Clause-Based Reordering Constraints to Improve Statistical Machine Translation

Ananthakrishnan Ramanathan[1], Pushpak Bhattacharyya[2], Karthik Visweswariah[1],
Kushal Ladha[2] and Ankur Gandhe[1]

[1]IBM Research India
{aramana2,v-karthik,ankugand}@in.ibm.com

[2]Department of CSE, IIT Bombay
{pb,kush}@cse.iitb.ac.in

Abstract

We demonstrate that statistical machine translation (SMT) can be improved substantially by imposing clause-based reordering constraints during decoding. Our analysis of clause-wise translation of different types of clauses shows that it is beneficial to apply these constraints for finite clauses, but not for non-finite clauses. In our experiments in English-Hindi translation with an SMT system (DTM2), on a test corpus containing around 850 sentences with manually annotated clause boundaries, BLEU improves to 20.4 from the baseline score of 19.4. This statistically significant improvement is also confirmed by subjective (human) evaluation. We also report preliminary work on automatically identifying the kind of clause boundaries appropriate for enforcing reordering constraints.

1 Introduction

It has been recognized widely that reordering is the Achilles' heel of most SMT models (Birch et al., 2009; Hoang and Koehn, 2009), especially when applied to languages with large syntactic differences. In our experiments in English-Hindi SMT, we observe that it is quite frequent in multi-clause sentences for phrases to move out of their respective clauses due to incorrect reordering. While such mistakes can be avoided by restricting reordering over clause boundaries, this paper demonstrates that such a strategy works well only for finite clauses, and not for non-finite clauses.

Clause-wise or part-by-part translation has been a standard approach in traditional transfer-based systems. In Systran, as described by Hutchins and Somers (1992), conjunct and relative clauses were segmented in a preprocessing step. Similar methods were used in the the Stanford Machine Translation project (Wilks, 1973). Chandrasekar (1994) applies a sentence simplification method to machine translation, where sentences are split at conjunctions, relative pronouns, etc., before translation. Rao *et al.* (2000) describe a clause-wise translation strategy within an English-Hindi transfer-based MT system.

In the context of SMT, Koehn and Knight (2003) use a dedicated noun phrase (NP) translation subsystem to obtain significant improvements in German-English translation. Other similar work includes (Watanabe et al., 2003) for Japanese-English SMT and (Hewavitharana et al., 2007) for Arabic-English. Sudoh et al. (2010) propose methods to perform clause-level alignment of the parallel corpus, and to translate clauses (all clauses identified by a syntactic parser) as a unit to improve long-distance reordering in a specialized domain – English-Japanese translation of research paper abstracts in the medical domain.

While our approach draws from many of the above, it is novel in the following ways: (i) We provide an analysis of different clause types in translation – we show that only some kinds of clauses benefit from the use of reordering constraints, and (ii) We demonstrate significant improvements using this strategy for English-Hindi SMT in a general domain.

2 Problem Statement

We analyzed a set of 225 sentences translated through the baseline system. Of the 120 sentences in this set which had more than one clause, 45 sentences were found to have inter-clause reordering problems; that is, some words or phrases are wrongly placed in a clause where they do not belong. Such long-range reordering problems obviously have a serious detrimental effect on transla-

tion quality, and it is quite likely that even limiting the reordering problems to the respective clauses will aid comprehensibility.

3 Analysis of Clause Types in Translation

We will now look at how finite and non-finite clauses behave in translation. Finite clauses are basically tensed clauses, while non-finite clauses are untensed.

- **Finite clauses**: Finite clauses appear most commonly in conjunct or relative constructions. In such cases, each finite clause can be translated separately and glued together using the correct translation of the conjunction or the relative pronoun. While this strategy works very well for conjunct clauses, in relative clauses the resulting translation is sometimes not completely natural, but it is nevertheless always clearly understandable. Consider the following example:

  Input: The boy, who stays in Delhi, won the match.

  Using a clause-wise translation approach, this would be translated as,

  लडका, जो दिल्ली में रेहता है, मैच जीत गया।

  ladakaa, jo dillii men rehtaa hai, match jiita gayaa

  Gloss:boy, who Delhi in stays, match won

  However, a more natural translation would be using correlatives:

  जो लडका दिल्ली में रेहता है, वो मैच जीत गया।

  jo ladakaa dillii men rehtaa hai, vo match jiita gayaa

  Gloss:which boy Delhi in stays, he match won

  Certain kinds of nominal clauses also result in similar disfluencies. For example, in the sentence "The playground is where the statue is situated", putting a reordering constraint around the clause "where the statue is situated" results in an unnatural translation.

- **Non-finite clauses**: Compared to finite clauses, the translation of a non-finite clause is much more dependent on its role within the

sentence. There are two main issues: (i) all or part of the non-finite clause could get reordered with the surrounding clause, or (ii) the overall meaning is conveyed by a phrase or group of words from the non-finite clause and a surrounding or neighbouring clause.

Taking *to-infinitives* as the example category, consider the following constructions:

a. *To-infinitive clause in raising constructions*: Here the embedded clause with the raised element is translated as a finite clause with *that* as a complementizer. For example,

Input: John is certain to win,

which is translated as "It is certain that John will win":

यह तय है कि जॉन जीतेगा।

yaha tay hai ki John jiitega

Gloss:it certain is that John will-win

It is the combination of "certain" (or similar word) and the to-infinitive that results in this translation.

b. *To-infinitive with a copular verb*: In such cases, the infinitive is inflected with a कर (*kara*) ending to indicate sequentiality:

Input: John was happy to see him

जॉन उसे देखकर खुश हुआ।

John use dekhakara khusha huaa

Gloss:John him-to seen-having happy was/became

Here, the non-finite clause "to see him", gets reordered into the main clause, something that will not be possible if reordering constraints are applied.

Similarly, there are differences in the way other non-finite clauses (with -ed and -ing participles) are treated. The point being made here is that handling these differences is crucial to the correct meaning being conveyed. In other words, simply translating non-finite clauses separately with reordering constraints around them, will not lead to good translation, because the translation of these clauses is often dependent on the superordinate clause, and also there is reordering between these clauses and the superordinate clause.

	BLEU	Adequacy	Fluency
baseline	19.4	2.04	2.41
finite	20.4^{δ}	2.32^{δ}	2.67^{δ}
non-finite	19.6	2.17^{ψ}	2.5
finite + non-finite	19.8^{ψ}	2.17	2.51^{ψ}

Table 1: Manually identified clauses. δ: 99% statistical significance; ψ: 95% statistical significance

Method	ACI accuracy	BLEU	Adequacy	Fluency
parser	0.42	19.3	-	-
CRF – word and pos	0.69	19.8^{ψ}	2.27^{δ}	2.59^{δ}

Table 2: Automatically identified finite clauses

Based on the above analysis, we are encouraged to test the hypothesis that it is helpful to put reordering constraints around finite clauses, but not around non-finite clauses.

Finiteness is one broad dimension along which clauses may be categorized. Subordinate clauses can be further classified based on the position in which they occur in the sentence – that is, whether they occur in the complement position, the specifier position, or the adjunct position. We may also classify clauses based on whether they play an adjectival, nominal, or adverbial role in the sentence. The focus of the present work, however, is only on the finiteness aspect.

4 Experiments

In this section, we first describe briefly our baseline system and our approach to handling clauses within this system. We then summarize the datasets used and our evaluation methodology, before describing the results of various experiments.

4.1 Approach

The baseline system we use is *DTM2* (a direct translation model) (Ittycheriah and Roukos, 2007). The word-alignments were done using an HMM aligner. We used the best-performing parameter setting in the decoder [1].

A beam search decoder (Ittycheriah and Roukos, 2007) similar to other phrase-based decoders (Tillmann and Ney, 2003) is used for translation. The reordering restriction is applied by treating the relevant clause-boundaries as *barriers*, and putting a hard constraint on reordering across

barriers – that is, during decoding, if a new hypothesis requires reordering over a barrier, the hypothesis is discarded.

We experiment with clause boundaries identified in three ways: (i) manually (section 4.3.1), (ii) automatically using a constituency parser (section 4.3.2), and (iii) automatically using a CRF-based clause-boundary classifer using part-of-speech and parser features (section 4.3.3).

4.2 Data and Evaluation

The system was trained on a parallel corpus with 289k sentences (combining the LDC English-Hindi parallel texts with internal datasets) consisting of various domains including news, and tested on 844 sentences. The language model was trained on around 1.5 million sentences.

Automatic evaluation was done using BLEU (Papineni et al., 2001) with a single reference translation per sentence. Statistical significance of the test results with BLEU was computed using paired bootstrap resampling (Koehn, 2004). Subjective evaluation was performed on hundred randomly selected multi-clause sentences, using a five-point scale.

4.3 Results

4.3.1 Manual annotation: finite vs. non-finite clauses

For this experiment, the sentences in the test sets were manually annotated with finite and non-finite clause boundaries. The results in table 1 indicate that reordering constraints around finite clauses work much better than around non-finite clauses.

---

[1]Specifically, the skip-length (distortion limit) was set to 8. Lower skip-lengths led to much poorer performance. For example, with a skip-length of 4, the BLEU score dropped by around 2 points compared to a skip-length of 8

	improved	degraded
finite (manual)	36	8
finite (auto)	35	17
non-finite (manual)	17	10
finite + non-finite (manual)	19	11

Table 3: Number of translations improved/degraded

4.3.2 Automatic clause identification using a parser

The goal of this experiment was to determine whether clauses obtained from a parser could be used for the purpose of imposing clause-based constraints. We performed the experiment by using a state-of-the-art maxent parser (Ratnaparkhi, 1999) to mark all clause boundaries (clause-level nodes based on the Penn Treebank II Style bracketing guidelines – S, S-BAR, SQ, SBARQ, and SINV). The results were negative (table 2 – row titled *parser*), indicating that straightforward use of a parser is not sufficient to help in identifying clauses suitable for reordering constraints. The column titled *ACI accuracy* has the F-measure for automatic clause-boundary identification measured over the entire test corpus.

4.3.3 Automatic clause identification using a CRF classifier

We annotated a set of 1500 English sentences with finite-clause boundaries, and used this to train a CRF-based clause-boundary classifier (Ram and Devi, 2008; Tjong et al., 2001). Unigram and bigram word features, unigram, bigram and trigram POS features, and the POS of the following verb group were used in the model (Kashioka et al., 2003) [2]. We see (table 2) that a reasonable gain is obtained using the classifier, though not nearly as much as with manual annotation.

5 Discussion

The subjective evaluation scores reveal that only a few translations degrade in quality when reordering constraints are used with finite clauses (largely due to fluency issues of the kind described in section 3). Table 3 shows the number of translations that improved (i.e., the average adequacy and fluency score increased) due to clause-based translation, and the number which degraded, among the hundred sentences taken up for subjective evalua-

tion. The following is an illustration of the kind of improvements clause-based translation brings:

Input: America claims that Iran wants to continue its nuclear programme, and secretly builds atomic weapons.

Baseline translation: अमरीका का दावा है कि उसके परमाणु कार्यक्रम रहना चाहते हैं और् ईरान परमाणु हथियार निर्माण करता है ।

amerika kaa daavaa hai ki usake paramaanu kaaryakrama rahanaa caahate hain aur iraana paramaanu hathiyaara nirmaana karataa hai

Gloss: America's claim is that their nuclear programme continue want and Iran nuclear weapons builds

Clause-based translation: अमरीका का दावा है कि ईरान अपने परमाणु कार्यक्रम को जारी रखना चाहता है और परमाणु हथियार निर्माण करता है ।

amerika kaa daavaa hai ki iran apane paramaanu kaaryakrama ko jaarii rakhanaa caahataa hai aura paramaanu hathiyaara nirmaana kartaa hai

Gloss: America's claim is that Iran its nuclear programme to continue wants and nuclear weapons builds

6 Conclusion and Future Work

We have shown that the quality of English-Hindi statistical machine translation can be improved by performing clause-wise translation. An important finding in our work is that, in general, finite clauses are more directly suited for such translation compared to non-finite clauses.

Possible directions of future work include: (i) using clause-based constraints also in training, (ii) learning from word-alignments, the kinds of clauses that can benefit from reordering constraints, (iii) analyzing the impact of finer-grained clause types (as mentioned in section 3) on translation, (iv) comparing with other SMT models (Galley and Manning, 2010; Chiang, 2005), and (v) using soft constraints instead of hard constraints for reordering (Marton and Resnik, 2008).

---

[2] Adding clause-boundary information from the parser as features in the classifier resulted in a lower F-measure of 0.65

References

Adwait Ratnaparkhi, Learning to parse natural language with maximum entropy models, *Machine Learning*, 34(1-3), 1999.

Birch, A., Blunsom, P., and Osborne, M., A quantitative analysis of reordering phenomena, *Proceedings of the Fourth Workshop on Statistical Machine Translation*, 2009.

Chandrasekar, R., *A Hybrid Approach to Machine Translation using Man Machine Communication*, Ph.D. thesis, Tata Institute of Fundamental Research, Mumbai, 1994.

Chiang, D., A hierarchical phrase-based model for statistical machine translation, *Proceedings of ACL*, 2005.

Galley, M. and Manning, C., Accurate Non-Hierarchical Phrase-Based Translation, *Proceedings of HLT-NAACL*, 2010.

Hewavitharana, S., Lavie, A., and Vogel, S., Experiments with a noun-phrase driven statistical machine translation system, *Proceedings of the MT Summit XI*, 2007.

Hoang, H. and Koehn, P., Improving mid-range re-ordering using templates of factors, *Proceedings of EACL*, 2009.

Hutchins, J., and Somers, H., *An Introduction to Machine Translation*, pages 175–189, Academic Press, 1992.

Ittycheriah, A., and Roukos, S., Direct translation model 2, *Proceedings of NAACL-HLT*, 2007.

Kashioka, H., Maruyama, T., and Tanaka, H., Building a parallel corpus for monologue with clause alignment, *Proceedings of the MT Summit IX*, 2003.

Koehn, P., and Knight, K., Feature-rich statistical translation of noun phrases, *Proceedings of ACL*, 2003.

Koehn, P., Statistical significance tests for machine translation evaluation, *Proceedings of EMNLP*, 2004.

Marton, Y. and Resnik, P., Soft syntactic constraints for hierarchical phrase-based translation, *Proceedings of ACL-HLT*, 2008.

Papineni, K., Roukos, S., Ward, T., and Zhu, W., BLEU: a method for automatic evaluation of machine translation, *IBM Research Report*, Thomas J. Watson Research Center, 2001.

Ram, V., S. and Devi, S., L., Clause boundary identification using conditional random fields, *Proceedings of CICLing*, 2008.

Rao, D., Mohanraj, K., Hegde, J., Mehta, V., and Mahadane, P., A practical framework for syntactic transfer of compound-complex sentences for English-Hindi machine translation, *Proceedings of KBCS*, 2000.

Sudoh, K., Duh, K., Tsukada, H., Hirao, T., and Nagata, M., Divide and translate: improving long distance reordering in statistical machine translation, *Workshop on Statistical Machine Translation and Metrics*, 2010.

Tillmann, C. and Ney, H., Word reordering and a dynamic programming beam search algorithm for statistical machine translation *Computational Linguistics*, 29(1), 2003.

Tjong, E., F., Sang, K., and, Dejean, H., Introduction to the CoNLL-2001 shared task: clause identification *Proceedings of CoNLL*, 2001.

Watanabe, T., Sumita, E., and Okuno, H., Chunk-based statistical translation, *Proceedings of ACL*, 2003.

Wilks, Y., The Stanford Machine Translation project, *Natural Language Processing*, Algorithmics Press, pages 243-290, 1973.