Baltic J. Modern Computing, Vol.4 (2016), No.2, 115-128

Graphonological Levenshtein Edit Distance: Application for Automated Cognate Identification

Bogdan BABYCH

Centre for Translation Studies, University of Leeds, Leeds, LS2 9JT, UK

b.babych@leeds.ac.uk

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 can be used 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. It discusses the advantages of the proposed method, which can be used for morphology induction, for robust transliteration across different alphabets (Latin, Cyrillic, Arabic, etc.) and robust identification of words with non-standard or distorted spelling, e.g., in user-generated content on the web such as posts on social media, blogs and comments. Software for calculating the modified feature-based Levenshtein distance, and the corresponding graphonological feature representations (vectors and the hierarchies of graphemes' features) are released on the author's webpage: http://corpus.leeds.ac.uk/bogdan/phonologylevenshtein/. Features are currently available for Latin and Cyrillic alphabets and will be extended to other alphabets and languages.

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, examples include fast creation of morphological

and syntactic taggers exploiting similarities between closely related languages (Hana et al., 2006), statistical learning of preferred edits for detecting regular orthographic correspondences in closely related languages (Ciobanu and Dinu, 2014). Applications of Levenshtein's metric for the translation technologies and specifically for 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 and Pekar, 2006; Bergsma and Kondrak, G. 2007), improving document alignment by using cognate translation equivalents as a seed lexicon (Enright, J and Kondrak, G., 2007), automated MT evaluation (e.g., Niessen et al., 2000; Leusch et al., 2003).

Levenshtein distance metrics has been modified and extended for applications in different areas; certain ideas have yet not been tested in MT context, but have a clear potential for benefiting MT-related tasks. This paper develops and evaluates one of such ideas for a linguistic extension of the metric proposed in the area of computational modelling of dialectological variation and measuring 'cognate' lexical distance between languages, dialects and different historical periods in development of languages, e.g., using cognates from the slow-changing part of the lexicon – the Swadesh list (Swadesh, 1952; Serva and Petroni, 2008; Schepens et al., 2012).

In this paper the suggestion is explored of calculating the so called Levenshtein's 'phonological edit distance' between phonemic transcriptions of cognates, rather than the traditional string edit distance (Nerbonne and Heeringa 1997; Sanders and Chin, 2009). This idea is based on the earlier linguistic paradigm of describing phonemes as systems of their phonological features, formulated in its modern form by Roman Jacobson – see (Anderson, 1985) for the development of the theory; later it was introduced into generative and computational linguistic paradigms by Chomsky and Halle (1968). The idea is that each phoneme in a transcription of a cognate is represented as a structure of phonological distinctive features, such as:

[a] = [+vowel, +back; +open; –labialised]

1.1. Distinctive phonological features: the background

In phonology, sounds of a language form a system of phonemes (i.e., minimal segments of speech that can be used in the same context and distinguish meanings in minimal word pairs, which differ only by one such segment (i.e., a phoneme). For example, English phonemes */p/* and */b/* distinguish meaning in *pull vs. bull; pill vs. bill*; phonemes /*v*/ and /*w*/ distinguish meanings of *vary vs. wary*. However, Ukrainian sounds /*v*/ and /*w*/ are positional variants, or allophones, of the same phoneme, since they are never used in the same position or distinguish meanings: /w/ is restricted to a word-final position after a vowel: *вийшов /vyjšow/ 'entered'*). There is evidence that phonemes are not simply linguistic constructs, but have a psychological reality, e.g., for native speakers they form cognitive pronunciation targets; non-native speakers often confuse phonemes that are not separated in their first language (e.g., native Ukrainian speakers would confuse /*v*/ and /*w*/ when speaking English). In languages where writing systems and pronunciation are close to each other, e.g., Ukrainian or Georgian, the written characters usually correspond to phonemes (much less often – to allophones).

Phonemes and allophones are characterised by a further internal structure, which consists of a system of distinctive phonological features (Jakobson et al., 1958). These features are typically based on differences in their acoustic properties and the way of how they are pronounced (their articulation). For example, /*v*/ and /*w*/ are both *consonants*, i.e., they are formed with a participation of noise (unlike vowels */u, o, a/*, etc., which are

formed with an unobstructed sound); both are *fricative* consonants, i.e., they are formed with a constant air friction against an obstacle in the vocal tract (unlike plosive consonants, such as /*b, p, d, t, g, k*/ that include a build up of air behind some obstacle during an initial silence, followed by its instant release); the difference between /*v*/ and /*w*/ is that */v/* is *labio-dental*, i.e., the air friction is created with the teeth and the lower lip, while */w/* is bilabial, i.e., the source of friction is the upper and lower lip, while the teeth are not involved.

However, not all acoustic or articulatory differences become distinctive phonological features. The necessary condition is that these features should capture *phonological* distinctions, i.e., those needed for differentiation between phonemes: e.g., *long* vs. *short* pairs of vowels in Dutch differ primarily by their length; however, they have further qualitative differences as well, which are visible on their spectrograms, but are not perceived by speakers as features that make phonemic distinctions; therefore, these qualitative differences are not part of their distinctive phonological features. Similarly, the same Ukrainian vowels in stressed and unstressed positions are very different qualitatively, but these differences are not perceived as phonological, i.e., the ones that distinguish different phonemes, so both stressed and unstressed variants have the same set of distinctive features.

Some distinctive phonological features are in *correlated oppositions*, i.e., they distinguish sets of phonemes that only differ by a single feature, e.g., *+voiced* vs *–voiced* (i.e., formed with or without the vocal cords) distinguishes */d/~/t/; /z/~/s/; /b/~/p/; /v/~/f/, /g/~/k/*. These correlated features often switch their value in positional or historical alternations, and as a result, may distinguish cognates in closely related languages.

Nowadays there are standard description of phonemes and phonological features for most languages of the world, illustrated with sound charts, e.g., by the International Phonetic Association (IPA) (Ladefoged and Halle, 1988). These charts group sounds along several dimensions of their distinctive phonological features, such as *place, manner* of articulation, *voiced/voiceless* for consonants; *high/low, back/front, roundness* for vowels, with finer-grained sub-divisions. Sound charts for individual languages can be found in standard language references. For the experiments described in this paper the systems of phonological distinctive features for Ukrainian and Russian has been adapted from (Comrie and Corbett, Eds., 1993: 949, 951, 829).

1.2. Application of phonological features for calculating the edit distance

For using phonological distinctive features in calculation of the Levenshtein edit distance, the idea is to replace the operation of substitution of a whole character by the substitution of its constituent phonological feature representations, which would be sufficient to convert it into another character: so rewriting [o] into [a] (which, e.g., is a typical vowel alternation pattern in Russian and distinguishes some of its major dialects) would incur a smaller cost compared to the substitution of the whole character, since only two of its distinctive phonological features need to be rewritten:

[o] = [+vowel, +back; *+mid; +labialised*]

On the other hand, the cost of rewriting the vowel [a] into the consonant [t] (the change which normally does not happen as part of the historical language development or dialectological variation) would involve rewriting all the phonological features in the representation, so the edit cost will be the same as for the substitution of the entire character:

[t] = [*+consonant; –voiced; +plosive; +fronttongue; +alveolar*]

According to Nerbonne and Heeringa (1997:2) the feature-based Levenshtein distance makes it "…possible to take into account the affinity between sounds that are not equal, but are still related"; and to "…show that *'pater'* and *'vader'* are more kindred then *'pater'* and *'maler'*." This is modelled by the fact that phonological feature representations for pairs such as [t] and [d] (both front-tongue alveolar plosive consonants, which only differ by 'voiced' feature), as well as [p] and [v] (both labial consonants), share greater number of phonological features compared to the pairs [p] and [m] (which differ in sonority, manner and passive organ of articulation) or [t] and [l] (which differ in sonority and the manner of articulation). However, the authors point out to a number of open questions and problems related to their modified metric, e.g., how to represent phonetic features of complex phonemes, such as diphthongs; what should be the structure of feature representations: Nerbonne and Heeringa use feature vectors, but are these vectors sufficient or more complex feature representations are needed; how to integrate edits of individual features into the calculation of a coherent distance measure (certain settings are not used, whether to use Euclidian or Manhattan distance, etc.).

Linguistic ideas behind the suggestion to use Levenshtein phonological edit distance are intuitively appealing and potentially useful for applications beyond dialectological modelling. However, to understand their value for other areas, such as MT, there is a need to develop a clear evaluation framework for testing the impact of different possible settings of the modified metric and different types of feature representations, to compare specific settings of the metric to alternatives and the classical Levenshtein's baseline. Without a systematic evaluation framework the usefulness of metrics remain unknown.

This paper proposes an evaluation framework for testing alternative settings of the modified Levenshtein's metric. This framework is task-based: it evaluates the metric's alternative settings and feature representations in relation to its success on the task of automated identification of cognates from non-parallel (comparable) corpora. The scripts for calculating the modified feature-based Levenshtein distance, and the corresponding graphonological feature representations (vectors and the hierarchies of features) are released on the author's webpage[1]. Features are currently available for Latin and Cyrillic alphabets, new alphabets will be added in future.

Graphonological Levenshtein distance can also be applied, calibrated and evaluated for other tasks, beyond the task of cognate identification, e.g., to robust transliteration, reconstruction of diacritics or recognition of words with distorted, non-standard or variable spelling, e.g.: the names *Osama/ Usama/ Ousamma /Осама/ Усама/ Усамма* are closer to each other in terms of their underlying phonological feature sequences than their plain character-based distances. Evaluation on these tasks may lead to alternative preferred settings and feature representations for the graphonological Levenshtein metric, compared to evaluation on the cognate identification task described here.

The paper is organised as follows: Section 2 presents the set-up of the experiment, the application of automated cognate identification; the design and feature representations for the metric and the evaluation framework. Section 3 presents evaluation results of different metric settings and comparison with the classical Levenshtein distance; Section 4 presents conclusion and future work.

[1] http://corpus.leeds.ac.uk/bogdan/phonologylevenshtein/

2. Set up of the experiment

2.1. Application of automated cognate identification for MT

Automated cognate identification is important for a range of MT-related tasks, as mentioned in Section 1. Our project deals with rapid creation of hybrid MT systems for new translation directions into and from a range of under-resourced languages, many of which are closely related, or 'cognate', such as Spanish and Portuguese, German and Dutch, Ukrainian and Russian. The systems combine rich linguistic representations used by a backbone rule-based MT engine with statistically derived linguistic resources and statistical disambiguation and evaluation techniques, which work with complex linguistic data structures for morphological, syntactic and semantic annotation (Eberle et al., 2012). While there is a potential in using a better-resourced pivot language for creating linguistic resources for MT and building pivot systems (e.g., Babych et al., 2007), in our project the translation lexicon for the hybrid MT systems is derived mainly via two routes:

1. Translation equivalents for a smaller number of highly frequent words, which under empirical observations of Zipf's and Menzerath's laws (Koehler, R. 1993; 49) tend to be shorter (Zipf, 1935:38; Sigurd et al., 2004:37) and more ambiguous (Menzerath, 1954, Hubey, 1999; Babych et al., 2004: 7), are generated as statistical dictionaries from sentence-aligned parallel corpora. However, as only small number of parallel resources is available for under-resourced languages, there remain many out-of-vocabulary lexical items.
2. The remaining 'long tail' in Zipfian distribution containing translation equivalents for a large number of low-frequent and usually unambiguous lexical items (as they typically have only one correct translation equivalent) is derived semi-automatically from much larger non-parallel comparable corpora, which are usually in the same domain for both languages. We use a number of different techniques depending on available resources and language pairs (Eberle et al., 2012: 104-106). For closely related languages (depending on the degree of their 'relatedness') the 'long tail' contains a large number of cognates. In the experiments described here, for Ukrainian / Russian language pair this number reached 60% of the analysed sample of the lexicon selected from different frequency bands (see Section 3).

In order to cover this part of the lexicon, the automated cognate identification from non-parallel corpora is used for generating draft ranked lists of candidate translation equivalents. The candidate lists are generated using the following procedure:

1. Large monolingual corpora (in my experiments – about 250M for Ukrainian and 200M for Russian news corpora) are PoS tagged and lemmatised.
2. Frequency dictionaries are created for lemmas. A frequency threshold is applied (to keep down the 'noise' and the number of hapax legomena.
3. Edit distances for pairs of lemmas in a Cartesian product of the two dictionaries are automatically calculated using variants of the Levenshtein measure.
4. Pairs with edit distances below a certain threshold are retained as candidate cognates (in the experiments I used the threshold value of the Levenshtein edit distance normalised by the length of the longest word <=0.36, intuitively: 36% of edits per character)
5. Candidate cognates are further filtered by part-of-speech codes (cognates with non-matching parts of speech are not ranked).

6. Candidate cognates are filtered by their frequency bands: if the TL candidate is beyond the frequency band threshold of the SL candidate, the TL candidate is not ranked (in the experiment I used the threshold *FrqRange* > 0.5 for the difference in natural logarithms of absolute frequencies – see formula (1), intuitively: candidates should not have frequency difference several orders of magnitude apart.
7. Candidate cognate lists are ranked by the increasing values of the edit distance.

$$FrqRange = \frac{\min(\ln(FrqB), \ln(FrqA))}{\max(\ln(FrqB), \ln(FrqA))} \quad (1)$$

These ranked lists are presented to the developers, candidate cognates are checked and either included into system dictionaries, or rejected. Developers' productivity of this task crucially depends on the quality of automated edit distance metric that generates and ranks the draft candidate lists.

The task of creating parallel resources and dictionaries from comparable corpora is not exclusive to hybrid or rule-based MT. Similar ideas are used in SMT framework for enhancing SMT systems developed for under-resourced languages via identification of aligned sentences and translation equivalents in comparable corpora, which generally reduces the number of out-of-vocabulary words not covered by scarce parallel corpora (Pinnis et al., 2012). In these settings, dictionaries of cognate lists can become an additional useful resource, so achieving a higher degree of automation for the process of cognate identification in comparable corpora is equally important for the SMT development. Under these settings an operational task-based evaluation for Levenshtein edit distance metrics will be the performance parameters of the developed SMT systems.

2.2. Development of Levenshtein graphonological feature-based metric

For the task of automated cognate identification a feature-based edit distance will need further adjustments, which go beyond the metric used in modelling dialectological variation. The metric is designed to work directly with orthography rather than with phonetic transcriptions; alternative ways of representing phonological features (feature vectors vs. feature hierarchies) are evaluated, and a method of calculation of rewriting cost for feature-based representations is selected.

2.2.1. Phonological distance: phonetic transcription vs. raw orthographic strings

The metric works directly with word character strings, not via the intermediate stage of creating a phonological transcription for each word. While for modelling of dialects (many of which do not capture pronunciation differences in their own writing systems) the transcription may be a necessary step, MT systems normally deal with languages with their own established writing systems. There are practical reasons for extracting features from orthography rather than phonological transcriptions: automated phonological transcription of the orthographic strings may create an additional source of errors; resources for transcribing may be not readily available for many languages; for the majority of languages very little can be gained by replacing the orthography by

transcription (apart from more adequate representation of digraphs and phonologically ambiguous characters, which can be addressed also on the level of orthography).

However, there are more important theoretical reasons for preferring original orthographic representations. For instance, orthography of languages is usually based on a combination of three principles: *phonetic* (how words are pronounced), *morphological* (keeping the same spellings for morphemes – minimal meaning units, such as affixes, stems, word routes, irrespective of any pronunciation variation caused by their position, phonological context, regular sound alternations, etc.) and *historic* (respecting traditional spelling which reflects an earlier stage of language development, even though the current pronunciation may have changed; often orthography reflects the stage when cognate languages have been closer together). Example 2 illustrates the point why orthography might work better for cognate identification:

	Russian	*Ukrainian*
Orthography	s**o**baka (собака) 'dog'	sobaka (собака) 'dog'
Phonological transcription	[s**a**baka] (с[а]бака)	[s**o**baka] (с[о]бака)
Change	[o] -> [a]	[o] -> (no change)

(2)

The pronunciation change [o] -> [a], which in some (at that time) marginal Russian dialects dates back to the 7th-8th century AD (Pivtorak, 1988: 94) (one of the explanations for this change is the influence the Baltic substratum), was not reflected in Russian educated written tradition, even at the later time when those dialects received much more political prominence and influenced the pronunciation norm of the modern standard Russian. In many cases such historic orthography principle makes the edit distance between cognates in different languages much shorter, and the phonological transcription in these cases may obscure innate morphological and historical links between closely related languages reflected in spelling. Therefore, using orthography to directly generate phonological feature representations has a theoretical motivation.

One specific issue in using the orthography-based phonological metric is dealing with digraphs – the two letter combinations denoting one sound (c.f., similarly, diphthongs need special treatment in the transcription-based metric), especially in cases when the two languages use different writing systems. This problem, however, is much smaller if the alphabets are similar or the same. On the other hand, treating historic digraphs as two separate letters with two feature sets may be beneficial in some cases, e.g., *Thomas* vs. *Хома (Homa)*, where the first letter of the Ukrainian word (*h*) is historically a closer match to one of the letters of the English digraph *th*.

In this paper the term *graphonological features* is used to refer to representations of phonological features that are directly derived from graphemes. The approach adopted in my experiment is that each orthographic character in each language is unambiguously associated with a set of phonological features, even though its pronunciation may be different in different positions.

2.2.2. Graphonological representations: feature vectors vs. feature hierarchies

Features in graphonological representations of characters can be organized in different ways. In my initial experiments the problems with structuring them as flat feature vectors became apparent. Even though in some examples there has been improvement in the rate

of cognate identification caused by richer feature structures, as compared to the baseline Levenshtein metric, in many more cases (and often counter to the earlier intuition) these feature structures caused unnecessary noise and lower ranking for true cognates, while non-cognates received smaller feature-based edit distance score. This unwanted overgeneration issue has been traced back to the use of feature vectors as graphonological feature structures.

The example (3) illustrates the reason for such overgeneration. If the feature vector representations are used, the proposed graphonological metric (GrPhFeatLev) calculates that the following edit distances should be the same, which is a counter-intuitive result (especially given that the traditional Levenshtein's metric (Lev) clearly shows that the character-based edit distance is shorter):

robitnyk (робітник) 'worker' (uk) & *rabotnik* (работник) 'worker' (ru)
GrPhFeatLev =1.2 Lev=2.0

---

robitnyk (робітник) 'worker'(uk) & *rovesnik* (ровесник) 'age-mate, of the same age' (ru) (3)
GrPhFeatLev =1.2 Lev=3.0

There is a specific problem when intuitively unrelated consonants (at least among Ukrainian-Russian lexical cognates) [b] and [v], or [t] and [s] – still receive very small rewriting scores. Figure 1 and Tables 1 and 2 show overlapping graphonological features for these words. In both cases, while one of the more essential features was not matched – *manner of articulation*, but instead the smaller edit distance resulted from matching less important features: [*active* and *passive* articulation organs] and [*voice*]. The problem with using feature vector representation is that all of the features stay on the same level, there is no way of indicating that certain features are more important for cognate formation and perception.

		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

Figure 1. GPhFeatLev Levenshtein: Edit distance matrix with *feature vectors* for *robitnyk* (робітник) 'worker'(uk) & *rovesnik* (ровесник) 'age-mate, of the same age' (ru)

b (б)	**['type:consonant'**, **'voice:voiced'**, 'maner:plosive', **'active:labial'**, 'passive:bilabial']
t (т)	**['type:consonant', 'voice:unvoiced'**, 'maner:plosive', **'active:fronttongue'**, **'passive:alveolar'**]

Table 1: Phonological feature vectors in Ukrainian word 'robitnyk' (робітник) – 'worker' overlapping features in intuitively unrelated characters are highlighted

v (в)	**['type:consonant', 'voice:voiced'**, 'maner:fricative', **'active:labial'**, 'passive:labiodental']
s (с)	**['type:consonant', 'voice:unvoiced'**, 'maner:fricative', **'active:fronttongue'**, **'passive:alveolar'**]

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

To address this problem, instead of feature vectors *hierarchical representations of features* are used, where a set of central features at the top of the hierarchy needs to be matched first, to allow lower level features to be matched as well (Figure 2).

Figure 2 shows that for the feature hierarchy of the grapheme [b] to match the hierarchy of the grapheme [v] there is a need to match first the grapheme type: consonant (which is successfully matched), and then – a combination of *manner* of articulation and *active* articulation organ (which is not matched, since [b] is plosive and [v] is fricative), and only after that – low level features such as voice may be tried (not matched again, because the higher level feature structure of *manner* + *active* did not match). Note that the proposed hierarchy applies to Ukrainian–Russian language pair, and generalizing it to other translation directions may not work, as relations may need rearrangements of the hierarchy to reflect specific graphonological relations between other languages.

Consonant feature hierarchy	**Example (pl- prefix on lower level features enforces feature hierarchy)**
Type {Manner+Active} Voice Passive	[b]: ['type:consonant', {'maner:**pl**-plosive', 'active:**pl**-labial',} 'voice:**pl**-voiced', 'passive:**pl**-bilabial'

Figure 2. Hierarchical feature representations for consonants: non-matching higher levels prevent from matching at the lower levels: [**pl**-voiced] will not match before [plosive, labial] match

2.2.3. Calculating combined substitution cost for variable length feature sets

As the number of features for different graphemes may vary, the edit distance is computed between partially matched feature sets as an F-measure between Precision and Recall of their potentially overlapping feature sets, and subtracting it from 1. As a result the measure is symmetric, (4):

$$Prec = len(FeatOverlap) \,/\, len(NofFeatA)$$

$$Rec = len(FeatOverlap) \,/\, len(NofFeatB)$$

$$OneMinusFMeasure = 1 - (2 * Prec * Rec) / (Prec + Rec)$$

$$\begin{aligned} matrix[zz+1][sz+1] \\ = min(matrix[zz+1][sz] + 1, matrix[zz][sz \\ + 1] + 1, matrix[zz][sz] \\ + OneMinusFMeasure) \end{aligned} \quad (4)$$

In these settings lower cost is given to substitutions; while insertion and deletions incur a relatively higher cost. As a result, cognates that have different length are much harder to find using the graphonological Levenshtein edit distance, and in these cases the baseline character-based Levenshtein metric performs better. A general observation is that the feature-based metric can often find cognates inaccessible to character-based metrics when the main differences are in *substitution*, but it misses cognates that involve more *insertions*, *deletions* and changing *order* of graphemes, as shown in Table 3.

uk	***ru***	***GPhFeatLev***	***Baseline Lev***
рішення rishennia 'decision'	решение resheniye 'decision'	**Found**	Missed
сьогодні s'ogodni 'today'	сегодня segodnia 'today'	**Found**	Missed
колгосп kolgosp 'collective farm'	колхоз kolhoz 'collective farm'	**Found**	Missed
коментар komentar 'commentary'	комментарий kommentariy 'commentary'	Missed	**Found**
перерва pererva 'break'	перерыв pereryv 'break'	Missed	**Found**

Table 3. Examples of missed and found cognates for each metric

2.3. Evaluation sample

Evaluation is performed for the baseline Levenshtein metric and the proposed feature-based metric with two settings: one using flat feature vectors for graphonological representations, and the other – using hierarchically organised features. Evaluation was done on a sample of 300 Ukrainian words selected from 6 frequency bands in the frequency dictionary of lemmas (ranks 1-50, 3001-3050, 6001-6050, 9001-9050, 12001-12050, 15001-15050), Russian cognates were searched in the full-length frequency dictionary of 16,000 entries automatically derived from the Russian corpus (as described in Section 2.1). For 274 out of the 300 Ukrainian words either the baseline Levenshtein metric, or the experimental feature metric returned Russian candidate cognates (with the threshold of

$$\frac{LevDist}{\max(len(W1), len(W2))} \leq 0.36$$

applied across all the metrics, as mentioned in Section 2.1. Different settings for modifications of Levenshtein edit distance can be systematically evaluated in this scenario by using human annotation of the candidate cognate lists.

3. Evaluation results

The 274 lists of cognate candidates provided by each metric were then labelled according to the following annotation scheme: Table 4:

Label	**Interpretation**
NC	No cognate: a word in source language (SL) does not have a cognate in the target language (TL)
0D	Zero difference: absolute cognates there is no difference in orthographic strings in the SL and TL
FF	'False friends' cognates with different meaning in the SL and TL
CL	Cognate **wins** in the *baseline* (string-based Levenshtein) – having a higher rank
CF	Cognate **wins** in the *tested* approach (feature-based Levenshtein)
WL	Cognate **looses** in the *baseline* (string-based Levenshtein)
WF	Cognate **looses** in the *tested* approach (feature-based Levenshtein)
ML	Cognate is **missed** by the *baseline* (string-based Levenshtein)
MF	Cognate is **missed** by the *tested* approach: (feature-based Levenshtein)

Table 4. Labels used for candidate cognate annotation

Counts of annotation labels for each of the categories are shown in Table 5 and Table 6.

	per cent	**count**
Have no cognates (*NC*)	34.31%	94
False Friends (*FF*)	1.82%	5
0 Difference cognates (*0D*)	16.42%	45
Cognates with +/– differences (existence, rank)	41.6%	114
All cognate candidates in sample	**100%**	**274**

Table 5. Parameters of evaluation sample

	Lev *(baseline character-based)*		***GPFeat Vectors*** *(feature-based flat vectors)*		***GPFeat Hierarchy*** *(feature-based hierarchical)*		***Difference****: GPFeatHierarchy - Lev*
	per cent	**#**	***per cent***	***#***	**per cent**	**#**	**per cent**
correct, higher is better: *CL* vs *CF* (+exclude 0 differences, *0D*)	47.08% (36.68%)	129 (84)	*46.72%*	*128*	**51.09%** **(41.48%)**	140 (95)	**+4.01%** **(+4.80%)**
present, but lost on rank (*WL* vs *WF*; lower better)	**2.19%**	6	*10.58%*	*29*	2.55%	7	-0.36%
cognates missing (*ML* vs *MF*; lower is better)	13.87%	38	*10.58%*	*29*	**9.85%**	27	**+4.02%**

Table 6. Comparative performance of distance measures for the task of ranking cognates

It can be seen from the tables that while the baseline Levenshtein metric (Table 6, column *Lev*) outperforms the feature-based metric that uses feature vector graphonological representations (column *GPFeat Vectors*), but the feature-based metric outperforms the baseline when hierarchical graphonological feature representations are used (column *GPFeat Hierarchy*). The improvement is about 4% (or nearly 5%, if trivial examples of absolute cognates are discounted). There is no improvement in ranking of found equivalents, which may be due to the noise related to a relatively higher cost of insertions, deletions and reordering of characters.

4. Conclusion and future work

Even though the traditional character-based Levenshtein metric gives a very strong baseline for the task of automated cognate identification from non-parallel corpora, the proposed graphonological Levenshtein edit distance measure outperforms it. Hierarchically structured feature representations, proposed in this paper, capture linguistically plausible correspondences between cognates much more accurately compared to traditionally used feature vectors. These representations are essential components of the proposed graphonological metric. Feature-based metric often identifies cognates which are missed by the baseline Levenshtein character-based metric.

Different settings of the metrics were compared under the proposed task-based evaluation framework, which requires a relatively small amount of human annotation and can calibrate further developments of the metric and refinements of the feature representation structures. This framework tests the metric directly for its usefulness for the task of creating cognate dictionaries for closely related languages.

For practical tasks both the traditional and feature-based Levenshtein metrics can be used in combination, supporting each other strengths, especially if boosting recall in the cognate identification task is needed.

Future work will include extending evaluation to other languages and larger evaluation sets, measuring improvements in MT systems enhanced with automatically extracted cognates, learning optimal feature representations and optimising feature weights for specific translation directions from data, extending character-based frameworks, such as (Beinborn et al., 2013). However, the graphonological Levenshtein distance metric may find applications beyond the task of cognate identification, e.g., for robust transliteration,

identification of spelling variations or distortions, for integrating feature-based representations into algorithms for learning phonological and morphosyntactic correspondences between closely related languages and into algorithms for automatically deriving morphological variation models for automated grammar induction tasks, with a goal of building large-scale morphosyntactic resources for MT.

Acknowledgements

I thank the reviewers of this paper for their insightful, detailed and useful comments.

Bibliography

Anderson, S. R. (1985). Phonology in the twentieth century: Theories of rules and theories of representations. University of Chicago Press.

Babych, B., Elliott, D., Hartley, A. (2004, August). Extending MT evaluation tools with translation complexity metrics. In Proceedings of the 20th international conference on Computational Linguistics (p. 106). Association for Computational Linguistics.

Babych, B., Hartley, A., Sharoff, S. (2007). Translating from under-resourced languages: comparing direct transfer against pivot translation. Proceedings of MT Summit XI, Copenhagen, Denmark.

Beinborn, L., Zesch, T., Gurevych, I. (2013). Cognate Production using Character-based Machine Translation. In IJCNLP (pp. 883-891).

Bergsma, S., Kondrak, G. (2007, September). Multilingual cognate identification using integer linear programming. In RANLP Workshop on Acquisition and Management of Multilingual Lexicons.

Chomsky, N., Halle, M. (1968). The sound pattern of English. Harper & Row Publishers: New York, London.

Ciobanu, A. M., Dinu, L. P. (2014). Automatic Detection of Cognates Using Orthographic Alignment. In ACL (2) (pp. 99-105).

Comrie, B. , Corbett, G., Eds. (1993). The Slavonic Languages. Routledge: London, New York.

Eberle, K., Geiß, J., Ginestí-Rosell, M., Babych, B., Hartley, A., Rapp, R., Sharoff, S & Thomas, M. (2012, April). Design of a hybrid high quality machine translation system. In Proceedings of the Joint Workshop on Exploiting Synergies between Information Retrieval and Machine Translation (ESIRMT) and Hybrid Approaches to Machine Translation (HyTra) (pp. 101-112). Association for Computational Linguistics.

Enright, J., Kondrak, G. (2007) A fast method for parallel document identification. Proceedings of Human Language Technologies: The Conference of the North American Chapter of the Association for Computational Linguistics companion volume, pp 29-32, Rochester, NY, April 2007.

Hana, J., Feldman, A., Brew, C., Amaral, L. (2006, April). Tagging Portuguese with a Spanish tagger using cognates. In Proceedings of the International Workshop on Cross-Language Knowledge Induction (pp. 33-40). Association for Computational Linguistics.

Hubey, M. (1999). Mathematical Foundations of Linguistics. Lincom Europa, Muenchen.

Jakobson, R., Fant, G., Halle, M. (1951). Preliminaries to speech analysis. The distinctive features and their correlates.

Koehler. R. (1993). Synergetic Linguistics. In: Contributions to Quantitative Linguistics, R. Koehler and B.B. Rieger (eds.), pp. 41-51.

Koehn, P., Knight, K. (2002). Learning a Translation Lexicon from Monolingual Corpora, , ACL 2002, Workshop on Unsupervised Lexical Acquisition

Ladefoged, P., Halle, M. (1988). Some major features of the International Phonetic Alphabet. Language, 64(3), 577-582.

Leusch, G., Ueffing, N., Ney, H. (2003, September). A novel string-to-string distance measure with applications to machine translation evaluation. In Proceedings of MT Summit IX (pp. 240-247).

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. Soviet Physics Doklady 10 (8): 707–710.

Menzerath, P. (1954). Die Architektonik des deutchen Wortschatzes. Dummler, Bonn.

Mulloni, A., Pekar, V. (2006). Automatic detection of orthographic cues for cognate recognition. Proceedings of LREC'06, 2387, 2390.

Nerbonne, J., Heeringa, W. (1997). Measuring dialect distance phonetically. In Proceedings of the Third Meeting of the ACL Special Interest Group in Computational Phonology (SIGPHON-97).

Nießen, S.; F. J. Och; G. Leusch, and H. Ney. (2000) An evaluation tool for machine translation: Fast evaluation for MT research. In Proc. Second Int. Conf. on Language Resources and Evaluation, pp. 39–45, Athens, Greece, May

Pinnis, M., Ion, R., Ştefănescu, D., Su, F., Skadiņa, I., Vasiļjevs, A., Babych, B. (2012) Toolkit for Multi-Level Alignment and Information Extraction from Comparable Corpora // Proceedings of ACL 2012, System Demonstrations Track, Jeju Island, Republic of Korea, 8-14 July 2012.

Pivtorak, H. P. (1988). Forming and dialectal differentiation of the old Ukrainian language. (Formuvannya i dialektna dyferentsiatsiya davn'orus'koyi movy – Формування і діалектна диференціація давньоруської мови). Naukova Dumka, Kyiv. (in Ukrainian).

Sanders, N. C., Chin, S. B. (2009). Phonological Distance Measures. Journal of Quantitative Linguistics, 16(1), 96-114.

Schepens, J., Dijkstra, T., Grootjen, F. (2012). Distributions of cognates in Europe as based on Levenshtein distance. Bilingualism: Language and Cognition, 15(01), 157-166.

Serva, M., Petroni, F. (2008). Indo-European languages tree by Levenshtein distance. EPL (Europhysics Letters), 81(6), 68005.

Sigurd, B., Eeg-Olofsson, M., Van Weijer, J. (2004). Word length, sentence length and frequency–Zipf revisited. Studia Linguistica, 58(1), 37-52.

Swadesh, M. (1952). Lexico-statistic dating of prehistoric ethnic contacts: with special reference to North American Indians and Eskimos. Proceedings of the American philosophical society, 96(4), 452-463.

Zipf, G. K. (1935). The psycho-biology of language.

Received May 3, 2016, accepted May 4, 2016