NTT System Description for the WMT2006 Shared Task

Taro Watanabe **Hajime Tsukada** **Hideki Isozaki**
NTT Communication Science Laboratories
2-4 Hikaridai, Seika-cho, Soraku-gun,
Kyoto, Japan 619-0237
{taro,tsukada,isozaki}@kecl.ntt.co.jp

Abstract

We present two translation systems experimented for the shared-task of "Workshop on Statistical Machine Translation," a phrase-based model and a hierarchical phrase-based model. The former uses a phrasal unit for translation, whereas the latter is conceptualized as a synchronous-CFG in which phrases are hierarchically combined using non-terminals. Experiments showed that the hierarchical phrase-based model performed very comparable to the phrase-based model. We also report a phrase/rule extraction technique differentiating tokenization of corpora.

1 Introduction

We contrasted two translation methods for the Workshop on Statistical Machine Translation (WMT2006) shared-task. One is a phrase-based translation in which a phrasal unit is employed for translation (Koehn et al., 2003). The other is a hierarchical phrase-based translation in which translation is realized as a set of paired production rules (Chiang, 2005). Section 2 discusses those two models and details extraction algorithms, decoding algorithms and feature functions.

We also explored three types of corpus preprocessing in Section 3. As expected, different tokenization would lead to different word alignments which, in turn, resulted in the divergence of the extracted phrase/rule size. In our method, phrase/rule translation pairs extracted from three distinctly word-aligned corpora are aggregated into one large phrase/rule translation table. The experiments and the final translation results are presented in Section 4.

2 Translation Models

We used a log-linear approach (Och and Ney, 2002) in which a foreign language sentence $f_1^J = f_1, f_2, ... f_J$ is translated into another language, i.e. English, $e_1^I = e_1, e_2, ..., e_I$ by seeking a maximum likelihood solution of

$$\hat{e}_1^I = \operatorname*{argmax}_{e_1^I} Pr(e_1^I | f_1^J) \quad (1)$$

$$= \operatorname*{argmax}_{e_1^I} \frac{\exp\left(\sum_{m=1}^{M} \lambda_m h_m(e_1^I, f_1^J)\right)}{\sum_{e'^{I'}_1} \exp\left(\sum_{m=1}^{M} \lambda_m h_m(e'^{I'}_1, f_1^J)\right)} \quad (2)$$

In this framework, the posterior probability $Pr(e_1^I | f_1^J)$ is directly maximized using a log-linear combination of feature functions $h_m(e_1^I, f_1^J)$, such as a ngram language model or a translation model. When decoding, the denominator is dropped since it depends only on f_1^J. Feature function scaling factors λ_m are optimized based on a maximum likelihood approach (Och and Ney, 2002) or on a direct error minimization approach (Och, 2003). This modeling allows the integration of various feature functions depending on the scenario of how a translation is constituted.

In a phrase-based statistical translation (Koehn et al., 2003), a bilingual text is decomposed as K phrase translation pairs $(\bar{e}_1, \bar{f}_{\bar{a}_1}), (\bar{e}_2, \bar{f}_{\bar{a}_2}), ...$: The input foreign sentence is segmented into phrases $\bar{f}_1^K$,

Proceedings of the Workshop on Statistical Machine Translation, pages 122–125,
New York City, June 2006. ©2006 Association for Computational Linguistics

mapped into corresponding English $\bar{e}_1^K$, then, reordered to form the output English sentence according to a phrase alignment index mapping $\bar{a}$.

In a hierarchical phrase-based translation (Chiang, 2005), translation is modeled after a weighted synchronous-CFG consisting of production rules whose right-hand side is paired (Aho and Ullman, 1969):

$$X \rightarrow \langle \gamma, \alpha, \sim \rangle$$

where X is a non-terminal, γ and α are strings of terminals and non-terminals. $\sim$ is a one-to-one correspondence for the non-terminals appeared in γ and α. Starting from an initial non-terminal, each rule rewrites non-terminals in γ and α that are associated with $\sim$.

2.1 Phrase/Rule Extraction

The phrase extraction algorithm is based on those presented by Koehn et al. (2003). First, many-to-many word alignments are induced by running a one-to-many word alignment model, such as GIZA++ (Och and Ney, 2003), in both directions and by combining the results based on a heuristic (Och and Ney, 2004). Second, phrase translation pairs are extracted from the word aligned corpus (Koehn et al., 2003). The method exhaustively extracts phrase pairs (f_j^{j+m}, e_i^{i+n}) from a sentence pair (f_1^J, e_1^I) that do not violate the word alignment constraints a.

In the hierarchical phrase-based model, production rules are accumulated by computing "holes" for extracted contiguous phrases (Chiang, 2005):

1. A phrase pair $(\bar{f}, \bar{e})$ constitutes a rule:

$$X \rightarrow \left\langle \bar{f}, \bar{e} \right\rangle$$

2. A rule $X \rightarrow \langle \gamma, \alpha \rangle$ and a phrase pair $(\bar{f}, \bar{e})$ s.t. $\gamma = \gamma' \bar{f} \gamma''$ and $\alpha = \alpha' \bar{e} \alpha''$ constitutes a rule:

$$X \rightarrow \left\langle \gamma' \; X_{\boxed{k}} \; \gamma'', \alpha' \; X_{\boxed{k}} \; \alpha'' \right\rangle$$

2.2 Decoding

The decoder for the phrase-based model is a left-to-right generation decoder with a beam search strategy synchronized with the cardinality of already translated foreign words. The decoding process is very similar to those described in (Koehn et al., 2003): It starts from an initial empty hypothesis. From an existing hypothesis, new hypothesis is generated by consuming a phrase translation pair that covers untranslated foreign word positions. The score for the newly generated hypothesis is updated by combining the scores of feature functions described in Section 2.3. The English side of the phrase is simply concatenated to form a new prefix of English sentence.

In the hierarchical phrase-based model, decoding is realized as an Earley-style top-down parser on the foreign language side with a beam search strategy synchronized with the cardinality of already translated foreign words (Watanabe et al., 2006). The major difference to the phrase-based model's decoder is the handling of non-terminals, or holes, in each rule.

2.3 Feature Functions

Our phrase-based model uses a standard pharaoh feature functions listed as follows (Koehn et al., 2003):

- Relative-count based phrase translation probabilities in both directions.
- Lexically weighted feature functions in both directions.
- The supplied trigram language model.
- Distortion model that counts the number of words skipped.
- The number of words in English-side and the number of phrases that constitute translation.

For details, please refer to Koehn et al. (2003).

In addition, we added three feature functions to restrict reorderings and to represent globalized insertion/deletion of words:

- Lexicalized reordering feature function scores whether a phrase translation pair is monotonically translated or not (Och et al., 2004):

$$h_{lex}(\bar{a}_1^K | \bar{f}_1^K, \bar{e}_1^K) = \log \prod_{k=1}^{K} p_r(\delta_k | \bar{f}_{\bar{a}_k}, \bar{e}_k) \qquad (3)$$

  where $\delta_k = 1$ iff $\bar{a}_k - \bar{a}_{k-1} = 1$ otherwise $\delta_k = 0$.

- Deletion feature function penalizes words that do not constitute a translation according to a

Table 1: Number of word alignment by different preprocessings.

	de-en	es-en	fr-en	en-de	en-es	en-fr
lower	17,660,187	17,221,890	16,176,075	17,596,764	17,237,723	16,220,520
stem	17,110,890	16,601,306	15,635,900	17,052,808	16,597,274	15,658,940
prefix4	16,975,398	16,540,767	15,610,319	16,936,710	16,530,810	15,613,755
intersection	12,203,979	12,677,192	11,645,404	12,218,997	12,688,773	11,653,242
union	23,186,379	21,709,212	20,760,539	23,066,052	21,698,267	20,789,570

Table 2: Number of phrases extracted from differently preprocessed corpora.

	de-en	es-en	fr-en	en-de	en-es	en-fr
lower	37,711,217	61,161,868	56,025,918	38,142,663	60,619,435	55,198,497
stem	46,550,101	75,610,696	68,210,968	46,749,195	75,473,313	67,733,045
prefix4	53,429,522	78,193,818	70,514,377	53,647,033	78,223,236	70,378,947
merged	80,260,191	111,153,303	103,523,206	80,666,414	110,787,982	102,940,840

lexicon model $t(f|e)$ (Bender et al., 2004):

$$h_{del}(e_1^I, f_1^J) = \sum_{j=1}^{J} \left[\max_{0 \leq i \leq I} t(f_j|e_i) < \tau_{del} \right] \quad (4)$$

The deletion model simply counts the number of words whose lexicon model probability is lower than a threshold τ_{del}. Likewise, we also added an insertion model $h_{ins}(e_1^I, f_1^J)$ that penalizes the spuriously inserted English words using a lexicon model $t(e|f)$.

For the hierarchical phrase-based model, we employed the same feature set except for the distortion model and the lexicalized reordering model.

3 Phrase Extraction from Different Word Alignment

We prepared three kinds of corpora differentiated by tokenization methods. First, the simplest preprocessing is lower-casing (lower). Second, corpora were transformed by a Porter's algorithm based multilingual stemmer (stem) [1]. Third, mixed-cased corpora were truncated to the prefix of four letters of each word (prefix4). For each differently tokenized corpus, we computed word alignments by a HMM translation model (Och and Ney, 2003) and by a word alignment refinement heuristic of "grow-diag-final" (Koehn et al., 2003). Different preprocessing yields quite divergent alignment points as illustrated in Table 1. The table also shows the numbers for the intersection and union of three alignment annotations.

The (hierarchical) phrase translation pairs are extracted from three distinctly word aligned corpora. In this process, each word is recovered into its lower-cased form. The associated counts are aggregated to constitute relative count-based feature functions. Table 2 summarizes the size of phrase tables induced from the corpora. The number of rules extracted for the hierarchical phrase-based model was roughly twice as large as those for the phrase-based model. Fewer word alignments resulted in larger phrase translation table size as observed in the "prefix4" corpus. The size is further increased by our aggregation step (merged).

Different induction/refinement algorithms or preprocessings of a corpus bias word alignment. We found that some word alignments were consistent even with different preprocessings, though we could not justify whether such alignments would match against human intuition. If we could trust such consistently aligned words, reliable (hierarchical) phrase translation pairs would be extracted, which, in turn, would result in better estimates for relative count-based feature functions. At the same time, differently biased word alignment annotations suggest alternative phrase translation pairs that is useful for increasing the coverage of translations.

4 Results

Table 3 shows the open test translation results on 2005 and 2006 test set (the development-test set and the final test set) [2]. We used the merged (hierarchical) phrase tables for decoding. Feature function scaling factors were optimized on BLEU score using the supplied development set that is identical to the 2005's development set. We observed that our

[1] We used the Snowball stemmer from `http://snowball.tartarus.org`

[2] We did not differetiated in-domain or out-of-domain for 2006 test set.

Table 3: Open test on the 2005/2006 test sets (BLEU [%]).

		de-en	es-en	fr-en	en-de	en-es	en-fr
test2005	Phrase	25.72	30.97	30.97	18.08	30.48	32.14
	Rule	25.14	30.11	30.31	17.96	27.96	31.04
	2005's best	24.77	30.95	30.27			
test2006	Phrase	23.16	29.90	27.89	15.79	29.54	29.19
	Rule	22.74	28.80	27.28	15.99	26.56	27.86

results are very comparable to the last year's best results in test2005. Also found that our hierarchical phrase-based translation (Rule) performed slightly inferior to the phrase-based translation (Phrase) in both test sets. The hierarchically combined phrases seem to be too flexible to represent the relationship of similar language pairs. Note that our hierarchical phrase-based model performed better in the English-to-German translation task. Those language pair requires rather distorted reordering, which could be represented by hierarchically combined phrases.

We also conducted additional studies on how differently aligned corpora might affect the translation quality on Spanish-to-English task for the 2005 test set. Using our phrase-based model, the BLEU scores for lower/stem/prefix4 were 30.90/30.89/30.76, respectively. The differences of translation qualities were statistically significant at the 95% confidence level. Our phrase translation pairs aggregated from all the differently preprocessed corpora improved the translation quality.

5 Conclusion

We presented two translation models, a phrase-based model and a hierarchical phrase-based model. The former performed as well as the last year's best system, whereas the latter performed comparable to our phrase-based model. We are going to experiment new feature functions to restrict the too flexible reordering represented by our hierarchical phrase-based model.

We also investigated different word alignment annotations, first using lower-cased corpus, second performing stemming, and third retaining only 4-letter prefix. Differently preprocessed corpora resulted in quite divergent word alignment. Large phrase/rule translation tables were accumulated from three distinctly aligned corpora, which in turn, increased the translation quality.

References

Alfred V. Aho and Jeffrey D. Ullman. 1969. Syntax directed translations and the pushdown assembler. *J. Comput. Syst. Sci.*, 3(1):37–56.

Oliver Bender, Richard Zens, Evgeny Matusov, and Hermann Ney. 2004. Alignment templates: the RWTH SMT system". In *Proc. of IWSLT 2004*, pages 79–84, Kyoto, Japan.

David Chiang. 2005. A hierarchical phrase-based model for statistical machine translation. In *Proc. of ACL 2005*, pages 263–270, Ann Arbor, Michigan, June.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proc. of NAACL 2003*, pages 48–54, Edmonton, Canada.

Franz Josef Och and Hermann Ney. 2002. Discriminative training and maximum entropy models for statistical machine translation. In *Proc. of ACL 2002*, pages 295–302.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51, March.

Franz Josef Och and Hermann Ney. 2004. The alignment template approach to statistical machine translation. *Comput. Linguist.*, 30(4):417–449.

Franz Josef Och, Daniel Gildea, Sanjeev Khudanpur, Anoop Sarkar, Kenji Yamada, Shankar Fraser, Alex a nd Kumar, Libin Shen, David Smith, Katherine Eng, Viren Jain, Zhen Jin, and Dragomir Radev. 2004. A smorgasbord of features for statistical machine translation. In *HLT-NAACL 2004: Main Proceedings*, pages 161–168, Boston, Massachusetts, USA, May 2 - May 7.

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *Proc. of ACL 2003*, pages 160–167.

Taro Watanabe, Hajime Tsukada, and Hideki Isozaki. 2006. Left-to-right target generation for hierarchical phrase-based translation. In *Proc. of COLING-ACL 2006 (to appear)*, Sydney, Australia, July.