A POS Tagger analysed in collaboration environments and literary texts

Dumitru-Clementin Cercel*
Ştefan Trăuşan-Matu**

*University *"Politehnica"* of Bucharest,
Department of Computer Science and Engineering
Splaiul Independenţei Bd., No. 313, Bucharest, Romania
clementin.cercel@gmail.com
**Romanian Academy Research Institute for Artificial Intelligence
13 Septembrie Street, No. 13, Bucharest, Romania
stefan.trausan@cs.pub.ro

Abstract. Part-of-speech (POS) tagging is often used in other modules of natural language processing and therefore the results of this process should be as precise as possible. Many different types of taggers have been developed to improve the accuracy of the results in the field of literature or newspapers. Nowadays when the internet is widespread, the environments for online collaboration as chats, forums, blogs, wikis have become important means of communication. The purpose of this research is to analyse the results of tagging the words obtained from the labelling of the words from the online collaboration environments and literary texts with the corresponding parts of speech. In the case of POS tagging, the ambiguities arise due to the fact that a word may have multiple morphological values depending on context.

1 Introduction

Part-of-speech (POS) tagging is the process of grammatical labelling of each word inside a text with its appropriate part of speech. Labelling may also contain extra information related to the morphological characteristics of the respective language like number, gender, person, tense and aspect of the verb.

Many different types of taggers have been developed to improve the accuracy of the results in the field of literature or newspapers. Nowadays when the internet is widespread, the environments for online collaboration as chats, forums, blogs, wikis have become important means of communication. The purpose of the research presented in this paper is performing a comparative analysis of POS tagging in collaborative corpora (specifically for chats, Wikipedia and Twitter as an example of microbloggins) and literature (the texts from Brown corpus). In this aim we implemented a trigram HMM tagger according to Jurafsky and Martin (2000) and Brants (2000).

The rest of this paper is structured as follows. In section 2 we briefly review: the state of the art approaches to POS tagging, the Markov Hidden Model, the implementation of the