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.

1. Introduction

Text

Text2

1. Section2

Text

Text2

1. Section3

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |
|  |  |

|  |  |
| --- | --- |
| r (р) | ['type:consonant', 'voice:sonorant', 'maner:thrill', 'active:fronttongue', 'passive:palatal'] |
| o (о) | ['type:vowel', 'backness:back', 'height:mid', 'roundedness:rounded', 'palate:nonpalatalizing'] |
| b (б) | **['type:consonant'**, **'voice:voiced'**, 'maner:plosive', **'active:labial'**, 'passive:bilabial'] |
| i (і) | ['type:vowel', 'backness:front', 'height:close', 'roundedness:unrounded', 'palate:nonpalatalizing'] |
| t (т) | **['type:consonant',** **'voice:unvoiced'**, 'maner:plosive', **'active:fronttongue'**, **'passive:alveolar'**] |
| n (н) | ['type:consonant', 'voice:sonorant', 'maner:nasal', 'active:fronttongue', 'passive:alveolar'] |
| y (и) | ['type:vowel', 'backness:front', 'height:closemid', 'roundedness:unrounded', 'palate:nonpalatalizing'] |
| k (к) | ['type:consonant', 'voice:unvoiced', 'maner:plosive', 'active:backtongue', 'passive:velar'] |

Table 1: Phonological feature vectors for Ukrainian word ‘robitnyk’ (робітник) – ‘worker’

|  |  |
| --- | --- |
| r (р) | ['type:consonant', 'voice:sonorant', 'maner:thrill', 'active:fronttongue', 'passive:palatal'] |
| o (о) | ['type:vowel', 'backness:back', 'height:mid', 'roundedness:rounded', 'palate:nonpalatalizing'] |
| v (в) | **['type:consonant',** **'voice:voiced'**, 'maner:fricative', **'active:labial'**, 'passive:labiodental'] |
| e (е) | ['type:vowel', 'backness:front', 'height:mid', 'roundedness:unrounded', 'palate:palatalizing'] |
| s (с) | **['type:consonant',** **'voice:unvoiced'**, 'maner:fricative', **'active:fronttongue'**, **'passive:alveolar'**] |
| n (н) | ['type:consonant', 'voice:sonorant', 'maner:nasal', 'active:fronttongue', 'passive:alveolar'] |
| i (и) | ['type:vowel', 'backness:front', 'height:close', 'roundedness:unrounded', 'palate:nonpalatalizing'] |
| k (к) | ['type:consonant', 'voice:unvoiced', 'maner:plosive', 'active:backtongue', 'passive:velar'] |

Table 2: Phonological feature vectors for Russian word ‘rovesnik (ровесник) – ‘age-mate’, ‘of the same age’

|  |  |
| --- | --- |
| r (р) | ['type:consonant', 'voice:sonorant', 'maner:thrill', 'active:fronttongue', 'passive:palatal'] |
| a (а) | ['type:vowel', 'backness:back', 'height:open', 'roundedness:unrounded', 'palate:nonpalatalizing'] |
| b (б) | ['type:consonant', 'voice:voiced', 'maner:plosive', 'active:labial', 'passive:bilabial'] |
| o (о) | ['type:vowel', 'backness:back', 'height:mid', 'roundedness:rounded', 'palate:nonpalatalizing'] |
| t (т) | ['type:consonant', 'voice:unvoiced', 'maner:plosive', 'active:fronttongue', 'passive:alveolar'] |
| n (н) | ['type:consonant', 'voice:sonorant', 'maner:nasal', 'active:fronttongue', 'passive:alveolar'] |
| i (и) | ['type:vowel', 'backness:front', 'height:close', 'roundedness:unrounded', 'palate:nonpalatalizing'] |
| k (к) | ['type:consonant', 'voice:unvoiced', 'maner:plosive', 'active:backtongue', 'passive:velar'] |

Table 2: Phonological feature vectors for Russian word ‘rabotnik (работник) – ‘worker’

Compare b ~ v

Compare t ~ n

r(р) o(о) b(б) i(і) t(т) n(н) y(и) k(к)

0.0 1.0 2.0 3.0 4.0 5.0 6.0 7.0 8.0

r(р) 1.0 0.0 1.0 2.0 3.0 4.0 5.0 6.0 7.0

o(о) 2.0 1.0 0.0 1.0 2.0 3.0 4.0 5.0 6.0

v(в) 3.0 2.0 1.0 1.0 2.0 3.0 4.0 5.0 6.0

e(е) 4.0 3.0 2.0 2.0 2.0 3.0 4.0 5.0 6.0

s(с) 5.0 4.0 3.0 3.0 3.0 3.0 4.0 5.0 6.0

n(н) 6.0 5.0 4.0 4.0 4.0 4.0 3.0 4.0 5.0

i(и) 7.0 6.0 5.0 5.0 5.0 5.0 4.0 3.0 4.0

k(к) 8.0 7.0 6.0 6.0 6.0 6.0 5.0 4.0 3.0

r(р) o(о) b(б) i(і) t(т) n(н) y(и) k(к)

0.0 1.0 2.0 3.0 4.0 5.0 6.0 7.0 8.0

r(р) 1.0 0.0 1.0 2.0 3.0 4.0 5.0 6.0 7.0

o(о) 2.0 1.0 0.0 1.0 2.0 3.0 4.0 5.0 6.0

v(в) 3.0 2.0 1.0 0.4 1.4 2.4 3.4 4.4 5.4

e(е) 4.0 3.0 2.0 1.4 0.8 1.8 2.8 3.8 4.8

s(с) 5.0 4.0 3.0 2.4 1.8 1.0 2.0 3.0 4.0

n(н) 6.0 5.0 4.0 3.4 2.8 2.0 1.0 2.0 3.0

i(и) 7.0 6.0 5.0 4.4 3.4 3.0 2.0 1.2 2.2

k(к) 8.0 7.0 6.0 5.4 4.4 3.8 3.0 2.2 1.2

Annotation scheme: ranking

Practical session for automated annotation of cognates

|  |  |  |
| --- | --- | --- |
|  | **per cent** | **count** |
| 0 Difference cognates | 16.42% | 45 |
| Of which proper nouns | 5.84% | 16 |
| Have no cognates | 34.31% | 94 |
| False Friends | 1.82% | 5 |
| **All sample** | **100%** | **274** |

Table 1. Parameters of evaluation sample

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Lev** | | ***GPFeat Vectors*** | | **GPFeat Hierarch** | | **Difference: GPFeatHierarchy - Lev** |
|  | **per cent** | **#** | ***per cent*** | ***#*** | **per cent** | **#** | **per cent** |
| **correct, higher better (+exclude 0 differences)** | 47.08% (36.68%) | 129 (84) | *46.72%* | *128* | **51.09% (41.48%)** | 140 (95) | **4.01% (4.80%)** |
| missing (lower better) | 13.87% | 38 | *10.58%* | *29* | **9.85%** | 27 | **4.02%** |
| lower rank (lower better) | **2.19%** | 6 | *10.58%* | *29* | 2.55% | 7 | -0.36% |

Table 2. Comparative performance of distance measures for the task of ranking cognates

Text

Text2