Can Cognate Prediction Be Modelled as a Low-Resource Machine Translation Task?

Clémentine Fourrier **Rachel Bawden** **Benoît Sagot**
Inria, France
firstname.lastname@inria.fr

Abstract

Cognate prediction is the task of generating, in a given language, the likely cognates of words in a related language, where cognates are words in related languages that have evolved from a common ancestor word. It is a task for which little data exists and which can aid linguists in the discovery of previously undiscovered relations. Previous work has applied machine translation (MT) techniques to this task, based on the tasks' similarities, without, however, studying their numerous differences or optimising architectural choices and hyper-parameters. In this paper, we investigate whether cognate prediction can benefit from insights from low-resource MT. We first compare statistical MT (SMT) and neural MT (NMT) architectures in a bilingual setup. We then study the impact of employing data augmentation techniques commonly seen to give gains in low-resource MT: monolingual pretraining, backtranslation and multilinguality. Our experiments on several Romance languages show that cognate prediction behaves only to a certain extent like a standard low-resource MT task. In particular, MT architectures, both statistical and neural, can be successfully used for the task, but using supplementary monolingual data is not always as beneficial as using additional language data, contrarily to what is observed for MT.

1 Introduction

The Neogrammarians (Osthoff and Brugmann, 1878) formalised one of the main hypotheses of the then recent field of comparative linguistics, the regularity of sound changes: if a phone in a word, at a given moment in the history of a given language, evolves into another phone, then all occurrences of the same phone in the same phonetic context in the same language evolve in the same way.

Sound changes are usually identified by looking at the attested (or hypothesised) phonetic form of specific sets of words, called *cognates*, whose definition varies in the literature depending on the field.[1] We use an extension of the customary definition used in historical linguistics, as described for instance in (Hauer and Kondrak, 2011; List et al., 2017), which is the following: given two languages with a common ancestor, two words are said to be cognates if they are an evolution of the same word from said ancestor, having undergone the sound changes characteristic of their respective languages' evolution. We extend it by also allowing the ancestor word (from the parent language) to also be considered a cognate. For example, Latin *bonus* 'good' gave Italian *buono* 'id.', Spanish *bueno* 'id.' and Spanish *bono* 'id.' by inheritance, and they are all cognates, whereas Spanish *abonar* 'to fertilise', obtained by derivation, is related but not a cognate (Figure 1).

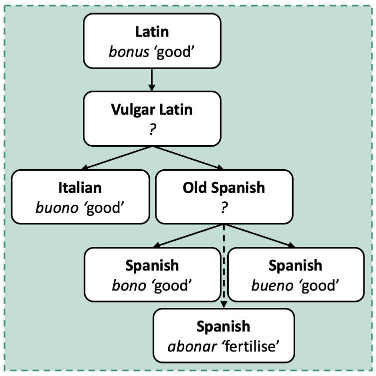

Figure 1: Related words in Italian and Spanish, both the outcome of Latin *bonus* 'good'. Plain arrows represent inheritance, dotted arrows derivation. "*?*" indicates that the word is not present in our database.

Cognate identification (finding cognate pairs in a multilingual word set) and prediction (producing

[1] Cognates have for example been defined as words sharing spelling and meaning, regardless of their etymology (Frunza and Inkpen, 2006, 2009), or words etymologically related no matter the relation (Hämäläinen and Rueter, 2019).

likely cognates in related languages) are two of the fundamental tasks of historical linguistics. Over the last three decades, automatic cognate identification has benefited from advances in computational techniques, first using dictionary-based methods (Dinu and Ciobanu, 2014) and purely statistical methods (Mitkov et al., 2007; McCoy and Frank, 2018), then statistical methods combined with clustering algorithms (Hall and Klein, 2010, 2011; List et al., 2017; St Arnaud et al., 2017), statistical methods combined with neural classifiers (Inkpen et al., 2005; Frunza and Inkpen, 2006, 2009; Hauer and Kondrak, 2011; Dinu and Ciobanu, 2014) and neural networks only (Ciobanu and Dinu, 2014; Rama, 2016; Kumar et al., 2017; Soisalon-Soininen and Granroth-Wilding, 2019).

Automatic cognate prediction is less studied despite its interesting applications, such as predicting plausible new cognates to help field linguists (Bodt et al., 2018) and inducing translation lexicons (Mann and Yarowsky, 2001). In the last few years, it has been approached as an MT task, as it can be seen as modelling sequence-to-sequence correspondences. Using neural networks has been promising (Beinborn et al., 2013; Wu and Yarowsky, 2018; Dekker, 2018; Hämäläinen and Rueter, 2019; Fourrier and Sagot, 2020a), although in most works the hyper-parameters of the neural models were not optimised. Moreover, the differences between MT and cognate prediction have not been studied.

In this paper, we choose to study the application of MT approaches to the cognate prediction task. Our aim is to investigate whether the task can benefit from techniques commonly seen to improve standard low-resource MT. We first highlight the specific characteristics of cognate prediction, and (to our knowledge) provide the first detailed analysis of the expected differences with standard MT. We then compare MT architectures (bilingual SMT vs. bilingual and multilingual NMT) when applied to cognate prediction. We study how to leverage extra data in our NMT models, either monolingual (via backtranslation or pretraining) or multilingual (introducing new languages). We experiment with Latin and its Romance descendants Spanish and Italian for all our experiments, as well as added French and Portuguese in a data augmentation setting. We find that cognate prediction is only similar to standard MT to a certain extent: the task can be modelled well using standard MT architectures (adjusted for a low-resource setting), and extending neural architectures to a multilingual setting significantly improves the results. In such multilingual settings, further improvements can be obtained by leveraging data from extra languages. However, using extra monolingual data via backtranslation or pretraining is not always as beneficial as it is in standard MT settings.[2]

2 Related Work

2.1 Cognate Prediction

Cognate prediction is the task that aims to produce from words in a source language plausible cognates in a target language (according to the aforementioned definition of cognates). It is a lexical task that models regular, word-internal sound changes that transform words over time. It has been approached with phylogenetic trees combined with stochastic sound change models (Bouchard et al., 2007; Bouchard-Côté et al., 2009; Bouchard-Côté et al., 2013), purely statistical methods (Bodt et al., 2018), neural networks (Mulloni, 2007), language models (Hauer et al., 2019) and character-level MT techniques (Beinborn et al., 2013; Wu and Yarowsky, 2018; Dekker, 2018; Hämäläinen and Rueter, 2019; Fourrier and Sagot, 2020a; Meloni et al., 2021), because of its similarity to a translation task (modelling sequence-to-sequence crosslingual correspondences between words).

2.2 Low-resource MT

Since data is scarce, we postulate that cognate prediction could benefit from low-resource MT settings techniques and architectural choices.

2.2.1 Architecture Comparison

Several papers comparing SMT with NMT (recurrent neural networks (RNNs) with attention) in low-resource settings conclude that SMT performs better, being more accurate and less prone to overfitting (Skadiņa and Pinnis, 2017; Dowling et al., 2018; Singh and Hujon, 2020). However, as Dowling et al. (2018) themselves note, they did not optimise hyper-parameters for NMT. Sennrich and Zhang (2019) analysed and reproduced previous comparisons, to conclude that SMT can actually be outperformed by NMT when architectures and hyper-parameters are carefully chosen, but only above a certain quantity of data.

[2] Both our code and data are freely available at http://github.com/clefourrier/CopperMT.

2.2.2 Leveraging Extra Data

Several techniques are commonly used in low-resource MT to mitigate the lack of parallel data: monolingual pretraining, backtranslation and using data from additional languages.

Monolingual pretraining (unsupervised) has, as in other NLP tasks, been highly beneficial to MT (Song et al., 2019; Conneau and Lample, 2019; Devlin et al., 2019; Liu et al., 2020). Before training on a translation task, model parameters are first pretrained using a language modelling objective, which enables the exploitation of monolingual data, more freely available than bilingual data.

Backtranslation originated in SMT (Bertoldi and Federico, 2009; Bojar and Tamchyna, 2011), and has been standard in NMT for several years (Sennrich et al., 2016; Edunov et al., 2018). Its goal is to artificially create larger quantities of parallel data from monolingual datasets, which are often more readily available. Target-side monolingual data is provided to a bilingual model trained in the opposite direction (target-to-source), which produces synthetic source-side data. The data is then filtered to keep the highest quality sentences. The newly generated dataset, made of synthetic source-side data parallel to real target-side data is then combined with the original bilingual set to train a new model.

Training multilingual NMT models has been shown to help low-resource scenarios by providing data in other languages and constraining the hidden representations to a shared, language-independent space. The amount of sharing between languages varies according to the approach, from multi-encoder, multi-decoder architectures (Luong et al., 2016), optionally sharing attention mechanisms (Firat et al., 2016a), to approaches with a single shared encoder and decoder (Ha et al., 2016; Johnson et al., 2017).

3 Differences between Cognate Prediction and MT

Cognate prediction and MT both focus on learning sequence-to-sequence correspondences. However, amongst the works using MT techniques for cognate prediction, little attention has been paid to their differences; the underlying linguistic assumptions and aims are quite distinct, which could impact the transferability of choices and techniques from MT.

Representation Units MT processes sentences split into individual (graphemic) units that can be of diverse granularity levels (characters, subwords or words). Cognate prediction, on the other hand, involves predicting sound correspondences from one cognate word to another, and so is best modelled using sequences of phones (like character-level MT).

Reordering and Alignment In MT, the correspondence between source and target sentences can involve long-distance reorderings, whereas the reorderings sometimes found in the correspondence between cognates are almost always local (e.g. metatheses). We therefore expect SMT, which is somewhat limited with respect to the modelling of long-distance context, to be less penalised in the cognate prediction setting than it usually is a standard MT setting.

Sample Length The input sequence to MT is the sentence, whereas for cognate prediction it is the word. Even with different segmentation granularities for MT, the average sequence length is generally much shorter for cognate prediction than for MT. Again, this could mean that SMT is less penalised than it is in the standard MT setup.

Modelled Relations MT involves symmetrical relations between sentences, whereas cognate prediction, as defined above, is inherently ambiguous in a counter-intuitive way (especially because it is structurally different from the usual MT ambiguity, where many valid translations exist for the same input). The cognate task models both symmetrical and asymmetrical relationships between cognates: parent-to-child (e.g. LA→ES), i.e. modelling sequences of regular sound changes, is non-ambiguous, whereas child-to-parent (e.g. ES→LA) and as a result, child-to-child (e.g. IT↔ES) is intrinsically ambiguous, as two distinct sounds in the parent language can result in the same outcome in the child language. When two distinct sounds in the child language are the outcome of the same sound in the parent language, it is always because their (word-internal) phonetic contexts were different in the parent language. In other words, the parent-to-child direction is (virtually) non-ambiguous, but might require taking the phonetic context into account. However, the child-to-parent direction is intrinsically ambiguous, which results from the fact that a sound in the child language can be the regular outcome of more than one sound in the parent language: for instance Spanish /b/ comes from Latin /p/ in *abría* (from Latin *aperīre*) but from Latin /b/ in *habría* (from Latin *habeō*).

Ambiguity Management When using cognate prediction as a tool to aid linguists, as in (Bodt and List, 2019), our aim is not to predict **the** single correct answer, but to provide a list of plausible candidates. In MT however, while many translations can be produced by the model (some better than others—including poor ones), it is possible to simply use the best ranked translation. In cognate prediction, as a consequence of the inherent ambiguity of the task discussed above, at most one prediction is correct, other predictions could have been correct (i.e. they are compatible with the phonetic laws involved), while other predictions are incorrect. A linguist would be interested in all correct or plausible predictions, not just the best ranked one, and there is therefore a need for n-best prediction.

Relevance of Leveraging Extra Data Whereas MT models could theoretically be trained on any sentence pair that are translations of each other, cognate prediction is far more limited in terms of which data can be used; cognacy relations only link a limited number of words in specific language pairs, limiting not only available parallel data but also the potential for synthetic data (e.g. via backtranslation). Using generic translation lexicons may help, but, as they do not only contain cognate pairs, all non-cognate pairs they contain (parallel borrowings from a third language and etymologically unrelated translations) are effectively noise for our task (Fourrier, 2020).

4 Experimental setup

Bearing in mind these differences, we seek to determine whether MT architectures and techniques are well suited to tackling the task of cognate prediction, paying attention to avoid the pitfalls raised by Sennrich and Zhang (2019) by carefully selecting architecture sizes and other hyper-parameters.

For our baselines, we train several character-level[3] MT models (SMT vs. RNNs and Transformers) in a bilingual setup, training a single model for each language pair.

We then assess the impact of techniques commonly used to improve MT in low-resource scenarios. We first investigate the impact of using monolingual data for all 3 architecture types, via pretraining and backtranslation,[4] then take advantage of the ability of NMT to accommodate multilingual architectures to experiment with a multi-encoder multi-decoder architecture (Firat et al., 2016b) involving all language directions.

Finally, we test whether there can be any benefit from combining multilinguality with either pretraining or backtranslation.

4.1 Data

Our datasets (detailed below) are bilingual cognate lexicons for all our experiments, extended with monolingual lexicons for backtranslation and pretraining (see Table 1). As we focus on sound correspondences, we phonetise our datasets. Each word is phonetised into IPA using `espeak` (Duddington, 2007-2015), then cleaned to remove diacritics and homogenise double consonant representations. For example, *conocer* 'to know' is phonetised as [konoθɛɾ], then split into phones (segmented into [k, o, n, o, θ, ɛ, ɾ]).

Bilingual	LA–IT	LA–ES	ES–IT
#words	5,109	4,271	1,804
#phones	77,771	63,131	24,576
#Unique phones	34	39	38
Avg. word length	7.62	7.40	6.81
Monolingual	**ES**	**IT**	**LA**
#words	78,412	99,949	18,639
#phones	626,175	815,562	142,955
#Unique phones	38	40	29
Word length	7.98	8.24	7.67

Table 1: Dataset statistics for our lexicons.

Bilingual Lexicons Our bilingual cognate lexicons contain cognate pairs (between 657 and 5,109 depending on the language pair), extracted from the etymological database EtymDB2 (Fourrier and Sagot, 2020b) following (Fourrier and Sagot, 2020a) for extraction and duplicate management, then phonetised as described above. We use the lexicons containing Latin, Italian and Spanish (LA–IT, LA–ES, ES–IT) in all experiments. They respectively contain 5,109, 4,271 and 1,804 words for a total of 77,771, 63,131 and 24,576 phones (ES–IT being considerably smaller), with on average 40 different and unique phones.

We run all experiments on three different train/dev/test splits in order to obtain confidence scores. For the bilingual (baseline) and multilingual setups, each split is obtained by sampling sentences 80%/10%/10% randomly.

[3] We use here the customary term "character-level MT," although in our case, characters correspond to phones.

[4] We detail what pretraining and backtranslation means for our SMT models in Section 4.4.

Monolingual Lexicons Monolingual datasets are used for the monolingual pretraining and back-translation experiments. They were extracted from a multilingual translation graph, YaMTG (Hanoka and Sagot, 2014), by keeping all unique words for each language of interest. To remove noise, words containing non-alphabetic characters were discarded (punctuation marks, parentheses, etc.). The final datasets (cleaned and phonetised) contain between 18,639 and 99,949 unique words (the LA set is more than 4 times smaller than the others).

4.2 MT Architectures

4.2.1 SMT

We train a separate SMT model for each language direction using the MOSES toolkit (Koehn et al., 2007). Our bilingual training data is aligned with `GIZA++` (Och and Ney, 2003). The target data for the pair is used to train a 3-gram language model using `KenLM` (Heafield, 2011). We tune our models using `MERT` based on BLEU on the dev set.

4.2.2 NMT

We compare two encoder-decoder NMT models: the RNN (bi-GRU) with attention (Bahdanau et al., 2015; Luong et al., 2015) and the Transformer (Vaswani et al., 2017). We use the multilingual Transformer implementation of `fairseq` (Ott et al., 2019), and extend the library with an implementation of the multilingual RNN with attention (following the many-to-many setting from (Firat et al., 2016a) but with separate attention mechanisms for each decoder).[5] Each model is composed of one encoder per input language, and one decoder (and its own attention) per output language.[6] We train each model for 20 epochs (which is systematically after convergence), using the Adam optimiser (Kingma and Ba, 2015), the cross-entropy loss, and dev BLEU as selection criterion.

4.2.3 Hyper-parameter Selection

We ran optimisation experiments for all possible bilingual and multilingual architectures, using three different data splits for each parameter combination studied, and choosing the models performing best across seeds. Our initial parameters were selected from preliminary experiments (in bold in Table 2).

Parameters	Values studied
1) Learning rate × Batch size	{0.01, **0.05**, 0.001} × {10, 30, **65**, 100}
2) Embed. dim. × Hidden dim.	{8, 12, 16, **20**, 24} × {18, 36, **54**, 72}
3) Number of layers	**1**, 2, 4
4) Number of heads	**1**, 2, 3, 4
4) Attention type	None, Bahdanau, Luong (**dot**, concat, general)

Table 2: Parameter exploration experiments for NMT models. In bold, the initial parameters at each step.

Table 2 contains the successive parameter exploration steps: at the end of a step, we automatically selected (according to average dev BLEU) the step-best value, used as input parameter for the next parameter exploration step.[7] The final best parameters are given in Appendix A.1. Smaller learning rates (0.005 and 0.001) are better, while there is no observable pattern to the best batch sizes or numbers of layers. Interestingly, however, for the RNNs, the best results are obtained with the highest hidden dimension irrespective of the embedding size (72 vs. 20 or 24), whereas, for the Transformers, best results are obtained with the largest embedding size irrespective of the hidden dimension (24 vs. 54 or 72). Increasing the number of layers or using more than 1 head almost always increases performance.

4.3 Evaluation

For our task, we use the most commonly used MT evaluation metric, BLEU (Papineni et al., 2002), using the `sacreBLEU` implementation (Post, 2018). It is based on the proportion of 1- to 4-grams in the prediction that match the reference.

In standard MT, BLEU can under-score the many valid translations that do not match the reference. For cognate prediction, however, we expect a single correct prediction in most cases (there are a few exceptions such as variants due to gender distinctions specific to the target language). This makes BLEU better suited to the cognate prediction task than it is to standard MT.[8]

[5]These implementations are used in all setups, bilingual (using one language as source and one as target) as well as multilingual.

[6]In a multilingual setup, encoders, decoders and attention mechanisms can either be shared between languages or be language-specific. In preliminary experiments, using independent items proved to be the most effective. We also observe that a coherent phonetic embedding space is learned during training (described in Appendix A.2).

[7]When looking at multilingual models, we chose the model performing best on most languages, as measured by comparing the sum of the ranks (according to their average performance per language) of each model over all language pairs.

[8]BLEU is also more adapted than an exact match, as it allows us to compare how close the prediction is to the reference and in cognate prediction does not suffer from the same problems as in standard MT.

4.4 Leveraging Extra Data

Monolingual pretraining For NMT, one way to take advantage of additional monolingual data is to teach the model to "map" each language to itself by using an identity function objective on the monolingual data for the model's target language.[9] Using monolingual target data during pretraining allows each target decoder to have access to more target data (which avoids overfitting), while we expect it to be beneficial to encoders too, since our source and target languages tend to share common sound patterns in cognate prediction, being closely related. In practice, we pretrain the model for 5 epochs[10] using the identity function objective together with the initial cognate prediction objective (on the original bilingual data) and then fine-tuned on the cognate task as before for 20 epochs.

For SMT, model parameters cannot be pretrained as in NMT, so in the guise of pretraining, we take the nearest equivalent: we use target-side monolingual data to train an extra language model.

For each language pair, the monolingual dataset we use is composed of 90% of the target monolingual data. The bilingual data is the same as before.

Backtranslation For each architecture type, we use the previously chosen models to predict 10-best results for each seed from the monolingual target-side data, and construct synthetic cognate pairs from monolingual lexicons and source-side predictions. For each word, we keep the first prediction of the 10 that also appears in the relevant monolingual source language lexicon as our new source, and the initial source as target (this is akin to filtering back-translated data (e.g. to in-domain data) in MT, a standard practice). We discard pairs with no prediction match.

This large back-translated bilingual dataset is extended with our original training set. For NMT, it is used to train a new model for 10 epochs,[10] which is then fine-tuned for 20 epochs with the original bilingual training set. For SMT, it is used (instead of the original bilingual data) to train a new phrase table.

Multilingual NMT We exploit the fact that NMT can readily be made multilingual by training a single model on all language directions at once.

[9]For the multilingual model, this means that every encoder will see data from all languages, whereas each decoder will only see data from its specific language.

[10]This number of epochs is systematically big enough to reach convergence.

5 Results

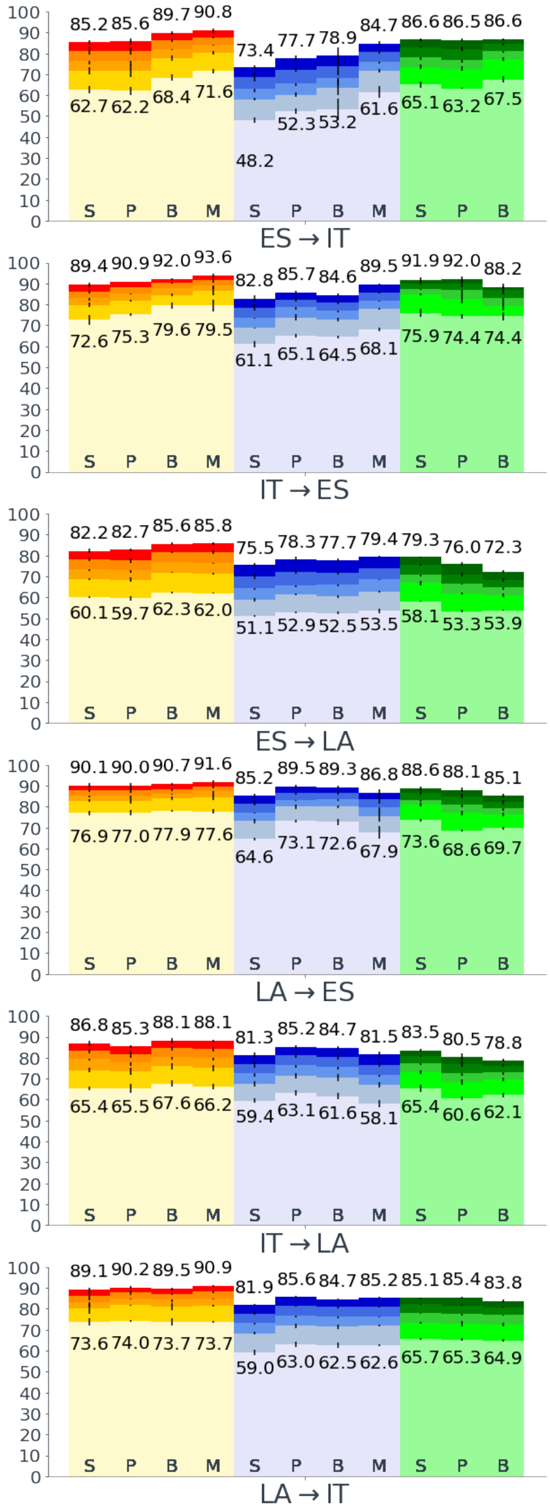

Figure 2: BLEU scores. Colours indicate the model type: RNNs in orange (col 1 to 4), Transformers in blue (col 5 to 8), SMT in green (col 9 to 11). Colour shades indicate the value of n in n-best predictions (1, 2, 3, 5 and 10 from bottom to top). The letters (x-axis) indicate the setup: S - standard/bilingual, P - with pretraining, B - with backtranslation, M - multilingual.

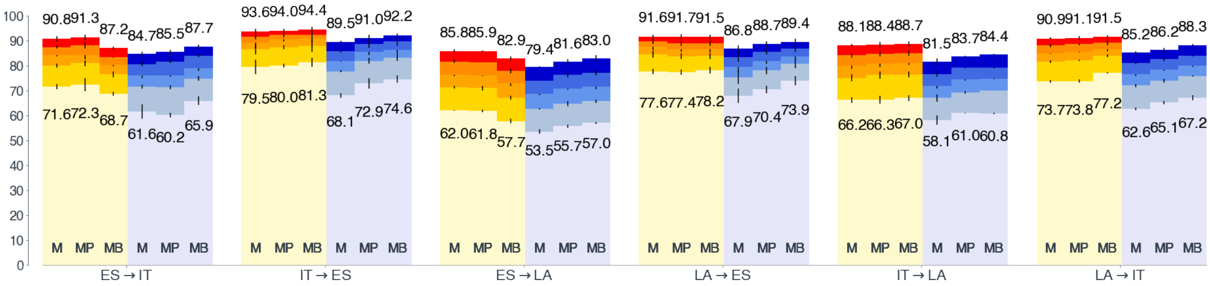

Figure 3: BLEU scores: RNNs in orange (col 1 to 3), Transformers in blue (col 4 to 6). Colour shades indicate the value of n in n-best predictions (1, 2, 3, 5 and 10 from bottom to top). On the x-axis, the letters indicate the setup: M - multilingual, MP - multilingual with pretraining, MB - multilingual with backtranslation.

5.1 Baseline: Bilingual setup

1-best Results At a first glance (Figure 2, "S" columns), SMT and RNN appear to have relatively similar results, varying between 58.1 and 76.9 BLEU depending on the language pair, outperforming the Transformer by 5 to 15 points on average. However, SMT performs better for IT↔ES (pair with the least data), and RNNs for the other pairs. This confirms results from the literature indicating that SMT outperforms NMT when data is too scarce, and seems to indicates that the data threshold at which NMT outperforms SMT (for our Romance cognates) is around 3,000 word pairs for RNNs, and has not been reached for Transformers.

n-best Results The BLEU scores for NMT and SMT increase by about the same amount for each new n ($n \leq 10$), reaching between 79.3 and 91.9 BLEU score at $n = 10$ for RNN and SMT. The Transformer, however, does not catch up.

5.2 Leveraging Extra Data

5.2.1 Pretraining, backtranslation

Both pretraining the models and using backtranslation (Figure 2, "P" and "B" columns) increase the results of the Transformer models by 1 to 9 points, though they are still below the RNN baseline. It is likely the added monolingual data mitigates the effect of too scarce bilingual sets. The impact on RNN performance is negligible for most language pairs, apart from the lowest resourced one (ES–IT), for which backtranslation increases results. Lastly, these methods seem to mostly decrease SMT performance, due to noisy data diluting the original (correct) bilingual data (cf. Section 3); this is less of a problem for NMT models, because they are then fine-tuned on the cognate task specifically.

5.2.2 Multilinguality

Data augmentation through a multilingual setup (Figure 2, "M" columns) seems to be the most successful data augmentation method for RNNs (increasing performance almost all the times), and allows them to finally outperform bilingual SMT for the least-resourced pair as well (ES↔IT). The Transformers benefit less from this technique than from adding extra monolingual data, apart for ES↔IT, most likely for the same reason as earlier: this dataset being the smallest, adding words in ES and IT from other language pairs helps to learn the translation and stabilises learning. This technique is not applicable to SMT.

Impact of the Translation Direction We observe that the for child-to-parent directions (ES or IT to LA), backtranslation is almost as successful as multilinguality for RNNs. For child-to-child directions (ES to IT and IT to ES), the SMT baseline performs best in the baseline setup (with parallel data only), although RNNs surpass it when backtranslation or multilinguality is used. For the parent-to-child directions (LA to ES or IT), multilinguality helps RNNs slightly more than backtranslation.

5.3 Combining data augmentation methods

We choose to combine the best performing data augmentation technique overall, multilinguality, with pretraining and backtranslation (Figure 3) for our NMT models.

Multilinguality + pretraining Combining multilinguality with pretraining has virtually no significant impact on the RNNs' results with respect to multilinguality only. For the Transformers, however, it increases the results by 2 to 3 BLEU on average.

Source context	Target context	Source	Target	SMT pred.	RNN pred.
(a) AMBIGUOUS SOUND CORRESPONDENCE NOT LEARNED WELL					
corvino 'raven'	*corvino* 'id.'	[kɔrβino]	[korviːno]	[kɔrviːno]	[kɔrbiːno]
liebre 'hare'	*lepre* 'id.'	[lieβre]	[lɛːpre]	[liebrɛː]	[liɛːvre]
(b) MEANING AND/OR FORM OF THE COGNATES CHANGED TOO MUCH					
calaña 'kind/sort'	*quale* 'what/which'	[kalaɲa]	[kwaːle]	[kalaɲː]	[kalaɲː]
pie 'foot'	*piede* 'id.'	[pje]	[pjɛːde]	[pɛ]	[pɛːre]
(c) DATA ERROR					
suspirar 'to sigh'	*squillan* '(it) rings'	[suspiɾar]	[skwilːan]	[sospiraːre]	[sospiraːre]
frenesí 'frenzy'	*frenetico* 'frenetic'	[frenesi]	[frenɛtiko]	[freːnezi]	[fronɛsː]
(d) MODEL MISTAKE					
licencioso 'licencious'	*licenciozo* 'id'	[liθɛnθjoso]	[litʃentsioːzo]	[litʃɛntsioːzo]	[litʃɛntso]

Table 3: Prediction errors examples across ES→IT datasets for both SMT and RNN.

Multilinguality + backtranslation Combining multilinguality with backtranslation provides the best results overall for Transformers (both being the best performing methods for these models). For the RNNs, however, the performance increase is smaller for most languages, and we even observe a decrease in performance when translating from ES (which was not the case with bilingual models).

6 Discussion

We discuss the results of the best performing models for the best seed across all architectures (SMT, *multilingual* + *pretraining* RNN and *multilingual* + *backtranslation* Transformer) from ES→IT. More than a third of the predicted words are above 90 BLEU[11] (resp. 35.4/46.4/38.1% for SMT/RNN/Transformer), and for error analysis, we study the words below this threshold. The observations generalise to other language pairs.

6.1 Predictions

Close Results We observe a lot of inaccurate but very close translations (e.g. Spanish *conveniente* 'convenient', phonetised [kɔmbenjɛnte], was predicted as corresponding to Italian [konvenjɛnte] instead of [konveniɛnte], with only one phone different, and coherently so). Sometimes these translations have a very bad score: Spanish *pulpito* 'pulpit', phonetised [pulpito], was predicted as [pʊlpiːto] instead of [pulpito], two close pronunciations, for a sentence BLEU score of only 20.

Analysis of wrong results Wrongly predicted cognates correspond to four cases, as defined in Table 3.[12] We carried out a manual error analysis, and observed that their distribution was similar across models (resp. SMT/RNN/Transformer):

(a) 84.6/81.4/79.5% were cognates with an ambiguous sound correspondence (e.g. Spanish [β] to Italian [b/v/p]).

(b) 10.3/13.4/11.6% were cognates that had either evolved too far away from one another or contain rare sound correspondences, such as *pie* 'foot', phonetised [pje], predicted [pe] and [peːre] instead of [pjɛːde] *piede* 'foot'.

(c) 0.9/0.9/0.9% corresponded to data errors, such as *suspirar* 'to sigh', phonetised as [suspiɾar], which was predicted as [sospiraːre] *sospirare* 'to sigh', its actual cognate, instead of its erroneous counterpart in our database ([skwilːan] *squillan* '(it) rings').

(d) 4.3/4.3/8.0% were model errors, such as "forgetting" part of a word during translation.

6.2 Usefulness of n-best results

The average position at which the best prediction (according to dev BLEU) occurs (in 10-best predictions) is between 1 and 3 (Table 7 in Appendix A.3). The lowest indices occur for Spanish (between 1 and 1.7) and Italian (between 1.6 and 2.2). The highest indices encountered occur when going for IT→LA or ES→LA (between 2 and 3). This illustrates the importance of n-best prediction when predicting cognates from child to parent languages, due to ambiguity. Standard deviations are between 2 and 3: for these languages, when studying cognate prediction, it is interesting to at least check the 5-best results.

[11] To study the BLEU of individual words, we use the sentenceBLEU function from `sacreBLEU` with its default parameters.

[12] Statistics are provided for the best models of the best seed, but examples are taken across seeds and models.

6.3 Language choice in a multilingual setup

To study the impact of the language pairs used in the multilingual setup, we train additional multilingual neural models on only 1000 pairs of ES–IT data (single set), complemented by either nothing (to act as baseline), an extra 600 pairs of ES–IT, or 600 pairs of ES–L and IT–L (L being either Latin, a parent language, French, a related language, or Portuguese, more closely related to Spanish than Italian). The rest of the data (Table 4) is split equally between dev and test.

BILINGUAL	FR–IT	FR–ES	PT–IT	PT–ES
#words	666	657	1,503	1,874
#phones	8,698	8,530	20,738	25,867
#Unique phones	40	43	36	41
Word length	6.53	6.49	6.90	6.90

Table 4: Supplementary bilingual lexicon statistics.

As we saw in Section 5.1, the Transformers' scores are far more affected by low resource settings than the RNNs. We therefore study the impact of adding extra languages with RNNs only.

BASELINE	ES→IT	IT→ES
1000 pairs	53.9 ± 3.4	66.6 ± 4.2
ADDED DATA	ES→IT	IT→ES
Same language pair	62.5 ± 2.5	71.8 ± 1.7
Latin	57.1 ± 1.8	67.4 ± 3.3
French	58.5 ± 2.0	67.0 ± 2.8
Portuguese	58.8 ± 1.1	66.9 ± 2.9

Table 5: BLEU for different multilingual settings.

Results on our new low-resourced baseline are lower than the our previous baselines by around 10 points (Table 5), which is expected, since we use less data for training.

Adding 600 pairs of ES–IT words has more effect on ES–IT performance than adding any other pair of related languages, which indicates that, unsurprisingly, the best possible extra data to provide is in the language pair of interest. When adding a related extra language, the results are better than with the initial data only. From Spanish, the performance is best when adding Portuguese, its most closely related language, then French, then Latin. From Italian, we observe the opposite trend. Adding an extra language seems to help most to translate from, and not to, the language it is most closely related to. For very low-resource settings, where extra pairs of the languages of interest might not be available, it will probably be interesting to explore using extra languages related to the source language.

7 Conclusion

We examined the differences between cognate prediction and MT, in terms of data as well as underlying linguistic assumptions and aims. We then observed that, above a certain training data size, SMT and multilingual RNNs provide the best BLEU scores for the task, SMT still being unrivalled when it comes to smaller datasets (which coincides with previous work comparing SMT and NMT for low-resource settings).

When studying how to increase the amount of training data seen by our models, we found that exploiting the multilinguality of NMT architectures consistently provided better results than adding monolingual lexicons (through pretraining or back-translation), which contain noise for our task; combining the methods provided a significant amelioration for Transformers only. Adding multilingual data by training with extra languages also proved interesting, and we found the best possible extra data to add in a multilingual setting is, first, data from the languages at hand, followed by pairs between them and a parent language, then finally data from additional languages as close as possible to the source language.

We conclude that cognate prediction can benefit from certain conclusions drawn in standard low-resource MT, but that its specificities (intrinsic ambiguity which requires n-best prediction, reliance on cognate data only) must be systematically taken into account. Computational cognate prediction using MT techniques is a field in its infancy, and the work in this paper can be extended along several axes: working on less studied language families, or using the method in collaboration with linguists to better understand the etymology and history of languages.

Acknowledgments

This work was granted access to the HPC resources of IDRIS under the allocation 2021-AD011011459R1 made by GENCI. This work was partly funded by the last author's chair in the PRAIRIE institute funded by the French national agency ANR as part of the "Investissements d'avenir" programme under the reference ANR-19-P3IA-0001.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proceedings of the 3rd International Conference on Learning Representations*.

Lisa Beinborn, Torsten Zesch, and Iryna Gurevych. 2013. Cognate production using character-based machine translation. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 883–891, Nagoya, Japan. Asian Federation of Natural Language Processing.

Nicola Bertoldi and Marcello Federico. 2009. Domain adaptation for statistical machine translation with monolingual resources. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 182–189, Athens, Greece. Association for Computational Linguistics.

Timotheus A. Bodt, Nathan W Hill, and Johann-Mattis List. 2018. Prediction experiment for missing words in Kho-Bwa language data. *Open Science Framework Preregistration*.

Timotheus A. Bodt and Johann-Mattis List. 2019. Testing the predictive strength of the comparative method: an ongoing experiment on unattested words in Western Kho-Bwa languages. *Papers in Historical Phonology*, 4:22–44.

Ondřej Bojar and Aleš Tamchyna. 2011. Improving translation model by monolingual data. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 330–336, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Alexandre Bouchard, Percy Liang, Thomas Griffiths, and Dan Klein. 2007. A probabilistic approach to diachronic phonology. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 887–896, Prague, Czech Republic. Association for Computational Linguistics.

Alexandre Bouchard-Côté, Thomas L. Griffiths, and Dan Klein. 2009. Improved reconstruction of protolanguage word forms. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 65–73, Boulder, Colorado. Association for Computational Linguistics.

Alexandre Bouchard-Côté, David Hall, Thomas L. Griffiths, and Dan Klein. 2013. Automated reconstruction of ancient languages using probabilistic models of sound change. *Proceedings of the National Academy of Sciences*, 110(11):4224–4229.

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, Baltimore, Maryland. Association for Computational Linguistics.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems*, volume 32, pages 7057–7067.

Peter Dekker. 2018. Reconstructing language ancestry by performing word prediction with neural networks.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Liviu Dinu and Alina Maria Ciobanu. 2014. Building a dataset of multilingual cognates for the Romanian lexicon. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1038–1043, Reykjavik, Iceland. European Language Resources Association (ELRA).

Meghan Dowling, Teresa Lynn, Alberto Poncelas, and Andy Way. 2018. SMT versus NMT: Preliminary comparisons for Irish. In *Proceedings of the AMTA 2018 Workshop on Technologies for MT of Low Resource Languages (LoResMT 2018)*, pages 12–20, Boston, MA. Association for Machine Translation in the Americas.

Jonathan Duddington. 2007-2015. espeak text to speech.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 489–500, Brussels, Belgium. Association for Computational Linguistics.

Orhan Firat, Kyunghyun Cho, and Yoshua Bengio. 2016a. Multi-way, multilingual neural machine translation with a shared attention mechanism. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 866–875, San Diego, California. Association for Computational Linguistics.

Orhan Firat, Baskaran Sankaran, Yaser Al-onaizan, Fatos T. Yarman Vural, and Kyunghyun Cho. 2016b. Zero-resource translation with multi-lingual neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 268–277, Austin, Texas. Association for Computational Linguistics.

Clémentine Fourrier and Benoît Sagot. 2020a. Comparing statistical and neural models for learning sound correspondences. In *Proceedings of LT4HALA 2020 - 1st Workshop on Language Technologies for Historical and Ancient Languages*, pages 79–83, Marseille, France. European Language Resources Association (ELRA).

Clémentine Fourrier and Benoît Sagot. 2020b. Methodological aspects of developing and managing an etymological lexical resource: Introducing EtymDB-2.0. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3207–3216, Marseille, France. European Language Resources Association.

Clémentine Fourrier. 2020. Évolution phonologique des langues et réseaux de neurones : travaux préliminaires (sound change and neural networks: preliminary experiments). In *Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition). Volume 3 : Rencontre des Étudiants Chercheurs en Informatique pour le TAL*, pages 110–122. ATALA et AFCP.

Oana Frunza and Diana Inkpen. 2006. Semi-supervised learning of partial cognates using bilingual bootstrapping. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 441–448, Sydney, Australia. Association for Computational Linguistics.

Oana Frunza and Diana Inkpen. 2009. Identification and disambiguation of cognates, false friends, and partial cognates using machine learning techniques. *International Journal of Linguistics*, 1(1):1–37.

Thanh-Le Ha, Jan Niehues, and Alexander Waibel. 2016. Toward multilingual neural machine translation with universal encoder and decoder. In *Proceedings of the 13th International Conference on Spoken Language Translation. Seattle, US*.

David Hall and Dan Klein. 2010. Finding cognate groups using phylogenies. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1030–1039, Uppsala, Sweden. Association for Computational Linguistics.

David Hall and Dan Klein. 2011. Large-scale cognate recovery. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 344–354, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Mika Hämäläinen and Jack Rueter. 2019. Finding Sami cognates with a character-based NMT approach. In *Proceedings of the 3rd Workshop on the Use of Computational Methods in the Study of Endangered Languages Volume 1 (Papers)*, pages 39–45, Honolulu. Association for Computational Linguistics.

Valérie Hanoka and Benoît Sagot. 2014. An open-source heavily multilingual translation graph extracted from wiktionaries and parallel corpora. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3179–3186, Reykjavik, Iceland. European Language Resources Association (ELRA).

Bradley Hauer, Amir Ahmad Habibi, Yixing Luan, Rashed Rubby Riyadh, and Grzegorz Kondrak. 2019. Cognate projection for low-resource inflection generation. In *Proceedings of the 16th Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 6–11, Florence, Italy. Association for 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, Chiang Mai, Thailand. Asian Federation of Natural Language Processing.

Kenneth Heafield. 2011. KenLM: Faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 187–197, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Diana Inkpen, Oana Frunza, and Grzegorz Kondrak. 2005. Automatic Identification of Cognates and False Friends in French and English. In *Proceedings of Recent Advances in Natural Language Processing 2005*, pages 251–257, Borovets, Bulgaria.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference on Learning Representations*.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.

Shantanu Kumar, Ashwini Vaidya, and Sumeet Agarwal. 2017. Discovering Cognates Using LSTM Networks. In *Proceedings of the 4th Annual Conference of the Association for Cognitive Science*, Hyderabad, India.

Johann-Mattis List, Simon J. Greenhill, and Russell D. Gray. 2017. The potential of automatic word comparison for historical linguistics. *PLOS ONE*, 12(1):1–18.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Minh-Thang Luong, Quoc V. Le, Ilya Sutskever, Oriol Vinyals, and Lukasz Kaiser. 2016. Multi-task sequence to sequence learning. In *Proceedings of the 4th International Conference on Learning Representations*.

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421, Lisbon, Portugal. Association for Computational Linguistics.

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*.

Richard T. McCoy and Robert Frank. 2018. Phonologically informed edit distance algorithms for word alignment with low-resource languages. In *Proceedings of the Society for Computation in Linguistics (SCiL) 2018*, pages 102–112.

Carlo Meloni, Shauli Ravfogel, and Yoav Goldberg. 2021. Ab Antiquo: Proto-language Reconstruction with RNNs. In *Proceedings of the 2021 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, Online. Association for Computational Linguistics. To appear.

Ruslan Mitkov, Viktor Pekar, Dimitar Blagoev, and Andrea Mulloni. 2007. Methods for extracting and classifying pairs of cognates and false friends. *Machine Translation*, 21(1):29–53.

Andrea Mulloni. 2007. Automatic prediction of cognate orthography using support vector machines. In *Proceedings of the ACL 2007 Student Research Workshop*, pages 25–30, Prague, Czech Republic. Association for Computational Linguistics.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Hermann Osthoff and Karl Brugmann. 1878. *Morphologische Untersuchungen auf dem Gebiete der indogermanischen Sprachen*. Hirzel, Leipzig, Germany.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

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

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Rico Sennrich and Biao Zhang. 2019. Revisiting low-resource neural machine translation: A case study. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 211–221, Florence, Italy. Association for Computational Linguistics.

T. D. Singh and A. Vellintihun Hujon. 2020. Low Resource and Domain Specific English to Khasi SMT and NMT Systems. In *Proceedings of the 2020 International Conference on Computational Performance Evaluation (ComPE)*, pages 733–737, Shillong, India.

Inguna Skadiņa and Mārcis Pinnis. 2017. NMT or SMT: Case study of a narrow-domain English-Latvian post-editing project. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 373–383, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Eliel Soisalon-Soininen and Mark Granroth-Wilding. 2019. Cross-family similarity learning for cognate identification in low-resource languages. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 1121–1130, Varna, Bulgaria. INCOMA Ltd.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2019. MASS: masked sequence to sequence pre-training for language generation. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pages 5926–5936, Long Beach, California, USA. PMLR.

Adam St Arnaud, David Beck, and Grzegorz Kondrak. 2017. Identifying cognate sets across dictionaries of related languages. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2519–2528, Copenhagen, Denmark. Association for Computational Linguistics.

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.

Winston Wu and David Yarowsky. 2018. Creating large-scale multilingual cognate tables. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

A Appendix

A.1 Parameter Exploration Results

This section presents the hyperparameters which lead to the best dev BLEU across seeds, chosen during Section 4.2.3, and used for all subsequent experiments.

Model	Learning rate	Batch size	Embed. dim	Hidden dim	#layers	Model specific
RNN						Attention type
ES→IT	0.005	65	20	54	1	Luong-dot
IT→ES	0.005	65	20	72	1	Luong-dot
ES→LA	0.001	10	24	72	4	Luong-dot
LA→ES	0.005	100	20	72	1	Luong-dot
IT→LA	0.001	10	24	72	2	Bahdanau-dot
LA→IT	0.001	10	20	72	2	Luong-dot
Multilingual	0.001	10	24	72	2	Luong-general
Transformers						#heads
ES→IT	0.005	65	24	54	1	1
IT→ES	0.005	30	24	54	1	3
ES→LA	0.005	65	24	54	1	2
LA→ES	0.001	10	24	72	4	2
IT→LA	0.001	10	24	72	4	3
LA→IT	0.005	65	24	72	2	3
Multilingual	0.005	30	24	72	4	3

Luong-dot and Luong-general refer respectively to the dot and general attentions in (Luong et al., 2015), while Bahdanau-dot refers to our own implementation of the attention from (Bahdanau et al., 2015), simplified using the dot product to compute attention weights introduced in (Luong et al., 2015). See the code implementation with this paper for more detail.

Table 6: Results of parameter exploration experiments for RNN and Transformer models.

A.2 Learning Embeddings

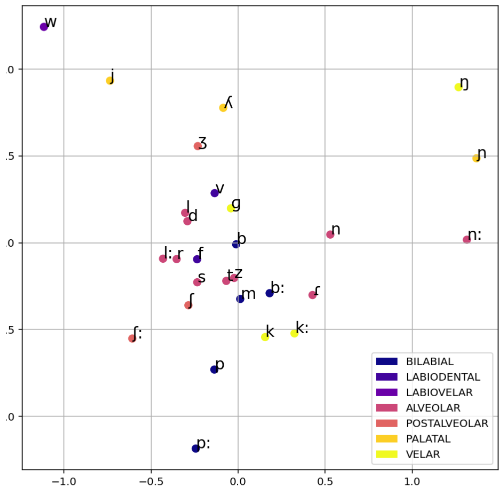

Figure 4: PCA of IT phonetic source embeddings (ES–IT RNN), coloured according to place of articulation.

Our learned embeddings seem to contain relevant phonetic information. For instance, applying a principal component analyses (PCA) and colouring consonants according to place, we can see that they are coherently divided, as illustrated for an ES–IT RNN model in Figure 4. More precisely, the figure shows the PCA of the learned source phonetic embeddings of one RNN model, for IT consonant phones, coloured according to place of articulation. It is radially organised, with a smooth transition between labio-dentals from the centre [b] to the bottom [pː], and from centre alveolar to left post alveolar.[13]

[13]We observed similar results for our other languages and architectures. They also hold when colouring consonants according to manner of articulation, and when colouring vowels according to backness or height.

A.3 Average position of the best result among the 10-best results.

We present here at which position the best prediction (according to sentenceBLEU, from `sacreBLEU`) occurs amongst the 10-best predictions. For example, when going from Spanish *terroso* 'muddy', phonetised [tɛroso], to Italian *terroso* 'muddy', phonetised [terroːzo], the RNN predicted [teroːzo], [terɔːzo], [tɛrroːzo], **[terroːzo]**, [tɛrrosː], [tɛrrɔːzo], [tɛrrɔsː], [terrɔːzo], [tɛrosː], and [tɛrosːo]: the correct result corresponds to the 4^{th} position.

For all multilingual models, we computed the sentence BLEU score for each of the 10-best predictions and saved the position of the highest scoring prediction. We averaged these positions for all words in the test set and calculated the standard deviation. Table 7 contains the full results, analysed in Section 6.2.

	IT→ES	ES→IT	IT→LA	LA→IT	ES→LA	LA→ES
SMT	1.08 ± 1.93	2.12 ± 2.55	2.01 ± 2.50	1.67 ± 2.30	2.49 ± 2.68	1.30 ± 2.14
Multilingual RNN	1.04 ± 2.03	1.67 ± 2.42	2.20 ± 2.54	1.63 ± 2.38	2.51 ± 2.68	1.38 ± 2.30
Multilingual Transformer	1.34 ± 2.17	1.94 ± 2.34	2.42 ± 2.65	2.17 ± 2.57	2.78 ± 2.73	1.64 ± 2.31

Table 7: Average position of the closest prediction to the reference amongst the 10-best predictions.