Morphology and Word Segmentation and their integrations within MT

Thuy Linh Nguyen (thuylinh@andrew.cmu.edu)

January 22, 2008

1 Introduction

Statistical machine translation translates text from one language into another language using available parallel corpora. Machine translation models extract phrase translation based on word alignment output. However, languages are different, in many language pairs, it is impossible to find an equivalent translation of a word of the source language into a word in the target language. In the coming weeks, we will review existing research on source language processing and their integrations in machine translation.

In the current survey, we would like to classify the problem into two dimensions: morphology processing for morphologically rich languages such as German, Spanish, Arabic, etc and word segmentation for Asian languages like Chinese, Thai, Japanese where word boundaries do not appear in written text.

1.1 Morphology Analysis for Machine Translation

Note that there is an active research on morphology disambiguation without considering the impact of the output on other applications. We limit our topic to the research focus on application of morphology for English translation and will only review the former when there is an overlap. Most of the work used lemmatization, tokenization, POS tagging language dependent techniques such as (Sonja [2004]) on German, (Goldwater and Mcclosky [2005]) on Czech, (Lee [2004]) on Arabic.

1.2 Word Segmentation

Approaches to word segmentation fall into two categories: heuristic dictionary-based methods (Wu [2003]) and statistical machine learning methods i.e (Peng et al. [2004]).

Instead of using single best segmentation for translation, there has been work (Xu [2005]) handled all segmentation alternatives by reading segmentation lattices for translation.

2 Preferred Presentation Dates

I want to have my presentation after the 5th of March.

References

Sharon Goldwater and David Mcclosky. Improving statistical mt through morphological analysis. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 676–683, Vancouver, British Columbia, Canada, October 2005. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology-new/H05/H05-1085.bib`.

Young S. Lee. Morphological analysis for statistical machine translation. In Daniel and Salim Roukos, editors, *HLT-NAACL 2004: Short Papers*, pages 57–60, Boston, Massachusetts, USA, May February - May July 2004. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology-new/N/N04/N04-4015.bib`.

Fuchun Peng, Fangfang Feng, and Andrew Mccallum. Chinese segmentation and new word detection using conditional random fields. In *Proceedings of Coling 2004*, pages 562–568, Geneva, Switzerland, Aug FebruaryMarch–Aug FebruaryJuly 2004. COLING. URL `http://www.aclweb.org/anthology-new/C/C04/C04-1081.bib`.

Sonja. Statistical machine translation with scarce resources using morpho-syntactic information. *Computational Linguistics*, 30(2), June 2004. URL `http://www.aclweb.org/anthology-new/J/J04/J04-2003.pdf`.

A. Wu. Chinese word segmentation in msr-nlp. In *Proceedings of the Second SIGHAN Workshop on Chinese Language Processing*, Japan, 2003. URL `http://acl.ldc.upenn.edu/W/W03/W03-1727.pdf`.

Xu. Integrated chinese word segmentation in statistical machine translation. In *Proceedings of the International Workshop on Spoken Language Translation (IWSLT)*, pages 141–147, Pittsburgh, PA, October 2005. URL `http://www-i6.informatik.rwth-aachen.de/ zens/publications/Xu+Matusov+Zens+Ney_Integrat`