Automatic Generation of Multilingual Lexicon by Using Wordnet

Nitin Verma and Pushpak Bhattacharyya

Department of Computer Science and Engineering, I.I.T. Bombay,
{nitinv,pb}@iitb.ac.in

Abstract. A lexicon is the heart of any language processing system. Accurate words with grammatical and semantic attributes are essential or highly desirable for any application- be it machine translation, information extraction, various forms of tagging or text mining. However, good quality lexicons are difficult to construct requiring enormous amount of time and manpower. In this paper, we present a method for automatically generating multilingual Universal Word (UW) dictionaries (for English, Hindi and Marathi) from an input document-making use of English, Hindi and Marathi WordNets. The dictionary entries are in the form of Universal Words (UWs) which are language words (primarily English) concatenated with disambiguation information. The entries are associated with syntactic and semantic properties- most of which too are generated automatically. In addition to the WordNet, the system uses a word sense disambiguator, an inferencer and the knowledge base (KB) of the Universal Networking Language which is a recently proposed interlingua. The lexicon so constructed is sufficiently accurate and reduces the manual labor substantially.

1 Introduction

Construction of good quality lexicons enriched with syntactic and semantic properties for the words is time consuming and manpower intensive. Also word sense disambiguation presents a challenge to any language processing application, which can be posed as the following question: *given a document **D** and a word **W** therein, which sense **S** of **W** should be picked up from the lexicon?*. It is, however, a redeeming observation that a particular **W** in a given **D** is mostly used in a single sense throughout the document. This motivates the following problem: *can the task of disambiguation be relegated to the background before the actual application starts? In particular, can one construct a **Document Specific Dictionary*** wherein single senses of the words are stored?

Such a problem is relevant, for example, in a machine translation context [1]. For the input document in the source language, if the *document specific dictionary* is available a-priory, the generation of the target language document reduces to essentially syntax planning and morphology processing for the pair of languages involved. The WSD problem has been solved before the MT process starts, by putting in place a lexicon with the document specific senses of the words.

© J. Cardeñosa, A. Gelbukh, E. Tovar (Eds.)
Universal Network Language: Advances in Theory and Applications.
Research on Computing Science 12, 2005, pp. 380–391.