The NiCT/ATR Speech Translation System for IWSLT 2007

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

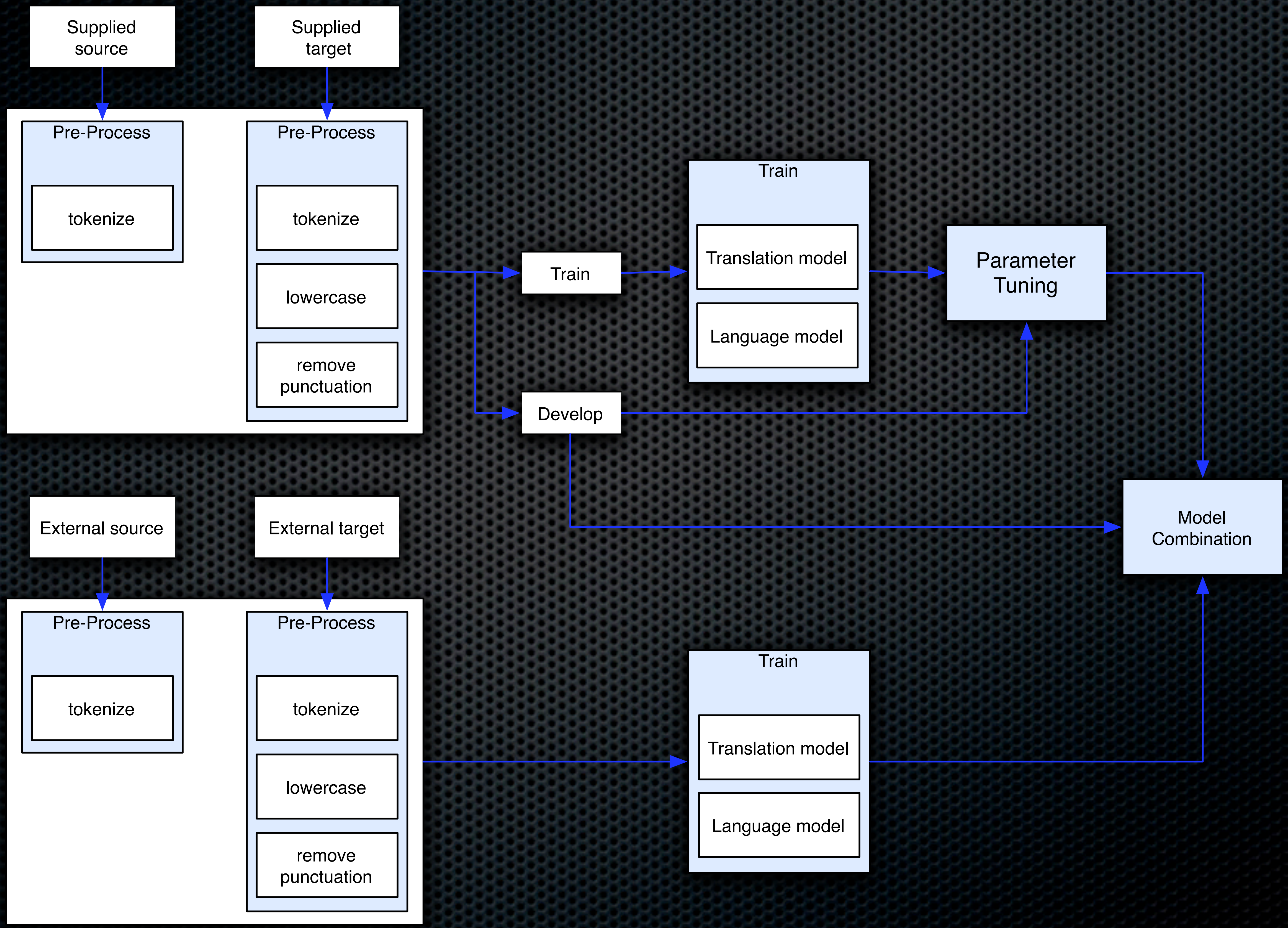
Overview

- Phrase-based SMT approach
  - In-house CleopATRa multi-stack decoder
- Participated in tracks CE, JE, IE
- Decoded from *n*-best lists
  - Tried decoding directly from confusion networks
- Focus was on the utilization of external resources

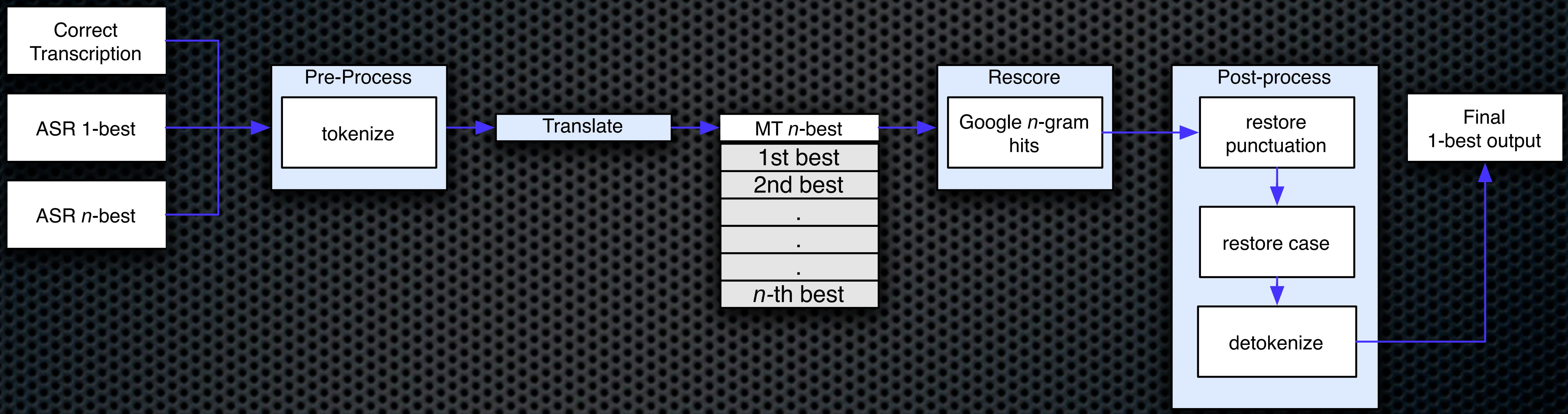
Translation System models

- Inverse phrase translation probability
- Lexical weighting probability from source to target
- Inverse lexical weighting probability
- Phrase penalty
- Language model probability
- Simple distance-based distortion model
- Word penalty

Translation System (training)



Translation System (decoding)



Division of the Tasks

- Post-processing (punctuation and case restoration) and rescoring handled in the same way for all language pairs
- Pre-processing to decoder output handled by independent teams, one team for each language pair
  - Therefore differing approaches are sometimes taken to solve the same tasks (e.g. sentence selection from the external corpora)

Punctuation and Case

- Large differences in BLEU can arise from different schemes of punctuation and casing
- Pilot experiments were conducted on Italian-English
  - Better to lowercase and remove punctuation
  - Recover case and punctuation in post-processing
  - The optimal scheme may depend on the language pair

Punctuation restoration

- Two approaches evaluated
  - ME model
  - SRI LM Toolkit's *hidden-ngram* tool
- *hidden-ngram* tool more effective
- Models built on supplied and external corpora were combined by linear interpolation

Case Restoration

- Hidden-ngram mode
- CRF tagging model
  - 3 tags (all upper, all lower, initial capital)
  - Mixed case words handled using a dictionary
  - Only lexical features
- CRF model superior
  - Used for all experiments

Hit-rate-based Skip *n*-gram Rescoring

- Huge set of 5-grams from Google Inc.
  - Hard to deal with the size
  - Use a technique based on *n*-gram hit counting
    - Use only 4-gram and 5-gram counts
    - Allow holes in the *n*-grams
  - Rescore using a weighted function of the count

Results

Data	Rescoring	BLEU	NIST	METEOR
dev5a	no	0.4288	9.1800	0.6944
	yes	0.4434	9.3165	0.7110
dev5b	no	0.2056	5.4001	0.5265
	yes	0.2089	5.4023	0.5351

* In the real evaluation this technique degraded performance

Chinese⇒English

source	# sentences	Description
IWSLT07 supplied corpus	40K	provided by IWSLT 2007
Chinese Olympic corpus	50K	part of the CLDC 2004-863-009
LDC	2.5M	LDC corpus LDC2002T01 LDC2004T07 LDC2004T08 LDC2003T17

Chinese⇒English

- Lemmatization
  - The English words ‘do’ ‘doing’ ‘did’ and ‘done’ should all map to the same word
  - Only used to improve word alignment (not used in the phrase table)
- External resources included by linearly interpolating their models (weights selected by hand by tuning on development data)

Results

TM	BLEU
IWSLT07 provided corpus	46.65
Provided+LDC	49.70
Provided+LDC (lemmatizing for alignment)	50.48
Provided+Olympic+LDC (lemmatizing)	51.78
Provided+Olympic+LDC+MERT (lemmatizing)	57.32

Italian⇒English

- 20K Supplied corpus
- 940K selected from EUROPARL data
  - Filtered: length ratio > 0.85 (based on pilot expts)

Italian⇒English

- Linearly interpolated translation models
  - Gains on dev5a, BUT no gain on dev5b
  - Therefore not used for primary system
- EUROPARL was helpful for language modeling
  - EUROPARL LM was interpolated with LM from supplied data

Japanese⇒English

- In addition to the supplied corpus we used:
  - The Tanaka corpus (203K sentence pairs)
  - The Yomiuri News corpus (202K sentence pairs)
  - The SLDB corpus (72K sentence pairs)
  - The Chinese Olympic corpus included in the Chinese-LDC (104K sentence pairs)

Japanese⇒English

- Tokenization - CHASEN (publicly available)
- Training sentences were selected from external corpora
  - Build tri-gram LM from supplied corpus
  - Select sentences based on LM perplexity W.R.T. the LM (perplexity < 100)
  - After selection 40K supplied and 117K external sentence pairs available for training

Japanese⇒English

- *n*-best decoding
  - 20-best ASR hypotheses decoded
  - Decoding directly from Confusion Network gave similar performance (within 0.002 BLEU)
    - *n*-best decoding simpler and more flexible
    - No tokenization issues (must accept ASR tokenization if using CN)
  - ASR scores added as a log-linear feature
    - Weight learned independently (maximize BLEU)

Additional Experiments

- Use longer phrases
  - Maximum phrase length 12 instead of 7
- Use lexical re-ordering model
  - The same model used in MOSES
- We do not use cluster-based models
- We decode from 1-best rather than *n*-best

Responsible for about 2 BLEU points

Results (BLEU)

	3-gram	4-gram	5-gram
Baseline	39.51	41.20	41.43
Long phrases	40.22	41.79	41.82
Long phrases + lexical reordering	40.68	42.04	$\boxed{42.24}$

Conclusions

- Case, punctuation and tokenization choices have a large impact on overall system performance
- Additional out-of-domain data can help, but can harm if not used carefully
  - Select sentences based on similarity to the in-domain corpus
  - Verify effectiveness on development data
- Longer phrases can be effective

The End

Thank you!