n-Best Reranking for the Efficient Integration of Word Sense Disambiguation and Statistical Machine Translation

Lucia Specia[1,2], Baskaran Sankaran[2] and
Maria das Graças Volpe Nunes[1]

[1] NILC/ICMC - Universidade de São Paulo
Trabalhador São-Carlense, 400, São Carlos, 13560-970, Brazil
lspecia@icmc.usp.br, gracan@icmc.usp.br
[2] Microsoft Research India
"Scientia", 196/36, 2nd Main, Sadashivanagar, Bangalore-560080, India
baskaran@microsoft.com

Abstract. Although it has been always thought that Word Sense Disambiguation (WSD) can be useful for Machine Translation, only recently efforts have been made towards integrating both tasks to prove that this assumption is valid, particularly for Statistical Machine Translation (SMT). While different approaches have been proposed and results started to converge in a positive way, it is not clear yet how these applications should be integrated to allow the strengths of both to be exploited. This paper aims to contribute to the recent investigation on the usefulness of WSD for SMT by using n-best reranking to efficiently integrate WSD with SMT. This allows using rich contextual WSD features, which is otherwise not done in current SMT systems. Experiments with English-Portuguese translation in a syntactically motivated phrase-based SMT system and both symbolic and probabilistic WSD models showed significant improvements in BLEU scores.

1 Introduction

The need for Word Sense Disambiguation (WSD) in Machine Translation (MT) systems has been discussed since the early research on MT back in the 1960's. While MT was primarily addressed by rule-based approaches, the consequences of the lack of semantic disambiguation was already emphasized in DARPA's report by Bar-Hillel [2], which resulted in a considerable reduction in the funding for research on MT that time. Meanwhile, WSD grew as an independent research area, without focusing on any particular application.

With the introduction in the 1990's of the Statistical Machine Translation (SMT) approach [3], it has been taken for granted that SMT systems can implicitly address the sense disambiguation problem, especially by their word alignment and target language models. However, the SMT systems normally consider a very short window as context and therefore lack richer information coming from larger contexts or other knowledge sources.