Word Alignment Revisited

Francisco Guzman
Centro de Sistemas Inteligentes
ITESM
Monterrey,Mexico
guzmanhe@gmail.com

Jan Niehues
Institut für Theoretische Informatik
Universität Karlsruhe (TH)
Karlsruhe, Germany
jniehues@ira.uka.de

Qin Gao
Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA, 15213, USA
qing@cs.cmu.edu

Stephan Vogel
Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA, 15213, USA
stephan.vogel@cs.cmu.edu

1 Introduction

Word alignments have been considered the backbones of Statistical Machine Translation. Even when Statistical Machine Translation has shifted from a word-based to a phrase-based paradigm, the word alignment has remained the base for most phrase-based [Koehn et al., 2003] and syntactic augmented phrase based SMT systems [Zollmann and Venugopal, 2006, Chiang, 2007, Marcu et al., 2006].

Most SMT systems use the freely available GIZA++ Toolkit [Al-Onaizan et al., 1999] to generate the word alignment. This toolkit implements the IBM Models and the HMM model introduced in [Brown et al., 1990, Vogel et al., 1996].

Generative models have the advantage that they are well suited for a noisy-channel approach. Unsupervised training can be used to align large amount of unlabeled parallel corpora. Nonetheless they have a major disadvantage: because these models are completely unsupervised, they can hardly make use of the increasingly available manual alignments. Also, given their complexity, to incorporate other sources of informations such as POS tags, word frequencies etc., is a non-trivial task. Moreover, because the IBM models are not symmetric, the alignments for different directions are quite different, which makes the search for a symmetrized combination of the word alignments a challenging procedure.

Given the fact that the word alignments serve as a starting point of the SMT pipeline, improving their quality has been a major focus of research in the SMT community. However, due to the amount of processing that a word alignment undergoes before being used in translation (for example, phrase extraction), the quality of word alignment is not directly related to the quality of translation. In fact, only weak correlation between alignment error rate (AER) and BLEU scores has been reported [Fraser and Marcu, 2006a]. The mismatch between the quality of word alignment models and that of phrase-based or syntactic based SMT may lead to the phenomenon of improved translation quality resulting from "degraded" alignment quality [Vilar et al., 2006]. This calls for more careful analysis of word alignment errors. There has been little effort doing a thorough error analysis of the alignment process. As a result, the role of the quality of word alignments in machine translation remains rather unclear.

Recently, different efforts have focused on the symmetrization of the word alignment models [Matusov et al., 2004, Liang et al., 2006], the inclusion of annotated data in the training of generative models [Fraser and Marcu, 2006b], and the use of discriminative models [Blunsom and Cohn, 2006, Taskar et al., 2005, Niehues and Vogel, 2008]. One of the advantages of the latter models is that the word alignment quality can be tuned towards a given word alignment quality measurement[1]. Moreover, their conditional probability model allows the inclusion of different features, enabling that any available knowledge source can be used to find the best alignment.

In this work, we present the results of an extensive error analysis of the alignments created by the generative models using GIZA++. By characterizing the errors, we hope to shed light on the behavior of the aligners, as well as to identify some opportunities for improvement. We also present our work on a discriminative word alignment framework, as presented in [Niehues and Vogel, 2008], which is easy to enhance with new features. We believe that with a proper analysis of the alignment behaviors, coupled with the use of discriminative word aligner, can help to overcome many of the weaknesses of the generative models.

The paper is organized as follows. In Section 2, the analysis of the alignment errors is presented. In Section 3 the discriminative aligner is introduced, along with proposed new features. The alignment experiments and analysis are presented in Section 4.

[1]For some measurements, smoothing is required.