Cognition-aware Cognate Detection

Diptesh Kanojia†, Prashant K. Sharma○, Sayali Ghodekar‡, Pushpak Bhattacharyya†, Gholamreza Haffari*, & Malhar Kulkarni†

†University of Surrey, United Kingdom; ‡IITB-Monash Research Academy, India; ○IIT Bombay, India; *Monash University, Australia; ●Hitachi CRL, Japan; †RingCentral, India

{dipresh,pb,malhar}@iitb.ac.in, †prashaantsharma@gmail.com, ‡sayalighodekar26@gmail.com, ●gholamreza.haffari@monash.edu

Introduction

• Cognates are word pairs, across languages, having a common etymological origin. For example, the French and English word pair, Liberté - Liberty, reveals itself to be a cognate through orthographic similarity.
• Automatic Cognate Detection (ACD) is a well-known task, explored for many languages; and has shown to help NLP sub-tasks of Cross-lingual Information Retrieval, Machine Translation (MT), and Phylogenetics.
• Cognitive features have also shown to improve various NLP tasks (Mishra et. al., 2016)
• We hypothesize that gaze behaviour data from human participants can improve the performance of the cognate detection task with cognitive features.
• Gaze features like fixation duration, fixation counts, & saccades, help provide important insights into how humans disambiguate cognates vs. non-cognates.

Dataset Statistics

<table>
<thead>
<tr>
<th></th>
<th>Cognates (1)</th>
<th>False Friends (0)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kanojia et. al. (2020)</td>
<td>15726</td>
<td>5826</td>
</tr>
<tr>
<td>D1</td>
<td>5826</td>
<td>5826</td>
</tr>
<tr>
<td>D2</td>
<td>100</td>
<td>100</td>
</tr>
</tbody>
</table>

We extract 100 pairs, at random, from each of the positive and negative labels for collecting gaze behaviour data, to construct what we call “D2”.

Key Questions

• “Can cognitive features be used to help the task of Cognate Detection?”
• “Using gaze features collected on a small set of data points, can we predict the same features on a larger set of data points to alleviate the need for collecting gaze data?”
• “Can cognitive features be used to help the task of Cognate Detection?”
• “Using gaze features collected on a small set of data points, can we predict the same features on a larger set of data points to alleviate the need for collecting gaze data?”

Motivation

Consider a scenario where an NLP task comes across a false friend pair: For e.g., the word “shikhsha” in Hindi and Marathi.
• False friends are similarly spelt words that have distinct, unrelated meanings.
• Good quality cross-lingual models need data, and Hindi and Marathi are scarce. Hence, we obtain gaze behaviour data over a small dataset of cognates & false-friends.

Gaze Behaviour Analysis

• Gaze data is collected with the help of nine native Marathi speakers, who can understand Hindi.
• The precision of similarity annotation lies between 98% to 99.5% for individual annotators.
• Out of the 1800 annotations (9 annotators/200 word-pairs), only 40 incorrect annotations.
• We observe statistically significant fixation duration amongst all participants (cognates fixated for 1.3 times more than false-friends.)

Observations

• On D1+D2, using the XLM-based features, we observe an improvement of 9% over the stronger baseline and 13% over the system by Rama et. al.
• It can be seen that MUSE and VecMap based features also perform better on the combined dataset. In terms of both precision and recall, cross-lingual features are shown to outperform all the baseline systems.
• Appending gaze features to our best reported system help our model outperform it by 3%.
• Cognate pair “uPann” (Hindi) - “uPAnDTo” (Marathi) (both meaning manufactured) is classified correctly by this system, but incorrectly by baselines, and cross-lingual systems.
• We were hopeful that the participants would focus only on important contextual clues and not the stop words. However, the sample points are not enough to concretely discuss this aspect.

Conclusion

• We harness cross-lingual embeddings and gaze-based features to help the cognate detection task, for the Indian languages, Hindi & Marathi.
• To answer our key questions, “Yes,” & “Yes!”.

For additional results, see our paper at: Paper Link

Dataset & Code Repository

https://www.cfilt.iitb.ac.in/eacl2021dipresh