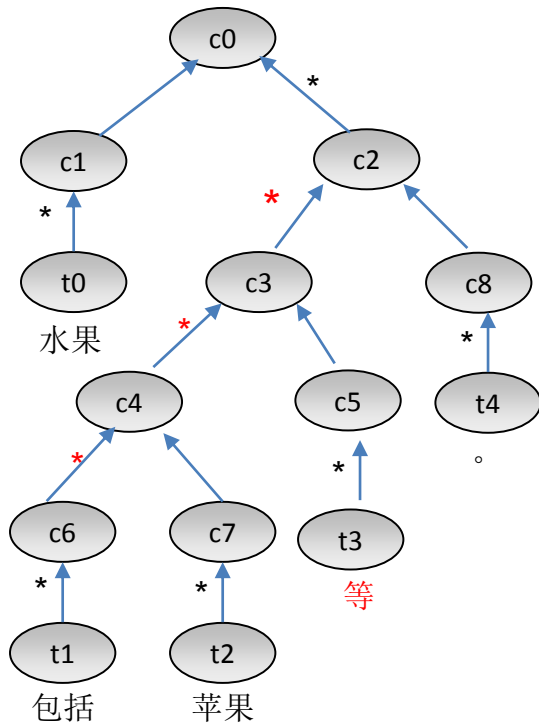
Sentence-final Particle



Example for Sentence-final Particle

Ch	天气是真好啊。			
En	It is good weather.			
HFC	啊	天气(weather)	真好(good)	是(is)
Ja	いい	天気	です	ね
R-HFC	天气	真好	是	啊

Et cetera



Example for Et cetera.

Ch 水果包括苹果等。

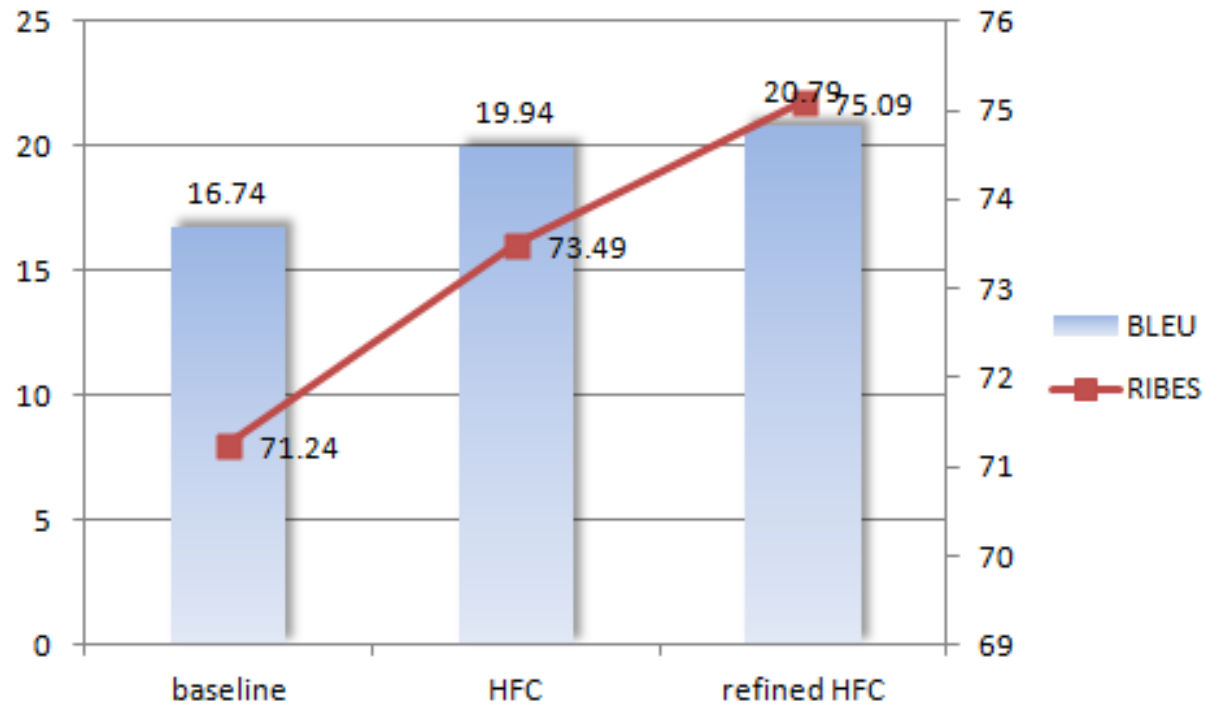
En Fruits include apples, etc.

HFC	水果(fruits)	等(etc.)	苹果(apples)	包括(include)
Ja	果物(は)	リンゴ	など(を)	含む
R-HFC	水果	苹果	等	包括

Experiments

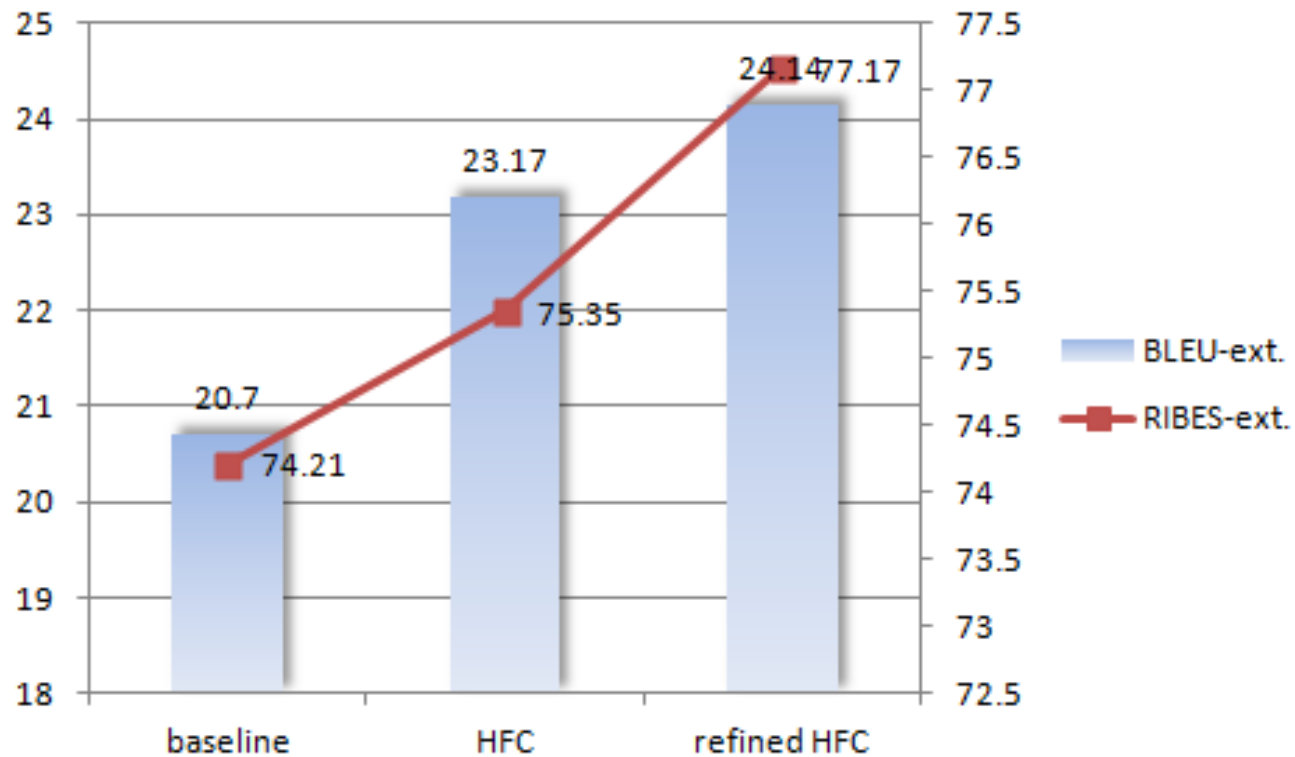
- Training data: CWMT data set (News domain, 282K sentences)
- Additional training data: XINHUA parallel data sets (News domain, 593K sentences) is used to compare the results.
- Dev & Test data: CWMT data set (1K sentences)
- Word alignment: GIZA++
- Decoder: Moses

Experiments (cont.)



BLEU and RIBES scores while CWMT corpus was used for training.

Experiments (cont.)

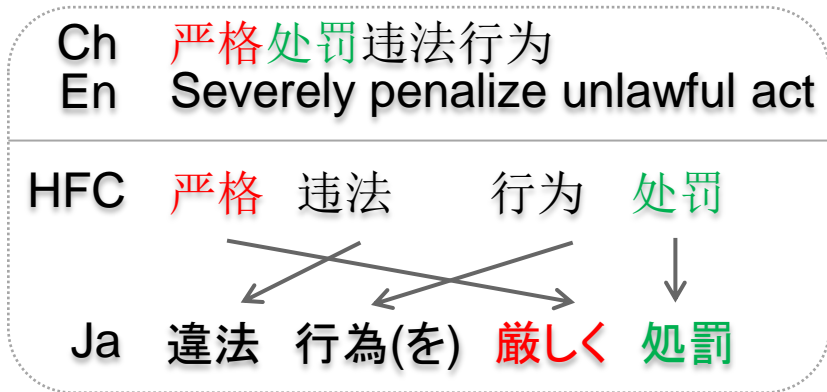


BLEU and RIBES scores while CWMT ext. corpus was used for training.

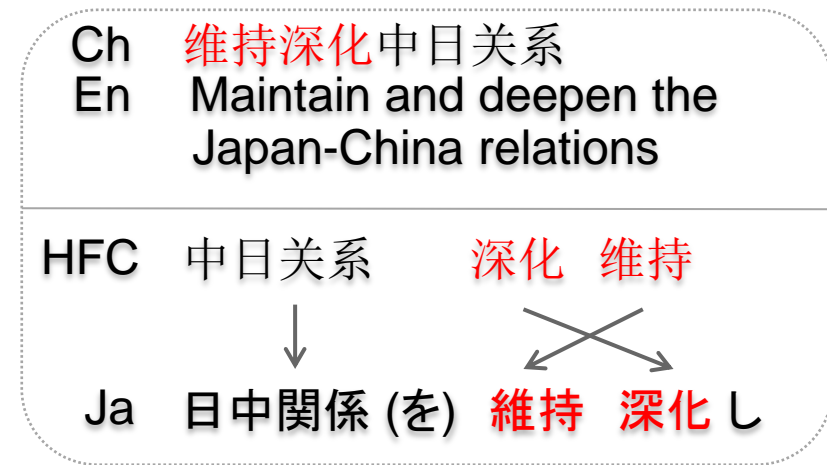
Other Reordering Issues

- Serial Verb Construction
- Complementizer
- Verbal Nominalization and Nounal Verbalization
- Adverbial Modifier
- POS tagging and Parsing Errors

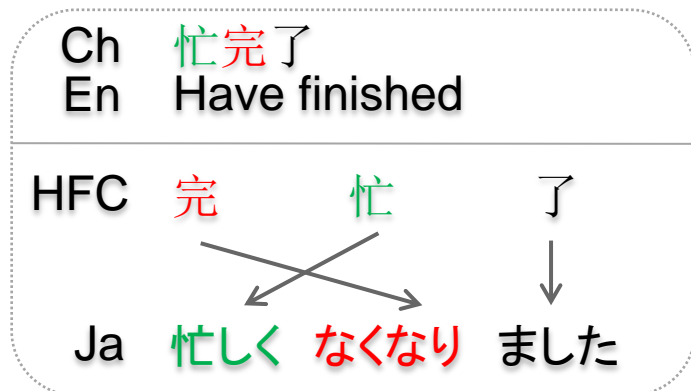
Example for Adverbial Modifier



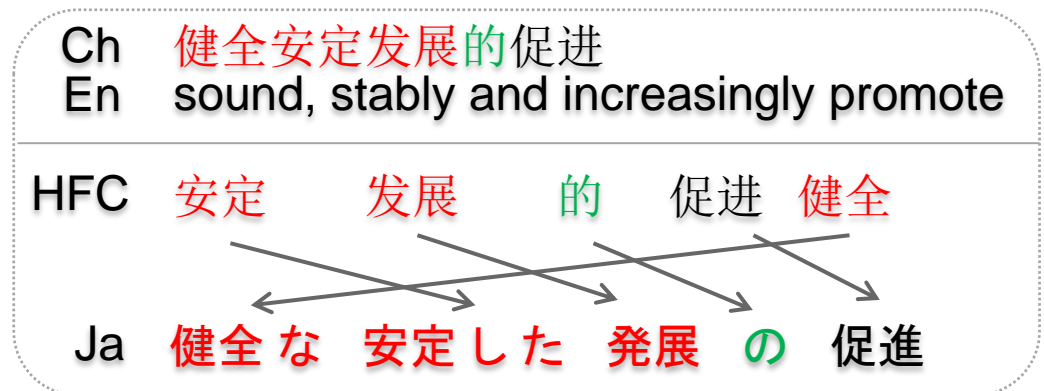
Example for Serial Verb Construction



Example for Complementizer



Example for Verbal Nominalization



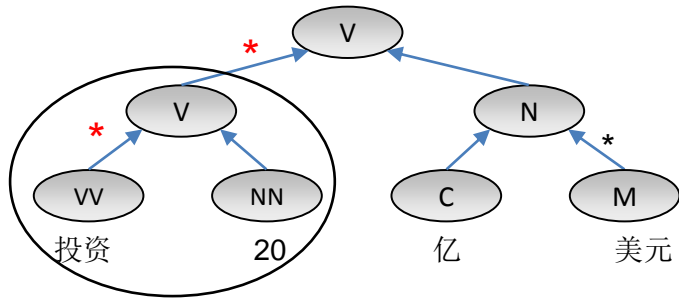
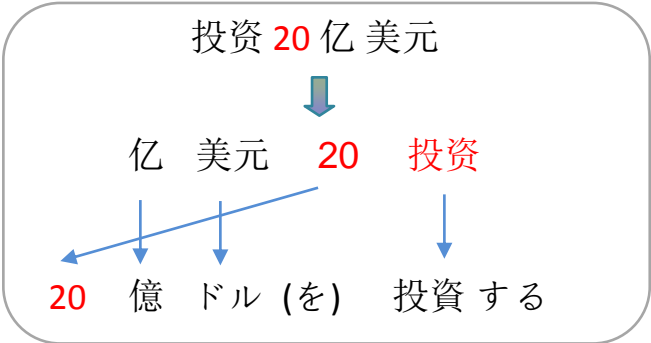
POS Tagging & Parsing Errors

- POS Tagging Errors
 - 「伊朗」(イラン, Iran)
 - POS = “VV” or “JJ” ➤ POS=“NR”
 - 「胡主席」(フー・チンタオ, Hu Jintao)
 - POS = “VV” ➤ POS=“NR”
 - 「実施」(実施する, Implement)
 - POS = “NN” ➤ POS=“VV”

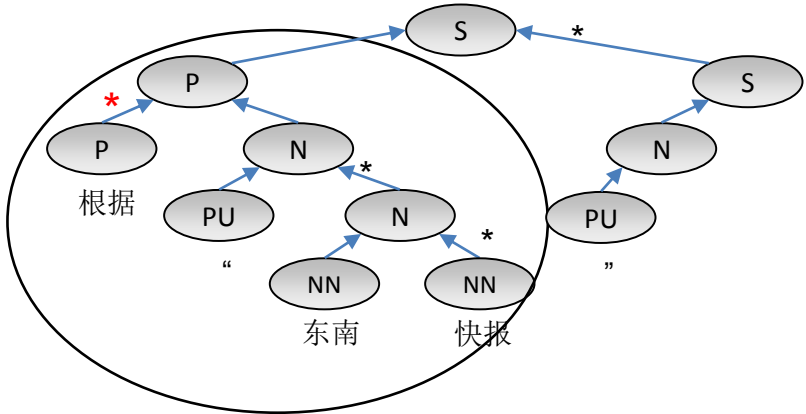
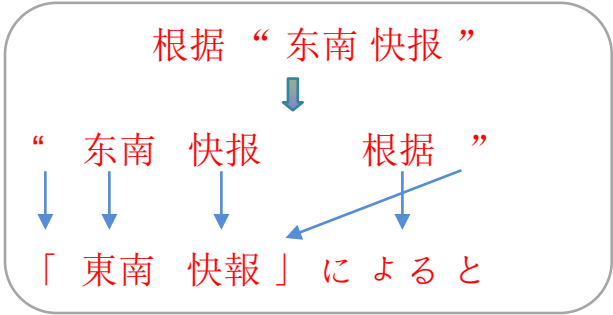
POS Tagging & Parsing Errors

- Parsing Errors

Invest 2 billion US dollars



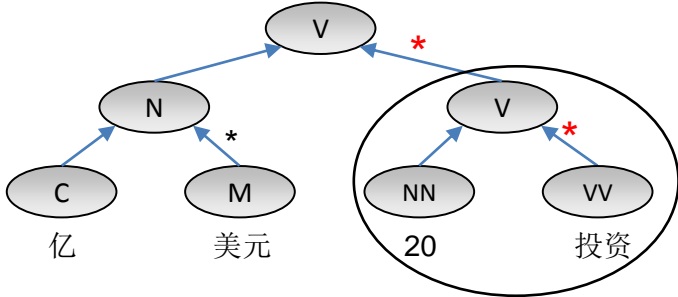
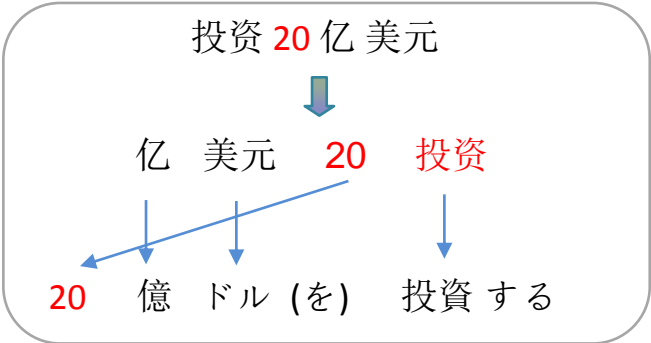
According to "TONAN NEWS"



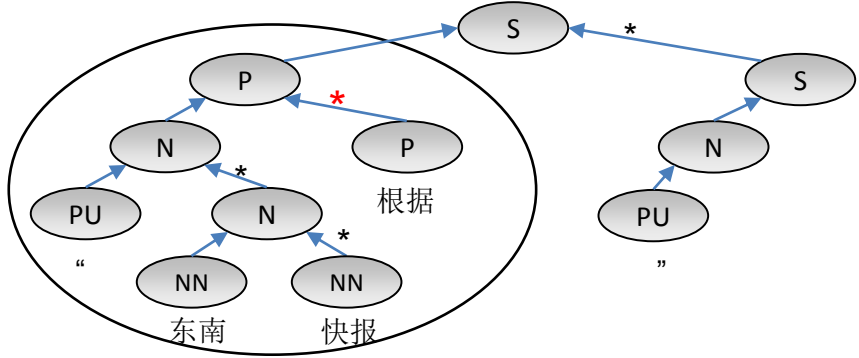
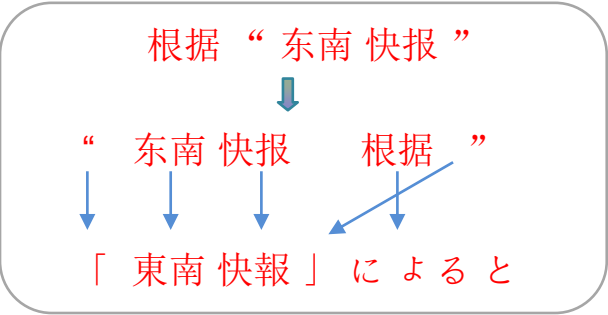
POS Tagging & Parsing Errors

- Parsing Errors

Invest 2 billion US dollars



According to "TONAN NEWS"

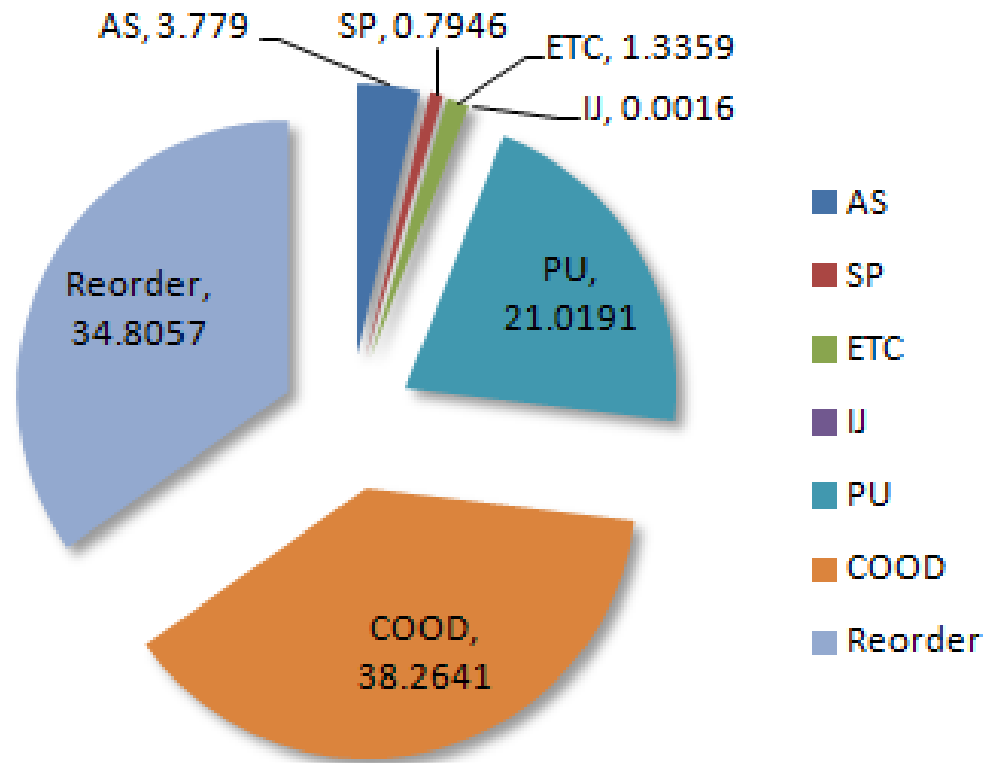


Conclusion

- Basic Head Finalization reordering rules improved the Chinese-to-Japanese machine translation quality.
- However, the refined-HFC substantially achieved further improvement.
 - Due to more monotonic word alignment.

Thank you for your attention!
Suggestions & Questions

Experiments (cont.)



The effect frequency of each exception rule during reordering on CWMT extended corpus.

	CWMT				CWMT ext.			
	BLEU	RIBES	TER	WER	BLEU	RIBES	TER	WER
baseline	16.74	71.24	70.86	77.45	20.70	74.21	66.10	72.36
HFC	19.94	73.49	65.19	71.39	23.17	75.35	61.38	67.74
refined HFC	20.79	75.09	64.91	70.39	24.14	77.17	59.67	65.31

AS	3.7790
SP	0.7946
ETC	1.3359
IJ	0.0016
PU	21.0191
COOD	38.2641
