Pattern Visualization for Machine Translation Output

Adam Lopez
Institute for Advanced Computer Studies
Department of Computer Science
University of Maryland
College Park, MD 20742
alopez@cs.umd.edu

Philip Resnik
Institute for Advanced Computer Studies
Department of Linguistics
University of Maryland
College Park, MD 20742
resnik@umd.edu

Abstract

We describe a method for identifying systematic patterns in translation data using part-of-speech tag sequences. We incorporate this analysis into a diagnostic tool intended for developers of machine translation systems, and demonstrate how our application can be used by developers to explore patterns in machine translation output.

1 Introduction

Over the last few years, several automatic metrics for machine translation (MT) evaluation have been introduced, largely to reduce the human cost of iterative system evaluation during the development cycle (Papineni et al., 2002; Melamed et al., 2003). All are predicated on the concept of *n*-gram matching between the sentence hypothesized by the translation system and one or more *reference translations*—that is, human translations for the test sentence. Although the formulae underlying these metrics vary, each produces a single number representing the "goodness" of the MT system output over a set of reference documents. We can compare the numbers of competing systems to get a coarse estimate of their relative performance. However, this comparison is holistic. It provides no insight into the specific competencies or weaknesses of either system.

Ideally, we would like to use automatic methods to provide immediate diagnostic information about the translation output—*what* the system does well, and what it does poorly. At the most general level, we want to know how our system performs on the two most basic problems in translation – word translation and reordering. Holistic metrics are at odds with day-to-day hypothesis testing on these two problems. For instance, during the development of a new MT system we may may wish to compare competing reordering models. We can incorporate each model into the system in turn, and rank the results on a test corpus using BLEU (Papineni et al., 2002). We might then conclude that the model used in the highest-scoring system is best. However, this is merely an implicit test of the hypothesis; it does not tell us anything about the specific strengths and weaknesses of each method, which may be different from our expectations. Furthermore, if we understand the relative strengths of each method, we may be able to devise good ways to combine them, rather than simply using the best one, or combining strictly by trial and error. In order to fine-tune MT systems, we need fine-grained error analysis.

What we would really like to know is how well the system is able to capture systematic reordering patterns in the input, which ones it is successful with, and which ones it has difficulty with. Word *n*-grams are little help here: they are too many, too sparse, and it is difficult to discern general patterns from them.

2 Part-of-Speech Sequence Recall

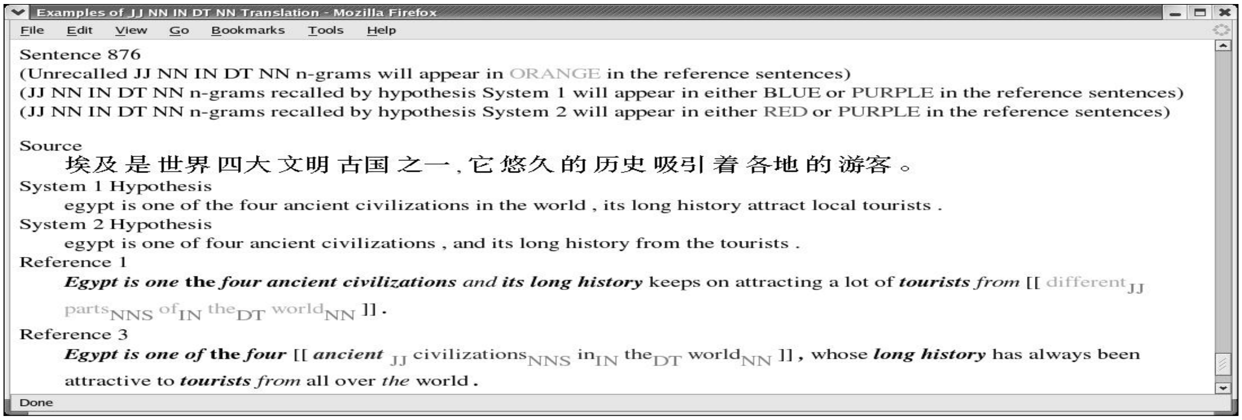
In developing a new analysis method, we are motivated in part by recent studies suggesting that word reorderings follow general patterns with respect to syntax, although there remains a high degree of flexibility (Fox, 2002; Hwa et al., 2002). This suggests that in a comparative analysis of two MT systems (or two versions of the same system), it may be useful to look for syntactic patterns that one system (or version) captures well in the target language and the other does not, using a syntax-based, recall-oriented metric.

As an initial step, we would like to summarize reordering patterns using part-of-speech sequences. Unfortunately, recent work has confirmed the intuition that applying statistical analyzers trained on well-formed text to the noisy output of MT systems produces unuseable results (e.g. (Och et al., 2004)). Therefore, we make the conservative choice to apply annotation only to the reference corpus. Word *n*-gram correspondences with a reference translation are used to infer the part-of-speech tags for words in the system output.

The method:

1. Part-of-speech tag the reference corpus. We used

Proceedings of HLT/EMNLP 2005 Demonstration Abstracts, pages 12–13,
Vancouver, October 2005.



Sentence 876
(Unrecalled JJ NN IN DT NN n-grams will appear in ORANGE in the reference sentences)
(JJ NN IN DT NN n-grams recalled by hypothesis System 1 will appear in either BLUE or PURPLE in the reference sentences)
(JJ NN IN DT NN n-grams recalled by hypothesis System 2 will appear in either RED or PURPLE in the reference sentences)

Source
埃及 是 世界 四大 文明 古国 之一 , 它 悠久 的 历史 吸引 着 各地 的 游客 。
System 1 Hypothesis
egypt is one of the four ancient civilizations in the world , its long history attract local tourists .
System 2 Hypothesis
egypt is one of four ancient civilizations , and its long history from the tourists .
Reference 1
Egypt is one **the** ***four ancient civilizations*** *and* ***its long history*** keeps on attracting a lot of ***tourists*** *from* [[different$_{JJ}$ parts$_{NNS}$ of$_{IN}$ the$_{DT}$ world$_{NN}$]] .
Reference 3
Egypt is one of **the** ***four*** [[***ancient*** $_{JJ}$ civilizations$_{NNS}$ in$_{IN}$ the$_{DT}$ world$_{NN}$]] , whose ***long history*** has always been attractive to ***tourists*** *from* all over *the* world .

Figure 1: Comparing two systems that differ significantly in their recall for POS n-gram JJ NN IN DT NN. The interface uses color to make examples easy to find.

MXPOST (Ratnaparkhi, 1996), and in order to discover more general patterns, we map the tag set down after tagging, e.g. NN, NNP, NNPS and NNS all map to NN.

2. Compute the frequency $\mathit{freq}(t_i \ldots t_j)$ of every possible tag sequence $t_i \ldots t_j$ in the reference corpus.
3. Compute the correspondence between each hypothesis sentence and *each* of its corresponding reference sentences using an approximation to maximum matching (Melamed et al., 2003). This algorithm provides a list of *runs* or contiguous sequences of words $e_i \ldots e_j$ in the reference that are also present in the hypothesis. (Note that runs are order-sensitive.)
4. For each recalled *n*-gram $e_i \ldots e_j$, look up the associated tag sequence $t_i \ldots t_j$ and increment a counter $\mathit{recalled}(t_i \ldots t_j)$

Using this method, we compute the recall of tag patterns, $R(t_i \ldots t_j) = \mathit{recalled}(t_i \ldots t_j)/\mathit{freq}(t_i \ldots t_j)$, for all patterns in the corpus.

To compare two systems (which could include two versions of the same system), we identify POS n-grams that are recalled significantly more frequently by one system than the other, using a difference-of-proportions test to assess statistical significance. We have used this method to analyze the output of two different statistical machine translation models (Chiang et al., 2005).

3 Visualization

Our demonstration system uses an HTML interface to summarize the observed pattern recall. Based on frequent or significantly-different recall, the user can select and visually inspect color-coded examples of each pattern of interest in context with both source and reference sentences. An example visualization is shown in Figure 1.

4 Acknowledgements

The authors would like to thank David Chiang, Christof Monz, and Michael Subotin for helpful commentary on this work. This research was supported in part by ONR MURI Contract FCPO.810548265 and Department of Defense contract RD-02-5700.

References

David Chiang, Adam Lopez, Nitin Madnani, Christof Monz, Philip Resnik, and Michael Subotin. 2005. The hiero machine translation system: Extensions, evaluation, and analysis. In *Proceedings of HLT/EMNLP 2005*, Oct.

Heidi J. Fox. 2002. Phrasal cohesion and statistical machine translation. In *Proceedings of the 2002 Conference on EMNLP*, pages 304–311, Jul.

Rebecca Hwa, Philip Resnik, Amy Weinberg, and Okan Kolak. 2002. Evaluating translational correspondence using annotation projection. In *Proceedings of the 40th Annual Meeting of the ACL*, pages 392–399, Jul.

I. Dan Melamed, Ryan Green, and Joseph P. Turian. 2003. Precision and recall of machine translation. In *HLT-NAACL 2003 Companion Volume*, pages 61–63, May.

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

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the ACL*, pages 311–318, Jul.

Adwait Ratnaparkhi. 1996. A maximum entropy model for part-of-speech tagging. In *Proceedings of the Conference on EMNLP*, pages 133–142, May.