AppTek's APT Machine Translation System for IWSLT 2010

Evgeny Matusov, Selçuk Köprü
{ematusov,skopru}@apptek.com

Applications Technology, Inc.

From Science to Solutions

Overview

- AppTek's new *APT* machine translation system
- Arabic-to-English and Turkish-to-English BTEC tasks
- Competitive translation quality is obtained with a system that can be seamlessly turned into a real-life product.

Baseline System

- State-of-the-art phrase-based SMT decoder similar to MOSES
- **New:** run-based reordering penalty model [5]
- Efficient translation of multiple input paths (lattices)
- Minimum-error-rate training (optimize BLEU on development data)

Morphological Analysis and Segmentation

- Morphological analysis and segmentation is needed because of the complex morphology of Arabic and Turkish.
- Morphological disambiguation is performed by syntactic analysis of the sentence [1, 2].
- Morphemes are detached from each other for a better correspondence with English words and to reduce the out-of-vocabulary (OOV) rate.
- Arabic: detach all prefixes, suffixes, and the definite article
- Turkish: words have a clear but complex morphotactic structure ⇒ rule-based segmentation

Arabic-to-English Experiments

Things which helped:

- using multiple word alignments for phrase extraction
  - heuristic combination (grow-diag-final, "ACL", intersection-union) of IBM model 4 and HMM alignments
- using several alternative morphological segmentation schemes for phrase extraction and translation
  - segmentation of all/infrequent words or no segmentation
  - efficient translation of a segmentation lattice
  - option to remove Arabic article/accusative marker in translation
- lexicalized run-based reordering model
  - probability of deviating from the monotonic path conditioned on the word that starts the new "run" and the last covered word
- sentence-level inverse IBM model 1

Things which did not help:

- "translation memory"
- character-level edit-distance based OOV correction
- POS-based OOV classification and phrase-level "translation"
  - assign POS tags to OOVs, translate using generalized phrase pairs
  - predict correct position of the OOV based on phrasal context

	BLEU and TER scores in % on the IWSLT 2010 evaluation set	BLEU	TER
1	no morphological segmentation	41.0	35.8
2	morphological segmentation for words w with $N(w) \leq 100$	45.3	32.5
3	full morphological segmentation	45.2	32.3
4	+ optionally remove article/accusative marker	45.7	32.0
5	+ multiple morphological segmentation paths	46.6	31.9
6	+ POS-based reordering model	46.0	31.8
7	+ sentence-level inverse IBM model 1	46.6	31.4
8	+ lexicalized reordering model	46.8	31.4
9	like 3., but single alignment (IBM model 4, ACL heuristic)	44.7	33.6
10	like 5., but monotonic translation	43.7	33.7
11	like 8., but no translation memory	**47.0**	**30.8**
12	like 11., but no restoration of contractions	46.3	32.5
	primary submission	45.7	32.9

Turkish-to-English Experiments

- `devset2` is used for the optimization of model scaling factors
- `devset1` is used as the test set.

Language Model Experiments

Different n-gram models are estimated and evaluated.

Word Alignment Experiments

The effect of different heuristics on symmetrization of word alignments is explored. Multiple alignments are used for phrase extraction.

LM Experiments		
n-gram	Opt. BLEU	Test BLEU
3	57.15	59.56
4	57.73	60.67
5	57.31	**61.55**
6	57.66	60.81
7	**57.76**	60.96
8	57.58	60.74
9	57.61	59.86

Alignment Experiments		
Method	Opt. BLEU	Test BLEU
ACL	57.18	59.87
grow-diag-final	53.75	56.41
intersection	55.80	59.27
inters.-union	57.30	59.68
unify	52.57	55.69
left	56.04	59.65
right	54.02	56.74
berkeley	55.12	57.81
subset merged	**57.70**	61.14
all merged	57.31	**61.55**

Other Experiments

- DESCRIBE THE EXPRERIMENTS
-

	Opt. BLEU	Test BLEU
mapping contractions	58.22	62.37
multiple segmentations	57.93	60.92
truecased opt devset1	63.26	59.46
truecased opt devset2	60.87	62.10
truecased opt devset1+2	61.84	N/A

Conclusions

- AppTek's new *APT* MT system is competitive: **ranked 3rd in Ar-En, 4th in Tr-En**.
- The system utilizes comprehensive morphological analysis components to deal with rich morphology languages.
- The biggest gains are achieved by considering morphological segmentation alternatives and by employing novel reordering models.
- The translation systems produce translations at a speed of 12 words per second or more.
- AppTek's IWSLT systems are efficient enough to be used in hand-held devices

References

[1] S. Köprü and J. Miller, "A Unification Based Approach to the Morphological Analysis and Generation of Arabic," in Proc. of the CAASL3, Ottawa, Ontario, Canada, 2009.

[2] S. Köprü, "AppTek Turkish-English Machine Translation System Description for IWSLT 2009," in *Proc. of the International Workshop on Spoken Language Translation*, Tokyo, Japan, 2009, pp. 19–23.

[3] R. Zens, "Phrase-based statistical machine translation: Models, search, training," Ph.D. dissertation, RWTH Aachen University, Aachen, Germany, February 2008.

[4] F. J. Och and H. Ney, "A systematic comparison of various statistical alignment models," *Computational Linguistics*, vol. 29, no. 1, pp. 19–51, March 2003.

[5] E. Matusov and S. Köprü, "Improving Reordering in Statistical Machine Translation from Farsi," in *AMTA 2010: The Ninth Conference of the Association for Machine Translation in the Americas*, Denver, Colorado, USA, November 2010.

[6] J. DeNero and D. Klein, "Tailoring Word Alignments to Syntactic Machine Translation," in *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*. Prague, Czech Republic: Association for Computational Linguistics, June 2007, pp. 17–24.