Prediction of Cognate Reflexes

Ishana Shinde, Kavya Sudha Kollu, Deeksha Gangadharan Srinivas, Varshaa Shree Bhuvanendar

{ishinde,kkollu,dgangadh,vshree}@gmu.edu

Highlights

The SIGTYP 2022 Shared Task on the Prediction of Cognate Reflexes

Mockingbird at the SIGTYP 2022 Shared Task: Two Types of Models for the Prediction of Cognate Reflexes

Highlights

- **AIM: To determine unknown cognate values based on similar cognate values in various cognate set using NLP models.**
- What is a cognate?
  - A word having the same linguistic derivation as another from the same original word or root
  - Example:

Cognate Set	German	English	Dutch
ASH	a ʃ ɛ	æʃ	ɑ s
BITE	b ai s ə n	b ai t	b ɛi t ə
BELLY	b au x	-	b œi k

Why is this task required?

- To anticipate the pronunciation of words in one language based on the pronunciation of cognate terms in related languages without the systematicity and regularity of sound change.
- To emphasize the value of classical research for computational applications

MockingBird Inpainting

- The goal of restoring corrupted parts of a 2D image is contrasted with the cognate reflex prediction task in this model. The dimensions of the 2D image correspond to languages and cognate phonemic representations. Convolutional neural networks are used to achieve the restoration.

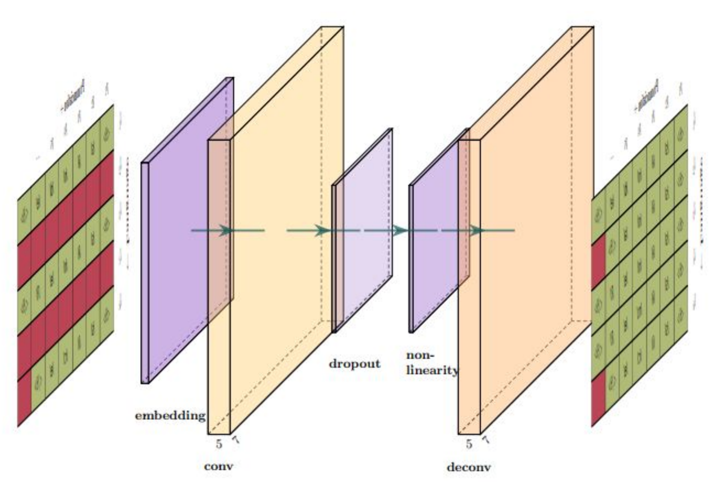


Figure 3: Simplified inpainting CNN architecture.

DATASET AND RESULTS

Mapping Dataset to languages

Training Data

Dataset	Source	Version	Family	Languages	Words	Cognates
⋆abrahammonpa	Abraham (2005)	v3.0	Tshanglic	8	2063	403
⋆allenbai	Allen (2007)	v4.0	Bai	9	5773	969
⋆backstromnorthernpakistan	Backstrom and Radloff (1992)	v1.0	Sino-Tibetan	7	1426	248
⋆castrosui	Castro and Pan (2015)	v3.0.1	Sui	16	10139	1048
davletshinaztecan	Davletshin (2012)	v1.0	Uto-Aztecan	9	771	118
felekesemitic	Feleke (2021)	v1.0	Afro-Asiatic	19	2583	340
⋆hantganbangime	Hantgan and List (2018)	v1.0	Dogon	16	4405	971
hattorijaponic	Hattori (1973)	v1.0	Japonic	10	1802	278
listsamplesize	List (2014)	v1.0	Indo-European	4	1320	512
mannburmish	Mann (1998)	v1.2	Sino-Tibetan	7	2501	576

Surprise Data

Dataset	Source	Version	Family	Languages	Words	Cognates
bantubvd	Greenhill and Gray (2015)	v4.0	Atlantic-Congo	10	1218	388
beidazihui	Běijīng Dàxué (1962)	v1.1	Sino-Tibetan	19	9750	518
birchallchapacuran	Birchall et al. (2016)	v1.1.0	Chapacuran	10	939	187
bodtkhobwa	Bodt and List (2022)	v3.1.0	Western Kho-Bwa	8	5214	915
⋆bremerberta	Bremer (2016)	v1.1	Berta	4	600	204
⋆deepadungpalaung	Deepadung et al. (2015)	v1.1	Palaung	16	1911	196
hillburmish	Gong and Hill (2020)	v0.2	Sino-Tibetan	9	2202	467
kesslersignificance	Kessler (2001)	v1.0	Indo-European	5	565	212
luangthongkumkaren	Luangthongkum (2019)	v0.2	Sino-Tibetan	8	2363	379
⋆wangbai	Wang and Wang (2004)	v1.0	Sino-Tibetan	10	4356	658

- **The datasets available, includes phonetic transcriptions produced by the Lexibank team and cognate sets provided by specialists.**
- **All singleton cognate sets were disregarded in every instance since we cannot use them in our prediction trials.**
- **Each dataset was divided into five training and test divisions with varying percentages of data maintained for testing, starting with 10% and increasing to 20%, 30%, 40%, and eventually 50%.**

Dataset Used: https://github.com/sigtyp/ST2022/tree/main/data

Results

Language	ED	ED (Normalized)	B-Cubed FS	BLEU
Amami	1.714	0.356	0.618	0.487
Hachijo	0.571	0.094	0.843	0.853
Kagoshima	1.429	0.340	0.653	0.502
Kochi	0.179	0.026	0.968	0.962
Kyoto	0.214	0.098	0.949	0.860
Miyako	1.607	0.381	0.596	0.481
Oki	0.643	0.135	0.820	0.802
Sado	0.214	0.028	0.937	0.961
Shuri	1.857	0.410	0.556	0.442
Tokyo	0.179	0.042	0.965	0.937
TOTAL	0.861	0.191	0.790	0.729

For each of the languages mentioned under dataset we applied our model and evaluated on Edit Distance , Edit Distance Normalized, B–Cubed FS , BLEU Score.

Reason why we choose BLEU Score as primary evaluation metric?

The main metric for evaluation was BLEU Score

Following are the reasons why other metrics were avoided:

B-Cubed F-Scores emphasize the systematicity of the prediction quality rather than the accuracy in individual cases

The classical edit distance was excluded in this overview, since it correlates highly with the normalized edit distance and would therefore artificially increase the overall ranks of systems performing well in this regard.

Model Used

PREVIOUS ATTEMPTS:

- **Support Vector Machine**
- **Transformer Model**

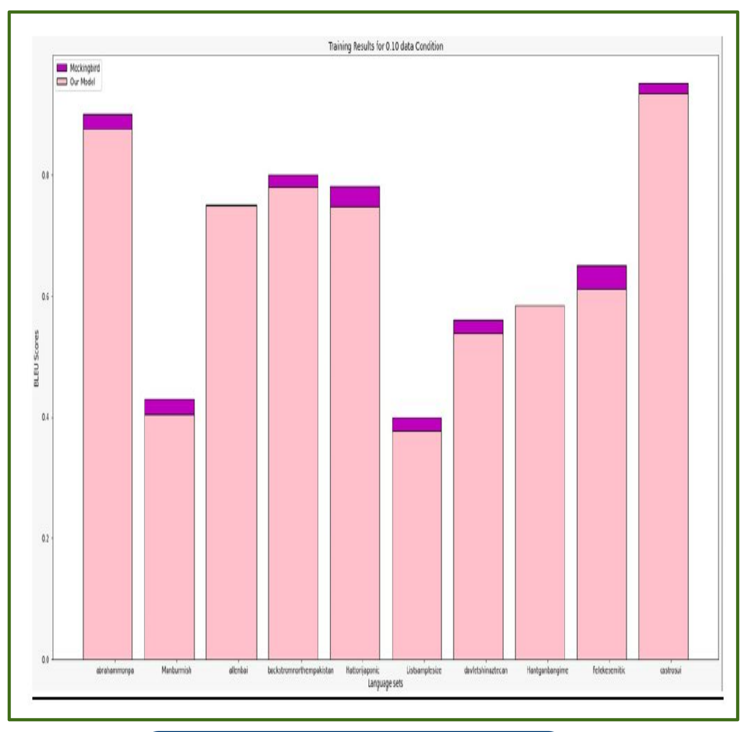
CURRENT SOTA MODEL:

- **Convolution Neural Network(CNN) this model was named as inpaint model in the baseline paper**

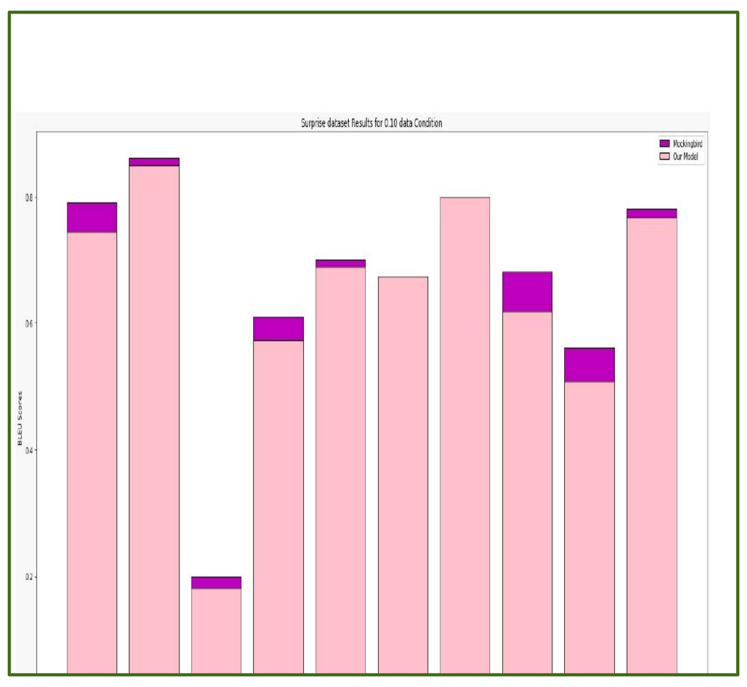
EXPERIMENTAL MODEL:

- **We are attempting to use graph based Convolution Neural Network(GCNN) to obtain predictive model**

How did our inpaint implementation perform in comparison to the original implementation?



TRAIN DATA PERFORMANCE



SURPRISE DATA PERFORMANCE

Baseline Model BLue Score

Our Model Blue Score

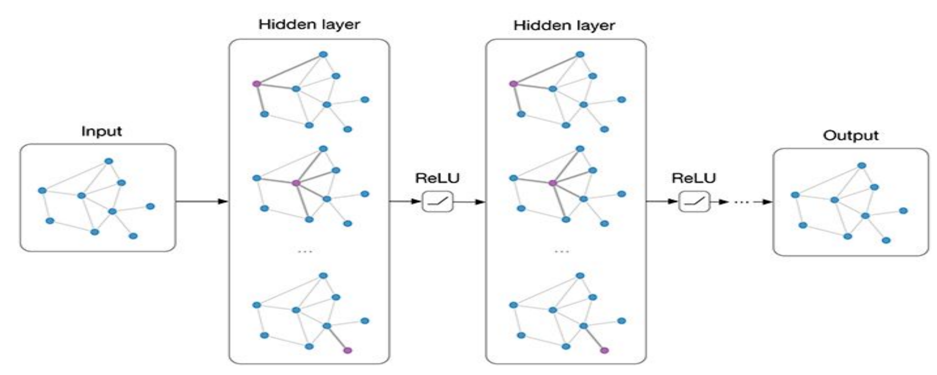
- **As we can see the our implementation of CNN gave a difference of 0.2-1.5 numerical value for BLEU Score in comparison to the original CNN implementation performed by the authors**

Graph Convolution Nueral Networks

Why use Graph Convolution -

Since we will be representing the data in the form of graphs, the nodes would be the cognate reflexes with their immediate neighbors being the other reflexes in the cogent set. Along with this, we aim to add relations across cogent sets which will help to get context over cogent sets globally across the dataset.

Because this representation would better capture the relations across cognate sets, we believe to get better scores for the cognate prediction task.



GCNN IMPLEMENTATION STEPS

THE FLOWCHART DEPICTS STEPS FOR GCNN IMPLEMENTATION

CREATE EMBEDDINGS → CREATE ADJACENCY MATRIX BASED ON EMBEDDINGS(THIS IS THE GRAPHICAL REPRESENTATION) → PASS THE ADJACENCY MATRIX TO Graph CNN → EVALUATE BLEU SCORES