NTT SMT System 2008 at NTCIR-7

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

Abstract

This paper describes NTT SMT System 2008 presented at the patent translation task (PAT-MT) in NTCIR-7. For PAT-MT, we submitted our strong baseline system faithfully following a hierarchical phrase-based statistical machine translation [2]. The hierarchical phrase-based SMT is based on a synchronous-CFGs in which a paired source/target rules are synchronously applied starting from the initial symbol. The decoding is realized by a CYK-style bottom-up parsing on the source side with each derivation representing a translation candidate. We demonstrate the strong baseline for the PAT-MT English/Japanese translations.

Keywords: *Statistical Machine Translation, Hierarchical Phrase-based SMT.*

1 Introduction

We present NTT Statistical Machine Translation System 2008 for the patent translation task (PAT-MT) in NTCIR-7. Our system has been successfully demonstrated for the numbers of evaluation tasks, including NIST[1], WMT[13] and IWSLT [10]. For PAT-MT Japanese/English translations, we employed a strong baseline system faithfully following a hierarchical phrase-based statistical machine translation [2].

Hierarchical phrase-based machine translation is formulated as a synchronous-CFG in which paired strings are simultaneously rewritten using a set of paired right-hand side rules. Decoding is realized as parsing on the source side with each target yield of a derivation representing translation. Specifically, we employed a variant of a CKY-based algorithm [2] with cube-pruning for efficient search [3]. The evaluation results indicate that our baseline implementation is very competitive to other SMT systems.

We introduce hierarchical phrase-based SMT in Section 2 followed by evaluations discussed in Section 3.

[1] http://www.nist.gov/speech/tests/mt/2008/doc/mt08_official_results_v0.html

2 Hierarchical Phrase-based SMT

Hierarchical phrase-based SMT is formulated as a probabilistic synchronous context-free grammar (PSCFG) [1] in which string pairs are generated. The system uses a set of source terminal symbols $\mathcal{T}_S$, a set of target terminal symbols $\mathcal{T}_T$ and a set of non-terminal symbols $\mathcal{N}$. Each production rule is realized as follows [2, 11]:

$$X \rightarrow \langle \gamma, \alpha, \sim, w \rangle \quad (1)$$

where $X \in \mathcal{N}$, $\gamma \in [\mathcal{N} \cup \mathcal{T}_S]^*$ and $\alpha \in [\mathcal{N} \cup \mathcal{T}_T]^*$. γ and α share the same number of non-terminals with each non-terminal mapped by $\sim$. $w \in \mathbb{R}$ is a real-valued weight associated with each rule. Starting from an initial non-terminal symbol, each non-terminal is recursively rewritten by the production rule's right hand side γ and α associated with $\sim$.

Based on the synchronous-CFG formalism, we adopted the hierarchical phrase-based modeling by introducing some constraints to each production rule [2].

1. A single non-terminal category X is used.
2. Each rule contains at most two non-terminals.

The set of production rules or grammar is automatically learned from word alignment annotated corpora. Specifically, given a bilingual data, we run GIZA++ [7] in two directions. Second, the word alignments are heuristically combined [8]. Finally, phrases are extracted that do not violate word alignment constraints [4]. At the same time, if there exists a phrase with potential embedded phrases, we treat the sub phrases as a non-terminal X [2]. In order to eliminate the spuriously extracted grammar, we further restrict the form of production rules as follows:

3. Each rule contains at most five terminals in each of the source and target sides.
4. No adjacent non-terminals exist in the source side.

In addition to the automatically acquired rules, monotonic rules are added to reduce the data sparseness problem:

$$X \rightarrow \langle X_{\boxed{1}} X_{\boxed{2}}, X_{\boxed{1}} X_{\boxed{2}} \rangle$$

where boxed indices denote one-to-one mapping of non-terminals between source and target sides.

Translation under PSCFG is regarded as the decoding problem which is cast as a parsing problem using the source side rules. Given a source sentence f, we perform CKY-based parsing using the source yield of the productions rules. The best translation is generated from the target yield $e(D)$ of the best derivation $\hat{D}$ according to the weight $w(D)$ [2].

$$\hat{e} = \underset{\{e: f(D)=f, e(D)=e\}}{\operatorname{argmax}} w(D) \quad (2)$$

The weight of a derivation $w(D)$ is a λ_i scaled linear combination of several (or many) feature functions ϕ_i decomposed by rules r in D:

$$w(D) = \sum_i \sum_{r \in D} \lambda_i \phi_i(r) \quad (3)$$

We employed a standard set of features, namely, relative count-based probabilities and lexical probabilities in two directions, various length penalties, and n-gram language models [2]. For an efficient intersection with n-gram language models, we introduce cube-pruning [3].

3 Evaluation

We exploited two set of data for each direction. For the official baseline system, we used only a set of aligned sentence pairs, namely PSD-1. For the contrastive runs, we employed additional data: PSD-2 for additional production rules and PPD-1,2 for larger n-gram language models. We have also included English Web-1T 5-grams and Japanese Web-1T 7-grams [2].

All the corpora were case-preserved but normalized according to NFKC, an unicode standard for encoding normalization. The Japanese corpus was tokenized by mecab [3]. The English corpus was tokenized by an in-house developed tool following the tokenization standard described in English Web 1T data.

We found that the formal run test data and the tuning/training data come from different epoch with totally different notations for non-ascii letters in English, such as symbols used in equations. Therefore, we convert all the old-style symbol notations into new styles by reverse engineering the publicly available tools [4].

Word alignment is annotated via an in-house spun tool which supports a variant of HMM alignment model [12] with various token factoring [13]. From the heuristically combined factored word alignment, hierarchical rules are extracted with each rule containing at most 5 terminals. The feature scaling factors are MERT tuned [6] using a combination of all the development data consisting of nearly 2,000 sentences with sentence length at most 40 words. The translation results in BLEU [9] are summarized in Table 1[5]. The single-reference BLEU with 1,381 sentences (sBLEU) indicated that our system using only a small subset of data (an official run) resulted in better BLEU. However, the multiple-reference BLEU with 300 sentences (mBLEU) gain a small increase by employing all the data.

Table 1. Evaluation results by single-reference BLEU (sBLEU) and multiple-reference BLEU (mBLEU).

	sBLEU	mBLEU
Japanese/English	27.20	35.93
+ Web 1T/PPD	26.88	36.05
English/Japanese	28.07	
+ Web 1T/PPD	27.20	

4 Conclusion

We presented our strong baseline system faithfully following hierarchical phrase-based machine translation [2]. The official results indicate that the performance is very competetive to the top ranked systems in terms of BLEU.

5 Acknowledgments

This work is partly supported by MEXT Grant-In-Aid for Scientific Research of Priority Areas: Cyber Infrastructure for the Information-Explosion Era.

References

[1] A. V. Aho and J. D. Ullman. Syntax directed translations and the pushdown assembler. *Journal of Computer and System Sciences*, 3(1):37–56, 1969.

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

[3] L. Huang and D. Chiang. Forest rescoring: Faster decoding with integrated language models. In *Proc. of ACL*, pages 144–151, Prague, Czech Republic, June 2007.

[4] P. Koehn, F. J. Och, and D. Marcu. Statistical phrase-based translation. In *Proc. of NAACL 2003*, pages 48–54, Edmonton, Canada, 2003.

[2] LDC2006T13 from LDC, and GSK2007-C [5] from GSK http://www.gsk.or.jp/, respectively.

[3] http://mecab.sourceforge.net/

[4] http://www.uspto.gov/web/offices/ac/ido/oeip/sgml/st32/redbook/grbv25x.html

[5] Table 1 shows the single reference BLEU-S with 1,381 sentences and the double reference BLEU-m300-DE with 300 sentences.

[5] T. Kudo and H. Kazawa. Web Japanese N-gram version 1. published by Gengo Shigen Kyokai.

[6] F. J. Och. Minimum error rate training in statistical machine translation. In *Proc. of ACL 2003*, pages 160–167, Sapporo, Japan, July 2003.

[7] F. J. Och and H. Ney. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51, 2003.

[8] F. J. Och and H. Ney. The alignment template approach to statistical machine translation. *Computational Linguistics*, 30(4):417–449, 2004.

[9] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. BLEU: a method for automatic evaluation of machine translation. In *Proc. of ACL 2002*, pages 311–318, Philadelphia, Pennsylvania, 2002.

[10] K. Sudoh, T. Watanabe, J. Suzuki, H. Tsukada, and H. Isozaki. NTT Statistical Machine Translation System for IWSLT 2008. In *Proc. of the International Workshop on Spoken Language Translation (to appear)*, 2008.

[11] A. Venugopal, A. Zollmann, and S. Vogel. An efficient two-pass approach to synchronous-CFG driven statistical MT. In *Proc. of NAACL*, pages 500–507, Rochester, New York, April 2007.

[12] S. Vogel, H. Ney, and C. Tillmann. Hmm-based word alignment in statistical translation. In *Proceedings of the 16th conference on Computational linguistics*, pages 836–841, Morristown, NJ, USA, 1996. Association for Computational Linguistics.

[13] T. Watanabe, H. Tsukada, and H. Isozaki. NTT system description for the wmt2006 shared task. In *Proc. of WMT*, pages 122–125, New York City, June 2006.