

Weakly-supervised Deep Cognate Detection Framework for Low-Resourced Languages Using Morphological Knowledge of Closely-Related Languages

Koustava Goswami^{1,2*}, Priya Rani², Theodorus Fransen^{2,3},
John P. McCrae²

¹ Adobe Research Bangalore, India

² Data Science Institute, University of Galway, Ireland

³ Università Cattolica del Sacro Cuore, Milan, Italy

koustavag@adobe.com, {p.rani1, john.mccrae}@universityofgalway.ie, theodorus.fransen@unicatt.it

Abstract

Exploiting cognates for transfer learning in under-resourced languages is an exciting opportunity for language understanding tasks, including unsupervised machine translation, named entity recognition and information retrieval. Previous approaches mainly focused on supervised cognate detection tasks based on orthographic, phonetic or state-of-the-art contextual language models, which under-perform for most under-resourced languages. This paper proposes a novel language-agnostic weakly-supervised deep cognate detection framework for under-resourced languages using morphological knowledge from closely related languages. We train an encoder to gain morphological knowledge of a language and transfer the knowledge to perform unsupervised and weakly-supervised cognate detection tasks with and without the pivot language for the closely-related languages. While unsupervised, it overcomes the need for hand-crafted annotation of cognates. We performed experiments on different published cognate detection datasets across language families and observed not only significant improvement over the state-of-the-art but also our method outperformed the state-of-the-art supervised and unsupervised methods. Our model can be extended to a wide range of languages from any language family as it overcomes the requirement of the annotation of the cognate pairs for training. The code and dataset building scripts can be found at https://github.com/koustavagoswami/Weakly_supervised-Cognate_Detection

1 Introduction

Cognates are etymologically related word pairs across languages (Crystal, 2011). However, cognates are defined in much broader terms in many different fields, including natural language processing (NLP) or psycholinguistics (Labat and Lefever,

2019). In these areas, word pairs with similar meanings and spelling are also considered as cognates. In the recent development of automatic machine translation (AMT), automatic cognate detection is found to be very effective for similar language translation (Kondrak, 2005; Kondrak et al., 2003). Moreover, it helps to efficiently perform cross-lingual information retrieval (Makin et al., 2007; Meng et al., 2001) from different sources. Very often, words that have similar spelling are recognised as cognates (Example: the Latin and the English word pair “cultūra” and “culture”). Nevertheless, there are word pairs which are false friends or partial cognates (Kanojia et al., 2020b). Partial cognates are similar words across languages but carry different meanings in different contexts (Kanojia et al., 2019b), thus making automatic cognate detection hard and challenging. Identifying these cognates requires extensive linguistic knowledge across languages, which is quite hard and expensive to annotate. While cross-lingual automatic cognate detection systems exist, they have primarily been supervised methods requiring labelled data or language-specific linguistic rules. For under-resourced languages, finding annotators or linguists is a challenging task. This highlights the need for an efficient unsupervised language-agnostic cognate detection framework. We show that our weakly-supervised and unsupervised approaches can better exploit the available data than existing supervised methods and thus produce better results for under-resourced languages. The method transfers the morphological knowledge of a shared encoder in the unsupervised cognate detection framework into a Siamese network setting, where the framework simultaneously learns word representation and cluster assignments in a self-learning setup.

Supervised cognate detection frameworks understand the relationship between word pairs by concentrating on their phonetic or lexical similar-

*Research work conducted during his time at the Data Science Institute, University of Galway, Ireland. Contact: koustavag@adobe.com

ities based on their annotated positive or negative labels (Jäger et al., 2017; Rama, 2016). Recently, researchers tried to exploit contextual multilingual word embedding techniques to identify cognates which produced better results than only concentrating on phonetic transcriptions (Kanojia et al., 2020a). Although such methods had good results, annotating labels is quite expensive and tedious for many under-resourced languages. Moreover, producing multilingual contextual word embeddings is challenging for these languages due to the need for more data sources on the web. Merlo and Rodriguez (2019a) highlighted that based on bilingual lexicon matching between two known languages, the similarity score produced by contextual word embeddings could differentiate between true or false cognate pairs. Though this technique is label independent, the framework depends on the bilingual lexicon availability of the known language pairs.

To alleviate the above challenges, in this paper, we propose a *language-agnostic weakly-supervised cognate detection framework based on Siamese architecture* with an iterative clustering approach (Xie et al., 2016) during back-propagation. Our encoder design is inspired by Goswami et al. (2020), where they learn the n-gram character features of a sentence with attention. We introduce a *positional encoder* on n-gram features, which, in combination with the attention mechanism, learns sub-word representations of a word. We also depict the performance of our *morphological knowledge-based weakly-supervised framework*. This variant gives a better understanding of the grammar and structural analysis of the words of a language. Thus, transferring this knowledge with the help of a shared encoder to closely-related languages enhances the understanding of structural and grammatical relatedness between cross-lingual word pairs. Moreover, our word encoding method helps to produce better supervised cognate detection results, which outperform the state-of-the-art supervised results on various language pairs. In this paper, we have presented a complete set up of results for supervised, weakly-supervised and unsupervised setups.

The extensive experiments (in Section 6) on three different cognate detection datasets across language families have showcased the efficacy of our weakly-supervised and supervised cognate detection framework. For example, on six different Indian language pairs, our weakly-supervised

model (with morphological knowledge) has outperformed the state-of-the-art supervised model proposed by Kanojia et al. (2020a) by an average of 9 points of F -score whereas, for Celtic language pairs, it outperformed by 8.6 points of F -score. At the same time, our supervised framework has produced a state-of-the-art performance by outperforming the existing supervised model by an average of 16 points of F -score. Thus, our model is robust across diverse language families for the supervised and weakly-supervised cognate detection task. Interestingly, our experiments show that on Indian language pairs like Hindi-Punjabi and Hindi-Marathi, an encoder with morphological knowledge of the Hindi language performed better than an encoder with morphological knowledge of their ancestral language, Sanskrit by an average of 1.5 points of F -score. However, the performance of the weakly-supervised framework for the language pair Marathi-Bengali has improved by 2 points of F -score.

In a nutshell, our contributions are:

- (i) a language-agnostic weakly-supervised cognate detection framework without the need for labels,
- (ii) efficiently transferring morphological knowledge of a low-resourced language to closely-related under-resourced languages with or without the need for the pivot language for better cognate detection,
- (iii) introduction of positional embedding along with attention to different n-grams of a word for better understanding of word structures,
- (iv) robustness in weakly-supervised and supervised cognate detection for low-resource languages across three different datasets of different language families, outperforming state-of-the-art supervised approaches.

2 Related Work

Algorithms for automatic cognate detection (ACD) are mostly based on phonetic or orthographic similarity measures and often language-dependent or supervised approaches. Covington (1996) developed an algorithm to align historical-comparative languages by their phonetic similarity. Kondrak (2000) released a novel algorithm for aligning phonetically similar sequences. Similarity-based al-

gorithms follow these works to identify cognates between a language pair of different families. The orthographic similarity-based works calculate distances or string similarities between word pairs and define similarity scores to identify cognates (Mulloni and Pekar, 2006; Melamed, 1999; Jäger et al., 2017). Rama (2016) has released a convolution-based model that also considers phonetic similarity-based scores into account to detect cognates for word pairs. Some researchers have also taken parallel datasets into account to identify cognates. Distance measurement based scores have become the feature set to identify cognates in these cases (Mann and Yarowsky, 2001; Tiedemann, 1999). Kanojia et al. (2019a) performed a cognate detection task on Indian languages, which includes a large amount of manual intervention during identification. Kanojia et al. (2019b) introduced a character sequence-based recurrent neural network for identifying cognates between Indian language pairs.

The influence of classical machine learning and dynamic programming-based approaches defines automatic cognate detection tasks as semi-supervised approaches. Hauer and Kondrak (2011) trained a linear SVM based on word similarity and language-pair features to detect cognates. Phonetic alignment based SVM models performed quite efficiently on different language families while detecting cognates (Jäger et al., 2017). Some researchers designed orthographic substring similarity measures based SVM models (Ciobanu and Dinu, 2014, 2015).

Another thread of research for cognate detection recently involved multilingual contextual knowledge injection methods. Merlo and Rodriguez (2019b) explored the effect of cross-lingual features on bilingual lexicon building, which was later implemented in Indian language cognate detection methods. Kanojia et al. (2020a) injected cross-lingual semantic features for cognate detection tasks using newly trained language models, which showed better results than state-of-the-art methods. Recently, Kanojia et al. (2021) proposed incorporating gaze features in context aware cognate detection tasks, which improved results for Hindi-Marathi language pairs. These methods may produce better results; however, the main disadvantages lie in training these models as they are data-hungry models. Thus, incorporating manually annotated gaze features and newly generated

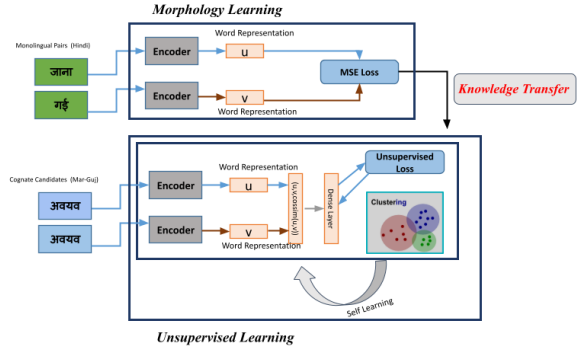


Figure 1: Weakly-supervised Cognate Detection Framework with Morphology Learner and Unsupervised Cognate Detector. For training, we pass monolingual word-pairs into morphology learners (coloured in green) and bilingual cognate candidates (coloured in blue) into unsupervised cognate detector.

cross-lingual contextual embeddings becomes an impossible solution for many under-resourced languages.

3 Cognate Detection Framework

In this section, we describe the components and working of the proposed *Language Agnostic Weakly-supervised Cognate Detection Architecture Using Morphological Knowledge*, trained using an unsupervised loss function and iterative clustering method. The iterative clustering process enhances the understanding of word representation and cluster assignment.

Figure 1 depicts the shared word encoder based framework. It consists of two parts - (i) a *Morphology Learner*, which gathers the morphological knowledge of a language and (ii) an *Unsupervised Cognate Detector*, which uses the morphological knowledge learnt using a shared word encoder to cluster the cognates between word pairs.

3.1 Word Encoder

The word encoder consists of an n-gram character-level CNN and a positional embedding layer, followed by a self-attention layer.

Character Encoding. The character level CNN layer generates word representation on n-gram ($n \in \{2,3,4,5,6\}$) characters, which helps to understand the representation of words on different sub-word levels (Goswami et al., 2020). A word’s different sub-word level representations are achieved using a 1-dimensional CNN (Zhang et al., 2015). The characters of a word are fed as input sequence $S = [w_1, w_2, \dots, w_m]$ to the 1-dimensional CNN,

where m is the number of characters present in the word and w_i is a character in the word. Considering the 1-dimensional CNN as a feature extractor, it slides over characters to create a window vector w_j with consecutive character vectors, as denoted in Equation 1.

$$w_j = [x_j, x_{j+1}, \dots, x_{j+k-1}] \quad (1)$$

where k is the size of the feature extractor filter and $x_i \in \mathbb{R}^d$ is the d -dimensional character embedding of the i -th character, where character vocabulary size is n . Thus this k -sized filters create the feature map s ($s \in \mathbb{R}^{m-k+1}$) from the window vector w_j according to Equation 2,

$$s_j = a(w_j \cdot m + b_j) \quad (2)$$

where the vector m is a filter for convolution operation, b_j is the bias for the j -th position and a is the non-linear function. Thus the new feature representation of the word F ($F \in \mathbb{R}^{(m-k+1) \times n}$) will be expressed as $F = [s_1, s_2, \dots, s_n]$, where n is the number of filters and m is the total input size. Observe the different filter size $k \in \{2,3,4,5,6\}$, representing the word's n -gram features, further represented as F_k .

Positional Encoding of Features. While getting n -gram features of the word representation, the character sequence order carries a significant role in word construction. We try to learn the different n -gram positions in a word with the help of our new introduction of positional encoding. The Transformer architecture (Vaswani et al., 2017) enforces the trainable positional embeddings on the input word pieces to understand the position of the words in a sentence. In our input words, the same character can appear in multiple positions, which makes positional embedding on each character irrelevant. Rather, our approach to learning the n -gram sequence in a word helps to understand the grammatical and morphological differences from the structural perspective of a word. Following the learning process of the positional embedding of transformer architecture, we encode the trainable positional encoding of different n -gram features.

Attention of Features. The new encoded feature representation F_k produces the ultimate word embedding, which is achieved by giving weight to n -grams according to their importance in word construction. A self-attention mechanism takes the feature representation as input and produces an out-

put weight vector a for every feature representation F using the following Equation 3.

$$a = \text{softmax}(\tanh(W_h \cdot F^T + b_h)) \quad (3)$$

The summation of feature representation F according to the weight vectors provided by a generates a vector representation r of a word by Equation 4

$$r = \sum_{i=1}^T a_i \cdot F_i \quad (4)$$

where a_i represents the attention weights, and \cdot represents the element-wise product between elements. The final vector representation r is the concatenation of different n -grams $\in \{2,3,4,5,6\}$, which is represented as $r = [r_2, r_3, r_4, r_5, r_6]$.

3.2 Morphology learner

We learn the morphological relationship between two words r_i and r_j of the same language in a Siamese setting (details of the morphological training dataset building with an example are given in Section 4). The encoded vector representations are then passed through a fully connected (FC) layer which gives two vector representations $z_l \in \mathbb{R}^{N \times K}$ and $z_r \in \mathbb{R}^{N \times K}$. The morphology learner model is then trained to minimize the mean-squared loss between two word representations such that their vector space distances reflect their degree of morphological relatedness in Equation 5

$$1/N \sum_{i=1}^N (z_{l_i} - z_{r_i})^2 \quad (5)$$

where N is the mini-batch size.

3.3 Weakly-supervised/Unsupervised Cognate Detector

The encoder with and without morphological knowledge for weakly-supervised and unsupervised methodology respectively accepts two word representations r_i and r_j from two different languages as input in a Siamese setting. The encoded vector representations are then passed through a fully connected (FC) layer which gives two vector representations $u \in \mathbb{R}^{N \times K}$ and $v \in \mathbb{R}^{N \times K}$. We concatenate the word representations u and v with their cosine similarity score and pass it through a sense layer to achieve the combined representation $z \in \mathbb{R}^{N \times K}$. It is then passed through a softmax

layer to get the probability distribution of all classes $p \in \mathbb{R}^{N \times K}$, as per Equation 6

$$p_{ij} = \frac{\exp(z_{ij})}{\sum_{t=1}^K \exp(z_{it})} \quad (6)$$

where k is the number of classes. We train the unsupervised model based on the maximum likelihood clustering loss proposed by Goswami et al. (2020), where they try to maximize the probability distribution function for each class and at the same time try to minimize the probability of all the datasets to be assigned to one class using Equation 7.

$$L_u = \sum_{i=1}^N \max_{j=1}^i p_{ij} - \max_{i=1}^N \sum_{j=1}^i p_{ij}^2 \quad (7)$$

While the unsupervised loss function helps us to get word embeddings and an initial cluster assignment, it is important to improve cluster purity according to datasets. Xie et al. (2016) proposed a self learning based deep clustering technique. The framework learns the clustering based on stochastic gradient descent (SGD) during backpropagation. We fine-tune our word embedding to learn better clustering using this iterative clustering technique. The initial sets of cluster centroids $u_{j=1}^k$ are obtained from the pre-training phase using Equation 7. In this self training phase, we assign word embeddings to initial cluster centroid and then fine-tune the word embeddings and cluster centroids using auxiliary target distribution.

The assignment of cluster centroids (u_j) and word embeddings (z_i) are calculated based on Student’s t-distribution (Maaten and Hinton, 2008) as per Equation 8

$$q_{ij} = \frac{\left(1 + \|z_i - u_j\|^2\right)^{-1}}{\sum_{j'} \left(1 + \|z_i - u_{j'}\|^2\right)^{-1}} \quad (8)$$

where q_{ij} is the probability of sample i to cluster j assignment.

We now refine the cluster learning from their high confidence assignments using auxiliary target distribution p_{ij} (Xie et al., 2016). This helps to improve cluster purity by putting more emphasis on data point assignment, as per Equation 9

$$p_{ij} = \frac{q_{ij}^2 / \sum_i q_{ij}}{\sum_{j'} \left(q_{ij}^2 / \sum_i q_{ij'}\right)} \quad (9)$$

where $\sum_i q_{ij}$ is the frequency of clusters.

Word1	Word2
nuachtán.	nuachtáin
eolas.	a eolais.
síceolaí.	leis an síceolaí.
Críostaí.	Críostaithe.

Table 1: Morphology Learning Dataset for the Irish.

Dataset	Hindi	Irish	Zulu	Sanskrit
Training Dataset.	42200	2579	49696	437675

Table 2: Morphological Training Dataset of Three Pivot Languages and Sanskrit

The cluster assignment self-learning process is trained based on KL divergence loss between assignments q_i and p_i , as shown in Equation 10.

$$L = KL(P||Q) = \sum_i \sum_j p_{ij} \log \frac{p_{ij}}{q_{ij}} \quad (10)$$

4 Training and Evaluation Dataset

The framework has two parts: **(i)** Morphology Learner and **(ii)** Weakly-supervised/Unsupervised Cognate Detector.

Morphological training is based on UniMorph (McCarthy et al., 2020) datasets. As shown in Table 1, our Siamese network accepts two words as input. The inputs are the monolingual word pairs of the pivot language in UniMorph data (as shown in Figure 1, we train the encoder of the morphological learner on the Hindi dataset in a supervised manner and transfer the knowledge to the unsupervised cognate detector for Marathi-Gujrati word pairs). Though our model is trained with the supervised dataset from Unimorph, we do not consider the annotated morphological class while training the morphological learner. The statistics of datasets for three pivot languages and Sanskrit are given in Table 2.

Cognate Detection Task We have evaluated our models on three different datasets for the cognate detection task: Indian, Celtic, and South African languages. For under-resourced Indian language pairs, we have followed the work of Kanojia et al. (2020b). As datasets for South African and Celtic languages are not easily available, we built the dataset from an open-source cognate database (Batsuren et al., 2019) and also have used the SigTyP 2023 shared task on cognate detection dataset (Rani et al., 2023). The true cognates are directly taken from the dataset and false cognate pairs are randomly shuffled word-pairs with a 60-40 split of the

Language-pairs	Cognates	Non-cognates
Hindi (Hi) - Marathi (Mr)	15726	15983
Hindi (Hi) - Gujrati (Gu)	17021	15057
Hindi (Hi) - Punjabi (Pa)	14097	15166
Hindi (Hi) - Bengali (Ba)	15312	16119
Hindi (Hi) - Tamil (Ta)	3363	4005
Hindi (Hi) - Assamese (As)	3478	4101
Irish (Ga) - Manx (Gv)	335	223
Irish (Ga) - Scottish Gaelic (Gd)	676	450
Zulu (Zu) - Xhosa (Xh)	2236	1490
Zulu (Zu) - Swati (Ss)	14	9

Table 3: Cognate dataset statistics across language-pairs.

total dataset available in the database for each language pair. We experiment with both supervised and unsupervised learning of these cognate classifiers based on the encoder that was learned in the previous step. During the training procedure of the unsupervised cognate detector, no word pair labels of cognate datasets are considered. The detailed statistics of the cognate datasets for the three language families can be found in Table 3.

For all of our experiments we have carried out *5-fold stratified cross-validation* which has helped us to get the train and test data randomly.

5 Training Details

We implemented our model using pytorch¹. The learning rate for the Indian, Celtic and South African datasets was hand-tuned to 1e-4, 2e-3 and 4e-3, respectively, for the morphological training. At the same time, for the unsupervised cognate detection tasks the learning rates were 1e-2, 1e-1, 1e-2, respectively. To stabilize the learning of the model, we have implemented LambdaLR² as the learning rate scheduler. For clustering, we have used *k*-means clustering (MacQueen, 1967) with mini-batch³.

6 Experimental Evaluation

We evaluated our framework on three different datasets in three different scenarios: (a) language pairs with pivot language and its morphological knowledge, (b) language pairs without the pivot language but with the shared encoder having morphological knowledge of the pivot language and (c) the effect of the historical language morphological knowledge transfer on the language pairs.

We compare our models with the following supervised state-of-the-art cognate detection frame-

¹<https://pytorch.org>

²<https://pytorch.org/docs/stable/optim.html>

³<https://scikit-learn.org/stable/modules/generated/sklearn.cluster.MiniBatchKMeans.html>

Approaches / Languages	Hi-Mr	Hi-Gu	Hi-Pa	Hi-Bn	Hi-Ta	Hi-As
Orthographic Similarity	0.21	0.23	0.21	0.36	0.20	0.34
(Rama, 2016)	0.69	0.67	0.47	0.65	0.53	0.71
(Kanojia et al., 2019b)	0.72	0.76	0.74	0.68	0.53	0.71
XLm-R + FFNN	0.73	0.76	0.73	0.78	0.56	0.71
Proposed Supervised Methods						
<i>Proposed-method</i>	0.81	0.79	0.80	0.79	0.69	0.78
<i>Proposed-method_{withknowledge}</i>	0.91	0.87	0.88	0.86	0.77	0.82
Proposed Weakly-supervised/Unsupervised methods						
<i>Proposed-method_{unspv}</i>	0.72	0.73	0.74	0.75	0.67	0.69
<i>Proposed-method_{wklysupv}</i>	0.85	0.84	0.81	0.82	0.74	0.79

Table 4: Results of supervised and weakly-supervised cognate detection task based on F-Score for Indian languages. The baseline performances are as reported in (Kanojia et al., 2020a).

Approaches / Languages	Zu-Ss	Zu-Xh	Ga-Gd	Ga-Gv
Orthographic Similarity	0.21	0.31	0.29	0.22
(Rama, 2016)	0.24	0.61	0.64	0.59
(Kanojia et al., 2019b)	0.23	0.74	0.72	0.61
Proposed Supervised Methods				
<i>Proposed-method</i>	0.65	0.76	0.77	0.69
<i>Proposed-method_{withknowledge}</i>	0.72	0.87	0.88	0.74
Proposed Weakly-supervised/Unsupervised methods				
<i>Proposed-method_{unspv}</i>	0.69	0.73	0.71	0.62
<i>Proposed-method_{wklysupv}</i>	0.78	0.79	0.81	0.71

Table 5: Results of supervised and weakly-supervised cognate detection task based on F-Score for South African and Celtic languages.

works: (i) **CNN based model** Siamese CNN based approach by Rama (2016), (ii) **Orthographic similarity** based approach from Labat and Lefever (2019), (iii) **Recurrent Neural Network** based approach proposed by (Kanojia et al., 2019b), (iv) **Contextual Word Embedding** based approach with XLM-R (Conneau et al., 2020) proposed by Kanojia et al. (2020a).

6.1 Pivot Language based Cognate Detection

Understanding word representation is the key to state-of-the-art deep learning frameworks for different cross-lingual cognate detection tasks. Supervised models rely on distributional learning based on annotated labels. In contrast, weakly-supervised and unsupervised frameworks should be able to learn the structural and syntactical representations of words to do clustering. We have evaluated our model on different language families, including Indo Aryan, Dravidian (Kanojia et al., 2020b), Celtic and South-African languages.

As shown in Table 4, we have evaluated our

model on six different language pairs. Our baseline model is based on the orthographic similarity approach. As expected, it does not perform well (Example: Word pair “Alankar (Ornament)” in Hindi, and “Alankaaram (Ornament)” in Tamil with similar word structures classified wrongly). The character-based CNN method proposed by Rama (2016) followed by the recurrent network-based solution proposed by Kanojia et al. (2019b) increases the model efficiency by quite a margin while detecting cognates. The contextual word embedding XLM-R based baseline model gives the best score compared to the previous models’ score. This model is capable of injecting contextual knowledge of words from a sentence. Our proposed approach has outperformed all of these baseline frameworks and achieved state-of-the-art results for supervised and weakly-supervised frameworks. As the language pairs Hindi-Marathi are very closely related, transferring the learnt Hindi morphological knowledge has increased the model’s efficiency by 18 points of F -score. We observed that our weakly-supervised framework outperformed the state-of-the-art supervised baseline system by 13 points of F -score. It is interesting to note that our supervised and unsupervised frameworks achieved state-of-the-art results by outperforming the baseline systems by 21 and 18 points of F -score, respectively, for the language pair Hindi-Tamil. Hindi and Tamil come from different language families, Indo Aryan and Dravidian, respectively, significantly boosting the efficacy of the cognate detection framework. On average, our supervised and weakly-supervised system has improved 16.8 and 9 points of F -score, respectively, on Indian language pairs.

Table 5 shows the model performance on South African and Celtic language pairs. For South African languages, we have Zulu (zu), Swati (ss) and Xhosa (xh). Irish (ga), Manx (gv), and Scottish Gaelic (gd) represent the Celtic languages. We transferred morphological knowledge of Zulu and Irish to other South African and Celtic languages, respectively. Our proposed supervised and weakly-supervised framework outperformed the state-of-the-art baseline models. We reported poor performance of the baseline models for the language pairs Zulu and Swati. Due to the lack of training data for Zulu-Swati (only 23 cognate pairs are available), the models cannot be trained effectively. On the other hand, our proposed approaches performed better than the baseline models by a large mar-

Approaches / Languages	Mr-Gu	Pa-Gu	Mr-Pa	Mr-Bn	Gd-Gv
Orthographic Similarity (Rama, 2016)	0.22	0.26	0.21	0.32	0.19
(Kanojia et al., 2019b)	0.64	0.65	0.69	0.61	0.49
	0.72	0.75	0.77	0.68	0.64
Proposed Supervised Methods					
<i>Proposed-method</i>	0.79	0.79	0.78	0.74	0.62
<i>Proposed-method_{withknowledge}</i>	0.91	0.88	0.87	0.86	0.75
Proposed Weakly-supervised/Unsupervised methods					
<i>Proposed-method_{unspv}</i>	0.72	0.71	0.74	0.70	0.59
<i>Proposed-method_{wklysupv}</i>	0.86	0.83	0.84	0.80	0.73

Table 6: Results of supervised and weakly-unsupervised cognate detection task based on F-Score for Indian and Celtic languages in the absence of Pivot Languages.

gin and very interestingly, our weakly-supervised model is better than the supervised model by 6 points of F -score. The relatively complex word pairs such as “umgqibelo (Saturday)” in Zulu and “úm-gcibélo (Saturday)” in Swati are correctly identified as cognate pairs.

These results show that the proposed cognate detection framework can efficiently detect cognates across language pairs with the morphological knowledge of the pivot language. Moreover, with little training data, both the proposed weakly-supervised and unsupervised frameworks are an efficient solution for cognate detection.

6.2 Absence of Pivot Language

We now evaluate the robustness of our transfer learning approach on the cross-lingual language pairs in the absence of the pivot language. For Indian language pairs, our pivot language is Hindi (Hi), whereas for the Celtic language pairs, the pivot language is Irish (Ga). Table 6, shows that the transfer learning approach is still very efficient when the pivot language is absent. As we can see, for the language pairs Gd-Gv, without knowledge transfer in supervised learning, the recurrent neural network approach Kanojia et al. (2019b) is better than our approach. However, with the morphological knowledge encoded for both supervised and weakly-supervised methods, our model outperforms by 11 and 9 points of F -score, respectively. On average, our transferred knowledge-based weakly-supervised method has outperformed the baseline method by 8.6 points of F -score. Thus, we can see a steady performance across all language pairs, showing the stability of the proposed morphological knowledge transfer supervised and weakly-supervised framework.

Approaches / Languages	Hi-Mr	Hi-Pa	Mr-Bn
Proposed Supervised Methods			
<i>With Hindi Knowledge</i>	0.91	0.88	0.86
<i>With Sanskrit Knowledge</i>	0.90	0.86	0.87
Proposed Weakly-supervised methods			
<i>With Hindi Knowledge</i>	0.85	0.81	0.80
<i>With Sanskrit Knowledge</i>	0.83	0.80	0.82

Table 7: Results of supervised and weakly-unsupervised cognate detection task based on F-Score for Indian languages transferring knowledge from Hindi and Sanskrit.

6.3 Knowledge of Historical Languages

In this section, we will discuss the effect of transferring knowledge from the historical language Sanskrit. Sanskrit is the historical ancestor of almost all the Indo-Aryan languages, thus making it one of the potential pivot languages to transfer the knowledge for the cognate detection task. We have studied the model’s efficiency while transferring the knowledge of Sanskrit to modern languages. In this experiment, we have taken the Indian language pairs Hindi-Punjabi, Hindi-Marathi and Marathi-Bengali. Comparing the results of the models given in Table 7, we can observe a slight dip in model efficacy in both supervised and weakly-supervised frameworks while transferring the knowledge from Sanskrit compared to transferring the knowledge from Hindi to the language pairs Hindi-Punjabi and Hindi-Marathi. However, the performance for the language pair Marathi-Bengali has improved 1.5 points in F -score on average.

We believe its performance can be attributed to the closeness and preserving more similarities in the characteristics of the language pairs Marathi and Bengali to Sanskrit than Hindi. Sanskrit is considered a highly agglutinating and morphologically rich language (Chatterji, 1926); thus, it is hard to parse it computationally. Though Marathi and Bengali are not as morphologically complex as Sanskrit, the languages in this pair are more agglutinating and morphologically richer than Hindi and Punjabi.

6.4 Statistical Significance

In this work, we hypothesise that transferring morphological knowledge of the pivot language to the closely-related languages helps to identify cognates in both supervised and weakly-supervised settings. To compute the performance of each language-pairs from Table 4, 5, 6 we run the models on

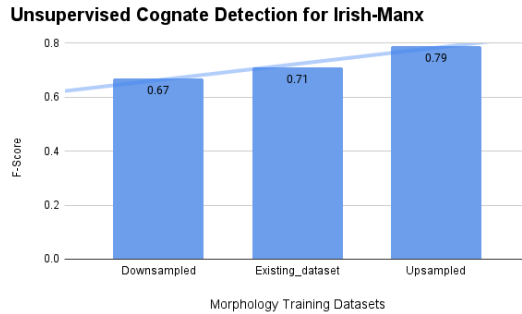


Figure 2: F-Score for Irish-Manx language pairs transferring knowledge from different morphological learners

two different settings and obtain the distribution of the performance scores: (i) we have run 5-fold cross-validations two times (which makes a total of 10 sets of results), and (ii) we kept 1-fold for a single test set and ran it 10 times for 10 different sets of results. Two sample t-tests showed that our results are statistically significant in both the cases over the baseline models ($p < 0.01$).

7 Ablation Study

Effect of morphological training dataseize. One of the challenges of morphological knowledge transfer is the efficient learning of word structure in the presence of a few morphological training datasets. As Irish has few training datasets, we have evaluated the proposed framework on different samples of the Irish-Manx dataset (Refer to Section 6). We have down-sampled and up-sampled the morphology training set to 30% compared to the existing datasets. From Figure 2, we can observe that the best F -score 0.79 is achieved when the morphological learner model is trained with 30% more data size than the original size. This emphasizes our claim of the model’s efficacy even with a slight increase in the morphological training datasets, which opens up the opportunity of implementing the weakly-supervised cognate detection framework on diverse under-resourced languages.

8 Conclusion

This paper proposed a novel *language agnostic weakly-supervised cognate detection framework based on Siamese architecture*. Experiments on three different datasets consisting of Indian languages, Celtic languages and South-African languages showcase the efficacy of our framework in understanding the structural relations between

cross-lingual words across languages. We also show that transferring morphological knowledge to closely-related word pairs with the help of a shared encoder improves the model’s efficacy in different scenarios. Our study on knowledge transfer from historical languages depicts changes in the word structures of modern languages. We demonstrate that our approach outperforms the existing supervised and semi-supervised frameworks and establishes state-of-the-art results for the cognate detection task. We also showcase the stability of learning morphology on a small number of training datasets, which opens up the possibility of deploying the system across language families.

Our future work will design and compare a semi-supervised framework based on labelled and unlabelled training sets. We will study whether the semi-supervised framework improves efficiency while detecting the cognates.

Limitations

During the evaluation, we have not experimented with choosing multiple languages as pivot languages in the same language family. So, the performance of the transfer learning framework may change depending on the choice of the pivot language. Also, during the evaluation, we mostly conducted our experiments on modern language pairs. Thus, the performance of the framework may differ for studies of historical linguistic. From the training perspective, more fine-tuning may improve the performance of the models but we have compared the results produced with the settings described in our work.

Acknowledgements

This publication was supported by a research grant from Irish Research Council: Grant IR-CLA/2017/129 (CARDAMOM-Comparative Deep Models of Language for Minority and Historical Languages) co-funded by Science Foundation Ireland (SFI) under Grant SFI/18/CRT/6223 (CRT-Centre for Research Training in Artificial Intelligence) and Science Foundation Ireland (SFI) under Grant SFI/12/RC/2289_P2 (Insight_2).

References

Khuyagbaatar Batsuren, Gabor Bella, and Fausto Giunchiglia. 2019. [CogNet: A large-scale cognate database](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*,

pages 3136–3145, Florence, Italy. Association for Computational Linguistics.

Suniti Kumar Chatterji. 1926. *Origin and development of the Bengali language*.

Alina Maria Ciobanu and Liviu P Dinu. 2014. Automatic detection of cognates using orthographic alignment. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 99–105.

Alina Maria Ciobanu and Liviu P Dinu. 2015. Automatic discrimination between cognates and borrowings. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 431–437.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. [Unsupervised cross-lingual representation learning at scale](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 8440–8451. Association for Computational Linguistics.

Michael A Covington. 1996. An algorithm to align words for historical comparison. *Computational linguistics*, 22(4):481–496.

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

Koustava Goswami, Rajdeep Sarkar, Bharathi Raja Chakravarthi, Theodorus Franssen, and John P. McCrae. 2020. [Unsupervised deep language and dialect identification for short texts](#). In *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, pages 1606–1617. International Committee on Computational Linguistics.

Bradley Hauer and Grzegorz Kondrak. 2011. Clustering semantically equivalent words into cognate sets in multilingual lists. In *Proceedings of 5th international joint conference on natural language processing*, pages 865–873.

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, Raj Dabre, Shubham Dewangan, Pushpak Bhattacharyya, Gholamreza Haffari, and Malhar Kulkarni. 2020a. [Harnessing cross-lingual features to improve cognate detection for low-resource languages](#). In *Proceedings of the 28th International*

- Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, pages 1384–1395. International Committee on Computational Linguistics.
- Diptesh Kanojia, Malhar Kulkarni, Pushpak Bhattacharyya, and Gholamreza Haffari. 2020b. [Challenge dataset of cognates and false friend pairs from indian languages](#). In *Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020*, pages 3096–3102. European Language Resources Association.
- Diptesh Kanojia, Malhar Kulkarni, Pushpak Bhattacharyya, and Gholamreza Haffari. 2019a. [Cognate identification to improve phylogenetic trees for indian languages](#). In *Proceedings of the ACM India Joint International Conference on Data Science and Management of Data, CoDS-COMAD '19*, page 297–300, New York, NY, USA. Association for Computing Machinery.
- Diptesh Kanojia, Kevin Patel, Malhar Kulkarni, Pushpak Bhattacharyya, and Gholamreza Haffari. 2019b. Utilizing wordnets for cognate detection among indian languages. In *Proceedings of the 10th Global Wordnet Conference*, pages 404–412.
- 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. 2000. A new algorithm for the alignment of phonetic sequences. In *1st Meeting of the North American Chapter of the 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.
- Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of machine learning research*, 9(Nov):2579–2605.
- J MacQueen. 1967. [Some methods for classification and analysis of multivariate observations](#). In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*, pages 281–297, Berkeley, Calif. University of California Press.
- Ranbeer Makin, Nikita Pandey, Prasad Pingali, and Vasudeva Varma. 2007. Experiments in cross-lingual ir among Indian languages. In *Proceedings of the International Workshop on Cross Language Information Processing (CLIP)*.
- Gideon S. Mann and David Yarowsky. 2001. [Multipath translation lexicon induction via bridge languages](#). In *Second Meeting of the North American Chapter of the Association for Computational Linguistics*.
- Arya D. McCarthy, Christo Kirov, Matteo Grella, Amrit Nidhi, Patrick Xia, Kyle Gorman, Ekaterina Vylomova, Sabrina J. Mielke, Garrett Nicolai, Miikka Silfverberg, Timofey Arkhangelskiy, Nataly Krizhanovsky, Andrew Krizhanovsky, Elena Klyachko, Alexey Sorokin, John Mansfield, Valts Ernštreits, Yuval Pinter, Cassandra L. Jacobs, Ryan Cotterell, Mans Hulden, and David Yarowsky. 2020. [UniMorph 3.0: Universal Morphology](#). In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3922–3931, Marseille, France. European Language Resources Association.
- I. Dan Melamed. 1999. [Bitext maps and alignment via pattern recognition](#). *Computational Linguistics*, 25(1):107–130.
- Helen M Meng, Wai-Kit Lo, Berlin Chen, and Karen Tang. 2001. Generating phonetic cognates to handle named entities in English-Chinese cross-language spoken document retrieval. In *IEEE Workshop on Automatic Speech Recognition and Understanding, 2001. ASRU'01.*, pages 311–314. IEEE.
- Paola Merlo and Maria Andueza Rodriguez. 2019a. [Cross-lingual word embeddings and the structure of the human bilingual lexicon](#). In *Proceedings of the 23rd Conference on Computational Natural Language Learning, CoNLL 2019, Hong Kong, China, November 3-4, 2019*, pages 110–120. Association for Computational Linguistics.
- Paola Merlo and Maria Andueza Rodriguez. 2019b. Cross-lingual word embeddings and the structure of the human bilingual lexicon. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 110–120.
- Andrea Mulloni and Viktor Pekar. 2006. [Automatic detection of orthographics cues for cognate recognition](#). In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy. European Language Resources Association (ELRA).
- 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.

Priya Rani, Koustava Goswami, Adrian Doyle, Theodorus Fransen, Bernardo Stearns, and John P. McCrae. 2023. [Findings of the SIGTYP 2023 shared task on cognate and derivative detection for low-resourced languages](#). In *Proceedings of the 5th Workshop on Research in Computational Linguistic Typology and Multilingual NLP*, pages 126–131, Dubrovnik, Croatia. Association for Computational Linguistics.

Jorg Tiedemann. 1999. [Automatic construction of weighted string similarity measures](#). In *1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. [Attention is all you need](#). In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Junyuan Xie, Ross Girshick, and Ali Farhadi. 2016. Unsupervised deep embedding for clustering analysis. In *International conference on machine learning*, pages 478–487.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Advances in neural information processing systems*, pages 649–657.