SMT Systems in The University of Tokyo for NTCIR-9 PatentMT

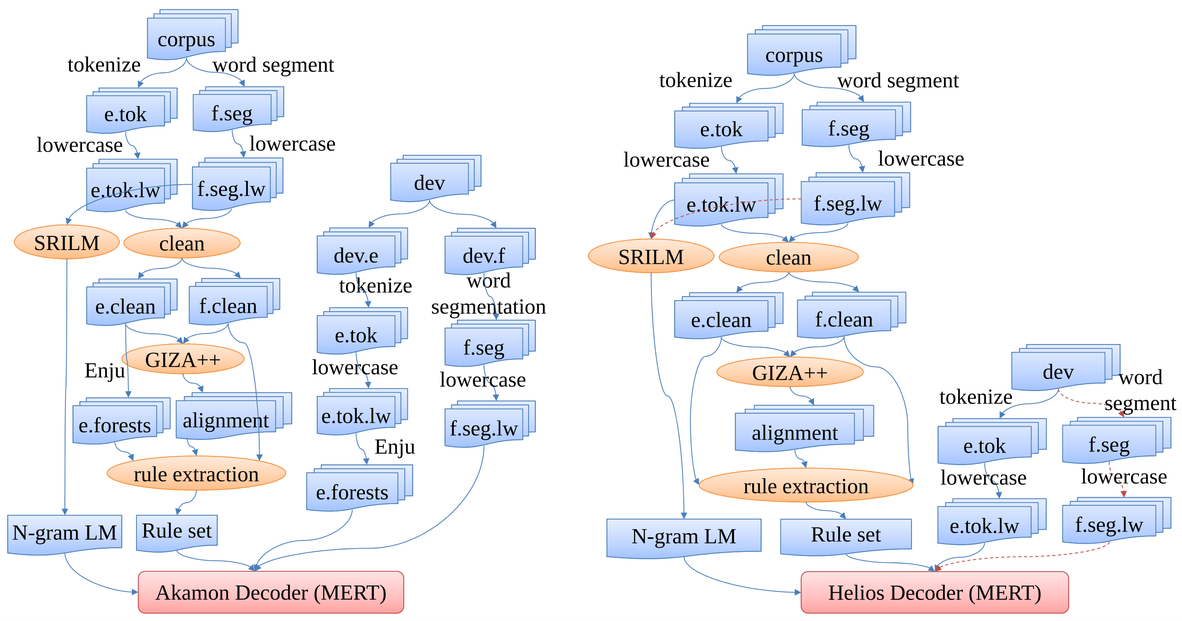
Xianchao Wu*, Takuya Matsuzaki, Jun'ichi Tsujii+

The University of Tokyo

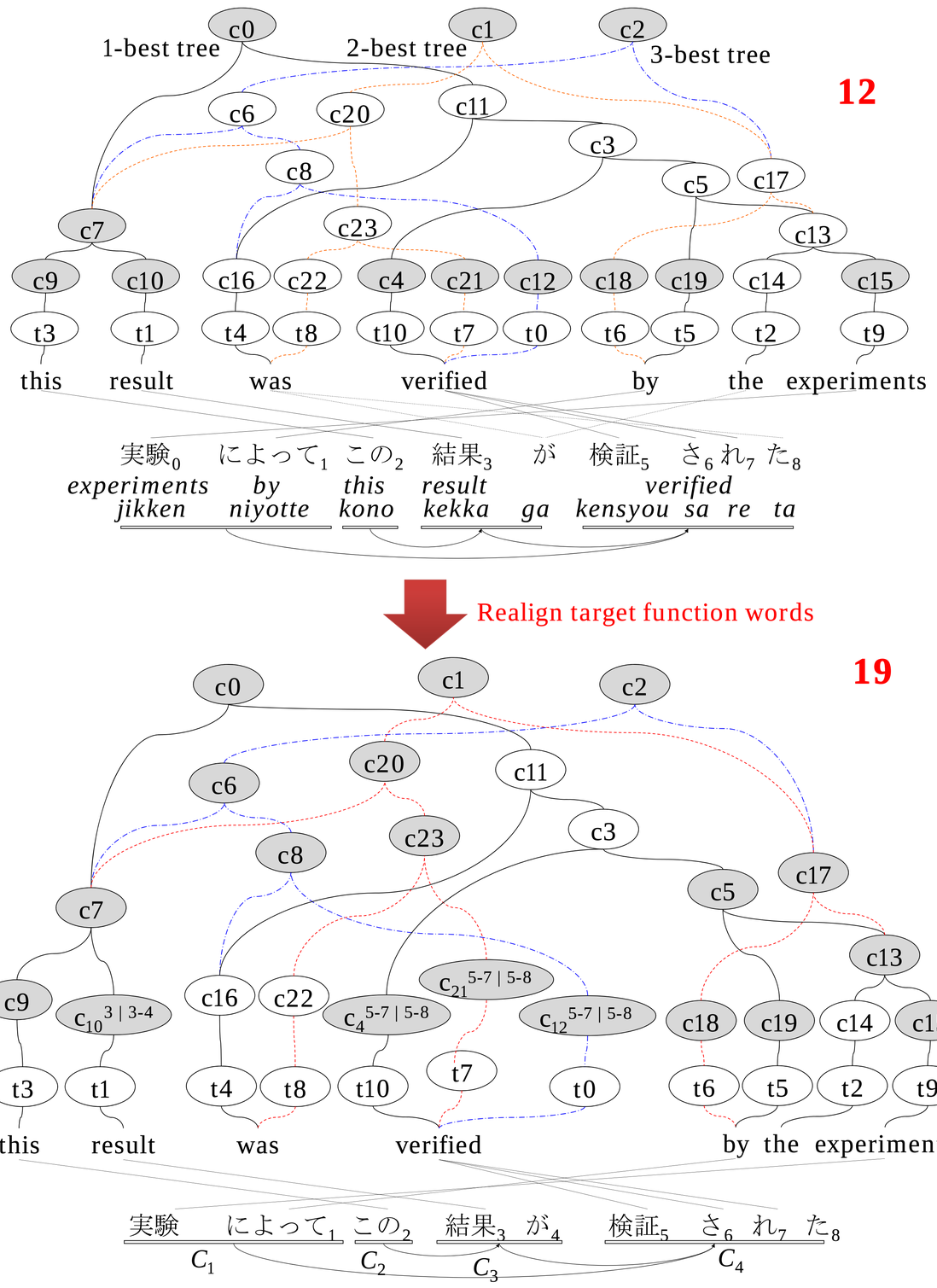
Overview

- We participated all the three PatentMT tasks:
  - English-to-Japanese (E2J),
  - Japanese-to-English (J2E) and
  - Chinese-to-English (C2E)
- **Akamon** (a forest-to-string decoder):
  - E2J (Wu+, HLT-ACL 2011)
- **Helios** (a hierarchical phrase based decoder):
  - E2J, J2E, C2E (Wu+, JaNLP 2011)

Akamon and Helios: training and tuning



Akamon (E2J): rule generalization (Wu+, 2011)



	Feature	Description
c-node	CAT	phrasal category
	XCAT	fine-grained phrasal category
	SCHEMA	name of the schema applied in the node
	HEAD	*pointer* to the head daughter
	SEM_HEAD	*pointer* to the semantic head daughter
t-node	CAT	syntactic category
	POS	Penn Treebank-style part-of-speech tag
	BASE	base form
	TENSE	tense of a verb (past, present, untensed)
	ASPECT	aspect of a verb (none, perfect, progressive, perfect-progressive)
	VOICE	voice of a verb (passive, active)
	AUX	auxiliary verb or not (minus, modal, have, be, do, to, copular)
	LEXENTRY	lexical entry, with supertags embedded
	PRED	type of a predicate
	ARG$\langle x \rangle$	*pointer* to semantic arguments, $x = 1..4$

Setup

- Mecab v0.98 (Helios) and Chasen v2.4.4 (Akamon) for Japanese word segmentation
- Stanford Chinese Segmenter (CTB standard) for Chinese word segmentation
- GIZA++ for word alignment
- SRILM for 5-gram LM training and managing
- Enju v2.4.2 for English parsing
- Cabocha v0.53 for Japanese dependency parsing

Data (# words <= 40 for Akamon, <=64 for Helios)

	Train	Dev (MERT)	Test
EJ: # parallel sentences (Helios)	2,963,963	2,000	2,000
# En words	86,048,310	63,825	E2J: 70,624 J2E: 69,521
# Ja words (% particles)	98,923,854	72,987	E2J: 78,587 J2E: 74,070
EJ: # parallel sentences (Akamon)	2,018,214	2,000	2,000
Enju parse success rate	98.5%	98.8%	98.3%
# En words	49,474,332	63,825	70,624
# Ja words (% particles)	53,271,286	73,462	73,984
CE: # parallel sentences	999,950	2,000	2,000
# Ch words	37,656,651	73,318	54,228
# En words	42,347,290	77,547	58,172

Experiment Results

		BLEU-4 (v11b)	BLEU-4 (v12)
E2J	Top-1	0.3948 (official: NTT-UT)	-
	Akamon	0.2799 (2M/3M)	-
	Helios	**0.3204**	0.3199
	Helios (bug on data prep.)	0.2781 (official)*	-
J2E	Top-1	0.3169 (official)	-
	Helios	0.3061	**0.3089**
	Helios (bug on data prep.)	0.2697 (official)*	-
C2E	Top-1	0.3944 (official)	-
	Helios	0.3227	**0.3243**
	Helios (bug on data prep.)	0.3074 (official)*	-

* Wu is now in NTT Communication Science Laboratories

+ Tsujii is now in Microsoft Research Asia