A Semantics-Enhanced Language Model for Unsupervised Word Sense Disambiguation

Shou-de Lin[1] and Karin Verspoor[2]

[1] National Taiwan University, sdlin@csie.ntu.edu.tw
[2] Los Alamos National Laboratory, verspoor@lanl.gov

Abstract. An N-gram language model aims at capturing statistical word order dependency information from corpora. Although the concept of language models has been applied extensively to handle a variety of NLP problems with reasonable success, the standard model does not incorporate semantic information, and consequently limits its applicability to semantic problems such as word sense disambiguation. We propose a framework that integrates semantic information into the language model schema, allowing a system to exploit both syntactic and semantic information to address NLP problems. Furthermore, acknowledging the limited availability of semantically annotated data, we discuss how the proposed model can be learned without annotated training examples. Finally, we report on a case study showing how the semantics-enhanced language model can be applied to unsupervised word sense disambiguation with promising results.

1 Introduction

Syntax and semantics both play an important role in language use. Syntax refers to the grammatical structure of a language whereas semantics refers to the meaning of the symbols arranged with that structure. To fully comprehend a language, a human must understand its syntactic structure, the meaning each symbol represents, and the interaction between the two. In most languages, syntactic structure conveys something about the semantics of the symbols, and the semantics of symbols may constrain valid syntactic realizations. As a simple example: when we see a noun following a number in English (e.g. "one book"), we can infer that the noun is countable. Conversely, if it is known that a noun is countable, a speaker of English knows that it can plausibly be preceded by a numeral. It is therefore reasonable to assume that for a computer system to successfully process natural language, it has to be equipped with capabilities to represent and utilize both the syntactic and semantic information of the language simultaneously.

The n-gram language model (LM) is a powerful and popular framework for capturing the word order information of language, or fundamentally syntactic information. It has been applied successfully to a variety of NLP problems such as machine translation, speech recognition, and optical character recognition. As described in equation (1), an n-gram language model utilizes conditional probabilities to capture word order information, and the validity of a sentence