SIGTYP 2022 Shared Task on the Prediction of Cognate Reflexes

Johann-Mattis List, Ekaterina Vylomova, Robert Forkel, Nathan W. Hill, and Ryan Cotterell

Introduction

The Reflex Prediction Task

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

Background on Reflex Prediction

- Quite some studies have been published in the past.
- They are outlined in more detail in our paper.
- The shared task itself showed, that there is additional potential for new methods to be developed in the future.

Difficulties of Reflex Prediction

- sounds without counterpart in target language
- sparsity and patchiness of data

Materials and Methods

Materials

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

Materials

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

Materials

- data comes from the Lexibank collection of CLDF datasets
- all wordlists in standard IPA, segmented by phonemes (https://clts.clld.org)
- concepts linked to Concepticon (https://concepticon.clld.org)
- languages linked to Glottolog (https://glottolog.org)
- cognate sets were either present or inferred with state-of-the-art methods

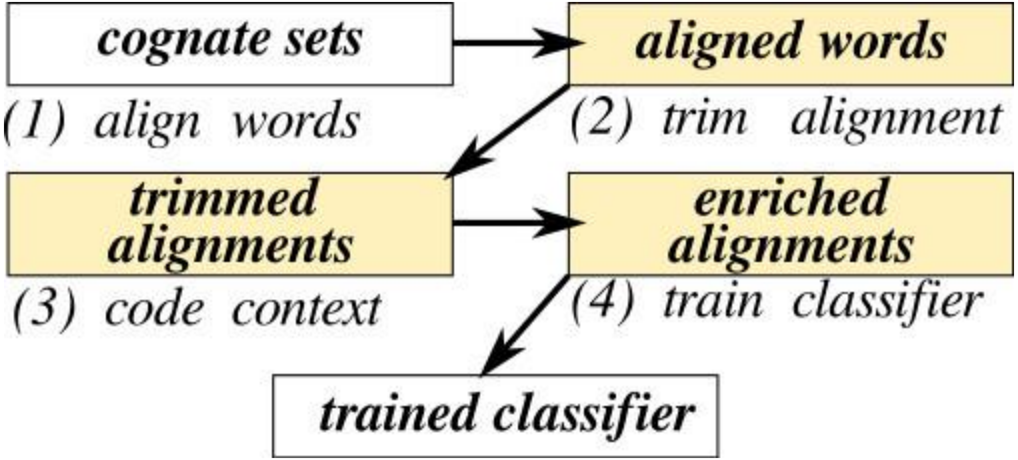
Methods: Evaluation

- edit distances (also Levenshtein distance)
- normalized edit distance (divide edit distance by length of alignment of two strings)
- B-Cubed F-Scores (structural measure, first proposed in List 2019)
- BLEU scores (suggested by participants as alternative to edit distance)

Methods: Baselines

- CorPaR baseline (correspondence pattern detection method first developed in List 2019 then extended in List et al. 2022)
- CorPaR baseline with SVM classifier

Methods: Baselines



Methods: Implementation

- Python package sigtypst2022 (https://github.com/sigtyp/ST2022)
- Evaluation methods in part taken from LingRex (https://github.com/lingpy/lingrex)
- Python package offers command line access to the evaluation and other routines (creating the baseline results, preparing files, etc.), and can also be accessed from within Python scripts
- Versions contain different parts of the data, version 1.4, the final version, contains all training and surprise data along with detailed results by all systems

Systems

Team CrossLingference

- Gerhard Jäger (Tübingen University)
- implemented in Julia
- uses Bayesian phylogenetic inference in combination with Hidden Markov Models for word prediction

Team Mockingbird

- Christo Kirov, Richard Sproat, Alexander Gutkin (Google)
- two systems, Neighbor Transformer and Image Inpainting
- Neighbor Transformer uses extended training data (based on random sampling)

Team Leipzig

- Giuseppe G. A. Celano (Leipzig University)
- Transformer-Based Architecture
- character and position embeddings
- prediction of one reflex from several candidate predicted from language pairs done by averaging target tensors

Team CEoT

- Tiago Tresoldi (Uppsala University)
- workflow very similar to baseline
- no trimming procedure
- different alignment enrichment
- random forest classifier

Results

General Results for Varying Held-Out Proportions

Proportion in Test: 0.1

System	ED	NED	B-Cubes	BLEU
Baseline	1.2095	0.3119	0.7231	0.5716
Baseline-SVM	**1.0189**	**0.2625**	**0.7626**	0.6387
CEoT-Extalign-RF	1.0377	0.2763	0.7475	0.6243
CrossLingference-Julia	1.4804	0.3929	0.7251	0.4793
Leipzig-Transformer	1.3901	0.3687	0.6489	0.5114
Mockingbird-I1	0.9201	0.2431	0.7673	0.6633
Mockingbird-N1-A	1.0223	0.2568	0.7604	**0.6479**
Mockingbird-N1-B	1.0437	0.2625	0.7572	0.6398
Mockingbird-N1-C	1.1263	0.2867	0.7302	0.6115
Mockingbird-N2	1.2095	0.3135	0.7054	0.5744

General Results for Varying Held-Out Proportions

Proportion in Test: 0.2

System	ED	NED	B-Cubes	BLEU
Baseline	1.3253	0.3361	0.6680	0.5412
Baseline-SVM	1.1723	0.2928	**0.7067**	0.5985
CEoT-Extalign-RF	1.2208	0.3175	0.6798	0.5709
CrossLingference-Julia	1.4954	0.3912	0.6882	0.4760
Leipzig-Transformer	1.5787	0.4046	0.5683	0.4646
Mockingbird-I1	1.0413	0.2648	0.7120	0.6326
Mockingbird-N1-A	**1.1512**	**0.2825**	0.7011	**0.6138**
Mockingbird-N1-B	1.1726	0.2901	0.6910	0.6054
Mockingbird-N1-C	1.2196	0.3051	0.6669	0.5841

General Results for Varying Held-Out Proportions

Proportion in Test: 0.3

System	ED	NED	B-Cubes	BLEU
Baseline	1.4354	0.3556	0.6372	0.5195
Baseline-SVM	1.3713	0.3310	**0.6565**	0.5554
CEoT-Extalign-RF	1.4038	0.3525	0.6331	0.5286
CrossLingference-Julia	1.6116	0.4130	0.6508	0.4503
Leipzig-Transformer	1.7746	0.4467	0.5129	0.4207
Mockingbird-I1	1.1762	0.2899	0.6717	0.6059
Mockingbird-N1-A	**1.2565**	**0.3119**	0.6557	0.5779
Mockingbird-N1-B	1.2712	0.3103	0.6531	**0.5792**
Mockingbird-N1-C	1.3009	0.3215	0.6343	0.5636

General Results for Varying Held-Out Proportions

Proportion in Test: 0.4

System	ED	NED	B-Cubes	BLEU
Baseline	1.6821	0.4011	0.6001	0.4717
Baseline-SVM	1.6159	0.3891	0.5990	0.4903
CEoT-Extalign-RF	1.5695	0.3960	0.5805	0.4773
CrossLingference-Julia	1.6059	0.4112	**0.6411**	0.4473
Leipzig-Transformer	1.9221	0.4800	0.4736	0.3893
Mockingbird-I1	1.2725	0.3162	0.6428	0.5724
Mockingbird-N1-A	**1.4542**	**0.3521**	0.6294	0.5293
Mockingbird-N1-B	1.3618	0.3349	0.6212	**0.5466**
Mockingbird-N1-C	1.4353	0.3547	0.5999	0.5228

General Results for Varying Held-Out Proportions

Proportion in Test: 0.5

System	ED	NED	B-Cubes	BLEU
Baseline	1.8889	0.4445	0.5617	0.4265
Baseline-SVM	1.9330	0.4619	0.5371	0.4204
CEoT-Extalign-RF	1.8434	0.4576	0.5194	0.4128
CrossLingference-Julia	1.6794	0.4274	0.6193	0.4296
Leipzig-Transformer	2.1036	0.5257	0.4306	0.3438
Mockingbird-I1	1.4170	0.3518	**0.6050**	0.5337
Mockingbird-N1-A	1.5527	0.3800	0.5959	0.4934
Mockingbird-N1-B	**1.5066**	**0.3734**	0.5864	**0.4989**
Mockingbird-N1-C	1.5818	0.3950	0.5610	0.4749

Ranks

System	Rank	NED	B-Cubes	BLEU	Aggregated
Mockingbird-I1	1	1	1.2	1	1.1 ± 0.3
Mockingbird-N1-A	2	2.6	3	2.6	2.7 ± 0.4
Mockingbird-N1-B	3	2.4	4	2.4	2.9 ± 0.9
Baseline-SVM	4	5.2	4	5	4.7 ± 1.9
Mockingbird-N1-C	5	4.6	6.6	4.6	5.3 ± 1.3
CEoT-Extalign-RF	6	6	7	6.2	6.4 ± 1.1
CrossLingference-Julia	7	7.6	4	7.6	6.4 ± 2.5
Baseline	8	6.8	6.2	6.8	6.6 ± 0.8
Leipzig-Transformer	9	8.8	9	8.8	8.9 ± 0.4

Discussion

General System Performance

- Interestingly, all systems performed reasonably well, confirming that the task is interesting and worth to be further pursued by computational methods.
- Best performance was by the Inpainting model, followed by the Neighor-Transformer
- Not all methods could beat the baselines (SVM baseline quite successful and fast)
- With lower amounts of data, some systems could gain some ground.

Open Questions

- Machine Learning vs. Targeted Algorithms
- Individual differences between machine learning systems
- Individual differences between similar systems (CEoT vs. Baselines)

Organization

What went well:

- Fully replicable baseline
- Clear evaluation routine
- Interesting new methods

Organization

What could be improved:

- Earlier announcement and propagation
- Organization (clearer instructions from the beginning)
- No sharing of test results
- Testing of all systems by the organizers

Thanks for Your Attention!

And Special Thanks to All Participants!