EAMT 2016

Graphonological Levenshtein Edit Distance: Application for automated cognate identification

Authors

Address

email@institution

Abstract: This paper presents a methodology for calculating a modified Levenshtein edit distance between character strings and applies it to the task of automated cognate identification from non-parallel (comparable) corpora. This task is an important stage in developing MT systems and bilingual dictionaries beyond the coverage of traditionally used aligned parallel corpora, which is especially useful for finding translation equivalents for the 'long tail' in Zipfian distribution: low-frequency and usually unambiguous lexical items in closely-related languages (many of those often under-resourced).

Graphonological Levenshtein edit distance relies on editing hierarchical representations of phonological features for graphemes (graphonological representations) and improves on phonological edit distance proposed for measuring dialectological variation. Graphonological edit distance works directly with character strings and does not require an intermediate stage of phonological transcription, exploiting the advantages of historical and morphological principles of orthography, which are obscured if only phonetic principle is applied. Difficulties associated with plain feature representations (unstructured feature sets or vectors) are addressed by using linguistically-motivated feature hierarchy that restricts matching of lower-level graphonological features when higher-level features are not matched. The paper presents an evaluation of the graphonological edit distance in comparison with the traditional Levenshtein edit distance from the perspective of its usefulness for the task of automated cognate identification and discusses the advantages of the proposed method.

Keywords: cognates; Levenshtein edit distance; phonological features; comparable corpora; closely-related languages; under-resourced languages; Ukrainian; Russian; Hybrid MT

1. Introduction

Levenshtein edit distance proposed in (Levenshtein, 1966) is an algorithm that calculates the cost (normally – the number of operations such as deletions, insertions and substitutions) needed to transfer a string of symbols (characters or words) into another string. This algorithm is used in many computational linguistic applications that require some form of the fuzzy string matching. Applications of this metric for the translation technologies and specifically Machine Translation include automated identification of cognates for the tasks of creating bilingual resources such as electronic dictionaries (e.g., Koehn and Knight, 2002; Mulloni & Pekar, 2006; Bergsma & Kondrak, G. 2007)