IITP English-Hindi Machine Translation System at WAT 2016

Sukanta Sen, Debajyoty Banik, Asif Ekbal, Pushpak Bhattacharyya
Department of Computer Science and Engineering
Indian Institute of Technology Patna, India

Methods

- Two Hierarchical En-Hi SMT
  - First: Without external data
  - Second: With external data (bilingual dictionary)
- Source-side reordering: to conform the syntactic structure of the target.
- Baseline: Phrase-based SMT

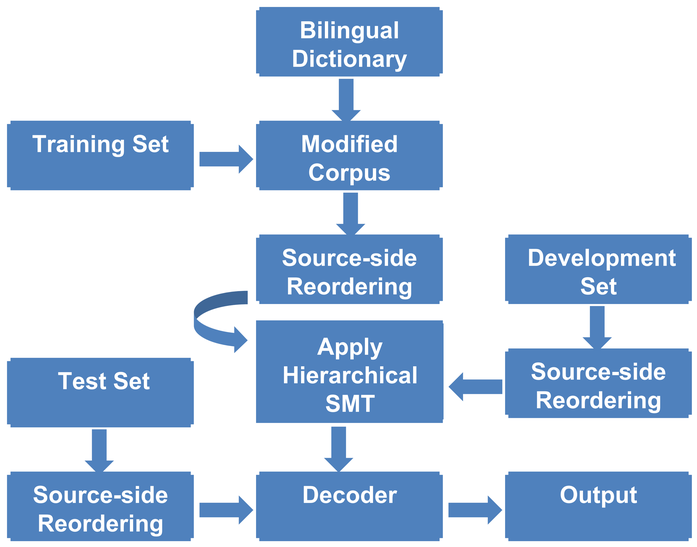
Source-side Reordering

- English → subject-verb-object (SVO)
- Hindi → subject-object-verb (SOV)
- Syntactic reordering of the source helps (Collins et al., 2005; Ramanathan et al., 2008)
- Reordering example
  - English: The president of America visited India in June.
  - Reordered: America of the president June in India visited.
  - Hindi: अमेरिका के राष्ट्रपति ने जून में भारत की यातरा की।

Augmenting Bilingual Dictionary

- Motivation: It improves word alignment
- English-Hindi bilingual mapping[3] → extract parallel words
- Augment parallel words with training corpus

IITP Machine Translation Pipeline



Experimental Setup

- Training using Moses Toolkit
- Stanford parser for parsing source sentences
- Source-side reordering using CFILT pre-ordering tool[4]
- GIZA++ with grow-diag-final-and heuristic for word-alignment
- 4-gram language model with modified Kneser-Ney smoothing using KenLM
- Distortion limit 6 (only for baseline Phrase-based SMT)
- Minimum-Error-Rate-Training (MERT) tuning

Results

Built SMT systems
- Phrase-based(Phr)
- Hierarchical model(Hie)
- Hierarchical model after reordering source(HieRe)
- Hierarchical model with dictionary and reordering (HieReDict)

	Phr	Hie	HieRe	HieReDict
BLEU	11.79	13.18	13.57	13.71

Error Analysis

An example

Source: *the rain and cold wind on Wednesday night made people feel cold.*

Hie: बुधवार रात को बरसात और ठंडी हवा **ने** लोगों को ठंड लग रह थी।
(budhavAra rAta ko barasAta aura ThaMDI havA **ne** logoM ko ThaMDa laga rahI thI.)

Observations → wrong post-position ने (ne)

HieRe: बुधवार की रात को बरसात और ठंडी हवा **से** ठंडक महसूस हुई।
(budhavAra kI rAta ko barasAta aura ThaMDI havA **se** ThaMDaka mahasUsa huI.)

Observations → 1. Correct post-position से (se)
2. लोगों (logoM) (Gloss: people) is dropped

HieReDict: बुधवार रात बारिश और ठंडी हवा से लोगों को ठंड लगने लगा।
(budhavAra rAta bArisha aura ThaMDI havA se logoM ko ThaMDa lagane lagA.)

Observations → लोगों(logoM) (Gloss: people) brought back

Conclusion

We found that hierarchical SMT model, when augmented with bilingual dictionary along with syntactic reordering of English sentences produced better translation score.

References

1. Ananthakrishnan Ramanathan, Jayprasad Hegde, Ritesh M Shah, Pushpak Bhattacharyya, and M Sasikumar. 2008. Simple syntactic and morphological processing can help english-hindi statistical machine translation. In IJCNLP, pages 513–520.
2. Michael Collins, Philipp Koehn, and Ivona Kučerová. 2005. Clause restructuring for statistical machine translation. In Proceedings of the 43rd annual meeting on association for computational linguistics, pages 531–540. ACL.
3. http://www.cfilt.iitb.ac.in/ sudha/bilingual_mapping.tar.gz
4. http://www.cfilt.iitb.ac.in/ moses/download/cfilt_preorder