On the Robustness of Cognate Generation Models

Winston Wu David Yarowsky
Center for Language and Speech Processing, Johns Hopkins University
wswu.github.io

Introduction

Cognate generation is the task of translating a word into an etymological relative, for example

English *tomato* → Finnish *tomaatti*

Our paper investigates the following research question: **How robust are cognate generation systems to noisy input?**

Scenario

A speaker of a low-resource language is searching the web. To provide a larger set of results, the search engine performs query expansion by translating cognates from a low-resource language to a high-resource language.

Robustness

There is lots of existing work in many areas of NLP, examining models' robustness to noisy inputs and adversarial attacks. However, these perturbations are largely at the *word level*. Cognate generation is done at the character level.

Human-Plausible Noise Models

Noise	Error Type	Example
Deletion	Typing	*hispan*
Duplication	Typing	*hiispana*
Swap	Typing	*hisapna*
Keyboard	Typing	*hispanz*
Phonological Substitution	Cognitive	*hisbana*

Phonological substitutions are generated by randomly selecting one of the top 3 character alignments computed over cognate pairs.

Cognate Generation Models

We experiment with two popular cognate generation models: LSTM encoder-decoder with attention (Bahdanau+ 2014) and Transformer (Vaswani+ 2017). We train a single large multilingual setup with the following input/output format:

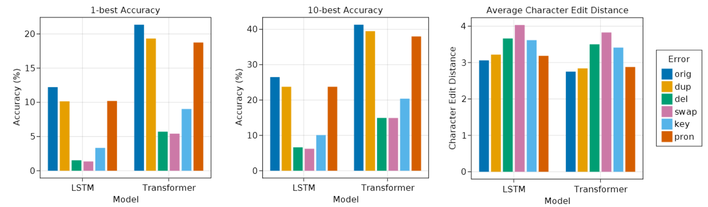
Input: `eng afr l i n e a r`
Output: `l i n e ê r`

Data

We use CogNet 2.0 (Batsurien+ 2022), a database of cognate pairs in 338 languages. We split this into 80-10-10 train-dev-test splits.

Results

Metrics: accuracy and average character edit distance



Analysis and Takeaways

We find that certain types of errors (deletions, swaps, keyboard errors) are more harmful than others. Why? They remove and distort phonological information necessary for translating cognates.

Perturbing vowels are less harmful than perturbing consonants.

Cognates are a natural source of noisy data that enable models to be robust to misspellings.

Much more in-depth analysis and examples in the paper!

LREC 2022