Findings of the SIGTYP 2023 Shared task on Cognate and Derivative Detection For Low-Resourced Languages

Priya Rani[1], **Koustava Goswami**[1,2], **Adrian Doyle**[1],
Theodorus Fransen[1], **Bernardo Stearns**[1], **John P. McCrae**[1]
[1] Data Science Institute, University of Galway, Ireland
[2] Adobe Research Bangalore, India
koustavag@adobe.com,
{priya.rani, adrian.doyle, theodorus.fransen, bernardo.stearns, john.mccrae}@insight-centre.org

Abstract

This paper describes the structure and findings of the SIGTYP 2023 shared task on cognate and derivative detection for low-resourced languages, broken down into a supervised and unsupervised sub-task. The participants were asked to submit the test data's final prediction. A total of nine teams registered for the shared task where seven teams registered for both sub-tasks. Only two participants ended up submitting system descriptions, with only one submitting systems for both sub-tasks. While all systems show a rather promising performance, all could be within the baseline score for the supervised sub-task. However, the system submitted for the unsupervised sub-task outperforms the baseline score.

1 Introduction

Cognates and derivatives have been studied in various fields of linguistics with different purposes (Labat and Lefever, 2019). In historical linguistics, cognates are useful in the reconstruction of proto-languages and can aid in establishing the relationship between languages; in lexicography, cognates are helpful in the development of multilingual dictionaries. Moreover, in recent years, NLP researchers have shown interest in using cognates to enhance the performance of multilingual tasks such as machine translation, lexical induction, word embeddings and many more (Kondrak, 2005; Kondrak et al., 2003).

As there has been little work on automatic cognate identification, it is still a challenging task, especially for less-resourced languages (Jäger et al., 2017; Rama, 2016). Supervised identification of cognates and derivatives is requires a substantial amount of annotated linguistic data, which may need to be manually annotated (Kanojia et al., 2021). At the same time, finding linguists and annotators for less-resourced languages is impractical. Thus we propose a shared task which aims to provide a new benchmark for differentiating between cognates and derivatives and introduce new unsupervised approaches for cognate and derivative detection in less-resourced languages.

Cognates are etymologically related word pairs across languages which may or may not have similar spelling, pronunciation and meaning (Crystal, 2011). Cognates can be traced back to a single ancestral word form in a common earlier language stage. On the other hand, derivatives are words which have been adopted into a language either from an earlier stage of the same language, or as a borrowing from a different language. To give an example, the Spanish *libro* and French *livre*, are each derived from Latin *liber* "book", and are cognates with each other because they share this common ancestor. By contrast, the Irish word *leabhar* is derived from Latin *liber* because it was borrowed into Irish from Latin, but *leabhar* is a cognate with Spanish *libro* because *libro* has been derived from an earlier developmental language stage, i.e. *leabhar* was not borrowed from Spanish, but from Latin, a precursor to Spanish. Where multiple stages of direct derivation occur, each successive stage is considered a derivation from the last, but also from any earlier stages. For example, *leabhar* in Modern Irish is derived from Old Irish *lebor*, but also from Latin *liber*.

As will be discussed in section 3, data used in this shared task has been drawn from Wiktionary. Apart from cognates (`cog`), Wiktionary distinguishes between derived (`der`), inherited (`inh`), and borrowed (`bor`). This distinction is not maintained in this shared task, and all three are treated broadly as derivation. Languages are distinguished from one another in the shared task based on ISO-639 codes. Anything which has a discrete ISO-639 code is considered a separate language. Therefore, Irish with the code `ga` is a completely separate language from Old Irish as this has a separate code, `sga`. This prevents any confusion as to the point

Proceedings of the 5th Workshop on Research in Computational Linguistic Typology and Multilingual NLP (SIGTYP 2023), pages 126–131
May 6, 2023 ©2023 Association for Computational Linguistics

at which something ceases to be a derivative and becomes a cognate. Such confusion may occur in speech, for example, where one may say that a term was borrowed into English from French. Such a statement could lead to the supposition that a Modern English (en) word is derived from the Modern French (fr) term, however, if the borrowing took place between earlier language stages, say into Middle English (enm) from Old French (fro), the Modern English term is only derived from Old French and precursors to it, like Latin (la), not from Modern French. This is the case with the English word *liberal*. It was borrowed into Middle English from Old French, and is ultimately derived from Latin *liber* "free". Hence, the Modern English, *liberal*, would be considered a derivative from both Old French and Latin in our data, however, it would be a cognate with Modern French *libéral* because *liberal* is not derived directly from Modern French.

The rest of this paper is organised as follows. Section 2 describes the setup and schedule of the shared task. Section 3 presents the dataset used for the competition. Section 4 describes the evaluation methods and the baselines. Section 5 describes the systems submitted by the teams in the competition, and Section 6 presents and analyses the results obtained by the competitors. Lastly, in Section 7, we conclude the whole findings of the shared task.

2 Shared task setup and schedule

The section describes how the shared task was organized. The shared tasks involve two sub-tasks to perform multiclass classification tasks, which require that the relationship between pairs of words be identified as either a cognate relationship, a derivative relationship, or no relationship. The sub-tasks are:

- Supervised: Cognate and Derivative Detection
- Unsupervised: Cognate and Derivative Detection

The shared task started with the registration process through Google Forms. The participants were asked to register their team along with their affiliation, team member and the sub-tasks they wanted to participate in. Registered participants were sent a link to access the training and development data. The participants were allowed to use additional data to train their system with the condition that any additional data used should be made publicly available and to provide a proper citation of the data used to develop their model. The schedule for the release of training data and release of test data, along with notification and submission, are given in Table 1.

Date	Event
9 January 2023	Release of training data
27 Feburary 2023	Release of test data
15 March 2023	Submission of the systems
27 March 2023	Submission of system description paper
31 March 2023	Camera-ready

Table 1: SIGTYP 2023 Shared Task schedule

3 Cognate Datasets

In this section, we present the characteristics and the statistics of the dataset used for the task of cognate and derivative prediction.

3.1 Training Data

We provide annotated word pairs for cognate and derivative prediction in a format given in Table 2 in which the first column represents the first word of the word pair and the second column represents the language of the given word through the ISO code. The third and the fourth column represent the second word and its language code, respectively. Lastly, the fifth column represents the relation between the two words in each pair; cognate, derivative or none. The detailed statistics of the words pairs according to the labels are given in Table 3.

Word_1	ISO	Word_2	ISO	Label
Yannick	en	Yannig	br	der
creta	ca	creta	la	der
roh	de	raw	en	cog
gnit	en	gnit	is	cog
erudit	oc	ergueito	gl	none

Table 2: Format of the dataset

The data consists of word pairs from 34 languages including both high-resourced and less-resourced languages. Table 4 gives an overview of the languages involved and statistics of each language. This data was collected and annotated using Wiktionary.

Labels	Train	Test
Cognate	11869	98
Derivatives	39205	340
None	181408	438
Total	232482	876

Table 3: Statistics of the dataset in each category.

Languages	Count in word_1	Count in word_2
en	22883	13414
es	14921	11996
it	12528	9804
nb	12473	9390
nn	12139	9415
pt	12118	9759
ca	11946	9434
fr	10944	12573
nl	10895	9670
gl	10437	9026
da	10280	9048
oc	8119	7904
sv	7823	7588
la	7757	37217
de	7340	9105
ro	7063	6664
pl	6346	5744
af	5465	5205
ga	4384	3872
cs	4342	4058
is	4136	4237
lb	3230	2754
no	2833	2904
gd	2833	2710
cy	2684	2742
sk	2680	2487
lv	2576	2549
sl	2481	2448
gv	1764	1651
fy	1759	1797
wa	1584	1562
br	1259	1255
kw	1244	1220
lt	1216	1280

Table 4: Statistics of the languages in the training data

In the later stages of the shared task we came across a number of false negatives in the training data. Specifically, some word pairs were labelled none, indicating that they shared no relationship, however, upon investigation they were found to be either cognates or derivatives. As we were close to releasing the test data, we decided not to make any changes in the training data, but instead to simply inform the participants. This was expected to cause the least disruption to participants for a couple of reasons. Firstly, participants had already been given the freedom to manipulate the data as they saw fit, in order for them to optimise their systems. Secondly, as discussed in section 2, the participants were allowed to use datasets other than those provided. If participants had already attempted to overcome the problem by editing or removing erroneous entries from the provided training data, it was perceived that providing all participants with cleaned training data at such a late stage would have unfairly benefited those who had not adapted the training data.

3.2 Test Dataset

Similar to training data, test data for the given task consists of word pairs from 34 languages, including high-resourced and less-resourced. Table 5 provides an overview of the languages involved and statistics of each language. Though the test data was collected using Wiktionary, it was annotated manually by the experts using the Wiktionary template.

4 Methods

4.1 Evaluations

The standard evaluation metrics for evaluating and ranking the teams was F1-Score for supervised classification. For unsupervised methods, we followed the standard cluster performance evaluation process. The number of clusters will be same as the number of original classes and evaluated with the cluster accuracy using the equation shown in Equation 1,

$$ACC = max_m \frac{\sum_{i=1}^{n} 1(l_i = m(c_i))}{n} \qquad (1)$$

where l_i is the ground truth label, c_i is the cluster assignment produced by the algorithm and m ranges over all possible one-to-one mappings between clusters and labels.

4.2 Baselines

This section gives a short description of the baselines used to compare the submitted systems.

Supervised: The system was a multi-layer LSTM-based network. The framework has two major stages, they are:

- Data preparation: In this stage pre-processing was carried out to remove punctuation, undesirable Unicode, conversion of cases and building one-hot vectors of both word and language information.

Languages	Count in word_1	Count in word_2
en	120	44
pt	55	17
nn	50	14
es	49	28
it	46	35
ca	44	16
nb	40	20
fr	29	38
da	27	23
ro	25	14
gl	25	21
nl	25	29
oc	25	22
de	23	42
fy	19	13
sv	18	24
pl	17	17
lb	17	06
af	17	07
is	17	18
lt	16	09
cy	16	12
no	16	17
cs	15	07
wa	15	15
sk	14	07
gv	14	08
kw	13	13
sl	13	18
lv	13	12
gd	12	10
la	12	276
br	11	09
ga	8	15

Table 5: Statistics of the languages in the test data

- Model training: After converting the data to one-hot vector various RNN model were trained. However, the best model was chose to be the baseline of the shared task. This model consisted of two hidden layers of 100 LSTM cells with only a single dense layer and softmax activation function. It uses Adam optimisation and categorical cross-entropy to calculate loss. The model was set to train for 250 epochs on a randomised selection of 90% of the training data. The other 10% was set aside for validation during training. Early stopping was applied to ensure overfitting did not occur, with the result that the actual number of epochs during training was less than 100. The input format for the model was a 34x50 matrix where 34 represents the number of languages (this was higher than the total number of unique characters), and 50 represents the buffered word-size (24) doubled as words were fed in in pairs, plus 2 as the language of each word also took up a vector each.

Unsupervised: A simple Levenshtein edit distance (Levenshtein, 1965) model was trained to perform the clustering task with the cluster set of 3.

5 Systems

A total of 9 teams registered for the shared task: 7 teams registered to participate in both the supervised and unsupervised tasks while 2 teams registered for only the supervised task. Out of these, only two teams submitted systems. Both teams submitted for the supervised task and one team submitted for both the supervised and unsupervised task. The teams who submitted their systems were invited to submit system description papers describing their experiments in the proceedings of the workshop (Beinborn et al., 2023). Since these systems are described in individual papers, we will only briefly present the main features here.

ÚFAL_supervised: The system submitted by team ÚFAL, represented by Tomasz Limisiewicz from Charles University, provided gradient boosted tree classifier trained on linguistic and statistical features. The features used by the team to train the classifiers were language model embeddings, typological information which included language identity and language group identity and orthographical information (Limisiewicz, 2023).

CoToHiLi_supervised: Team CoToHiLi, represented by Liviu Dinu from University of Bucharest, experimented with a few different multi-class classification algorithms such as Support Vector Machine, Naive Bayes, and SGD with the combination of three features graphic features, phonetic features and language features. At the end they selected the best performing classifiers to train a stackable ensemble classifier (Liviu P. Dinu, 2023).

CoToHiLi_unsupervised: The unsupervised system submitted by team CotoHiLi employed a set of features including graphic, phonetic and language encoding to KMeans algorithms (Liviu P. Dinu, 2023).

6 Results

The participants were asked to submit the final test results in the format of the training data files, with comma-separated fields for word pairs, language codes, and relationship labels. Files had to be named **team name_unsupervised/supervised** to indicate both the team's name and the sub-task in question.

Teams	F1-Score	Precision	Recall
Baseline	0.91	0.99	0.84
ÚFAL	0.87	0.89	0.86
CoToHiLi	0.83	0.87	0.81

Table 6: Results of submitted supervised systems for the SIGTYP 2023 Shared Task.

Teams	Accuracy
Baseline	0.38
CoToHiLi	0.49

Table 7: Results of submitted unsupervised systems for the SIGTYP 2023 Shared Task.

7 Conclusion

We have reported the findings of the SIGTYP 2023 Shared Task on cognate and derivative detection for less-resourced languages as part of the fifth edition of SIGTYP workshop. With the two teams that participated, we have seen different and interesting non-neural and neural systems that deal with cognate and derivative prediction task. While the baseline for supervised sub-task were based on neural networks, team ÚFAL used a gradient boosted tree classifier and team CoToHiLi came up with an ensemble classifier. However, neither team could beat the baseline set for the supervised task: the difference in the F1-Score was -0.04 for team ÚFAL and -0.08 for team CoToHiLi. Although, team ÚFAL's entry ranked first among the two supervised systems submitted, with an F1-Score of 0.87, the unsupervised system submitted by team CoToHiLi based on a KMeans algorithm beat the baseline for the unsupervised task with with an improvememt of 0.11 in accuracy.

Acknowledgements

This Shared Task was supported by the Irish Research Council as part of grant IRCLA/2017/129 (CARDAMOM-Comparative Deep Models of Language for Minority and Historical Languages) and co-funded by Science Foundation Ireland (SFI) as part of grant SFI/18/CRT/6223 (CRT-Centre for Research Training in Artificial Intelligence) and grant SFI/12/RC/2289_P2 (Insight_2).

References

Lisa Beinborn, Koustava Goswami, Saliha Muradoğlu, Alexey Sorokin, Ritesh Kumar, Andrey Shcherbakov, Edoardo Ponti, Ryan Cotterell, and Ekaterina Vylomova, editors. 2023. *Proceedings of the 5th Workshop on Research in Computational Linguistic Typology and Multilingual NLP*. Association for Computational Linguistics, Dubrovnik, Croatia.

David Crystal. 2011. *A dictionary of linguistics and phonetics*. John Wiley & Sons.

Gerhard Jäger, Johann-Mattis List, and Pavel Sofroniev. 2017. Using support vector machines and state-of-the-art algorithms for phonetic alignment to identify cognates in multi-lingual wordlists. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 1205–1216, Valencia, Spain. Association for Computational Linguistics.

Diptesh Kanojia, Prashant Sharma, Sayali Ghodekar, Pushpak Bhattacharyya, Gholamreza Haffari, and Malhar Kulkarni. 2021. Cognition-aware cognate detection. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3281–3292, Online. Association for Computational Linguistics.

Grzegorz Kondrak. 2005. Cognates and word alignment in bitexts. In *Proceedings of Machine Translation Summit X: Papers*, pages 305–312.

Grzegorz Kondrak, Daniel Marcu, and Kevin Knight. 2003. Cognates can improve statistical translation models. In *Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics, HLT-NAACL 2003, Edmonton, Canada, May 27 - June 1, 2003*. The Association for Computational Linguistics.

Sofie Labat and Els Lefever. 2019. A classification-based approach to cognate detection combining orthographic and semantic similarity information. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 602–610, Varna, Bulgaria. INCOMA Ltd.

Vladimir Levenshtein. 1965. Binary codes capable of correcting spurious insertions and deletions of ones. *Russian Problemy Peredachi Informatsii*, 1:12–25.

Tomasz Limisiewicz. 2023. Ufal submission for sigtyp supervised cognate detection task. In *Proceedings of the 4th Workshop on Research in Computational Linguistic Typology and Multilingual NLP*, Dubrovnik, Croatia. Association for Computational Linguistics.

Ana Sabina Uban Liviu P. Dinu, Ioan-Bogdan Iordache. 2023. Cotohili at sigtyp 2023: Ensemble models for cognate and derivative words detection. In *Proceedings of the 4th Workshop on Research in Computational Linguistic Typology and Multilingual NLP*, Dubrovnik, Croatia. Association for Computational Linguistics.

Taraka Rama. 2016. Siamese convolutional networks for cognate identification. In *COLING 2016, 26th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, December 11-16, 2016, Osaka, Japan*, pages 1018–1027. ACL.