Final submission – removed and rewritten fragments:

06/05/2018

Proposed models work for languages with phonemic orthography, where characters roughly correspond to phonemes (units of pronunciation).

. In this case heuristics can be applied to extend term alignments, e.g., adding an

This section discusses applications of automated cognate identification for MT-related tasks, such as terminology extraction and transliteration.

**1.1. Identification of cognate terminology**

**1.2. Cognate identification and transliteration**

**1.3. Cognates and Levenshtein edit distance**

can contribute

are needed be used to address it

so higher accuracy of automated terminology identification and its proper integration into MT

These resources can improve term translation in MT and computer-assisted translation (CAT), which can improve the accuracy of

for using cognates as a resource

together with can then be extracted from sentence-aligned corpora using statistical word alignment and term extraction methods.

adopted in this is mapping a character set for a given language into language-independent articulatory and acoustic features, that

No reuse, universal space :: phonology

Which resembles a phonological interlingua

This creates

Different languages which Also, transliteration needs to be done

For all these tasks, a reliable, language-independent and script-independent methods of cognate identification are needed.

via statistical word alignment from sentence-aligned parallel corpora

Non-cognate terms can be identified via statistical word alignment, which, in turn, may be more accurate if cognates are supplied to

successfully identified and fed into alignment tools.

Useful first step...

These methods rely on the core.

for selection of optimal feature arrangements

does not allo

That experiment also showed the need for a more systematic automated evaluation of alternative arrangements and weights for phonological features.

// the problem

// solution

that maps pronounciation of characters in different alphabets and languages into a universal feature space.

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: <https://github.com/qumtie/cognates-phonology>

// understand exact amount of the

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.

greater performance of the phonologically aware Levenshtein metric using hierarchical feature structures, determine an optimal value of 0.8 for insertion/deletion distance for the metric and such as flat feature vectors.

In the next stage frequency lists of lemmas have been produced from these

Only the entries found both the dictionaries and in the gold standard were retained which exist in both words lists, so the cognate identification programme can in principle find them.

Frequency information was not used, since the monolingual word lists from both languages came from diverse corpora of different sizes.

This is because

is much faster for the baseline Levenshtein edit distance, as it involves only

/

/ cognates

// non-cognate equivalents would not be picked up….

in the following way

for each Ukrainian entry in the evaluation set

to all the entries from the Russian word list.

in the evaluation set evaluation entries in the Russian word list.

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.

translation Each ranked candidate cognate list was compared with the translation equivalents in the gold standard.

(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.