The CMU-Oxford Translation System for the NIST Open Machine Translation 2012 Evaluation

1 Site affiliation

Carnegie Mellon University and the University of Oxford

2 Contact information

- Chris Dyer, Carnegie Mellon University, cdyer@cs.cmu.edu
- Noah A. Smith, Carnegie Mellon University, nasmith@cs.cmu.edu
- Graham Morehead, University of Maine, gm@pangeon.com
- Phil Blunsom, University of Oxford, Phil.Blunsom@cs.ox.ac.uk
- Abby Levenberg, University of Oxford, Abby.Levenberg@cs.ox.ac.uk

3 Submissions

CMU_kor2eng_cn_primary

4 Primary system specs

We submitted a Korean-English system (CMU_kor2eng_cn_primary), which is described in the following sections.

4.1 Core MT engine algorithmic approach

The core of our system is the hierarchical phrase-based translation model (Chiang, 2007), as implemented by the cdec decoder (Dyer et al., 2010).[1] A 4-gram language model estimated using modified Kneser-Ney smoothing was included (Chen and Goodman, 1999). Translation model features include the log relative frequency, $\log f(\mathbf{e} \mid \mathbf{k})$, the log counts of $\mathbf{k}$ and $\mathbf{e}, \mathbf{k}$, the log "lexical translation" probabilities in both directions, indicator features for rule counts of 1. Translation model parameters were tuned using the dynamic programming variant of minimum error rate training for hypergraphs to maximize the BLEU score on a held-out development set with a single reference translation (Kumar et al., 2009; Papineni et al., 2002).

In our training, development, and test data, we added sentence-begin and sentence-end markers, represented with the symbols $\langle s \rangle$ and $\langle s \rangle$. This gives the model somewhat more flexibility distinguishing translations that occur at the beginning and endings of sentences.

[1] http://cdec-decoder.org

4.2 Significant pre/post-processing

Preprocessing and segmentation of Korean text into lexemes suitable for translation is a nontrivial task. On one hand, Korean is an agglutinative language featuring an extensive system of case markers, particles, and a verbal inflection marking valency, mood, aspect, tense, formality, and several other syntactic and semantic features. As a result, individual inflected forms can be extremely precise, making distinctions that are not part of English morphology. On the other hand, Korean orthographic convention further complicates matters. First, it permits several inflected lexemes to be written together as single whitespace-delimited units called *eojeol* (Choi et al., 2009). Second, morphophonological processes such as vowel harmony and resyllabification may cause the syllabic "blocks" that are the basis of Korean text encoding to change in form.[2] Thus, recovering underlying morphemes requires more substantial analysis than simple segmentation.

To recover the Korean word sequences used as input to our translation and alignment models, the Korean portions of the training, development, and test data were analyzed using a rule-based finite-state morphological analyzer (Park et al., 2010) that produces an unweighted list of possible analyses for each *eojeol*. We assume that the underlying morphemes $k_1, k_2, \ldots$ are generated by a generative process that encodes our prior beliefs about the distribution of morphemes in the language. *A priori* we believe that the morpheme bigrams should have a power-law distribution (a few very common morpheme bigrams and a long tail of less frequent morpheme bigrams), so we use the following hierarchical Pitman-Yor process, which produces outputs with this distribution with high probability, and has been shown to be an effective segmentation model in fully unsupervised cases (Goldwater et al., 2009):[3]

$$\begin{aligned}
P_0(\mathbf{y}) &= \left(\frac{1}{2}\right)^{|\mathbf{y}|} \times \left(\frac{1}{2}\right) \times \frac{1}{|\Sigma|^{|\mathbf{y}|}} \\
G_0 &\sim \mathrm{PY}(d_0, \theta_0, P_0) \\
G_y &\sim \mathrm{PY}(d, \theta, G_0) \\
k_i \mid k_{i-1} &\sim G_{k_{i-1}}
\end{aligned}$$

Posterior inference over analyses was carried out using a dynamic programming block sampler to resample the segmentations of an entire sentence all at once (Mochihashi et al., 2009). Figure 1 shows an example sentence from our training data, together with the MAP segmentation.

For word alignment and translation model training data, we select the sequence of morphemes with the highest posterior probability. For the test data, we consider a lattice weighted with the posterior probability of each segmentation and use the entire lattice as the input to the decoder (Dyer et al., 2008).

English data consisting of the target side of the parallel text and the Gigaword v5 was tokenized and lower-cased. After translation, output was recased using a HMM-based recasing model, as implemented by the SRILM's `disambig` tool (Stolcke, 2002).

4.3 Additional features and tools used

Previous work has demonstrated that using word alignments from several different alignment models (or multiple alternative alignments from a single model) when extracting translation grammars can

[2] While Hangul is an alphabetic script, it is written together in syllabic blocks.

[3] Unlike the Goldwater et al. (2009) work, we constrain the inferred word sequences by the output of the morphological analyzer. This lets us simultaneously rule out impossible analyses and consider underlying forms that are related to the surface form by processes other than simple concatenation.

Input: 북한 당국은 이들 전시회를 해마다 개최할 계획이다.

The North Korean authorities plan to hold these events annually.

Surface	Analyzer Outputs	Gloss
북한	**북한**	***North***
당국은	**당국 +은**	***authorities*** +NOM
이들	**이 +들**	***this*** +PL
전시회를	**전시회 +를**	***exhibition*** +ACC
	전시 +회 +를	*display* +*time* +ACC
해마다	해마다	*yearly*
	해마 +이 +다	*sea horse* +*this* +EX
	해 +마다	***year*** +***each***
	하 +어 +마다	*one* +*term* +*each*
개최할	**개최 +하 +ㄹ**	***hold*** +VB +ADNOM
계획이다	**계획 +이 +다**	***plan*** +***this*** +EX
	계획 +이 +이 +다	*plan* +*village* +*this* +EX

Figure 1: Morphological analyzer outputs and their glosses for an example Korean sentence. The analyses selected by the nonparametric language model are shown in **bold face**.

improve translation quality (Venugopal et al., 2008; Dyer et al., 2011). One possible explanation is that models with different biases make different systematic alignment errors, so getting grammar rules from different alignment types gives you a better coverage of the contents of limited training data. We therefore chose to include alignments generated from the following models, each of which make substantially different assumptions during learning:

- IBM Model 4, as implemented by GIZA++ (Och and Ney, 2002)
- A log-linear parameterization of the Model 2 alignment model. In our variant of Model 2, the prior alignment log probability is linear in the distance to the diagonal, normalized by the source length. Lexical translation parameters are learned with EM.
- A word-based ITG alignment model with MCMC inference carried out using an auxiliary variable sampling technique (Blunsom and Cohn, 2010).
- A Bayesian SCFG alignment model based on Pitman-Yor processes (Levenberg et al., in review).

4.4 Data used

Parallel data. Translation models were learned from the Korean-English portions of the FBIS corpus (LDC2003E14) and a parallel corpus of automatically extracted Wikipedia titles.[4] 125 sentences were removed from the training data and used for development.[5]

Monolingual data. A 3-gram language model was constructed from the target side of the parallel training data. A second 4-gram language model was constructed by interpolating two language

[4] `http://www.cs.cmu.edu/~cdyer/wikipedia-201201.ko-en.gz`

[5] Due to the poor quality of our parallel data, finding truly parallel sentences for development was a challenge. Ideally, a much larger set with multiple reference translations would have been used.

models so as to minimize perplexity on a held-out development set (the first English reference from the NIST MT06 Chinese-English was used as the development data). The first was constructed from the New York Times (NYT) portion of Gigaword v5, and the second was constructed from the remaining Gigaword data.[6]

5 Summary

Table 1 summarizes the performance of our Korean-English system under various conditions. We report scores on the development set used by minimum error rate training since we did not have enough data to produce a reliable test set. Without segmentation of the *eojeol*, the system performs quite poorly, due largely to large numbers of OOV items in the output.

Table 1: Korean-English development set BLEU score reached under various configurations explored.

Condition	BLEU
unsegmented + Model 4 only	1.4
segmented + Model 4 only	18.4
segmented + Model 4 only + explicit $\langle s \rangle$	18.8
segmented + all alignments + explicit $\langle s \rangle$	21.6

References

Blunsom, P. and Cohn, T. (2010). Inducing synchronous grammars with slice sampling. In *Proc. NAACL*.

Chen, S. F. and Goodman, J. (1999). An empirical study of smoothing techniques for language modeling. *Computer Speech & Language*, 13(4):359–393.

Chiang, D. (2007). Hierarchical phrase-based translation. *Computational Linguistics*, 33(2):201–228.

Choi, K.-S., Isahara, H., Kanzaki, K., Kim, H., Pak, S. M., and Sun, M. (2009). Word segmentation standard in Chinese, Japanese and Korean. In *Proc. of the Workshop on the Interaction between Linguistics and Computational Linguistics*.

Dyer, C., Clark, J., Lavie, A., and Smith, N. A. (2011). Unsupervised word alignment with arbitrary features. In *Proc. of ACL*.

Dyer, C., Lopez, A., Ganitkevitch, J., Weese, J., Ture, F., Blunsom, P., Setiawan, H., Eidelman, V., and Resnik, P. (2010). cdec: A decoder, alignment, and learning framework for finite-state and context-free translation models. In *Proc. of ACL*.

Dyer, C., Muresan, S., and Resnik, P. (2008). Generalizing word lattice translation. In *Proceedings of HLT-ACL*.

Goldwater, S., Griffiths, T. L., and Johnson, M. (2009). A Bayesian framework for word segmentation: Exploring the effects of context. *Cognition*, 112:21–54.

Kumar, S., Macherey, W., Dyer, C., and Och, F. (2009). Efficient minimum error rate training and minimum Bayes-risk decoding for translation hypergraphs and lattices. In *Proc. of ACL-IJCNLP*.

[6] While the NYT constitutes by far the largest portion of Gigaword, our previous experience suggests that data does not generalize well for the purposes of translation. Indeed, the interpolation weights we learned were $\lambda_{\text{NYT}} = 0.25$ and $\lambda_{\neg\text{NYT}} = 0.75$.

Levenberg, A., Dyer, C., and Blunsom, P. (in review). A Bayesian model for learning SCFGs with discontiguous rules. In *Proc. of EMNLP*.

Mochihashi, D., Yamada, T., and Ueda, N. (2009). Bayesian unsupervised word segmentation with nested Pitman-Yor language modeling. In *Proc. of ACL*.

Och, F. J. and Ney, H. (2002). Discriminative training and maximum entropy models for statistical machine translation. In *Proceedings of ACL*, pages 295–302.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). BLEU: a method for automatic evaluation of machine translation. In *Proc. of ACL*.

Park, S., Choi, D., Kim, E.-k., and Choi, K.-S. (2010). A plug-in component-based Korean morphological analyzer. In *Proceedings of HCLT 2010*, pages 197–201.

Stolcke, A. (2002). SRILM – an extensible language modeling toolkit. In *Intl. Conf. on Spoken Language Processing*.

Venugopal, A., Zollmann, A., Smith, N. A., and Vogel, S. (2008). Wider pipelines: *n*-best alignments and parses in MT training. In *Proc. of AMTA*.