Identification of Transliterated Foreign Words in Hebrew Script

Yoav Goldberg and Michael Elhadad

Computer Science Department
Ben Gurion University of the Negev
P.O.B 653 Be'er Sheva 84105, Israel
{yoavg,elhadad}@cs.bgu.ac.il

Abstract. We present a loosely-supervised method for context-free identification of transliterated foreign names and borrowed words in Hebrew text. The method is purely statistical and does not require the use of any lexicons or linguistic analysis tool for the source languages (Hebrew, in our case). It also does not require any manually annotated data for training – we learn from noisy data acquired by over-generation. We report precision/recall results of 80/82 for a corpus of 4044 unique words, containing 368 foreign words.

1 Introduction

Increasingly, native speakers tend to use borrowed foreign terms and foreign names in written texts. In sample data, we found genres with as many as 5% of the word instances borrowed from foreign languages. Such borrowed words appear in a transliterated version. Transliteration is the process of writing the phonetic equivalent of a word of language A in the alphabet of language B. Borrowed words can be either foreign loan words with no equivalent in language B, or words from language A used as slang in language B. Identifying foreign words is not a problem in languages with very similar alphabets and sound systems, as the words just stay the same. But this is not the case in words borrowed from languages that have different writing and sound systems, such as English words in Japanese, Hebrew and Arabic texts.

Transliterated words require special treatment in NLP and IR systems. For example, in IR, query expansion requires special treatment for foreign words; when tagging text for parts of speech, foreign words appear as unknown words and the capability to identify them is critical for high-precision PoS tagging; in Machine Translation, back transliteration from the borrowed language to the source language requires the capability to perform the inverse operation of transliteration; in Named Entity Recognition and Information Extraction, the fact that a word is transliterated from a foreign language is an important feature to identify proper names.

We focus in this paper on the task of identifying whether a word is a transliteration from a foreign language – and not on the task of mapping back the