Neural Transduction for Multilingual Lexical Translation

Dylan Lewis and **Winston Wu** and **Arya D. McCarthy** and **David Yarowsky**
Department of Computer Science
Johns Hopkins University
{dlewis77,wswu,arya,yarowsky}@jhu.edu

Abstract

We present a method for completing multilingual translation dictionaries. Our probabilistic approach can synthesize new word forms, allowing it to operate in settings where correct translations have not been observed in text (cf. cross-lingual embeddings). In addition, we propose an approximate Maximum Mutual Information (MMI) decoding objective to further improve performance in both many-to-one and one-to-one word level translation tasks where we use either multiple input languages for a single target language or more typical single language pair translation. The model is trained in a many-to-many setting, where it can leverage information from related languages to predict words in each of its many target languages. We focus on 6 languages: French, Spanish, Italian, Portuguese, Romanian, and Turkish. When indirect multilingual information is available, ensembling with mixture-of-experts as well as incorporating related languages leads to a 27% relative improvement in whole-word accuracy of predictions over a single-source baseline. To seed the completion when multilingual data is unavailable, it is better to decode with an MMI objective.

1 Introduction

Translation matrices, i.e. concept-aligned word lists across the world's languages (Buck, 1949; Swadesh, 1950; Swadesh, 1952; Swadesh, 1955; Swadesh, 1971; Nastase and Strube, 2013; Wu et al., 2018), enable several avenues of exploration in computational linguistics and human language technologies. They strengthen word alignment models, which can in turn be useful for machine translation (Garg et al., 2019), robust projection of morphosyntactic information across alignments (Yarowsky and Ngai, 2001), and interlinear glossing. Further, fuller word lists enable neogrammarians to better explore phylogeny and phonology across languages (Hewson, 1973; Lowe and Mazaudon, 1994).

This work is motivated by the tremendous capacity for humans to generalize during translation, producing forms for words that have not been seen before. This becomes valuable especially for lower-frequency words, which may not have been observed in training data but could be inferrable through regular processes such as cognate relationships with related languages (Mulloni, 2007; Beinborn et al., 2013), borrowing from neighboring or other influential languages, and even esoteric features like temporal similarity (Schafer and Yarowsky, 2002; Wijaya et al., 2017) or image similarity (Bergsma and Van Durme, 2011). In this work, we focus on these cognate relationships, because cognates form a large amount of both core vocabulary (Wu et al., 2020) and technical language (Mulloni, 2007).[1] Unlike conventional bilingual lexicon induction (Rapp, 1995), we do not wish to limit the predictions to words that have been previously seen in a corpus. Automated methods to induce plausible translations for lexical translation would significantly reduce the human effort needed for both elicitation (Chelliah, 2001) and building machine translation systems for less heavily supported languages.

[1]We use a pragmatic definition of cognacy based on orthographic or phonetic similarity across languages (Kondrak, 2001), which is adopted by other computational cognate research (Inkpen et al., 2005; Dinu and Ciobanu, 2014; Wu and Yarowsky, 2018). A stricter linguistic definition demands shared etymology. Ciobanu and Dinu (2015) distinguish between these computationally.

Proceedings of the 28th International Conference on Computational Linguistics, pages 4373–4384
Barcelona, Spain (Online), December 8-13, 2020

Our approach to the problem of translation matrix completion is a neural model (parameterized as a character-level sequence-to-sequence network) that handles multiple language pairs, along with an objective function that maximizes both forward and backward probabilities. By leveraging both probabilities, we try to maximize the flow of information between both source and target languages, leading to more accurate model predictions. We find that by both leveraging information about a concept's form from related languages and carefully combining language-pair-wise predictions of an unknown target word, we can improve accuracy by 27% relative to our baseline multilingual neural model.

	Train			Test
Concept	French	Spanish	Italian	Romanian
MONARCH	monarque	-	monarca	??? (monarh)
CAFFEINE	caféine	cafeína	caffeina	??? (cofeină)
KNEE	genou	-	-	??? (genunchi)

Table 1: Completion of cognate clusters. Given partially observed cognate clusters, the task is to infer missing values. "???" denotes words that are held out for testing purposes.

2 Related Work

The task of translation matrix completion, the filling-out of a universal conceptual inventory, has been approached by three broad classes of methods. The first is to manually construct concept inventories, as in Swadesh (1950) and followup work. The next is to automatically *identify* cognate relationships, e.g. in word lists (Kondrak, 2001; Wijaya et al., 2017; Jäger et al., 2017) or raw text (Koehn and Knight, 2002). The third, which is our focus, is to *generate* putative cognates by performing transduction in the form of sound or orthographic shifts. In this vein, Mann and Yarowsky (2001) generate cognates by a pipeline of dictionary lookup and probabilistic orthographic shifts. Mulloni (2007) uses an SVM to perform cognate generation. Ciobanu (2016) uses a CRF with reranking to the same end.

Beinborn et al. (2013) and Wu and Yarowsky (2018) perform translation matrix completion with extracted cognate lists in 6 and 60 language families respectfully, using character-level statistical machine translation systems trained on separate source-target language pairs. Wu et al. (2020) performed the same cognate transliteration task with a multi-source multi-target character-level variant of Johnson et al. (2017).

We adopt the single system multilingual setup of Wu et al. (2020), which allows sharing information across language pairs. We also take inspiration from recent successes in other generation tasks. Nishimura et al. (2018)'s multi-source missing data problem used multiple encoders and a single decoder to leverage multiple source language inputs, which we build on to employ the multiple sources simultaneously during inference. Further, we introduce introduce a maximum mutual information (MMI) objective to the problem, motivated by the translational equivalence of cognates (Hauer and Kondrak, 2020). MMI has been explored in speech recognition (Bahl et al., 1986; Brown, 1987) and dialog (Li et al., 2016).

Besides MMI, there are a few existing methods for incorporating backward probabilities into the task of translation. Yee et al. (2019) and Ng et al. (2019) follow a noisy channel approach, using Bayes' rule to integrate forward, backward and target language model probabilities. We follow Yee et al. (2019)'s approach and implement both the MMI objective and an ensemble MMI objective.

3 Translation Matrix Completion

The core of this work is the *translation matrix* T (partially realized in Table 1), whose m rows are indexed by concepts c and whose n columns are indexed by languages ℓ. The entries $\mathsf{T}_{c,\ell}$ in the matrix T are orthographic sequences drawn from each language's alphabet Σ_ℓ that each form a word or multi-word expression. These entries may be absent. The goal, then, is to predict a missing entry $\mathsf{T}_{c,\ell}$ from one or more translational equivalents $\mathsf{T}_{c,\neg\ell}$. In this work, we restrict our focus to cognate relationships; the translational equivalents that represent a concept share surface features.

4 MMI Reranking

The MMI objective provides a principled way to rerank predicted cognates. We motivate its use with the notion of *translational equivalence* (Hauer and Kondrak, 2020)—the idea that two words (and particularly when we constrain our focus to cognates) should translate to each other, regardless of direction. Particularly, cognacy is a symmetric relationship. The surface form in each language is a *view* into the (interlingual) concept. This is explicitly modeled in the MMI objective, which simultaneously optimizes for both the forward and backward direction translations. In the context of translation matrix completion, when filling in a single concept across multiple languages, the translations—and particularly cognate relationships—should all be equivalent across all languages (as opposed to sentence level translation, where a sentence in one language can have multiple interpretations in another language).

Predicting cognates *de novo* is a sequence transduction task akin to transliteration. To perform this task, it is common to use probabilistic models $p_\theta(T \mid S)$ of target sequences T given source sequence S, controlled by a set of *parameters* θ. (We will omit the parameter subscript for brevity.) In our case, T and S are entries in the translation matrix, character sequences from alphabets Σ_T and Σ_S. The typical decoding objective is to maximize the conditional log-likelihood, which we will use as a baseline:[2]

$$\hat{T} = \arg\max_{T \in \Sigma_T^*} \log p(T|S). \tag{1}$$

However, to encourage the symmetric relationship of translational equivalence, we can instead use the MMI objective:

$$\hat{T} = \arg\max_{T \in \Sigma_T^*} \log \frac{p(T,S)}{p(S)p(T)} \tag{2}$$

Li et al. (2016) show that this can be reformulated as:

$$\hat{T} = \arg\max_{T \in \Sigma_T^*} \log p(T|S) + \log p(S|T) \tag{3}$$

and then generalize it with a hyperparameter λ to control the weight of each term:

$$\hat{T} = \arg\max_{T \in \Sigma_T^*} \left[(1-\lambda) \log p(T|S) + \lambda \log p(S|T)\right]. \tag{4}$$

Because we wish to predict a word from its several cognates, we can further generalize Li et al. (2016). That is, given source words S_i for languages $i \in 1..n-1$, we can find the target translation that maximizes the mutual information between each source-target language pair:

$$\hat{T} = \arg\max_{T} \sum_{i=1}^{n-1} (1-\lambda) \log p(T|S_i) + \lambda \log p(S_i|T) \tag{5}$$

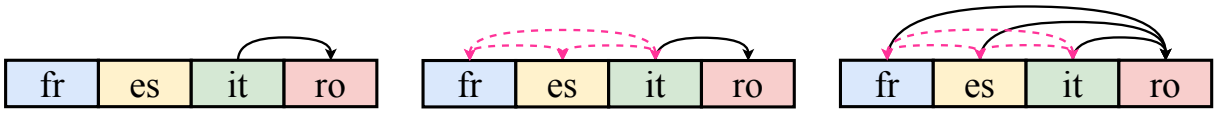
However, in our reformulated objective function, each term $\log p(S_i|T)$ in Equation (3) is intractable during decoding. It requires knowledge of the complete prediction T, which is unavailable until decoding has finished. Thus, we approximate this second term by rescoring k-best lists generated by the forward model $p(T|S_i)$. This approximation has previously been used successfully by Li et al. (2016).

5 Experimental Setup

We formulate lexical translation as sequence-to-sequence character translation, using one model capable of translating between any pair of languages. The input comprises the characters of the word along with tokens to identify the source and target language. Including the target language token in the input conditions the multilingual model to generate text in the target language (Johnson et al., 2017). The output is a character sequence in the target language.

We continue to use this single-model architecture even when using multiple known entries in a row of the translation matrix (that is, many translations of the word to be predicted). We do this by combining distributions from each source language in either a mixture or product of experts (Hinton, 2002). We also compare the two decoding objectives: conditional likelihood and maximum mutual information.

[2]More precisely, it is typical to decode with a approximate maximization to the argmax such as beam search.



(a) NOVEL (b) SINGLE (c) MULTI

Figure 1: The three experimental data settings, illustrated for a single row of the translation matrix. This figure assumes that Romanian is the held-out language. Black solid lines show conditioning during inference. Red dashed lines show relationships seen during training.

Dataset Cognate relationships among European languages are well-studied and broadly verifiable. To this end, we use a data set from Dinu and Ciobanu (2014) which contains cognates in six languages: French, Spanish, Italian, Romanian, Portuguese, and Turkish. All use the latin alphabet plus language specific diacritics. Except for Turkish, these are Romance languages. (In fact, Turkish is not even in the Indo-European family. Turkish is included because many French and Turkish words were imported into Romanian, as well as French words into Turkish, leading to cognates between these languages. The comparable performance thereon shows that the method is not limited to linguistic cognates.) Like Mulloni (2007), we group words by cognate cluster using an unsupervised clustering algorithm (Wu and Yarowsky, 2018), which results in 18K clusters covering 16K English concepts. Note that a single concept may have more than one cognate cluster. We hold out 500 concepts for a validation set and 500 concepts for a test set. The validation and test data are then used globally for our scenarios defined below.

Experimental scenarios Broadly, the test inputs available to our models at training and test time define three scenarios, illustrated in Figure 1. In NOVEL, only a single form is present in the row. The model must predict a cognate for a novel concept—one whose forms it has never seen in training. (This would be a first step to filling a completely new row.) In the other two scenarios, only a single entry is missing. In SINGLE, the model has seen other entries in the row during training, including the single source word, and it must generate the missing form which is a form of indirect supervision. During inference, we only test for the single (directly supervised) language pair. In MULTI, generation is conditioned on all known forms of the concept. Comparing NOVEL to SINGLE addresses whether exposure to the concept's forms is beneficial (they differ in data availability), and comparing SINGLE to MULTI shows whether the standard single-input, single-output sequence transduction framework is sufficient for cognate prediction (they differ in conditioning during inference).

In all scenarios, during training the model has seen all extant entries in the rows from the training set and no target-language words from the test set. While not all slots in the Cartesian product of data scenarios (NOVEL, SINGLE, MULTI), decoding objectives (conditional log-likelihood or MMI), and ensembling methods (mixture or product of experts) are plausible, this product subsumes the experiments we run.

Evaluation methods We report three metrics of cognate generation quality. For all, higher is better. The first is exact string match accuracy, following Wu and Yarowsky (2018): does the model's 1-best prediction correctly predict the unknown word? The others refine the notion of "inaccurate." Character-level BLEU using SacreBLEU (Papineni et al., 2002; Post, 2018), awards partial credit for inexact matches. Mean reciprocal rank (MRR), following Ciobanu (2016), answers: how far down the k-best list is the correct form?

Experimental details We use a sequence-to-sequence LSTM model with attention (Bahdanau et al., 2015) from the FAIRSEQ toolkit (Ott et al., 2019).[3] We train the model using a hidden size of 1024 in both the encoder and decoder, and embedding size of 512 with the NAG optimizer (Botev et al., 2017). In addition, we use a dropout of 0.25, clip gradients to 0.1, and use early stopping with a validation set after 5 epochs of no improvement. All MMI tradeoff λ values are 0.5 unless otherwise specified. We decode using a beam size of 10 and create k-best lists of length k=100.

[3]We focus on LSTM models as opposed to Transformer architectures due to their superior performance on grapheme-to-phoneme conversion, another monotonic sequence transduction task (Gorman et al., 2020).

	ro	es	fr	it	po	tr
ro	0	46	62	42	45	39
es	55	0	50	61	60	44
fr	58	56	0	53	43	44
it	54	50	67	0	50	30
po	58	66	65	61	0	38
tr	49	34	78	33	53	0

(a) NOVEL

	ro	es	fr	it	po	tr
ro	0	60	62	56	53	48
es	60	0	50	65	69	53
fr	83	62	0	67	48	56
it	56	62	67	0	57	37
po	62	73	65	65	0	49
tr	65	52	78	45	58	0

(b) SINGLE

Table 2: NOVEL (Left), SINGLE (Right), exact-match accuracy percentage without reranking. Source language on Y-axis. Target language on X-axis.

	ro	es	fr	it	po	tr	avg
NOVEL	55	50	64	50	50	39	51
SINGLE	65	62	64	60	57	49	59
NOVEL-MMI	60	53	63	51	53	47	55
SINGLE-MMI	66	62	63	58	59	63	60
MULTI	65	68	55	67	68	55	63

(a) MULTI exact-match accuracy percentage without reranking. Target language on X-axis.

	ro	es	fr	it	po	tr	avg
MULTI	65	68	55	67	68	55	63
MULTI-MMI	62	67	55	66	65	53	61
MULTI-LSE	67	69	55	67	68	55	63
MMI-LSE	64	67	55	66	68	53	62

(b) Comparison between the original objective and a mixture of experts approach on MULTI. MMI-LSE is the MULTI model with both MMI and LSE.

Table 4: Average exact match for SINGLE models, and exact-match accuracy for MULTI models.

6 Results

How much does seeing the concept help? (NOVEL vs. SINGLE) In Table 2, we report exact-match accuracy for both NOVEL and SINGLE without reranking. The model sees up to 25% absolute increases in accuracy between language pairs when tested on data for which it had prior knowledge of non-target-language concepts. Of particular note, performance increases when translating to and from Turkish once related language information is incorporated. This implies that the model can effectively leverage data from outside of the testing language pair, even though the concept has not been seen in the target language. While a similar finding has been shown in multilingual neural sentence-level translation, this is the first time that it has been shown for lexical translation.

By aggregating over languages (Table 3), we see that SINGLE increases performance in all metrics. SINGLE's higher MRR shows that the model is not only creating more accurate translations, but the quality of the translations is also higher due to gold predictions found higher in the k-best lists. This is also reflected in the increase in character-level BLEU.

	ACC	MRR	BLEU
NOVEL	51.47	.60	79.87
SINGLE	59.41	.67	82.61
NOVEL-MMI	54.55	.62	79.77
SINGLE-MMI	59.81	.66	82.28

Table 3: Comparing average Accuracy, MRR, and BLEU across multiple scenarios.

Does using multiple inputs help? (SINGLE vs. MULTI) The standard encoder-decoder architecture used for sequence transduction does not lend itself well to simple integration of multiple input sequences. Is it enough to pick one source language and use this model as-is (SINGLE), or does the invested effort in ensembling predictions conditioned on multiple sources (MULTI) pay off?

On average, the MULTI model, with its final predictions determined by highest sum of log-probabilities, $\sum_{i=1}^{n} \log p(T|S_i)$, outperforms both NOVEL and SINGLE (Table 4a). In every language besides French, the MULTI model does just as well if not much better translating into the target language.

How should we weight the ensemble? (MULTI vs MULTI-LSE) We observe one shortcoming in the MULTI model: in some cases, it chooses to ignore an answer very highly rated among many source

TARGET LANGUAGE: ROMANIAN TARGET TEXT: T E L E F O N I S T				
MULTI		PREDICTED: T E L E F O N I C		
Rank	Turkish	Portuguese	Italian	Spanish
1	t e l e f o n c	**t e l e f o n i s t**	**t e l e f o n i s t**	**t e l e f o n i s t**
2	t e l e f o n c ă	t e l l e f o n i s t	t e l l e f o n i s t	t e l l e f o n i s t
3	t e l e f o n c u	t h e l e f o n i s t	t h e l e f o n i s t	t h e l e f o n i s t
4	t e l e f o n c o	t e l e f o n i s t a	t e l e f o n i s t a	t e l e f o n i s t a
5	t e l e f o n c a	d e l e f o n i s t	d e l e f o n i s t	d e l e f o n i s t
6	t e l e f o n ț ă	t e l e p o n i s t	t e l e p o n i s t	t e l e p o n i s t
7	t e l e f o n c i e	t e l e f u n i s t	t e l e f u n i s t	t e l e f u n i s t
8	t e l e f o n c i c	t e l i f o n i s t	t e l i f o n i s t	t e l i f o n i s t
9	t e l e f o n c e	t e l e f o n i c	t e l e f o n i c	t e l e f o n i c
10	t e l e f o n c u e	t e l e f o n o s t	t e l e f o n o s t	t e l e f o n o s t
(17)	t e l e f o n i c			
MULTI-MMI		PREDICTED: T E L E F O N I S T		
Rank	Turkish	Portuguese	Italian	Spanish
1	t e l e f o n c u	**t e l e f o n i s t**	**t e l e f o n i s t**	**t e l e f o n i s t**
2	t e l e f o n c	t e l e f o n i s t a	t e l e f o n i s t a	t e l e f o n i s t a
3	t e l e f o n c o	t e l l e f o n i s t	t h e l e f o n i s t	t e l l e f o n i s t
4	t e l e f o n c ü	t h e l e f o n i s t	t e l e f o n i s t ă	t h e l e f o n i s t
5	t e l e f o n c ă	t e l e f f o n i s t	t e l e f o n i i s t	t e l e f f o n i s t
6	t e l e f o n c v	t e l e f o n i s t ă	t e l e p h o n i s t	t e l e f o n n i s t
7	t e l e f o n q u	t e l e f o n n i s t	t e l e f u n i s t	t e l e f o n i s t ă
8	t e l e f o n c v ă	t e l e f o n i i s t	t e l e f o n h i s t	t e l e f o n i i s t
9	t e l e f o n c a	t é l e f o n i s t	t e l e f o n i e s t	t e l e p h o n i s t
10	t e l e f o n c u e	t e l e f o n í s t	t é l e f o n i s t	t é l e f o n i s t

Table 5: Romanian example of shortcomings of base MULTI objective and how it is inherently corrected for in MULTI-MMI. The MULTI model incorrectly predicts "telefonic" due to it being the highest scored word across all languages; however, in MULTI-MMI this is corrected for by virtue of "telefonic" not being predicted by Turkish.

languages. This is due to the ensembling strategy, which determines a prediction's ensemble score by summing the log-probabilities of this prediction conditioned on each source language. This style of model is a *product of experts*, in which one low score can effectively 'veto' a prediction when it is low or absent on that language's k-best list. The high weights from other languages cannot salvage it. This motivates us to explore an alternative ensembling strategy, the *mixture of experts*. In a mixture of experts approach, we define the probability of a translation as $p(T|S_{1...n-1}) \propto \prod_{i=1}^{n} p(T|S_i)$.

For an explicit example, see Table 5, where we present a snippet of the k-best list for the target Romanian word *Telefonist*. In the MULTI model, despite having 3 languages agree that *Telefonist* is the correct translation, the model instead chooses *Telefonic*, which shows up much lower (rank 9 in the 3 languages) in their k-best lists. Here, *Telefonist* doesn't show up in the Turkish k-best list, so according to Equation (5), it is given 0% probability, so the combined probability of *Telefonist* is less than that of *Telefonic*. This flaw shows that trying to find a globally optimal solution via a product of experts approach may not be the correct way of leveraging multiple sources.

One possible solution is to instead sum the probabilities to form a mixture of experts, which is equivalent to taking the LogSumExp (LSE) of the log-probabilities. Unlike a product of experts, a single model's low probability cannot 'veto'. On average, using LSE improves overall accuracy, at worst it does nothing, and overall it results in our strongest model, MULTI-LSE (results in Table 4b).

How should we decode? (Log-likelihood vs. MMI objective) Above, we described two decoding strategies: maximum conditional log-likelihood and maximum mutual information. The latter infuses a bias toward translational equivalence. Because the model's training objective is the conditional log-likeihood, MMI amounts to a rescoring method of the objective. We can decode NOVEL, SINGLE, and MULTI with either strategy.

	ro	es	fr	it	po	tr
ro	0	44	54	42	48	45
es	58	0	50	60	65	53
fr	75	62	0	53	52	56
it	56	52	67	0	50	37
po	62	69	65	61	0	46
tr	51	39	78	36	52	0

(a) NOVEL-MMI

	ro	es	fr	it	po	tr
ro	0	51	54	53	56	48
es	59	0	50	68	71	47
fr	92	62	0	60	48	67
it	57	66	67	0	60	48
po	61	75	65	67	0	46
tr	63	55	78	42	60	0

(b) SINGLE-MMI

Table 6: NOVEL-MMI (Left), and SINGLE-MMI (Right), exact-match accuracy percentage without reranking. Source language on Y-axis. Target language on X-axis.

	Target:	biom	Target:	enterologie
Rank	MULTI	MMI	MULTI	MMI
1	**biom**	bioma	**enterologie**	enterology
2	bioma	**biom**	enterologic	**enterologie**
3	biomă	biomă	anterologie	enterologye
4	biomic	biome	enterologye	enterologgie
5	biomat	biomo	interologie	enterologic
6	bioman	biomic	enterology	enterologge
7	biome	biomas	entarologie	enterologiu
8	biomo	biomie	enterologist	enterologian
9	biomare	biomá	henterologie	anterologie
10	biomal	biomat	enterologia	interologie
	Target:	cinism	Target:	psiholog
Rank	MULTI	MMI	MULTI	MMI
1	**cinism**	kinism	**psiholog**	psicolog
2	kinism	**cinism**	psicolog	**psiholog**
3	chinism	kinizm	psicologic	psicologo
4	cinismo	cinismo	psicologă	psikolog
5	kinizm	kinismo	psicologo	psicholog
6	kinismo	chinism	psihologic	psicológ
7	scinism	cinismă	psicholog	psicologă
8	quinism	cinisma	psikolog	psicólogo
9	hinism	quinism	psicológ	psicologa
10	cinismă	kinismă	psicologa	psiológ

Table 7: Examples of Turkish to Romanian k-best lists for the SINGLE-MMI model.

On average, using rescoring helps NOVEL, but gives mixed results for SINGLE (Table 3). In particular for SINGLE, in terms of language-to-language accuracy, there are only a few language pairs that seem to benefit from rescoring (Table 6b). Table 7 contains explicit examples of how the model output is reranked for the Turkish–Romanian language pair under the SINGLE-MMI model. The MMI model takes the correct translation and swaps it with another translation (that often is rank 2 in the list). This would mean that the backward probabilities $p(T|S)$ of these words is larger than their forward counterparts, leading to wrong translations. In the Turkish–Romanian backward model's probabilities (which are the SINGLE results), we see that Romanian-Turkish language pair has very low accuracy, in addition to a very low BLEU score (Table 8), indicating that the generated translations are not only wrong but also far from correct. For this reason, rescoring in SINGLE tends to choose suboptimal candidates as the final translation.

	ro	es	fr	it	po	tr
ro	0	87	76	86	83	73
es	84	0	79	91	89	78
fr	89	84	0	88	77	76
it	81	88	82	0	86	70
po	86	90	77	90	0	75
tr	87	78	89	71	84	0

Table 8: SINGLE BLEU without reranking. Source language on Y-axis. Target language on X-axis.

	ro	es	fr	it	po	tr	AVG
MULTI	2.34	2.44	2.34	2.33	2.57	4.28	2.72
MMI	2.22	2.37	2.18	2.24	2.37	4.23	2.60

Table 9: Average Levenshtein distance between top 10 candidates and target word for MULTI and MULTI-MMI models

Looking at results from NOVEL-MMI, performance on most languages pairs is unchanged, but a few select language pairs have very large jumps in accuracy (Table 6a). For example, for French-Romanian, we see large increases in accuracy, amounting to an overall average gain of 17% against the base model (Table 4a). This implies that many correct translations are highly ranked in the k-best list, and the backward model merely gives these predictions a boost to the top. Further analysis confirms this: when we consider the MRR of only words the French Romanian model get wrong with NOVEL, the MRR is quite high: 0.875. This then indicates that the Romanian backward model is able to find these candidates and bring them to the top.

Initially, the MULTI-MMI results would suggest that MMI is not useful in an ensemble setting. On closer inspection though, there are a few cases that would suggest otherwise. We previously discussed a flaw with the MULTI model using a product of experts approach. However, MULTI-MMI inherently corrects for this (lower half of Table 5). Here, the MULTI-MMI model correctly chooses *Telefonist* as the correct translation. Despite the forward model not generating it as a candidate, the backward model scored it high enough to overcome the forward model. In addition, it is interesting that due to the nature of the backward model, the predictions that are towards the top of the list are more similar to the target candidate. We confirm this by computing the Levenshtein distance between the top 10 candidates and the target word (Table 9). Another example given in Table 10 shows a similar phenomenon. While the MULTI model predicts *Fonetista*, the MULTI-MMI model predicts PHONETISTA. We observe an interesting phenomenon: all words that start with "f" are no longer being scored as highly. Again, decoding is able to pick out the correct word, even when that word is not generated by one model.

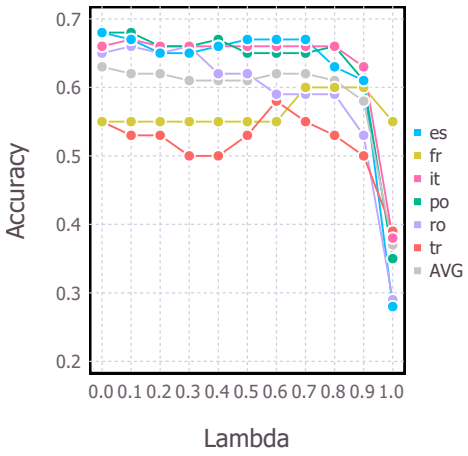


Figure 2: Effects of λ value on target language accuracy for MULTI-MMI. $\lambda = 0$ represents using only the forward probability, and $\lambda = 1$ represents using only the backward probability.

In the case of MULTI-MMI with LSE, we find that the gains over MULTI-MMI are greater than the gains between MULTI and MULTI-LSE. We believe that this is due to the backward model acting as a form of regularization that helps flatten the distribution so that one "expert" does not overpower the rest. Despite these gains, it still does not make up for the overall loss in accuracy due to the MMI objective.

Finally, we show the effects of λ on accuracy on a per-language basis (Figure 2). In most cases, a higher λ coincides with decreased performance. However, this does not hold for Romanian, French, and Turkish. First, for both Italian and Portuguese, we see changing λ does not greatly affect accuracy, implies that for these languages, the backward model is only acting as additional noise. In this case the backward distributions are too flat, so adding the backward term is the same as adding a constant to every candidate's score. For Turkish and French, as we increase λ, the model accuracy increases. This might lead one to believe that the backward model is doing all the work, and the forward model is not helping at all. However, this cannot be true: accuracy plummets when only using the backward term. We conclude that the forward model instead acts as a base reference, from which the backward model can then fine tune the results. In the case of Turkish, there seems to be a clear optimal λ of 0.6.

TARGET LANGUAGE: PORTUGUESE TARGET TEXT: p h o n e t i s t			
MULTI			PREDICTED: f o n e t i s t a
Rank	Italian	Spanish	Romanian
1	f o n e t i s t a	f o n e t i s t a	f o n e t i s t a
2	**p h o n e t i s t**	**p h o n e t i s t**	f o n e t i s t
3	p h o n e t i s t a	p h o n e t i s t a	f o n e t í s t a
4	f o n e t i s t	f o n e t i s t	f o n é t i s t a
5	f o n e t í s t a	f o n e t í s t a	f o n e t i s t o
6	p h o n e t i s t e	p h o n e t i s t e	f o n a t i s t a
7	p o n e t i s t a	p o n e t i s t a	f u n e t i s t a
8	p h o n e t i s t o	p h o n e t i s t o	f o n e t i s t e
9	f o n e t i s t o	f o n e t i s t o	f o n i t i s t a
10	p h o n é t i s t a	p h o n é t i s t a	f o n e d i s t a
MULTI-MMI			PREDICTED: P h o n e t i s t
Rank	Italian	Spanish	Romanian
1	**p h o n e t i s t**	**p h o n e t i s t**	f o n e t i s t a
2	p h o n e t i s t s	p h o n e t i s t s	f o n e t i s t
3	p h o n é t i s t	p h o n é t i s t	f o n é t i s t a
4	p h o n e t í s t	p h o n e t i s t o	f o n e t í s t a
5	p h ã o n e t i s t	p h o n e t í s t	f o n e t i s t e
6	p h o n e t i s t o	p h ã o n e t i s t	f o n ê t i s t a
7	p h o n e t i s t a	p h o n e t i s t a	f o n e t i s t m
8	p h o n e t i s t e	p h o n e t í s t o	f o n e t í s t
9	p h o n e t í s t o	p h o n e t i ş t	f o n e t i s t s
10	p h o n é t i s t a	p h o n e t i s t e	f o n e t i s t á

Table 10: Portuguese example of how MMI reorders k-best lists to improve accuracy.

7 Future Work

The work we presented here has particular applications to low-resource languages. As it is misguided to claim that our system is language-agnostic without verifying (Bender, 2009), we plan to expand this work to other language families, such as the Austronesian phonological cognate dataset of Bouchard-Côté et al. (2013). Another direction involves experimenting with non-uniform mixing weights that can adaptively give preference to certain languages, as in Wu and Yarowsky (2018). We would also like to extend this work to generate cognates of inflected forms, rather than lemmas and without explicit lemmatization and inflection subcomponents. Unlike existing cross-lingual morphological inflection tasks (McCarthy et al., 2019; Vylomova et al., 2020) the source and target are in different languages, rather than relying on transfer. Finally, to assess the downstream value of this linguistic tool, future work could populate a statistical translation model's phrase table with predictions from the model.

8 Conclusion

We present a single neural model to handle multilingual many-to-many translation of single words. In addition, by indirectly leveraging multi-lingual information in sequence-to-sequence models, we can improve accuracy in the matrix completion task (NOVEL vs SINGLE). By allowing knowledge of concepts that will eventually be tested on between non-target language pairs, the model indirectly learns the correct way to translate into an unseen word in the target language. In addition, directly leveraging multiple source languages improves accuracy on average (by 10% relative to our SINGLE model, which is equivalent to Wu et al. (2020)). A flaw in the ensemble scoring method is remedied in part by using LSE, and is also inherently corrected for in the MULTI-MMI model. In addition we show that the MMI objective is a feasible learning object and in some scenarios gives substantially better than baseline performance. One such scenario is when multilingual data is unavailable. When such data is available, our MULTI-LSE model tends to give best performance overall.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

L. Bahl, P. Brown, P. de Souza, and R. Mercer. 1986. Maximum mutual information estimation of hidden Markov model parameters for speech recognition. In *ICASSP '86. IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 11, pages 49–52.

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, October. Asian Federation of Natural Language Processing.

Emily M. Bender. 2009. Linguistically naïve != language independent: Why NLP needs linguistic typology. In *Proceedings of the EACL 2009 Workshop on the Interaction between Linguistics and Computational Linguistics: Virtuous, Vicious or Vacuous?*, pages 26–32, Athens, Greece, March. Association for Computational Linguistics.

Shane Bergsma and Benjamin Van Durme. 2011. Learning bilingual lexicons using the visual similarity of labeled web images. In *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence - Volume Volume Three*, IJCAI'11, page 1764–1769. AAAI Press.

Aleksandar Botev, Guy Lever, and David Barber. 2017. Nesterov's accelerated gradient and momentum as approximations to regularised update descent. In *2017 International Joint Conference on Neural Networks, IJCNN 2017, Anchorage, AK, USA, May 14-19, 2017*, pages 1899–1903. IEEE.

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.

Peter F. Brown. 1987. *The Acoustic-Modeling Problem in Automatic Speech Recognition*. Ph.D. thesis, USA. AAI8727170.

Carl Buck. 1949. *A Dictionary of Selected in the Principal Indo-European Languages*. University of Chicago Press.

Shobhana L. Chelliah. 2001. The role of text collection and elicitation in linguistic fieldwork. In Paul Newman and Martha Ratliff, editors, *Linguistic Fieldwork*, page 152–165. Cambridge University Press.

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, Beijing, China, July. Association for Computational Linguistics.

Alina Maria Ciobanu. 2016. Sequence labeling for cognate production. *Procedia Computer Science*, 96:1391 – 1399. Knowledge-Based and Intelligent Information & Engineering Systems: Proceedings of the 20th International Conference KES-2016.

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-2014)*, pages 1038–1043, Reykjavik, Iceland, May. European Languages Resources Association (ELRA).

Sarthak Garg, Stephan Peitz, Udhyakumar Nallasamy, and Matthias Paulik. 2019. Jointly learning to align and translate with transformer models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4453–4462, Hong Kong, China, November. Association for Computational Linguistics.

Kyle Gorman, Lucas F.E. Ashby, Aaron Goyzueta, Arya McCarthy, Shijie Wu, and Daniel You. 2020. The SIGMORPHON 2020 shared task on multilingual grapheme-to-phoneme conversion. In *Proceedings of the 17th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 40–50, Online, July. Association for Computational Linguistics.

Bradley Hauer and Grzegorz Kondrak. 2020. Synonymy = Translational Equivalence. *arXiv e-prints*, page arXiv:2004.13886, April.

John Hewson. 1973. Reconstructing prehistoric languages on the computer: The triumph of the electronic neogrammarian. In *COLING 1973 Volume 1: Computational And Mathematical Linguistics: Proceedings of the International Conference on Computational Linguistics*.

Geoffrey E. Hinton. 2002. Training products of experts by minimizing contrastive divergence. *Neural Computation*, 14(8):1771–1800.

Diana Inkpen, Oana Frunza, and Grzegorz Kondrak. 2005. Automatic identification of cognates and false friends in french and english. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, volume 9, pages 251–257.

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, April. Association for Computational Linguistics.

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.

Philipp Koehn and Kevin Knight. 2002. Learning a translation lexicon from monolingual corpora. In *Proceedings of the ACL-02 Workshop on Unsupervised Lexical Acquisition*, pages 9–16, Philadelphia, Pennsylvania, USA, July. Association for Computational Linguistics.

Grzegorz Kondrak. 2001. Identifying cognates by phonetic and semantic similarity. In *Second Meeting of the North American Chapter of the Association for Computational Linguistics*.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119, San Diego, California, June. Association for Computational Linguistics.

John B. Lowe and Martine Mazaudon. 1994. The reconstruction engine: A computer implementation of the comparative method. *Computational Linguistics*, 20(3):381–417.

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, Ekaterina Vylomova, Shijie Wu, Chaitanya Malaviya, Lawrence Wolf-Sonkin, Garrett Nicolai, Christo Kirov, Miikka Silfverberg, Sebastian J. Mielke, Jeffrey Heinz, Ryan Cotterell, and Mans Hulden. 2019. The SIGMORPHON 2019 shared task: Morphological analysis in context and cross-lingual transfer for inflection. In *Proceedings of the 16th Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 229–244, Florence, Italy, August. Association for Computational Linguistics.

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, June. Association for Computational Linguistics.

Vivi Nastase and Michael Strube. 2013. Transforming wikipedia into a large scale multilingual concept network. *Artificial Intelligence*, 194:62 – 85. Artificial Intelligence, Wikipedia and Semi-Structured Resources.

Nathan Ng, Kyra Yee, Alexei Baevski, Myle Ott, Michael Auli, and Sergey Edunov. 2019. Facebook FAIR's WMT19 news translation task submission. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 314–319, Florence, Italy, August. Association for Computational Linguistics.

Yuta Nishimura, Katsuhito Sudoh, Graham Neubig, and Satoshi Nakamura. 2018. Multi-source neural machine translation with missing data. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 92–99, Melbourne, Australia, July. Association for Computational Linguistics.

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, June. 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, July. 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, Belgium, Brussels, October. Association for Computational Linguistics.

Reinhard Rapp. 1995. Identifying word translations in non-parallel texts. In *33rd Annual Meeting of the Association for Computational Linguistics*, pages 320–322, Cambridge, Massachusetts, USA, June. Association for Computational Linguistics.

Charles Schafer and David Yarowsky. 2002. Inducing translation lexicons via diverse similarity measures and bridge languages. In *COLING-02: The 6th Conference on Natural Language Learning 2002 (CoNLL-2002)*.

Morris Swadesh. 1950. Salish internal relationships. *International Journal of American Linguistics*, 16(4):157–167.

Morris Swadesh. 1952. Lexico-statistic dating of prehistoric ethnic contacts: with special reference to North American Indians and Eskimos. *Proceedings of the American philosophical society*, 96(4):452–463.

Morris Swadesh. 1955. Towards greater accuracy in lexicostatistic dating. *International journal of American linguistics*, 21(2):121–137.

Morris Swadesh. 1971. *The origin and diversification of language*. Chicago: Aldine.

Ekaterina Vylomova, Jennifer White, Elizabeth Salesky, Sabrina J. Mielke, Shijie Wu, Edoardo Maria Ponti, Rowan Hall Maudslay, Ran Zmigrod, Josef Valvoda, Svetlana Toldova, Francis Tyers, Elena Klyachko, Ilya Yegorov, Natalia Krizhanovsky, Paula Czarnowska, Irene Nikkarinen, Andrew Krizhanovsky, Tiago Pimentel, Lucas Torroba Hennigen, Christo Kirov, Garrett Nicolai, Adina Williams, Antonios Anastasopoulos, Hilaria Cruz, Eleanor Chodroff, Ryan Cotterell, Miikka Silfverberg, and Mans Hulden. 2020. SIGMORPHON 2020 shared task 0: Typologically diverse morphological inflection. In *Proceedings of the 17th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 1–39, Online, July. Association for Computational Linguistics.

Derry Tanti Wijaya, Brendan Callahan, John Hewitt, Jie Gao, Xiao Ling, Marianna Apidianaki, and Chris Callison-Burch. 2017. Learning translations via matrix completion. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1452–1463, Copenhagen, Denmark, September. Association for Computational Linguistics.

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, May. European Languages Resources Association (ELRA).

Winston Wu, Nidhi Vyas, and David Yarowsky. 2018. Creating a translation matrix of the Bible's names across 591 languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*, Miyazaki, Japan, May. European Languages Resources Association (ELRA).

Winston Wu, Garrett Nicolai, and David Yarowsky. 2020. Multilingual dictionary based construction of core vocabulary. In *Proceedings of the Twelfth International Conference on Language Resources and Evaluation (LREC 2020)*, Marseilles, France, May. European Language Resources Association (ELRA).

David Yarowsky and Grace Ngai. 2001. Inducing multilingual POS taggers and NP bracketers via robust projection across aligned corpora. In *Second Meeting of the North American Chapter of the Association for Computational Linguistics*.

Kyra Yee, Yann Dauphin, and Michael Auli. 2019. Simple and effective noisy channel modeling for neural machine translation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5696–5701, Hong Kong, China, November. Association for Computational Linguistics.