The NiCT-ATR Speech Translation System for IWSLT 2006 Evaluation

Ruiqiang Zhang, Hirofumi Yamamoto, Michael Paul, Hideo Okuma, Keiji Yasuda, Yves Lepage, Etienne Denoual, Daichi Mochihashi, Andrew Finch,Eiichiro Sumita

Main techniques

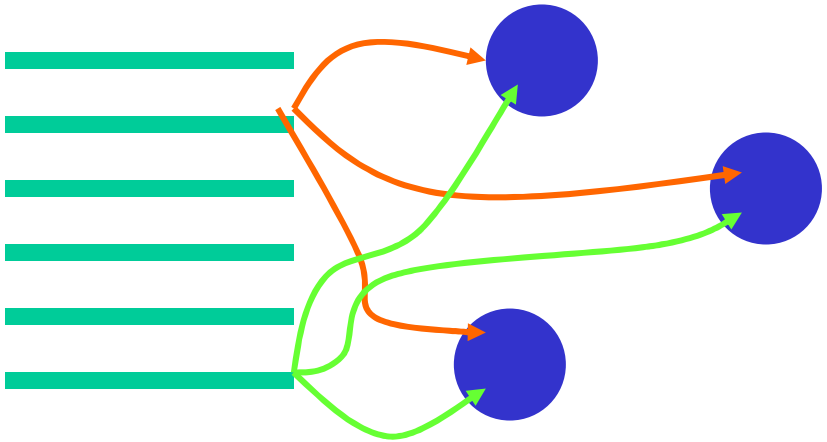
- Multiple engines: TATR, HPATR3, EM
- Language model adaptation: topic-based
- Subword-based unknown word (UW) translation
- Pre-processing:
  - Word segmentation
  - Sentence splitting
- Post-processing:
  - Higher-order LM for rescoring N-best translation results
  - Punctuation, Capitalization

Multi-engine SMT (Paul, 2005)

- TATR: phrase-based SMT
  - Multiple features:
    - Pr(e|f), Pr(f|e), Lex(e|f), Lex(f|e), LM
- HPATR3: SMT based on syntax-transfer
- EM: exact match based on translation memory
- Selector: multiple set of translation models(TM) and LMs.

Language model adaptation: clustering bilingual training data

1. The number of clusters (topics) is predetermined before clustering.
2. Initialization: assign a class for each pair randomly
3. Reassign the class for each pair
4. Repeat 3 until no entropy reduction is found.



Performance of LM adaptation

	BLEU	NIST	WER	PER
Baseline	0.222	6.68	0.681	0.517
Adapted	0.236(+1.4	6.81	0.672	0.505

IWSLT J-E translation

	BLEU	NIST	WER	PER
Baseline	0.212	6.78	0.709	0.513
Adapted	0.228(+1.6)	6.96	0.698	0.508

IWSLT C-E translation

Achilles – Chinese word segmentation (Zhang, 2006)

1. Dictionary based N-gram word segmentation
   黄 英 春 住 在 北京市
2. Subword-based IOB tagging
   黄/B英/I春/I 住/O 在/O 北京/B 市/I

- Combine the two by confidence measure
  黄/B/0.7 英/I/0.7 春/I/0.7 住/O/1.0 在/O/0.9 北京/B/1.0 市/I/0.9

the best balanced point between in vocabulary rate and out-of-vocabulary rate

Evaluation of Achilles

- The second Sighan bakeoff CWS evaluation
  - State-of-the-art CWS F-SCOREs (higher than the best-known)

- A higher translation performance (+1.1% BLEU) than the LDC default (using NIST 2005 data).
  - A plausible word segmenter.

	BLEU	NISt	WER	PER	METEOR
LDC default	0.226	7.62	0.895	0.642	0.528
OURs	0.237(+1.1)	7.93	0.867	0.614	0.525

Sentence splitting

- Shorter sentences are easier for translation.
- ASR output does not include punctuation.
- Sentence splitting following automatic punctuation:

  1. We added punctuation using SRI tools
  2. Sentence is split at the place of the added punctuation
  3. Sentence segments are translated individually, without using punctuation.
  4. Translations results are re-linked in the same order as the input.

Performance of sentence splitting

- Sentence splitting improved CE and JE: BLEU scores (+1)

	CE spont.	CE read	CE CRR	JE read	JE CRR
w/o splitting	0.1551	0.1756	0.2051	0.1817	0.2023
w/splitting	0.1591	0.1775	0.206	0.1899	0.2122

Performance of sentence splitting

- Sentence splitting does not work for AE and IE translation: BLEU scores

	AE read	AE CRR	IE read	IE CRR
w/o splitting	0.2122	0.2365	0.2991	0.3774
w/splitting	0.2117	0.2384	0.2989	0.3763

Unknown word (UW) translation – subword approach for CE translation

- Unknown words (out of phrase-translation table) cannot be translated by the phrase-based translation engines.
- For Chinese, longer UW split into translatable, shorter subwords. Training a subword translation model by
  - Same training approach as for full word translation model but
  - Use a subword lexicon (size=5000) for word segmentation

- **Effective** for UWs such as named entity, Chinese name, foreign words, and numbers. Translation obtained by combination of the meanings of subwords
- **Useless** if UW's meaning is unrelated to subwords.
- 0.5% BLEU improvement

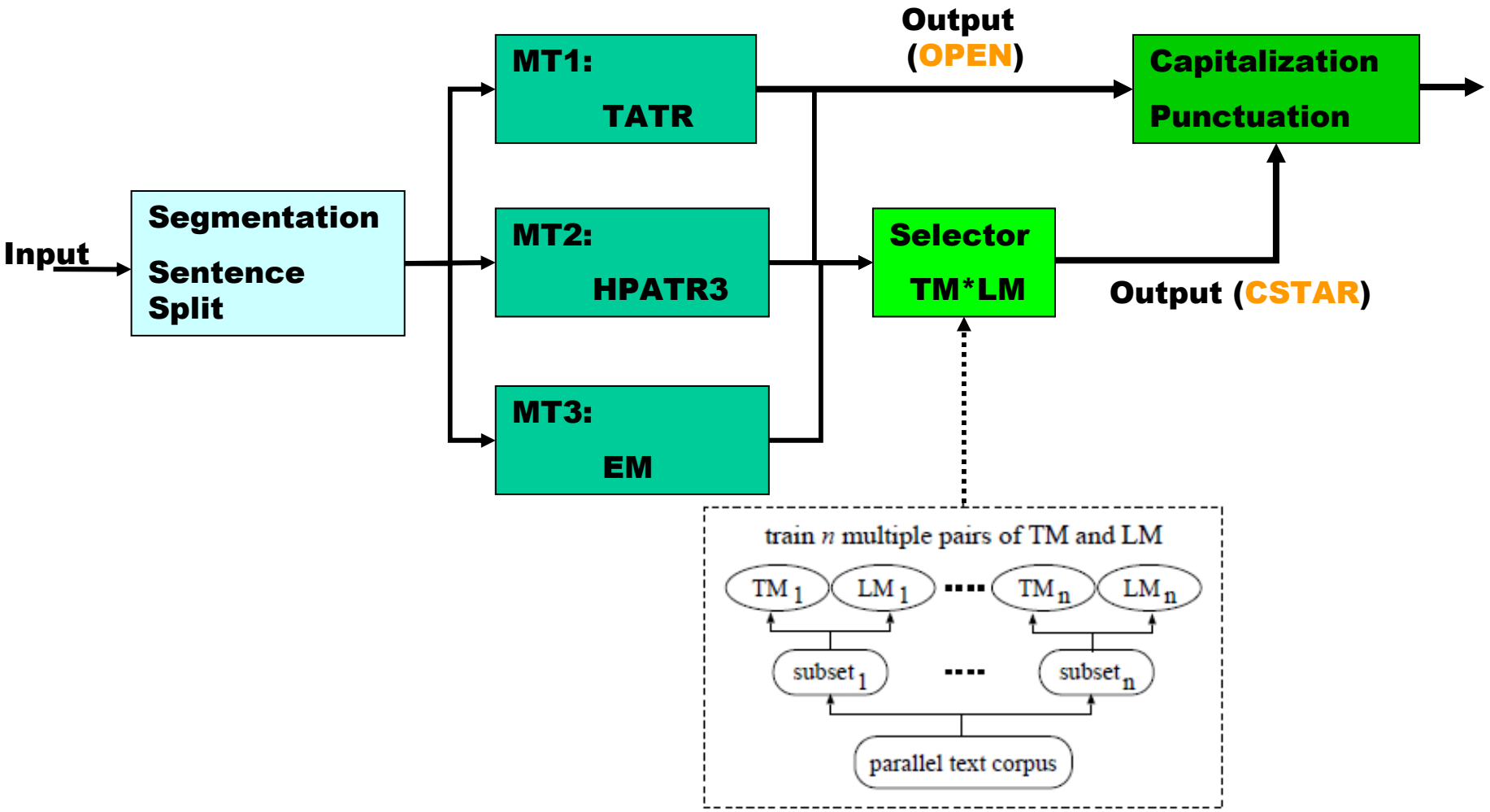
Punctuation and Capitalization

- Punctuation tool: Use SRI hidden-ngram
- Capitalization tool: an in-house CRF-based tagger
  - Capitalization as part-of-speech tagging
    1. AL(All Lowercase)
    2. AU(all uppercase)
    3. IU(initial uppercase)
    4. MX(mix case)
  - Label example:

    "McAdam is the CEO of a British company"

    "mcadam/MX is/AL the/AL ceo/AU of/AL a/AL british/IU company/AL"

System overview



Tools and Resources

- Morphological analysis
  - Achilles (CWS), BAMA(Arabic), SRI (LM, punctuation)
- Training translation model
  - GIZA++, Pharaoh
- Language data resources:
  - BTEC, LDC

		#Sentences		#Word Count	
		Source	English	Source	English
CE	OPEN	39,953(37,559)	39,953(39,633)	342,362	367,265
	CSTAR	678,748(399,527)	716,280(358,681)	4,606,373	5,756,026
JE	OPEN	39,953(37,173)	39,953(39,633)	398,498	367,265
	CSTAR	691,711(490,499)	651,558(444,859)	6,795,833	5,514,327
AE	OPEN	19,972(19,777)	19,972(19,880)	154,279	183,673
IE	OPEN	19,972(19,641)	19,972(19,880)	171,764	183,673

Evaluation tracks

- Open track
  - ASR spontaneous: CE
  - ASR read translation: CE, JE, AE,IE
  - Correct recognition result: CE, JE, AE, IE
- CSTAR track
  - ASR spontaneous: CE
  - ASR read translation: CE, JE
  - Correct recognition result: CE, JE

Translation results: Single engine vs. multiple engine

Table: BLEU score in the CSTAR track (with punctuation and case sensitive)

	CE spontaneous	CE read	JE read
SELECTOR	0.2008	0.2155	0.2487
TATR	0.2002	0.2189	0.2463
HPATR3			0.2177

Remark: Mixed results made by SELECTOR. Phrase-based TATR gives the best results.

Overall translation results: BLEU scores

	OPEN w/ca. punct.	OPEN w/o ca. punct.	CSTAR w/ca. punct.	CSTAR w/o ca. punct.
CE spont.	0.1591	0.1615	0.2008	0.2039
CE read	0.1775	0.1772	0.2155	0.2214
CE CRR	0.206	0.2123	0.2654	0.2751
JE read	0.1899	0.1832	0.2487	0.2466
JE CRR	0.2122	0.2077	0.2861	0.2867
AE read	0.2117	0.2164		
AE CRR	0.2365	0.2463		
IE read	0.2989	0.3194		
IE CRR	0.3763	0.412		

Result analysis

- BLEU scores: CRR (+4)> read speech(+1.5) > spont. speech
- BLEU scores: IE (+10)> AE(+2) >JE(+2) > CE
- For AE and IE, without ca. and punct. > with ca. and punct.
- For CE and JE, without ca. and punct = with ca. and punct.
- Capitalization and punctuation is important for AE and IE.

Conclusions

- We used some pre-processing approaches.
  - Word segmentation
  - Language model adaptation
  - Sentence splitting
  - Subword-based translation
- Phrase-based translation engine (TATR)