The UOT System: Improve String-to-Tree translation Using Head-Driven Phrase Structure Grammar and Predicate-Argument Structures

Xianchao Wu[1] , Takuya Matsuzaki[1], Naoaki Okazaki[1], Yusuke Miyao[1], Jun'ichi Tsujii[1,2,3]

1. The University of Tokyo 2. The University of Manchester 3. National Centre for Text Mining

OBJECTIVES

- Integrate deep syntactic information into current syntax-based SMT systems
- Extract compact translation rules guided by Predicate-Argument Structures (PASs)

BACKGROUND

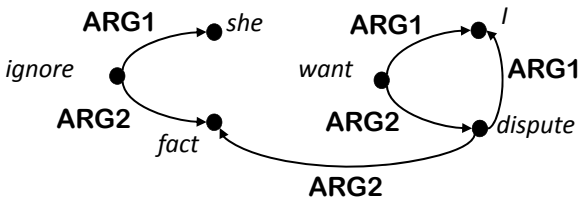
CFG vs. HPSG

- Both have tree structures
- POS/phrasal tags vs. Typed Feature Structures
- + PAS: semantic dependencies among a word and its arguments

Typed Feature Structure

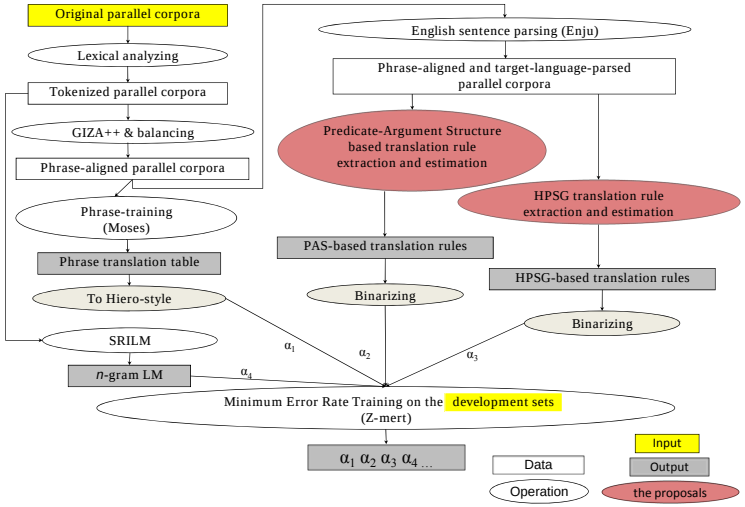
Feature	Description	Type of Node
CAT	phrasal category	non-terminal nodes
XCAT	fine-grained phrasal category	
SCHEMA	name of the schema applied in the node	
HEAD	pointer to the head daughter	
SEM_HEAD	pointer to the semantic head daughter	
POS	part-of-speech	terminal (leaf) nodes
TENSE	tense of a verb (past, present, etc.)	
VOICE	voice of a verb (passive/active)	
PRED	type of a predicate	
ARG<*n*>	pointer to semantic arguments (*n*=1,2,3,4)	

Predicate-Argument Structure

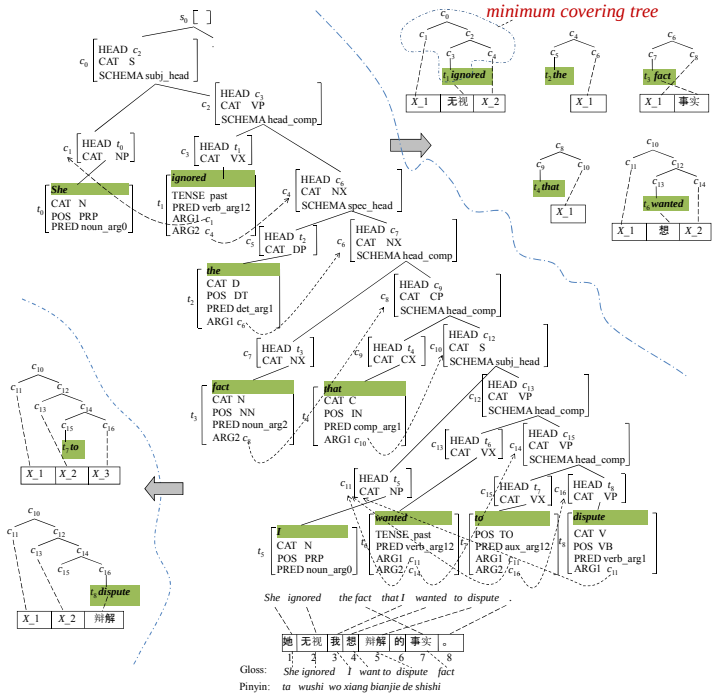


The UOT String-to-Tree System

Framework



Rule Extraction from PAS



EXPERIMENTS

Data

- IWSLT2009 Chinese-English training data
- IWSLT2003-2007 dev. sets
- IWSLT2008/2009 test sets

Rule Sets Used

- *ptt*: phrase translation table, maximum phrase length = 3, 4, or 5;
- *shpsg*: only use CAT and POS in Typed Feature Structures to approximate CFG trees;
- *hpsg*: normal HPSG tree based rules, using GHKM algorithm (Galley et al., 2004); and
- *pasr*: compact rules extracted from PASs.

Results

Rule Configuration	BLEU (%)	Dev.	Test
ptt_3	30.66	2003-2007	2008
ptt_4	30.75		
ptt_5	31.31		
ptt_5+shpsg	34.09		
ptt_5+hpsg	34.53		
ptt_5+pasr	32.51		
ptt_5+pasr+hpsg	34.65		
	36.14	2003-2008	2008
	35.38		2009

- 0.44 BLEU points gained by changing POS/phrasal tags to Typed Feature Structures;
- 1.20 BLEU points gained by appending *pasr* to *ptt_5*; and
- best result gained when using both *hpsg* and *pasr*.

FUTURE DIRECTIONS

- Test the proposals on written text sets, and
- extract composed rules (Galley et al., 2006) from HPSG tree structures.