Phonological models for cognate terminology identification

First Author
Affiliation / Address line 1
Affiliation / Address line 2
Affiliation / Address line 3
e-mail@domain

Second Author
Affiliation / Address line 1
Affiliation / Address line 2
Affiliation / Address line 3
e-mail@domain

Abstract

The paper presents a framework for the development and task-based evaluation of phonological linguistic models, which improve the accuracy of identifying cognate terminology, contributing to automated development of large term banks. Term translation remains a bottleneck even for Neural MT, especially for less-resourced languages and domains, so automated development of terminological resources may be useful for addressing this problem. Proposed phonological models can be applied to languages with predominantly phonemic orthography – the majority of world languages, including English, where graphemes (characters) roughly correspond to phonemes (intentionally pronounced sounds). Our models explicitly represent distinctive phonological features, such as the acoustic type (vowels, voiced and voiceless consonants, sonants), place and manner of articulation (closed/open, front/back vowels; plosive, fricative, or labial, dental, glottal, etc. consonants). The advantage of such representations is that they explicate information about characters' internal structure rather than treat them as elementary atomic units of comparison, placing graphemes into a feature space that provides additional information about their articulatory (pronunciation-based) or acoustic (sound-based) distances and similarity. The article presents experimental results of using our phonological models for extracting cognate terminology with the phonologically aware Levenshtein edit distance, which outperforms the baseline character-based distance measure. Project tools are released on[1]: https://github.com/qumtie/cognates-phonology

1 Introduction

This paper presents a methodology for the development and automated evaluation of a linguistic features set that extends a traditional Levenshtein edit distance metric used for the task of cognate terminology identification.

1.1. Identification of cognates and terminology

Cognate identification is important for a range of applications; in our experimental settings we use it for assisting MT developers in creating cognate term banks for computer-assisted translation, stand-alone dictionaries and for rule-based and hybrid MT systems between closely related languages, some of these languages are less-resourced (e.g., Ukrainian (Uk) and Russian (Ru), Portuguese (Pt) and Spanish (Es), Dutch (Nl) and German (De)). For many of such language pairs there are no electronic dictionaries, and there are only small parallel corpora available with limited lexical coverage. Typically these parallel corpora can provide translations for frequently used general words, but miss the 'long tail' of less frequent, often topic-specific or terminological words. However, in closely related languages these words are often cognates, which creates a possibility to rapidly extend bilingual lexicons in semi-automated way using non-parallel, comparable corpora and automated cognate identification techniques. In this task, cog-

[1] Our anonymised Github repository is live and the resources can be downloaded and evaluated in the reviewing stage: the code and the repository doesn't reveal the identity or affiliation of the authors.

nate candidates are generated from word lists created from large monolingual comparable corpora in both languages The assumption is that the developers have good linguistic intuition of both languages and work through lists of cognate candidates, checking which pairs can be added to the bilingual dictionary. Their productivity depends on whether cognates are presented high up in the list of candidates, ideally at the top, or at least in the top N items, where N lines should at most fit on a single screen.

Other uses of cognates for terminology identification include term extraction from parallel corpora. If boundaries of multiword source terms are known, some component words may not be necessarily cognate with the target. In this case heuristics can be applied to extend term alignments, e.g., adding an adjacent word of a cognate term, according to part-of-speech and word order patterns, e.g., En: *'**information** requirements'* ~ Uk: *'**інформаційні** потреби' ('**informatsijni** potreby')*; or splitting and extending compound terms which have cognate parts, e.g., En: 'multi-**national**' ~ Uk: 'багато**національний**'-('bahato**natsionalnyj**').

Yet another application of cognate identification is sentence alignment of parallel corpora, where statistical alignment methods are more accurate if cognates are used as an additional data source (Lamraoui and Langlais, 2013:2). Inaccuracies in cognate identification, which are due to orthographic differences, often create unnecessary bottlenecks for this task (Varga et al., 2015: 249). In this scenario identified cognates are not necessarily terms, but they contribute to a more accurate alignment and extraction of non-cognate terminology, produced from word alignment and monolingual terminology detection.

1.2. Cognate identification and transliteration

An additional complication for the multilingual terminology extraction scenarios that rely on cognate identification is the use of different writing systems in the source and target (e.g., Cyrillic or Georgian vs. Latin script), which requires transliteration between those languages.

Transliteration is often non-trivial, because of differences in pronounciation of the same letters, the lack of direct graphemic equivalents across languages, contextual dependencies in transliteration rules, different historical conventions for different words (e.g., En/De "h" → Ru "x" (*hockey ~ хоккей,* since borrowed directly from En), or "г" (*hermeneutics ~ герменевтика*, since borrowed via Ukrainian, where En: h → Uk: г [ɣ] → Ru: г [g]). Also, even if languages use the same alphabet, pronounciation of letters and corresponding transliteration rules may differ (e.g., Cyrillic letter "и" = [i] in Ru and [y] in Uk, Latin letter "g" = [g] in En/De, and [ɣ] in Nl), so new transliteration mappings need to be created for each translation direction, each with their potential language-specific problems.

As a result, the complexity of transliteration in some cases is comparable to the complexity of MT, and it is often addressed not via simple character mappings, but via fully developed character-based MT models that require an aligned training corpus for each translation direction, and which are used in MT applications to cover out-of-vocabulary words, such as compounds, morphologically complex words, named entities and cognate terminology (Senrich et al., 2016: 1716)

Transliteration problem resembles a traditional "direct translation" bottleneck in MT: this approach cannot reuse any of the previously created mappings between languages if a new language pair or translation direction need to be covered. A more principled approach to the transliteration problem developed in this paper is mapping characters for each language into a language-independent ("interlingual") phonological feature space.

1.3. Cognates and Levenshtein edit distance

The Levenshtein edit distance (Levenshtein, 1966) is typically used to compare word pairs from different languages and determine if they are cognate candidates. For example, if cognate candidates are extracted from a non-aligned or non-parallel corpus, the Levenshtein distance is computed for every pair of words in the two word lists created for each language (Cartesian product of the lists), the search space may be restricted to comparing words with the same part-of-speech (PoS) codes, if PoS annotation is available for the corpus.

However, there are several problems with the traditional Levenshtein metric, one of which is that all characters in comparison are treated as atomic units that do not have any internal structure and therefore, can be substituted only as a whole character. Because of this the Levenshtein metric does not distinguish between the substitutions of characters that correspond to acoustically/articulatory similar sounds vs. the substitution

of phonologically distant letters. As a result, words that are intuitively close may receive a large distance score, e.g.,

Uk "*жовтий*" (*zhovtyj*)='yellow'
Ru "*жёлтый*" (*zheltyj*) = 'yellow'
(Lev distance = 3),

where, for historical reasons, articulatory similar sounds are represented by different characters: the sound [o] – by 'o' in Uk and 'ë' in Ru, the sound [y] – by 'и' in Uk and 'ы' in Ru. On the other hand, words that are not cognates and are phonologically and intuitively far apart, still receive the same distance scores, such as:

Uk "*жовтий*" (*zhovtyi*) = 'yellow' and
Ru "*жутκий*" (*zhutkiy*) = 'dismal' (Lev = 3).

For example, here no distinction is made between, on the one hand, the substitution "*o*" (o) → "*ë*" ([i]o) of phonologically similar sounds (which differ only in a peripheral feature – triggering palatalization of the preceding consonant (Uk: -- Ru: +; in addition, this feature is neutralised after the sibilant "*ж*" (zh)), and on the other hand – the substitution "*o*" (o) → "*y*" (u), where sounds differ in core articulatory features of the place of vowel articulation (Uk: middle; Ru: close/high).

Some existing modifications and extensions of the Levenshtein metric introduce weightings for different character mapping, but these weights need to be set or empirically determined for each specific mapping: compared characters still do not have internal structure and there is no way to predict the weights in advance for any possible pair in a principled way.

1.4. Phonological models for cognates

In this paper we present an automated task-based evaluation framework for an extension to the Levenshtein edit distance metric, which explicitly represents linguistic phonological features of compared characters, so the metric can use information about characters' internal feature structure rather than treat them as elementary atomic units of comparison. Similar sets of distinctive features have been used for comparing transcriptions of spoken words in modeling dialectological variation and historical changes in languages (Nerbonne and Heeringa, 1997). In our project, phonological feature representations are applied to cognate identification and terminology extraction tasks, transliteration, and as well as modeling morphological variation. Previously we have shown (Anonymous, 2016) that there are multiple ways of identifying, representing, structurally arranging and comparing these features in a phonological feature space, so there is a need for a methodology for evaluating alternative feature configurations. The results of our pilot experiment, using a small-scale manual evaluation, indicated the need to use hierarchical phonological feature structures for consonants rather than flat feature vectors previously used in dialectological research.

However, the manual evaluation framework cannot be used for systematic testing and optimization of the metric parameters and weights, or for systematically comparing alternative phonological feature structures.

In this paper we present a new automated framework for evaluating different arrangements of phonological features using the task of cognate identification. Apart from practical applications mentioned above, this task can be used for feature engineering for character-based model in a wider range of machine learning methods and tools, since it now uses standard automatically computed evaluation metrics, e.g., a cognate in top-N candidates. Evaluation is performed on a larger data set generated on a high performance computing cluster. Our automated task-based evaluation methodology allows us to design and calibrate feature structures in a systematic way. Our evaluation results show greater improvement for hierarchical feature structures on a number of automatically computed parameters, and also allow us to optimize parameters of the phonologically aware Levenshtein metric, such as insertion/deletion weights, and to rule out some unproductive modifications.

2 Phonological distinctive features and their application for cognate identification

A theory of phonological distinctive features, which was first proposed by Roman Jacobson (Jakobson and Halle, 1956: 46; Anderson, 1995: 116), associates each phoneme (an elementary segmental unit of speech which can distinguish meanings) with a unique set of values for categories, which may apply to larger classes of sounds. For example, the phoneme [t] has the following values for the categories:

'type': consonant
'voice': unvoiced
'maner of articulation': plosive;
'active articulation organ': front of the tongue
'passive articulation organ': alveolar

Phoneme [d] has the same articulation, but is pronounced with the use of vocal cords, so it differs only in the value of one distinctive feature,

'voice': voiced,

with all other categories and values remaining the same.

In historical development of languages and in morphological variation within a language the sound changes more often apply only to values of certain distinctive features within characters, but not to the whole category-value system, e.g.: Ge "*Tag*" = Nl "*dag*" ('day); Ge: "*machen*" = Nl "*maken*" ('make'). Therefore, in languages where the writing system is at least partially motivated by pronunciation, it would be useful to represent the phonological distinctive features, in order to differentiate between varying degrees of closeness for different classes of characters, e.g., vowels, sonants and consonants, or sounds with identical or similar articulation. Greater closeness between characters in terms of their phonological features has important linguistic and technical applications, such as modelling dialectal variation, historical change, morphological and derivational changes in words, such as stem alternations in inflected forms.

Phonological distinctive features have been integrated into the Levenshtein distance metric in the following way (c.f. Anonymous, 2016: 123):

(1) Substitution cost for two characters is calculated as 1 [minus] F-measure between Precision and Recall of their feature overlap; this allows calculating the cost for character with different numbers of features, and the metric remains symmetric. Intuitively, to substitute [t] with [d] in "Tag" → "dag" we need to re-write only one feature out of 5, so the cost is 0.2 rather than 1.

(2) The order of matching the distinctive features was found to be important. Section 3 describes an experiment on comparing two different arrangements of features: as flat feature vectors and as feature hierarchies, where matching lower level features is a pre-condition for attempting to match lower level features. Hierarchical organization achieved better performance compared to traditional flat feature vectors. Intuitively this means that not all feature categories should be treated equally, some of them are more central, have higher priority and license comparison of lower level features, which form the periphery of the feature system.

(3) Insertion and deletion costs have been calibrated for the range between 0.2 and 1 using the proposed evaluation framework, described in this paper in Section 3. Optimal performance on cognate identification was achieved for cost of insertion = deletion = 0.8.

For the task of cognate identification, the introduction of these features distinguishes different types of character substitutions and gives more accurate prediction of the degree of closeness between compared characters and words, e.g., for the word pairs discussed above, where the baseline Levenshtein distance =3 for both (matching features, which do not need to be rewritten, are highlighted in bold):

Graphemic-Phonological (graphonological) feature Uk "*жовтий*" (*zhovtyj*) = 'yellow'

ж (zh) 'type:consonant', 'voice:ff-voiced', 'maner:ff-fricative', 'active:ff-fronttongue', 'passive:ff-palatal'
о (o) **'type:vowel'**, **'backness:back',** **'height:mid'**, **'roundedness:rounded'**, 'palate:nonpalatalizing'
в (v) **'type:consonant'**, 'voice:fl-voiced', 'maner:fl-fricative', 'active:fl-labial', 'passive:fl-bilabial'
т (t) 'type:consonant', 'voice:pf-unvoiced', 'maner:pf-plosive', 'active:pf-fronttongue', 'passive:pf-alveolar'
и (y) **'type:vowel'**, 'backness:front', **'height:closemid'**,**'roundedness:unrounded'**, **'palate:nonpalatalizing'**
й (j) 'type:consonant', 'voice:xm-sonorant', 'maner:xm-approximant','active:xm-midtongue', 'passive:am-palatal'

Feature representations for corresponding Ru characters in "*жёлтый*" (*zheltyj*) = 'yellow'.

ё (io) **'type:vowel', 'backness:back', 'height:mid', 'roundedness:rounded'**, 'palate:palatalizing'
л (l) **'type:consonant'**, 'voice:lf-sonorant', ' maner:lf-lateral', 'active:lf-fronttongue', 'passive:lf-alveolar'
…
ы (y) **'type:vowel'**, 'backness:central', **'height:closemid','roundedness:unrounded', 'palate:nonpalatalizing'**

Note that 'substitution in the graphonological metric only involves replacement of some distinctive feature in a character feature set, turning it into the target character. In this cases, the substitution cost is < 1 and corresponds to the proportion of features which needs to be rewritten.

Calculation of the Graphonological Levenshtein metric for Uk "*жовтий*" (*zhovtyj*) = 'yellow' and (Ru) "жёлтый" (zheltyi) = 'yellow':

0.0	1.0	2.0	3.0	4.0	5.0	6.0
1.0	**0.0**	1.0	2.0	3.0	4.0	5.0
2.0	1.0	**0.2**	1.2	2.2	3.2	4.2
3.0	2.0	1.2	**1.0**	2.0	3.0	4.0
4.0	3.0	2.2	2.0	**1.0**	2.0	3.0
5.0	4.0	3.2	3.0	2.0	**1.2**	2.2
6.0	5.0	4.2	4.0	3.0	2.2	**1.2**

cf.: Metric calculated for Uk "*жовтий*" (*zhovtyj*) = 'yellow' with Ru "*жутκий*" (*zhutkij*) 'dismal':

0.0	1.0	2.0	3.0	4.0	5.0	6.0
1.0	**0.0**	1.0	2.0	3.0	4.0	5.0
2.0	1.0	**0.2**	1.2	2.2	3.2	4.2
3.0	2.0	1.2	**1.0**	1.2	2.2	3.2
4.0	3.0	2.2	2.0	**1.8**	2.2	3.0
5.0	4.0	3.2	3.0	2.8	**2.0**	3.0
6.0	5.0	4.2	4.0	3.8	3.0	**2.0**

While the baseline Levenshtein distance Lev=2 for both pairs, the phonologically-aware distance, GLev = 2.0 for non-cognates, which is > 1.2 for cognates.

An additional advantage of using of phonological feature representations for graphemes is a more natural "interlingual" transliteration between different scripts and languages. The phonological models, presented in this paper, map characters from any given language into a universal space of acoustic and articulatory phonological features, which is independent of any specific writing system or a language-pair. This space can be seen as a phonological "interlingua", which shares some advantages with the idea of interlingual MT: graphonological mappings enable implicit cross-lingual transliteration, where mappings from individual languages into the common phonological feature space can be reused when new translation directions are added.

3 Set-up of the experiment and results

The proposed methodology is an automated performance-based evaluation for testing different settings of phonological categories and values for the extended Levenshtein metric. The experiment is set up in the following way:

(1) A small Ukrainian – Russian and Russian-Ukrainian dictionary was used to compile a gold-standard translation glossary of 11k Ukrainian words, each having one or more Russian translation equivalents.

(2) Translation equivalents in the glossary have been crosschecked with monolingual Russian and Ukrainian word lists compiled from PoS-tagged corpora for these languages. Only those entries in the gold standard were retained which exist in both words lists, so the cognate identification programme can in principle find them.

(3) An additional requirement has been introduced that in both word lists the cognates should be tagged with the same part-of-speech. This reduced the search space for cognates and reduced computing time needed to calculate distances for the entries in the Russian word list. Frequency information was not used, since the monolingual word lists from both languages came from diverse corpora of different sizes. Still calculation time is a limiting factor in evaluation, so 809 entries from the gold standard were randomly chosen for computing different types of Levenshtein edit distance.

(4) For each Ukrainian entry in the gold standard, the top-N list has been calculated from the Russian word list for each of available translation equivalents.

(5) Importantly, as the evolution methodology is automatic, all translation equivalents available in the gold standard are treated equally: in this stage no distinction is made between cognates and non-cognate equivalents. This removes the need for the manual filtering of the gold standard and also naturally covers 'near-cognates' or words with cognate morphemes where only parts of words match. Since the baseline and the modified Levenshtein metric are evaluated on the same gold standard, performance figures are relative and show the difference in finding translation equivalents for any degree of 'cognateness'.

(6) Different methods are compared by the following parameters: Median top-N number for the metric; In top-1, top-5, top-10 and top-25.

(7) The following settings were compared:

(a) Baseline Levenshtein edit distance;

(b) Levenshtein distance extended with phonological features with flat feature vectors;

(c) Levenshtein distance extended with hierarchical phonological features (where manner and active place of articulation are treated as top-level features, which need to be matched in order for other features to match;

(d) Variants of the (b) and (c) metric with different insertion / deletion values – between 0.2 and 0.8.

The results of the evaluation experiment are presented in Table 1, where:

BaseL Lev = baseline Levenshtein metric

Phon Lev H = Phonological extension to Levenshtein metric with feature hierarchy

Phon Lev V = Phonological extension to Levenshtein metric with flat feature vectors

PhonLevi=0.X = Phonological extension to Levenshtein metric with modified insertion / deletion cost: i0.2 = the cost of insertion deletion is set to 0.2, i0.8 = is set to 0.8 (it is set to 1 in the Phon Lev metrics.

4 Discussion of the results, conclusion

It can be seen from Table 1 that: (1) Hierarchical phonological Levenshtein metric outperforms the baseline on the Top 1 and Top 2 measures, the median rank improvements is +5%

(2) Flat phonological feature vector metric on all measures performs worse than the baseline. This can be interpreted as the need to take into account the order of matching higher-level features. Match of low-level features is not meaningful if higher-level features are not matched.

(3) The metric with insertion / deletion cost set to 0.8 outperforms both the baseline and the Hierarchical phonological Levenshtein metric, especially on the Median Top N, Top 1 and Top5 measures. This can be interpreted as the need to scale down insertion cost moderately, since the average substitution cost is down.

The results show that phonological extension to the Levenshtein edit distance metric on the task of cognate identification outperforms the character-based baseline. The proposed framework also allows accurate calibration of the feature arrangement and other parameter settings of the metric.

The modified Levenshtein metrics, phonological features sets for several alphabets and sample input files are released as an open-source software on our (anonymised) github repository (Qumitie, 2018: https://github.com/qumtie/cognates-phonology).

Future work will include systematic evaluation of different possible feature hierarchies and costs, and metrics application to other tasks, such as transliteration.

Experiment	**Median topN**	**Top 1**	**Top 5**	**Top 10**	**Top 25**
BaseL Lev	50	206	328	360	382
Phon Lev V	87.5	**215**	289	319	349
DiffBase L	-75%	+4.4%	-10%	-11%	-9%
PhLev Hierarchy:					
PhLev i=0.2	125.5	216	291	315	342
PhLev i=0.4	54.5	230	307	334	367
PhLev i=0.6	*48*	235	328	354	385
PhLev i=0.8	**40**	**240**	**337**	**359**	**391**
Ph Lev i=1.0	47.5	**240**	334	359	385
Best BaseL Improv	**+20%**	**+16.5%**	**+3%**	**0%**	**2%**

Table 1: Automated evaluation of metric settings.

References

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

Anonymous, Author. 2016. Paper anonymized.

Jakobson, Roman, and Moris Halle. 1956. *Fundamentals of language.* Vol. 1. Walter de Gruyter. URL: http://pubman.mpdl.mpg.de/pubman/item/escidoc:2350620/component/escidoc:2350619/Jakobson_Halle_1956_fundamentals.pdf

Lamraoui, Fethi, and Philippe Langlais. 2013. Yet another fast, robust and open source sentence aligner. time to reconsider sentence alignment. *XIV Machine Translation Summit*. URL: http://rali.iro.umontreal.ca/rali/sites/default/files/publis/MTSummit-2013-Fethi.pdf

Levenshtein, Vladimir I. 1966. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet physics doklady*. Vol. 10. No. 8.

Nerbonne, John, and Wilbert Heeringa. 1997. Measuring dialect distance phonetically. *Computational Phonology: Third Meeting of the ACL Special Interest Group in Computational Phonology.*

Qumtie, Systems (Anonymized). 2017. Phonological models for cognate terminology identification. GitHub repository, https://github.com/qumtie/cognates-phonology

Sennrich, Rico, Barry Haddow, and Alexandra Birch. 2016. Neural Machine Translation of Rare Words with Subword Units. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics* (Volume 1: Long Papers). Vol. 1. URL: http://www.aclweb.org/anthology/P16-1162

Varga, Dániel, Peter Hal´acsy, Andras Kornai, Viktor Nagy, Laszl´o N´emeth and Viktor Tron. 2007. Parallel corpora for medium density languages. *Amsterdam Studies In The Theory And History Of Linguistic Science Series 4 292*: pp. 247-253. URL: http://eprints.sztaki.hu/7902/1/Kornai_1762382_ny.pdf