NTT

NTT-UT SMT in NTCIR-9 PatentMT

Katsuhito Sudoh, Kevin Duh Hajime Tsukada, Masaaki Nagata
NTT Communication Science Laboratories, Japan

Xianchao Wu*, Takuya Matsuzaki, Jun'ichi Tsujii†
The University of Tokyo, Japan

* Currently with NTT Communication Scisence Labs., † Curently with Microsoft Research Asia

Results Overview .... ranked 1st in En-Ja!

Subtask	System	BLEU	Δbaseline	Δcompetitor	RIBES	Δbaseline	Δcompetitor	Adequacy	Δbaseline	Δcompetitor	Accept.	Δbaseline	Δcompetitor
EnJa	Primary	39.48	+7.58(PBMT)	+5.45(G05-1)	78.13	+6.13(HBMT)	+9.08(G05-1)	3.67	+1.07(HBMT)	+0.16(RBMT6-1)	0.69	+0.22(HBMT)	+0.03(RBMT6-1)
	Best Single	38.81	+6.91(PBMT)	+4.78(G05-1)	77.82	+5.82(HBMT)	+8.77(G05-1)						
JaEn	Primary	28.35	-0.60(HBMT)	-3.34(G01-1)	71.95	+1.31(HBMT)	-2.05(G01-1)	2.75	+0.13(HBMT)	-0.92(G04-1)	0.49	+0.02(HBMT)	-0.22(G04-1)
	Best Single	27.78	-1.17(HBMT)	-3.91(G01-1)	67.32	-3.32(HBMT)	-6.72(G01-1)						
ZhEn	Primary	30.26	-0.46(PBMT)	-9.28(G1-1)	76.47	-0.74(HBMT)	-6.80(G1-1)	3.23	-0.06(HBMT)	-0.80(G1-1)			
	Best Single	30.74	+0.02(PBMT)	-8.70(G1-1)	76.62	-0.59(HBMT)	-6.65(G1-1)						

GMBR-based System Combination (All subtasks)

Theory [Duh+ IJCNLP 2011]

MBR decision rule: $\mathrm{argmin}_{\delta(f)} \sum_{e} L(\delta(f)|e)\, p(e|f)$

loss function

hypothesis space for input *f*

1) N-best approximation
2) uniform assumption for p(e|f)
3) decomposing loss function

GMBR decision rule: $\mathrm{argmin}_{e' \in N(f)} \sum_{e \in N(f)} \sum_{k} \theta_k L_k(e'|e)$

N-best hypotheses for input *f*

decomposed loss-function (feature)
(e.g., n-gram precision, brevity penalty, rank correlation coeff.)

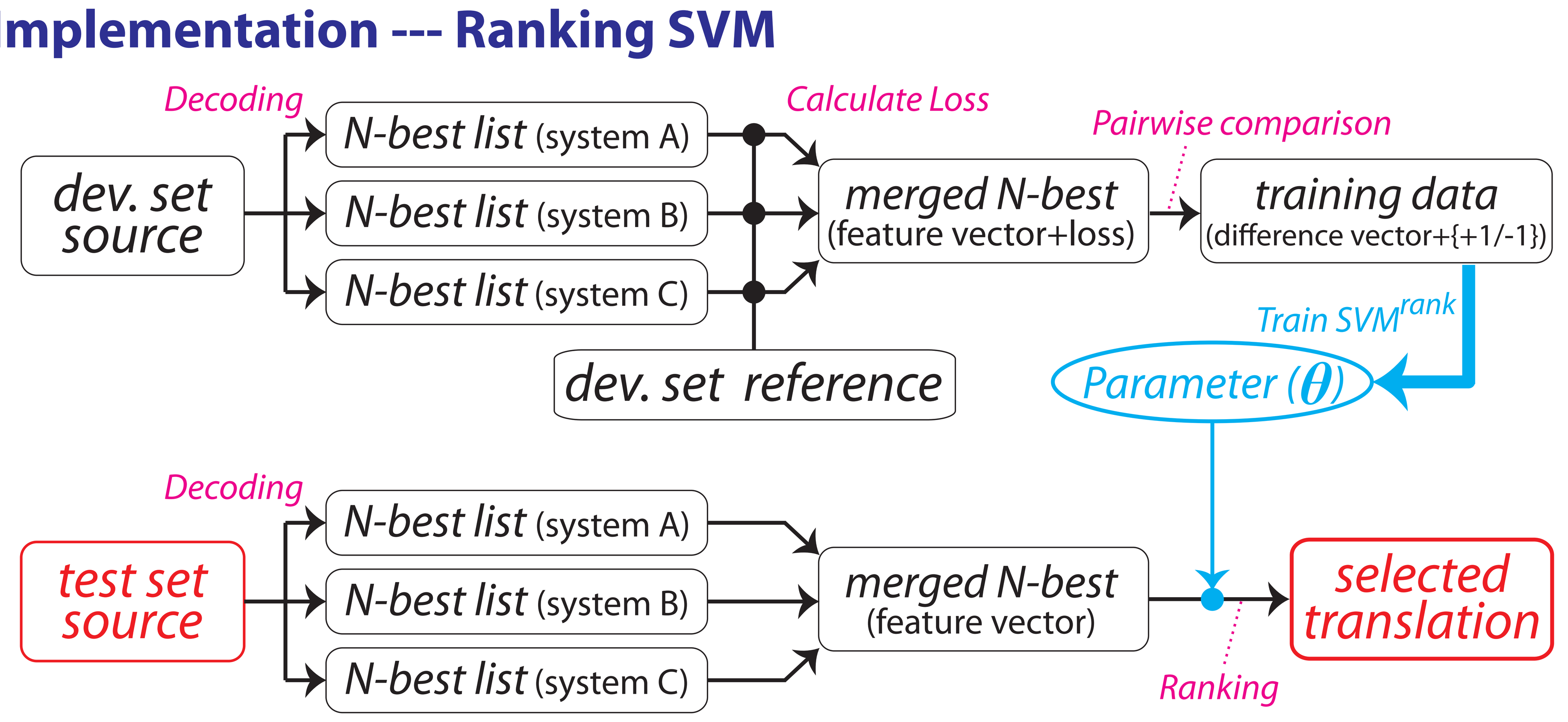
This is a *Learning to Rank* problem:

If *true loss* on dev. set reference: $L(e_1|e_r) < L(e_2|e_r)$ (i.e., e1 is better than e2)

Any sentence-wise metrics can be used: sentBLEU, RIBES, RIBES+sentBLEU, ...

we choose θ such that: $\sum_{e \in N(f)} \sum_{k} \theta_k L_k(e_1|e) < \sum_{e \in N(f)} \sum_{k} \theta_k L_k(e_2|e)$

Implementation --- Ranking SVM



Evaluation

Subtask	Opt. metric	Single systems	Comb.	Gain
EnJa	RIBES + BLEU	77.82, 77.54, 68.61	78.13	+0.25
		38.81, 36.83, 27.99	39.48	+0.67
JaEn	RIBES + BLEU	68.16, 67.32, 67.00	71.95	+3.79
		27.78, 26.75, 26.05	28.35	+0.57
ZhEn	BLEU	30.74, 27.39	30.74	±0

◀ overfit?

** One of the combined systems is from Univ. of Tokyo, others from NTT. See UT team presentation for details.*

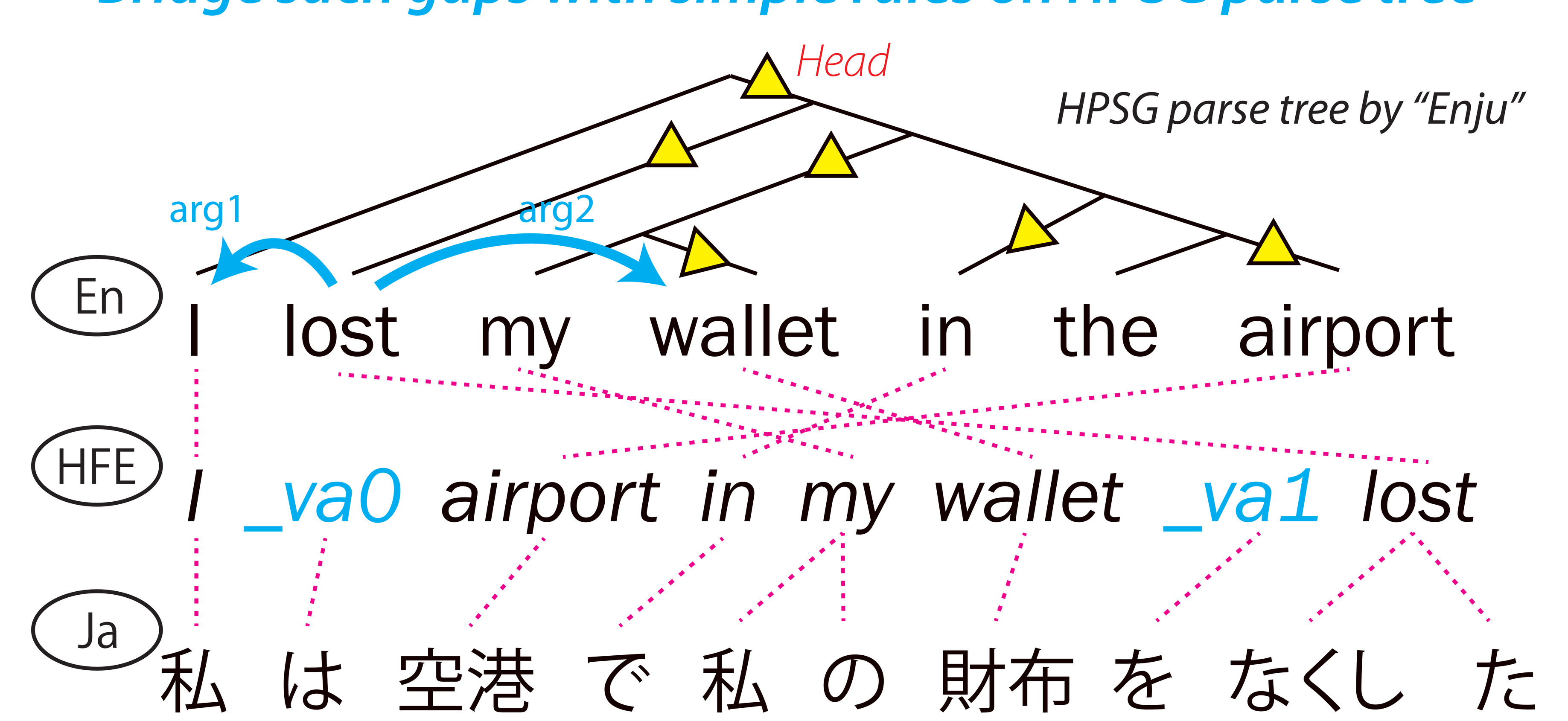
Syntactic Pre-ordering (EnJa, JaEn)

English-to-Japanese: Head Finalization [Isozaki+ WMT 2010]

(~5% BLEU gain)

- Japanese is a *head-final* language
- Articles (a, an, the) and noun plural form are not needed
- Verb arguments are specified by particles and case markers

Bridge such gaps with simple rules on HPSG parse tree



Japanese-to-English: Rule-based Chunk Reordering

(~1% BLEU gain)

- Series of hand-crafted reordering rules (see the paper for details)
- Japanese chunk-based dependency parser "Cabocha"

Fast monotone decoding by WFSTs (EnJa, JaEn)

- Decoding becomes (almost) monotone by pre-ordering
- Efficiently solved by WFSTs (~3x faster than Moses!)

$$d \bullet (P \bullet d(T \bullet d(S \bullet L_1))) \bullet L_{n-1}$$

$\bullet$: composition, $d()$: determinization
P : WFST for segmenting source sentences into source phrases
T : WFST for translating source phrases into target phrases
S : WFST for segmenting target phrases into target words
L_1 : 1-gram language model
L_{n-1} : (n-gram - 1-gram) language model

Domain Adaptation of Word Alignment (ZhEn)

[Duh+ MTSummit 2011]

- Use external bilingual corpora to improve word alignment
- NIST OpenMT 2008 data (107M words after cleaning)

References:

K. Duh+, "Generalized Bayes Risk System Combination for MT", in IJCNLP, 2011
H. Isozaki+, "Head Finalization: A Simple Reordering rule for SOV languages", in WMT, 2010
K. Duh+, "Bayesian Adaptation of Alignment Matrices for MT", in MT Summit XIII, 2011

Acknowledgments:

We'd like to thank the organizers for their great effort on this PatentMT task.
We also thank Dr. Takaaki Hori & Dr. Shinji Watanabe for their help on the use of WFSTs, and Prof. Hideki Isozaki for valuable discussion & suggestion about EJ patent translation.