Analyzing Hungarian webtext

Viktor Varga[1], Vilmos Wieszner[1], Hangya Viktor[1],
Veronika Vincze[2], Richárd Farkas[1]

[1] University of Szeged, Department of Informatics
{viktor.varga.1991,vilmos.wieszner,hangyav}@gmail.com,
rfarkas@inf.u-szeged.hu
[2] MTA-SZTE Research Group on Artificial Intelligence
vinczev@inf.u-szeged.hu

The Internet's role in people's lives is becoming more and more significant, especially due to its importance in modern communication. A large amount of data is generated by the users' communication through this medium, and this could be useful for a number of natural language processing applications, for example in information extraction and sentiment analysis. Thus analyzing webtext is gaining importance. Non-standard language use is the biggest difficulty in this context, which decreases the efficiency of language processing tools developed for standard texts.

In this paper, we focus on Hungarian webtexts. As Hungarian is the prototype of morphologically rich languages, we investigate the question whether the required adaptation techniques from standard texts to webtexts are similar to the ones introduced for English. We identified the most frequent error types of our linguistic analyzing toolchain for Hungarian (magyarlanc) and our Named Entity Recogniser on public facebook messages along with their comments and tweets. These tools were developed on the Szeged Treebank (i.e. on standard texts).

Imitating spoken language and therefore focusing on speed and the expression of emotions are part of the fundamental nature of social media texts. Speed is increased by quicker typing: diacritics, punctuations, whitespaces and capitals often disappear, abbreviations are used and typos are often made. Emotions may be expressed through the overuse of capitals and punctuations, or by emoticons. Explicit expression of hesitation, inventing words, and the use of English words and abbreviations are also frequent stylistic means. All these depend on the individual language use, registers and contexts.

Capitalization and punctuations cannot be used as guidelines in the segmentations of sentences, and the lack of whitespaces make word tokenization difficult. NER systems cannot handle lowercase names, while uppercase words are automatically detected as named entities. The morphological parser cannot analyze or assigns the wrong code to misspelt or unknown words, which affects the syntactic analysis as well. The differences between English and Hungarian make modifications based solely on English chat language insufficient, different solutions are required, e.g. phonetic transcription (*thru* instead of *through*) is more problematic for English texts due to the complexity of English orthography but the lack of accents (*kerek* vs. *kerék* vs. *kérek*) is only relevant for Hungarian. We propose the normalization of the input text, expansion of the lexica and domain-adaptation of current processing modules. We believe that the combination of all these methods could significantly increase performance.