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

Priya Rani[1], Koustava Goswami[2], Adrian Doyle[1], Theodorus Fransen[1], Bernado Stearns[1], John P. McCrae[1]

[1]Data Science Institute, University of Galway, Ireland
[2] Adobe Research Bangalore, India

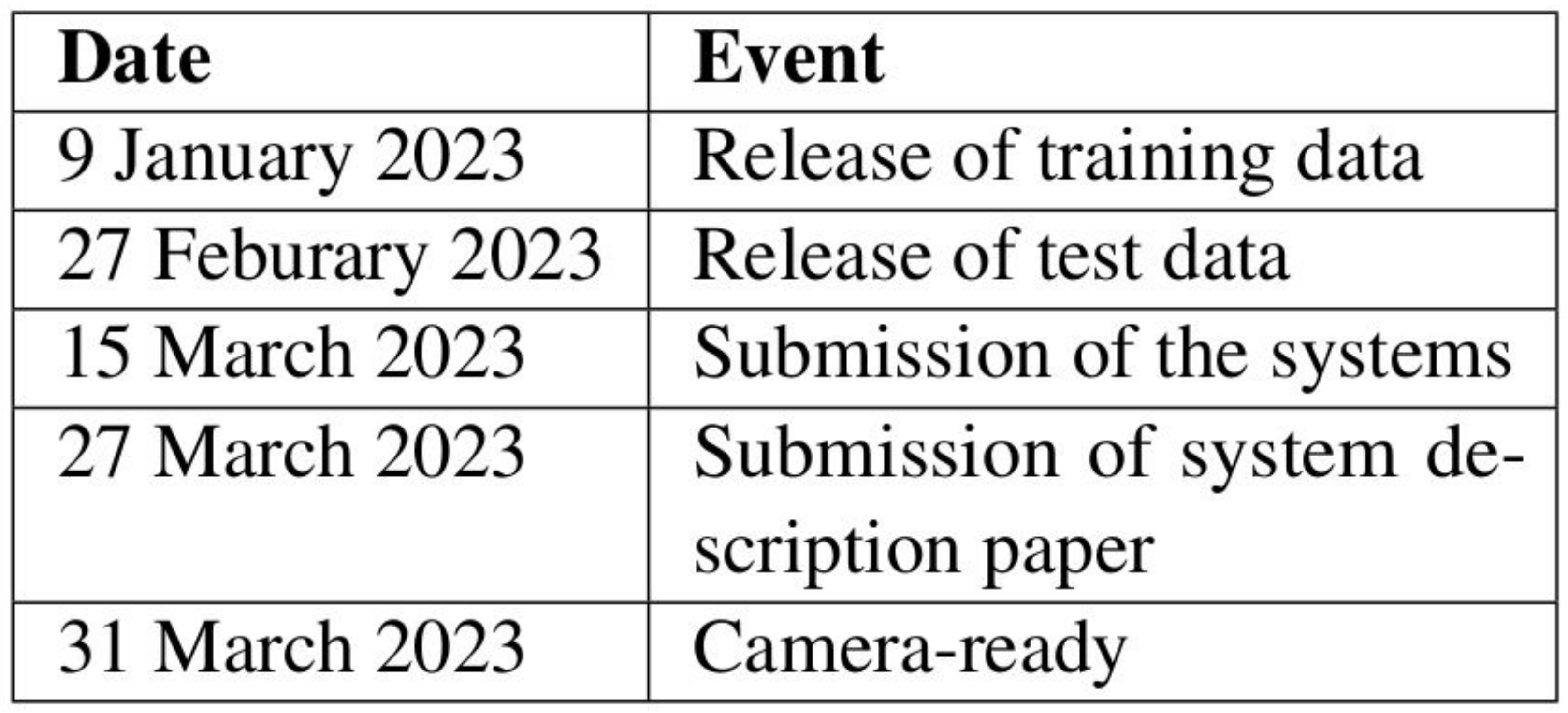
Introduction

- **Defining Cognates and Derivatives**
    - Libro (Spanish) and Livre (French) are cognate ----> Liber (Latin) ‘Book’
    - Leabhar (Irish) and Libro (Spanish) are cognate ------> Liber (Latin)
    - Leabhar (Irish) -----> Liber (Latin) are derivatives
    - Leabhar (New Irish) ------> Lebor (Old Irish) are derivatives

- **Motivation**
    - Reconstruction of proto languages
    - Multilingual dictionaries
    - NLP task such as MT, Lexical Induction
    - Annotation are expensive

Setup and Schedule

- Two Subtask
  - Supervised: Cognate and Derivatives Detection
  - Unsupervised: Cognate and Derivatives Detection
- Use of other additional data were allowed
- Schedule of the Shared task given in the Table



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

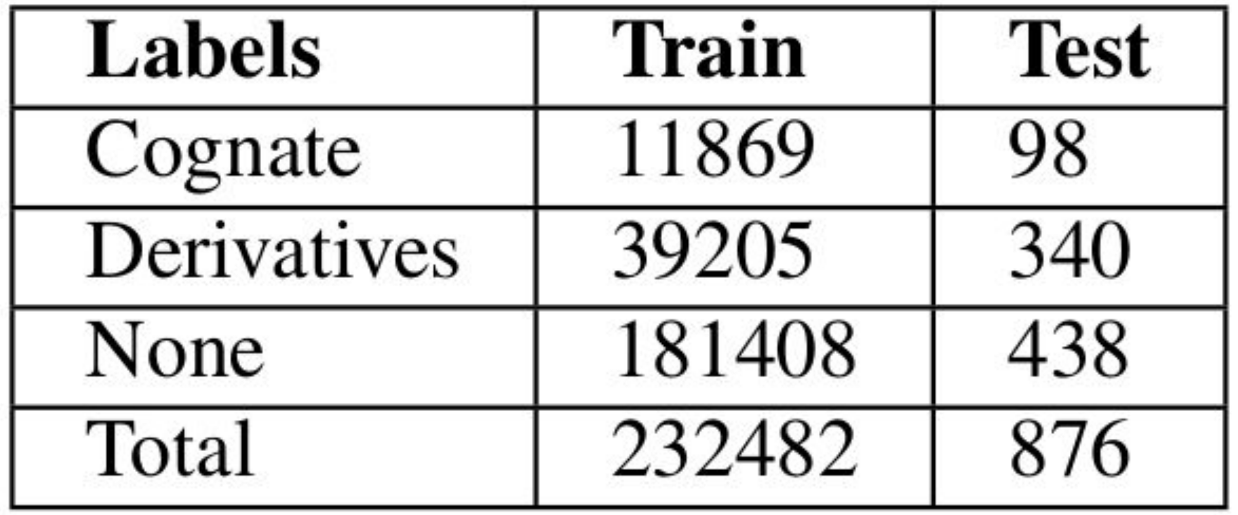
Table 1: Schedule of the Shared Task

OLLSCOIL NA GAILLIMHE
UNIVERSITY OF GALWAY

Adobe

Data Set

- Source of the Data : *Wiktionary*
- Annotated pairs of cognate, derivatives and none
- Data consists of word pairs of *34 languages*
    - High-resourced and low-resourced languages
- Test data were annotated manually using *Wikinationary template*

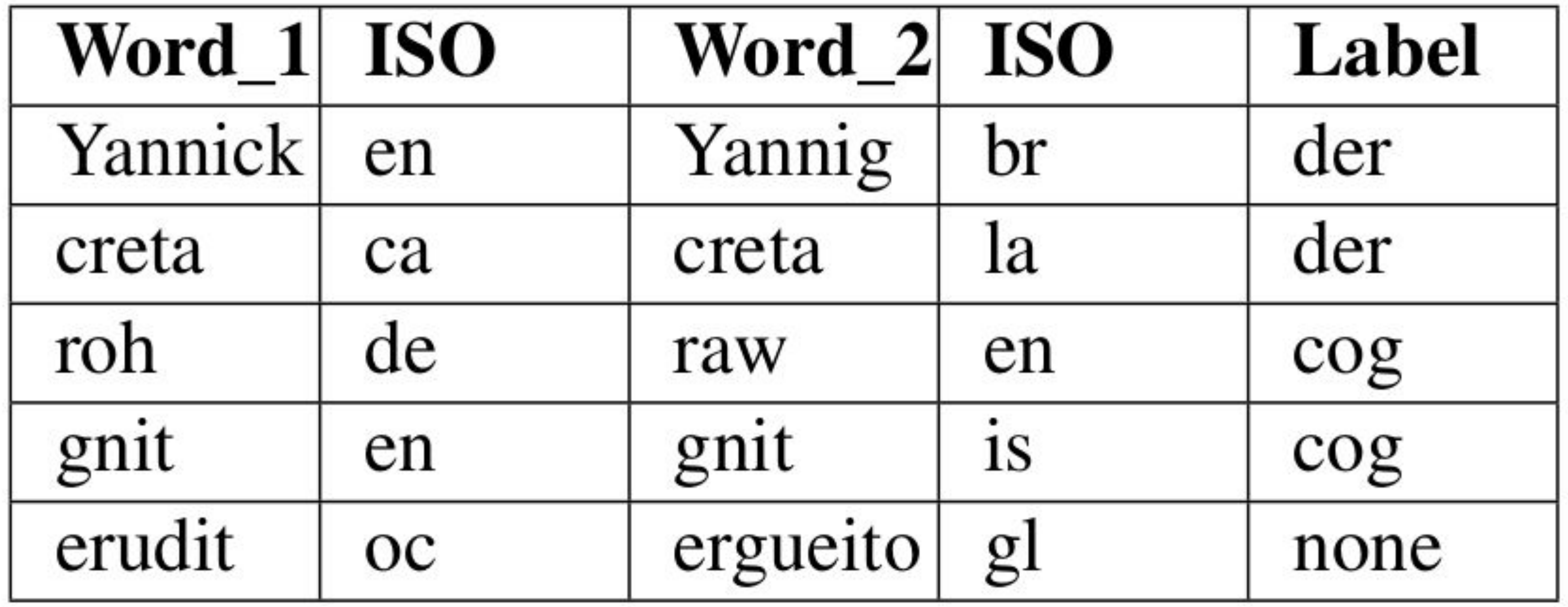


Labels	Train	Test
Cognate	11869	98
Derivatives	39205	340
None	181408	438
Total	232482	876

Table 2: Data Statistics

Data Set

- False negatives were found in training data set in the *none* category
- The distinction between *inherited* and *borrowed* are *not maintained*
- Languages are distinguished from each other using ISO-639
    - example New Irish with ISO *ga* is different from Old Irish with ISO *sga*



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 3: Format of the Data given to the participants

Methods

- **Evaluation Metrics:**
  - *F1-Score* for supervised Classification
  - For unsupervised standard cluster performance evaluation process using *Accuracy*
- **Baselines:**
  - Multilayered *LSTM* based network
    - Data Preprocessing
    - Model Training: input format for the model was a 34x50 matrix; 34 represents the no. of languages and 50 represents buffered word size.
  - *Levenshtein edit distance model* was trained to perform the clustering task with the cluster set of 3.

System Description

- Total 9 teams registered for the task
- 2 teams submitted for supervised task
- Only one team submitted for unsupervised task
- **Team CoToHiLi:**
  - Lead by Liviu Dinu from University of Bucharest
  - *Supervised System*
    - Trained stackable ensemble supervised classifier (SVM, Naive Bayes and SGD)
    - Using the three main features: graphic, phonetic and language
  - *Unsupervised*
    - Trained on K-Means Algorithm
    - With the features set of graphic, phonetic and language encoding

System Description

- **Team Ufal:**
  - Lead by Tomasz Limisiewicz from Charles University
  - Submitted for *Supervised task*
  - provided *gradient boosted tree* classifier
  - Classifiers trained on linguistic and statistical features
  - Features includes : *language model embeddings,* *typological information*
  - Typological features includes
    - language identity
    - language group identity
    - orthographic information

Results

- **Supervised Task**

Teams	**F1_SCore**
Baseline	0.91
Ufal	0.87
CoToHiLi	0.83

- **Unsupervised Task**

Teams	**Accuracy**
Baseline	0.38
CoToHiLi	0.49

Conclusion

- All the system provided a reasonable performance
- Both the teams came up with interesting though they can’t beat the baselines for supervised task
- Team CoToHiLi scored better than the baseline for unsupervised task
- Non- neural training could provide good results with selected feature sets

Ollscoil na Gaillimhe
University of Galway

Adobe

Thank You!