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

[†]{diptesh,pb,malhar}@iitb.ac.in, [◇]prashaantsharmaa@gmail.com [‡]sayalighodekar26@gmail.com , [⋆]gholamreza.haffari@monash.edu

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?"

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* cognate vs. non-cognates.**

Dataset Statistics

	Cognates (1)	False Friends (0)
Kanojia et. al. (2020)	15726	5826
D1	**5826**	**5826**
D2	**100**	**100**

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".

Motivation

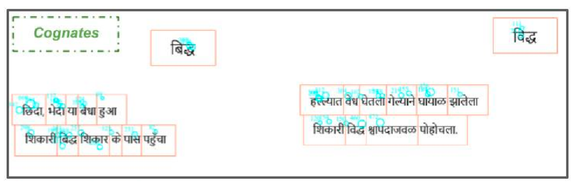
Consider a scenario where an NLP task comes across a *false friend pair*; For *e.g.,* the word "*shikhshA*" in Hindi and Marathi.

- *False friends are similarly spelt words that have distinct, unrelated meanings.*
- Good quality cross-lingual models need data, and **Hindi** and **Marathi** are data scarce.
- Hence, we obtain **gaze behaviour data** over a small dataset of cognates & false-friends.

Results

	P	R	F	P	R	F	P	R	F	P	R	F
Feature Set →	Phonetic			WLS								
Rama et. al., 2016 (D1+D2)	0.71	0.69	0.70	-	-	-						
Kanojia et. al., 2019 (D1+D2)	-	-	-	0.76	0.72	0.74						
Feature Set →	XLM			MUSE			VecMap					
Linear SVM (D1+D2)	0.83	0.71	0.77	0.72	0.68	0.70	0.70	0.65	0.67			
LogisticRegression (D1+D2)	0.85	0.74	0.79	0.80	0.71	0.75	0.70	0.66	0.68			
FFNN (D1 + D2)	0.82	0.84	**0.83**	0.83	0.79	0.81	0.75	0.76	0.75			
Feature Set →	XLM+Gaze			MUSE+Gaze			VecMap+Gaze			Gaze		
Linear SVM (D2)	0.81	0.69	0.75	0.72	0.73	0.72	0.70	0.75	0.72	0.77	0.76	
LogisticRegression (D2)	0.84	0.75	0.79	0.76	0.72	0.74	0.81	0.71	0.76	0.80	0.75	
FFNN (D2)	0.83	0.85	**0.84**	0.83	0.78	0.80	0.86	0.83	0.84	0.81	0.71	
Predicted Gaze Features On D1 (11652 samples) and Collected Gaze Features on D2 (200 samples)												
Feature Set →	XLM+Gaze			MUSE+Gaze			VecMap+Gaze			Gaze		
FFNN (D1 + D2)	0.84	0.88	**0.86**	0.85	0.78	0.81	0.83	0.85	0.84	0.77	0.76	
FFNN (D1) [Only Predicted Gaze]	0.83	0.84	0.83	0.82	0.79	0.80	0.80	0.86	0.83	0.76	0.77	

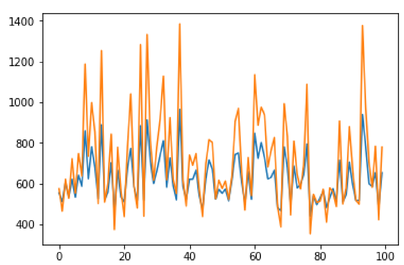
Data Annotation Screen



Gaze Behaviour Analysis

- Gaze data is collected with the help of nine native Marathi speakers, who can understand Hindi.
- The pecision 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.)

Predicting Gaze Fixation



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 "*uTpann*" (Hindi) - "*uTpADiTa*" (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!"**.

Full Paper

For additional results, see our paper at: Paper Link

Dataset & Code Repository

https://www.cfilt.iitb.ac.in/eacl2021diptesh

भारतीय भाषा प्रौद्योगिकी केन्द्र

IITB-Monash Research Academy
An Indian-Australian research partnership

Indian Institute of Technology Bombay | MONASH University