



The TÜBİTAK-UEKAE Statistical Machine Translation System for IWSLT 2007

Coşkun Mermer, Hamza Kaya, Mehmet Uğur Doğan

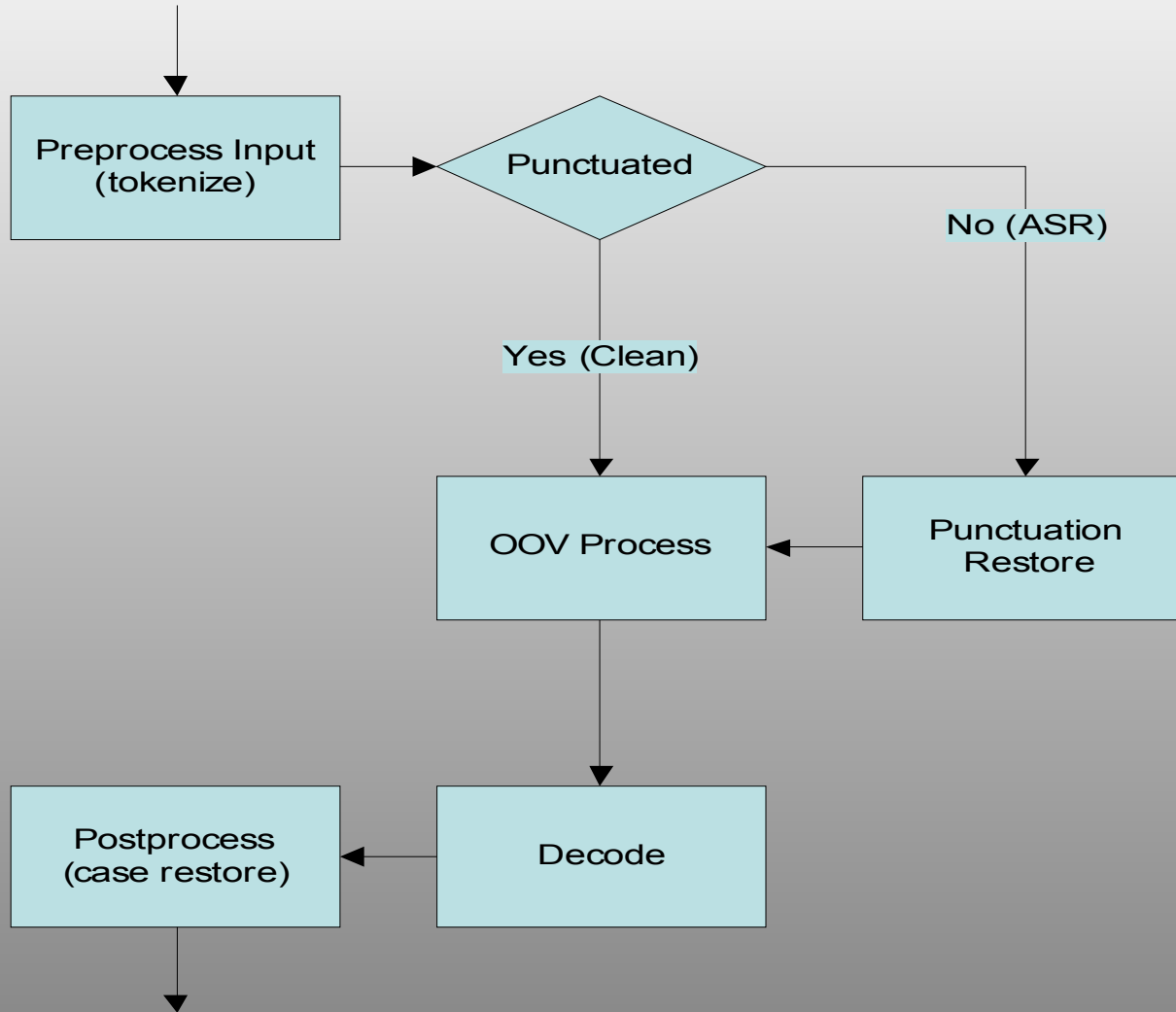
National Research Institute of Electronics and Cryptology (UEKAE)
The Scientific and Technological Research Council of Turkey (TÜBİTAK)
Gebze, Kocaeli, Turkey

{coskun, hamzaky, mugur}@uekae.tubitak.gov.tr

- System Description
- Training
 - Phrase table augmentation
- Decoding
 - Out of Vocabulary Words (OOV)
- Results
- Conclusion and Future Work

- Participated in translation tasks
 - Arabic-to-English
 - Japanese-to-English
- Built on phrase-based SMT software Moses
- Used only supplied data and Buckwalter Arabic Morphological Analyzer (BAMA)

System Description



- System Description
- Training
 - Phrase table augmentation
- Decoding
 - Out of Vocabulary Words (OOV)
- Results
- Conclusion and Future Work

- Devset1-3 are included in the training with all 16 reference segments
- Train and Devset1-3 are given equal weight
- Language models
 - 3-gram for AR-EN
 - 4-gram for JP-EN
 - Trained with modified Kneser-Ney discounting and interpolation

- Multi-sentence segments are split

Before splitting

After splitting

AR-EN

44,164 *

49,318

JP-EN

64,145 *

71,435

* train segments + 16 * dev1-3 segments

- Parameter tuning
 - Manually tested different set of parameters
 - Different data favored different parameters
 - Instead of selecting argmax , selected *mode* in a *desirable interval* to select a robust set of parameters

Phrase Table Augmentation

- Translation model is represented in a *phrase table*
- Bi-directional alignment and phrase extraction with *grow-diag-final-and* heuristics
- Source-language words without a one-word entry in phrase table are listed
- The words, which are in the list and have a lexical translation probability above a threshold in *GIZA++* word alignment, are added to phrase list

Phrase Table Augmentation

| Corpus | AR-EN | JP-EN |
|---|---------|---------|
| Source vocabulary size | 18,751 | 12,699 |
| Number of entries in the original phrase table | 408,052 | 606,432 |
| Number of source vocabulary words without a one-word entry in the original phrase table | 8,035 | 6,302 |
| Number of one-word bi-phrases added to the phrase table | 21,439 | 23,396 |
| Number of entries in the augmented phrase-table | 429,491 | 629,828 |

- System Description
- Training
 - Phrase table augmentation
- **Decoding**
 - Out of Vocabulary Words (OOV)
- Results
- Conclusion and Future Work



- Decoding is done on tokenized and punctuated data
 - Source-side punctuation insertion (for ASR data)
 - Target-side case restoration
- SRILM tools used for punctuation restoration



- Merged 10 sentences to train punctuation restorer with more internal sentence boundaries

| | N | Devset4 | Devset5 |
|-------|----|---------|---------|
| AR-EN | 1 | 24.32 | 20.23 |
| | 10 | 24.95 | 20.66 |
| JP-EN | 1 | 15.59 | 14.26 |
| | 10 | 17.82 | 16.12 |

- Lexical Approximation
 - Find a set of candidate approximations
 - Select the candidate with least edit distance
 - In case of a tie, more frequently used candidate is chosen



- Arabic lexical approximation (2 pass)
 - Morphological root(s) of the word found by feature function using BAMA
 - If not, *skeletonized* version of the word is found by feature function
- Japanese lexical approximation (1 pass)
 - Right-truncations of the word is found by feature function

- Run-time Lexical Approximation

| AR-EN | Devset4 | | Devset5 | |
|------------|-----------|-------|-----------|-------|
| | # of OOVs | BLEU | # of OOVS | BLEU |
| Original | 661 | 24.91 | 795 | 20.59 |
| After LA#1 | 185 | 25.33 | 221 | 21.22 |
| After LA#2 | 149 | 25.56 | 172 | 21.51 |

| JP-EN | Devset4 | | Devset5 | |
|----------|-----------|-------|-----------|-------|
| | # of OOVs | BLEU | # of OOVS | BLEU |
| Original | 119 | 23.68 | 169 | 20.44 |
| After LA | 10 | 23.84 | 17 | 20.69 |

Without lexical approximation

hl hw mjAny~ ?



Is it mjAny~ ?

mjAny~ is OOV

Out of Vocabulary Words

- Lexical approximation finds candidates
 - mjAnyP, mjAnY, mjAnA, kjm, mjAny, mjAnAF
- mjAny has an edit distance of 1, so it's selected

After lexical approximation

hl hw mjAny ?



Is it free ?

Outline

- System Description
- Training
 - Phrase table augmentation
- Decoding
 - Out of Vocabulary Words (OOV)
- **Results**
- Conclusion and Future Work



Clean Transcript

ASR Output

AR-EN

49.23

36.79

JP-EN

48.41

42.69



- Possible causes of performance drop in ASR condition
 - Recognition errors of ASR
 - Punctuation restorer performance
 - Parameter tuning for clean transcript but not for ASR output

- Possible causes of higher performance drop in AR-EN than JP-EN
 - Lower accuracy of Arabic ASR data than Japanese data
 - Higher difficulty of punctuation insertion due to higher number of punctuation types
 - Less reliable punctuation insertion caused by higher recognition error rate

- Lexical approximation is sensitive to recognition errors

| | Clean transcript | ASR output | Clean-to-ASR degradation |
|-----------------|------------------|------------|--------------------------|
| Original source | 38.48 | 31.82 | 17.3% |
| After LA | 49.23 | 36.79 | 25.3% |

Devset4-5 vs. Evaluation Set

- There is a dramatic variation in the improvement obtained with the lexical approximation technique on the evaluation and development sets

Devset4-5 vs. Evaluation Set

| | Devset4 | Devset5 |
|-----------------|------------------------------------|------------------------------|
| Original source | 24.91 | 20.59 |
| After LA#1 | 25.33 | 21.22 |
| After LA#2 | 25.56 | 21.51 |
| Improvment | 2.6% | 4.5% |
| | Evaluation set clean transcript | Evaluation set ASR output |
| Original source | 38.48 | 31.82 |
| After LA | 49.23 | 36.79 |
| Improvment | 27.9% | 15.6% |

Devset4-5 vs. Evaluation Set

- 167 of 489 evaluation set segments have at least one reference which is a perfect match with a training segment
- Only 19 of 167 have the source segment exactly the same as in the training set
- Remaining 148 segments represents a potential to obtain a perfect match



Devset4-5 vs. Evaluation Set



| Number of segments | Devset4 | Devset5 | Evaluation set |
|--|---------|---------|----------------|
| Exact match of at least one reference with a segment in the training set | 12 | 4 | 167 |
| Exact math of the source with a segment in the training set | 1 | 0 | 19 |
| Total | 489 | 500 | 489 |

Outline

- System Description
- Training
 - Phrase table augmentation
- Decoding
 - Out of Vocabulary Words (OOV)
- Results
- **Conclusion and Future Work**

- Make the system more robust to ASR output. For this goal:
 - Using n-best/lattice ASR output
 - Tuning system for ASR output
 - Better punctuation performance



Thank you for your attention!