عنوان مقاله:

Recognizing Transliterated English Words in Persian Texts

محل انتشار:

فصلنامه سیستم های اطلاعاتی و مخابرات, دوره 8, شماره 2 (سال: 1399)

تعداد صفحات اصل مقاله: 9

نویسندگان:

Ali Hoseinmardy - *Computer Engineering Department, Amirkabir University of Technology, Iran*

Saeedeh Momtazi - *Computer Engineering Department, Amirkabir University of Technology, Iran*

خلاصه مقاله:

One of the most important problems of text processing systems is the word mismatch problem. This results in limited access to the required information in information retrieval. This problem occurs in analyzing textual data such as news, or low accuracy in text classification and clustering. In this case, if the text-processing engine does not use similar/related words in the same sense, it may not be able to guide you to the appropriate result. Various statistical techniques have been proposed to bridge the vocabulary gap problem; e.g., if two words are used in similar contexts frequently, they have similar/related meanings. Synonym and similar words, however, are only one of the categories of related words that are expected to be captured by statistical approaches. Another category of related words is the pair of an original word in one language and its transliteration from another language. This kind of related words is common in non-English languages. In non-English texts, instead of using the original word from the target language, the writer may borrow the English word and only transliterate it to the target language. Since this kind of writing style is used in limited texts, the frequency of transliterated words is not as high as original words. As a result, available corpus-based techniques are not able to capture their concept. In this article, we propose two different approaches to overcome this problem: (۱) using neural network-based transliteration, (۲) using available tools that are used for machine translation/transliteration, such as Google Translate and Behnevis. Our experiments on a dataset, which is provided for this purpose, shows that the combination of the two approaches can detect English words with ۸۹.۳۹% accuracy.

کلمات کلیدی:

Transliteration; Text processing; Words Relation; Neural Network-Based Sequence۲Sequence Model; Google Translate; Behnevis