Cross-Lingual Word Embeddings for Low-Resource Language Modeling

Oliver Adams,[♠♡] **Adam Makarucha,**[♠] **Graham Neubig,**[♣] **Steven Bird,**[♡◇] **Trevor Cohn**[♡]
[♠]IBM Research – Australia
[♡]Computing and Information Systems, University of Melbourne
[♣]School of Computer Science, Carnegie Mellon University
[◇]International Computer Science Institute, University of California Berkeley
oliver.adams@gmail.com, adamjm@au1.ibm.com,
gneubig@cs.cmu.edu, {t.cohn,steven.bird}@unimelb.edu.au

Abstract

Most languages have no established writing system and minimal written records. However, textual data is essential for natural language processing, and particularly important for training language models to support speech recognition. Even in cases where text data is missing, there are some languages for which bilingual lexicons are available, since creating lexicons is a fundamental task of documentary linguistics. We investigate the use of such lexicons to improve language models when textual training data is limited to as few as a thousand sentences. The method involves learning cross-lingual word embeddings as a preliminary step in training monolingual language models. Results across a number of languages show that language models are improved by this pre-training. Application to Yongning Na, a threatened language, highlights challenges in deploying the approach in real low-resource environments.

1 Introduction

Most of the world's languages are not actively written, even languages with an official writing system (Bird, 2011). This limits the available textual data to small quantities of phonemic transcriptions prepared by linguists. Since phonemic transcription is time-consuming, such data is scarce. This makes language modeling, which is a key tool for facilitating speech recognition of these languages, a difficult challenge. One of the touted advantages of neural network language models (NNLMs) is their ability to model sparse data (Bengio et al., 2003; Gandhe et al., 2014). However, despite the success of NNLMs on large datasets (Mikolov et al., 2010; Sutskever et al., 2011; Graves, 2013), it remains unclear whether their advantages transfer to scenarios with extremely limited amounts of data.

Appropriate initialization of parameters in neural network frameworks has been shown to be beneficial across a wide variety of domains, including speech recognition, where unsupervised pre-training of deep belief networks was instrumental in attaining breakthrough performance (Hinton et al., 2012). Neural network approaches to a range of NLP problems have also been aided by initialization with word embeddings trained on large amounts of unannotated text (Frome et al., 2013; Zhang et al., 2014; Lau and Baldwin, 2016). However, in the case of extremely low-resource languages we do not have the luxury of this unannotated text.

As a remedy to this problem we focus on cross-lingual word embeddings (CLWEs), which learn word embeddings using information from multiple languages. Recent advances in CLWEs have shown that high quality embeddings can be learnt even in the absence of bilingual corpora by harnessing bilingual lexicons (Gouws and Søgaard, 2015; Duong et al., 2016). This is useful as some threatened and endangered languages have been subject to significant linguistic investigation, leading to the creation of high-quality lexicons, despite the dearth of transcriptions. For example, the training of a quality speech recognition system for Yongning Na, a Sino-Tibetan language spoken by approximately 40k people, is hindered by this lack of data (Do et al., 2014) despite significant linguistic investigation of the language (Michaud, 2008; Michaud, 2016).

In this paper we address two research questions. First, is the quality of CLWEs dependent on having large amounts of data in multiple languages, or can large amounts of data in a single *source* lan-

guage inform embeddings trained with little *target* language data? Secondly, can such CLWEs improve language modeling in low-resource contexts by initializing the parameters of an NNLM?

To answer these questions, we scale down the available monolingual data of the target language to as few as 1k sentences, while maintaining a large source language dataset. We assess intrinsic embedding quality by considering correlation with human judgment on the WordSim353 test set (Finkelstein et al., 2001). We then perform language modeling experiments where we initialize the parameters of a long short-term memory (LSTM) language model for low-resource language model training across a variety of language pairs.

Results indicate that CLWEs remain resilient when target language training data is drastically reduced in a simulated low-resource environment, and that initializing the embedding layer of an NNLM with these CLWEs consistently leads to better performance of the language model. In light of these results, we explore the method's application to Na, an actual low-resource language with realistic manually created lexicons and transcribed data. We present a discussion of the negative results found which highlights challenges and future opportunities.

2 Related Work

This paper draws on work in three general areas, which we briefly describe in this section.

Neural network language models and word embeddings Bengio et al. (2003) and Goodman (2001) introduce word embeddings in the context of an investigation of neural language modeling. One claimed advantage of such models is the ability to cope with sparse data by sharing information among words with similar characteristics. Neural language modeling has since demonstrated powerful capabilities at the word level (Mikolov et al., 2010) and character level (Sutskever et al., 2011). Notably, LSTM models (Hochreiter and Schmidhuber, 1997) for modeling long-ranging statistical influences have been shown to be effective (Graves, 2013; Zaremba et al., 2014).

Word embeddings have became more popular through the application of shallow neural network architectures that allow for training on large quantities of data (Mnih et al., 2009; Bengio et al., 2009; Collobert and Weston, 2008; Mikolov et al., 2013a), leading to many further investigations (Chen et al., 2013; Pennington et al., 2014; Shazeer et al., 2016; Bhatia et al., 2016). A key application of word embeddings has been in the initializing of neural network architectures for a wide variety of NLP tasks with limited annotated data (Frome et al., 2013; Zhang et al., 2014; Zoph et al., 2016; Lau and Baldwin, 2016).

Low-resource language modeling and language model adaptation Bellegarda (2004) review language model adaptation, and argue that small amounts of in-domain data are often more valuable than large amounts of out-of-domain data, but that adapting background models using in-domain data can be even better. Kurimo et al. (2016) present more recent work on improving large vocabulary continuous speech recognition using language model adaptation for low-resource Finno-Ugric languages.

Cross-lingual language modeling has also been explored with work on interpolation of a sparse language model with one trained on a large amount of translated data (Jensson et al., 2008), and integrated speech recognition and translation (Jensson et al., 2009; Xu and Fung, 2013).

Gandhe et al. (2014) investigate NNLMs for low-resource languages, comparing NNLMs with count-based language models, and find that NNLMs interpolated with count-based methods outperform standard n-gram models even with small quantities of training data. In contrast, our contribution is an investigation into harnessing CLWEs learnt using bilingual dictionaries in order to improve language modeling in a similar low-resource setting.

Cross-lingual word embeddings Cross-lingual word embeddings have also been the subject of significant investigation. Many methods require parallel corpora or comparable corpora to connect the languages (Klementiev et al., 2012; Zou et al., 2013; Hermann and Blunsom, 2013; Chandar A P et al., 2014; Kočiský et al., 2014; Coulmance et al., 2015; Wang et al., 2016), while others use bilingual dictionaries (Mikolov et al., 2013b; Xiao and Guo, 2014; Faruqui and Dyer, 2014; Gouws and Søgaard, 2015; Duong et al., 2016; Ammar et al., 2016), or neither (Miceli Barone, 2016).

In particular, we build on the work of Duong et al. (2016). Their method harnesses monolingual corpora in two languages along with a bilingual

lexicon to connect the languages and represent the words in a common vector space. The model builds on the continuous bag-of-words (CBOW) model (Mikolov et al., 2013a) which learns embeddings by predicting words given their contexts. The key difference is that the word to be predicted is a target language translation of a source language word centered in a source language context.

Since dictionaries tend to include a number of translations for words, the model uses an iterative expectation-maximization style training algorithm in order to best select translations given the context. This process thus allows for polysemy to be addressed which is desirable given the polysemous nature of bilingual dictionaries.

3 Resilience of Cross-Lingual Word Embeddings

Previous work using CLWEs typically assumes a similar amount of training data of each available language, often in the form of parallel corpora. Recent work has shown that monolingual corpora of two different languages can be tied together with bilingual dictionaries in order to learn embeddings for words in both languages in a common vector space (Gouws and Søgaard, 2015; Duong et al., 2016). In this section we relax the assumption of the availability of large monolingual corpora on the source and target sides, and report an experiment on the resilience of such CLWEs when data is scarce in the target language but plentiful in a source language.

3.1 Experimental Setup

Word embedding quality is commonly assessed by evaluating the correlation of the cosine similarity of the embeddings with human judgements of word similarity. Here we follow the same evaluation procedure, except where we simulate a low-resource language by reducing the availability of target English monolingual text while preserving a large quantity of source language text from other languages. This allows us to evaluate the CLWEs intrinsically using the WordSim353 task (Finkelstein et al., 2001) before progressing to downstream language modeling where we additionally consider other target languages.

We trained a variety of embeddings on English Wikipedia data of between 1k and 128k sentences from the training data of Al-Rfou et al. (2013). In terms of transcribed speech data, this roughly equates to between 1 and 128 hours of speech. For the training data, we randomly chose sentences that include words in the WordSim353 task proportionally to their frequency in the set. As monolingual baselines, we use the *skip-gram* (SG) and CBOW methods of Mikolov et al. (2013a) as implemented in the *Gensim* package (Řehůřek and Sojka, 2010). We additionally used off-the-shelf CBOW Google News Corpus embeddings with 300 dimensions, trained on 100 billion words.

The CLWEs were trained using the method of Duong et al. (2016) since their method addresses polysemy which is rampant in dictionaries. The same 1k-128k sentence English Wikipedia data was used but with an additional 5 million sentences of Wikipedia data in a source language. The source languages include Japanese, German, Russian, Finnish, and Spanish, which represent languages of varying similarity with English, some with great morphological and syntactic differences. To relate the languages, we used the *PanLex* lexicon (Kamholz et al., 2014). Following Duong et al. (2016), we used the default window size of 48 so that the whole sentence's context is almost always taken into account. This mitigates the effect of word re-ordering between languages. We trained with an embedding dimension of 200 for all data sizes as a larger dimension turned out to be helpful in capturing information from the source side.[1]

3.2 Results

Figure 1 shows correlations with human judgment in the WordSim353 task. The x-axis represents the number of English training sentences. Coloured lines represent CLWEs trained on different languages: Japanese, German, Spanish, Russian and Finnish.[2]

With around 128k sentences of training data, most methods perform quite well, with German being the best performing. Interestingly the CLWE methods all outperform GNC which was trained on a far larger corpus of 100 billion words. With only 1k sentences of target training data, all the CLWEs have a correlation around 0.5, with the exception of Finnish. Interestingly, no consis-

[1] Hyperparameters for both mono and cross-lingual word embeddings: iters=15, negative=25, size=200, window=48, otherwise default. Smaller window sizes led to similar results for monolingual methods.

[2] We also tried Italian, Dutch, German and Serbian, yielding similar results but omitted for presentation.

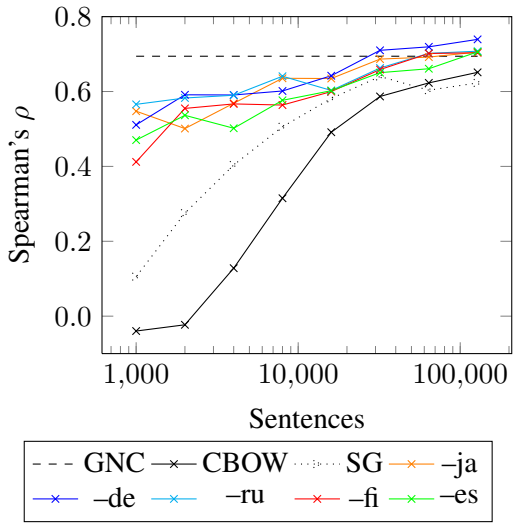

Figure 1: Performance of different embeddings on the WordSim353 task with different amounts of training data. *GNC* is the Google News Corpus embeddings, which are constant. *CBOW* and *SG* are the monolingual word2vec embeddings. The other, colored, lines are all cross-lingual word embeddings harnessing the information of 5m sentences of various source languages.

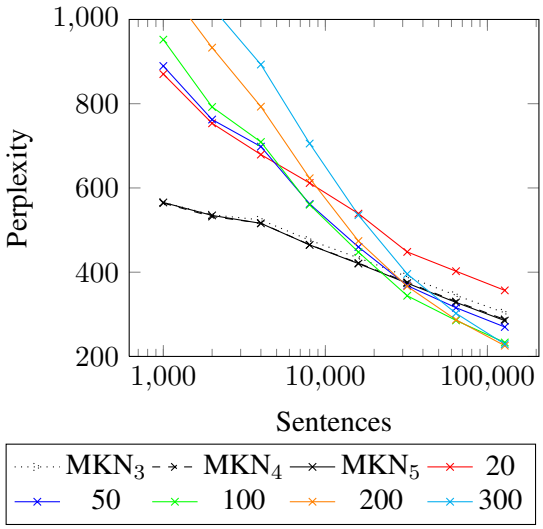

Figure 2: Perplexity of language models on the validation set. Numbers in the legend indicate LSTM language models with different hidden layer sizes, as opposed to Modified Kneser-Ney language models of order 3, 4 and 5.

tent benefit was gained by using source languages for which translation with English is simpler. For example, Spanish often under-performed Russian and Japanese as a source language, as well as the morphologically-rich Finnish.

Notably, all the CLWEs perform far better than their monolingual counterparts on small amounts of data. This resilience of the target English word embeddings suggests that CLWEs can serve as a method of transferring semantic information from resource-rich languages to the resource-poor, even when the languages are quite different. However, the WordSim353 task is a constrained environment, so in the next section we turn to language modeling, a natural language processing task of much practical importance for resource-poor languages.

4 Pre-training Language Models

Language models are an important tool with particular application to machine translation and speech recognition. For resource-poor languages and unwritten languages, language models are also a significant bottleneck for such technologies as they rely on large quantities of data. In this section, we assess the performance of language models on varying quantities of data, across a number of different source–target language pairs. In particular, we use CLWEs to initialize the first layer in an LSTM recurrent neural network language model and assess how this affects language model performance. This is an interesting task not simply for the practical advantage of having better language models for low-resource languages. Language modeling is a syntax-oriented task, yet syntax varies greatly between the languages we train the CLWEs on. This experiment thus yields some additional information about how effectively bilingual information can be used for the task of language modeling.

4.1 Experimental Setup

We experiment with a similar data setup as in Section 3. However, target training sentences are not constrained to include words observed in the WordSim353 set, and are random sentences from the aforementioned 5 million sentence corpus. For each language, the validation and test sets consist of 3k randomly selected sentences. The large vocabulary of Wikipedia and the small amounts of training data used make this a particularly challenging language modeling task.

For our NNLMs, we use the LSTM language model of Zaremba et al. (2014). As a count-based baseline, we use Modified Kneser-Ney (MKN) (Kneser and Ney, 1995; Chen and Goodman, 1999) as implemented in KenLM (Heafield,

2011). Figure 2 presents some results of tuning the dimensions of the hidden layer in the LSTM with respect to perplexity on the validation set,[3] as well as tuning the order of n-grams used by the MKN language model. A dimension of 100 yielded a good compromise between the smaller and larger training data sizes, while an order 5 MKN model performed slightly better than its lower-order brethren.[4]

Interestingly, MKN strongly outperforms the LSTM on low quantities of data, with the LSTM language model not reaching parity until between 16k and 32k sentences of data. This is consistent with the results of Chen et al. (2015) and Neubig and Dyer (2016) that show that n-gram models are typically better for rare words, and here our vocabulary is large but training data small since the data are random Wikipedia sentences. However these findings are inconsistent with the belief that NNLMs have the ability to cope well with sparse data conditions because of the smooth distributions that arise from using dense vector representations of words (Bengio et al., 2003). Traditional smoothing stands strong.

4.2 English Results

With the parameters tuned on the English validation set as above, we evaluated the LSTM language model when the embedding layer is initialized with various monolingual and cross-lingual word embeddings. Figure 3 compares the performance of a number of language models on the test set. In every case where pre-trained embeddings were used, the embedding layer was held fixed during training. However, we observed similar results when allowing them to deviate from their initial state. For the CLWEs, the same language set was used as in Section 3. The curves for the source languages (Dutch, Greek, Finnish, and Japanese) are remarkably similar, as were those for the languages omitted from the figure (German, Russian, Serbian, Italian, and Spanish). This suggests that the English target embeddings are gleaning similar information from each of the languages, information likely to be more semantic than syntactic, given the syntactic differences between the languages.

[3]We used 1 hidden layer but otherwise the same as the *SmallConfig* of models/rnn/ptb/ptb_word_lm.py available in Tensorflow.

[4]Note that all perplexities in this paper include out-of-vocabulary words, of which there are many.

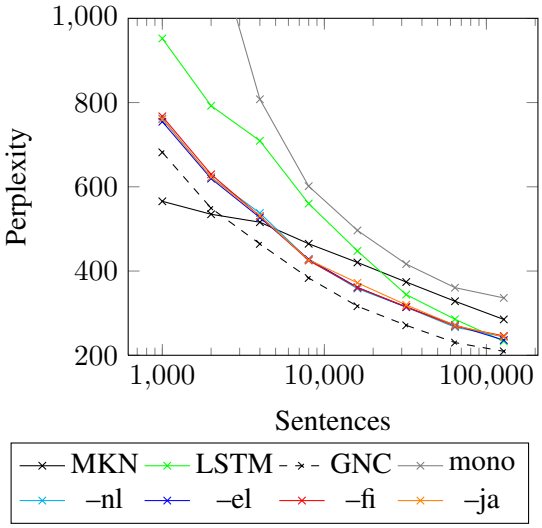

Figure 3: Perplexity of LSTMs when pre-trained with cross-lingual word embeddings trained on the same data. *LSTM* is a neural network language model with no pre-trained embeddings. *mono* is pre-trained with monolingual word2vec embeddings. *GNC* is pre-trained with Google News Corpus embeddings of dimension 300. The rest are pre-trained with CLWEs using information transfer from different source languages. *MKN* is an order 5 Modified Kneser-Ney baseline.

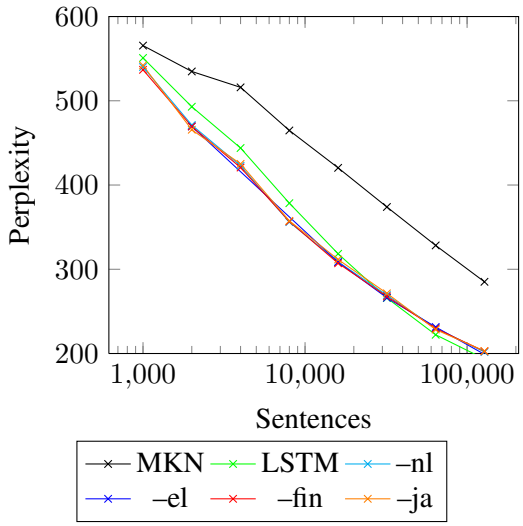

Figure 4: Perplexities when interpolating MKN with LSTMs pre-trained with various cross-lingual word embeddings. *LSTM* interpolates MKN with a neural network language model with no pre-trained embeddings. The rest are interpolations of MKN with LSTMs pre-trained with CLWEs using information transfer from different source languages. *MKN* is an order 5 Modified Kneser-Ney baseline without interpolation.

We compare these language models pre-trained with CLWEs with pre-training using other embeddings. Pre-training with the Google News Corpus embeddings of the method of Mikolov et al. (2013c) unsurprisingly performs the best due to the large amount of English data not available to the other methods, making it a sort of oracle. Monolingual pre-training of word embeddings on the same English data (*mono*) used by the CLWEs yields poorer performance.

The language models initialized with pre-trained CLWEs are significantly better than their un-pre-trained counterpart on small amounts of data, reaching par performance with *MKN* at somewhere just past 4k sentences of training data. In contrast, it takes more than 16k sentences of training data before the plain LSTM language model began to outperform *MKN*. The outperformance of LSTMs by MKN with the lowest amounts of training data motivated interpolation of MKN probabilities with LSTM language model probabilities, as shown in Figure 4. Such interpolation allows for consistent improvement beyond the performance of MKN or CLWE-pre-trained LSTMs alone.

4.3 Other Target Languages

In Table 1 we present results of language model experiments run with other languages used as the low-resource target. In this table English is used in each case as the large source language with which to help train the CLWEs. The observation that the CLWE-pre-trained language model tended to perform best relative to alternatives at around 8k or 16k sentences in the English case prompted us to choose these slices of data when assessing other languages as targets.

The pre-trained LSTM language model outperforms its non-pre-trained counterpart for all languages. There is competition between *MKN* and the CLWE-pre-trained models. The languages for which *MKN* tends to do better are typically those further from English or those with rich morphology, making cross-lingual transfer of information more challenging. There seems to be a degree of asymmetry here: while all languages helped English language modeling similarly, English helps the other languages to varying degrees. For all languages, interpolating MKN with the CLWE (*Interp.*) yields the best performance, corroborating the findings of Gandhe et al. (2014).

Neural language modeling of sparse data can be improved by initializing parameters with cross-lingual word embeddings. The consistent performance improvements gained by an LSTM using CLWE-initialization is a promising sign for CLWE-initialization of neural networks for other tasks given limited target language data.

5 First Steps in an Under-Resourced Language

Having demonstrated the effectiveness of CLWE-pre-training of language models using simulation in a variety of well-resourced written languages, we proceed to a preliminary investigation of this method to a low-resource, unwritten language, Na.

Yongning Na is a Sino-Tibetan language spoken by approximately 40k people in an area in Yunnan, China, near the border with Sichuan. It has no orthography and is tonal with a rich morphophonology. Given the small quantity of manually transcribed phonemic data available in the language, Na provides an ideal test bed for investigating the potential and difficulties this method faces in a realistic setting. In this section we report results in Na language modeling and discuss hurdles to be overcome.

5.1 Experimental Setup

The phonemically transcribed corpus[5] consists of 3,039 phonemically transcribed sentences which are a subset of a larger spoken corpus. These sentences are segmented at the level of the word, morpheme and phonological process, and have been translated into French, with smaller amounts translated into Chinese and English. The corpus also includes word-level glosses in French and English. The lexicon of Michaud (2016) contains example sentences for entries, as well as translations into French, English and Chinese.

The lexicon consists of around 2k Na entries, with example sentences and translations into English, French and Chinese. To choose an appropriate segmentation of the corpus, we used a hierarchical segmentation method where words were queried in the lexicon. If a given word was present then it was kept as a token, otherwise the word was split into its constituent morphemes.

We took 2,039 sentences to be used as training data, with the remaining 1k sentences split

[5] Available as part of the Pangloss collection at http://lacito.vjf.cnrs.fr/pangloss.

	8k sentences				16k sentences			
Lang	MKN	LSTM	CLWE	Interp.	MKN	LSTM	CLWE	Interp.
Greek	827.3	920.3	780.4	**650.6**	749.8	687.9	634.4	**549.5**
Serbian	492.8	586.3	521.3	**408.0**	468.8	485.3	447.8	**365.7**
Russian	1656.8	2054.5	1920.4	**1466.2**	1609.5	1757.3	1648.3	**1309.1**
Italian	777.0	794.9	688.3	**592.2**	686.2	627.7	559.7	**493.4**
German	997.4	1026.0	1000.9	**831.8**	980.0	908.8	874.1	**761.5**
Finnish	1896.4	2438.8	2165.5	**1715.3**	1963.3	2233.2	2109.9	**1641.2**
Dutch	492.1	491.3	456.2	**381.4**	447.9	412.8	378.0	**330.1**
Japanese	1902.8	2662.4	2475.6	**1866.7**	1816.8	2462.8	2279.6	**1696.9**
Spanish	496.3	481.8	445.6	**387.7**	445.9	412.9	369.6	**331.2**

Table 1: Perplexity of language models trained on 8k and 16k sentences for different languages. *MKN* is an order 5 Modified Kneser-Ney language model. *LSTM* is a long short-term memory neural network language model with no pre-training. *CLWE* is an LSTM language model pre-trained with cross-lingual word embeddings, using English as the source language. *Interp.* is an interpolation of MKN with CLWE.

	Types	Tokens
Tones	2,045	45,044
No tones	1,192	45,989

Table 2: Counts of types and tokens across the whole Na corpus, given our segmentation method.

	Tones	No tones
MKN	**59.4**	**38.0**
LSTM	74.8	46.0
CLWE	76.6	46.2
Lem	76.8	44.7
En-split	76.4	47.0

Table 3: Perplexities on the Na test set using English as the source language. *MKN* is an order 5 Modified Kneser-Ney language model. *LSTM* is a neural network language model without pretraining. *CLWE* is the same LM with pre-trained Na–English CLWEs. *Lem* is the same as CLWE except with English lemmatization. *En-split* extends this by preprocessing the dictionary such that entries with multiple English words are converted to multiple entries of one English word.

equally between validation and test sets. The phonemic transcriptions include tones, so we created two preprocessed versions of the corpus: with and without tones. Table 2 exhibits type and token counts for these two variations. In addition to the CLWE approach used in Sections 3 and 4, we additionally tried lemmatizing the English Wikipedia corpus so that it each token was more likely to be present in the Na–English lexicon.

5.2 Results and Discussion

Table 3 shows the Na language modeling results. Pre-trained CLWEs do not significantly outperform that of the non-pre-trained, and *MKN* outperforms both. Given the size of the training data, and the results of Section 4, it is no surprise that *MKN* outperforms the NNLM approaches. But the lack of benefit in CLWE-pre-training the NNLMs requires some reflection. We now proceed to discuss the challenges of this data to explore why the positive results of language model pre-training that were seen in Section 4 were not seen in this experiment.

Tones A key challenge arises because of Na's tonal system. Na has rich tonal morphology. Syntactic relationships between words influence the surface form tone a syllable takes. Thus, semantically identical words may take different surface tones than is present in the relevant lexical entry, resulting in mismatches with the lexicon.

If tones are left present, the percentage of Na tokens present in the lexicon is 62%. Removing tones yields a higher hit rate of 88% and allows tone mismatches between surface forms and lexical entries to be overcome. This benefit is gained in exchange for higher polysemy, with an average of 4.1 English translations per Na entry when tones are removed, as opposed to 1.9 when tones are present. Though this situation of polysemy is what the method of Duong et al. (2016) is designed to address, it means the language model fails to model tones and doesn't significantly help CLWE-pre-training in any case. Future work should investigate morphophonological

processing for Na, since there is regularity behind these tonal changes (Michaud, 2008) which could mitigate these issues if addressed.

Polysemy We considered the polysemy of the tokens of other languages' corpora in the PanLex dictionaries. Interestingly they were higher than the Na lexicon with tones removed, ranging from 2.7 for Greek–English to 19.5 for German–English. It seems the more important factor is the amount of tokens in the English corpus that were present in the lexicon. For the Na–English lexicon, this was only 18% and 20% when lemmatized and unlemmatized, respectively. However it was 67% for the PanLex lexicon. Low lexicon hit rates of both the Na and English corpora must damage the CLWEs modeling capacity.

Lexicon word forms Not all the forms of many English word groups are represented. For example, only the infinitive *'to_run'* is present, while *'running'*, *'ran'* and *'runs'* are not. The limited scope of this lexicon motivates lemmatization on the English side as a normalization step, which may be of some benefit (see Table 3). Furthermore, such lemmatization can be expected to reduce the syntactic information present in embeddings which does not transfer between languages as effectively as semantics.

Some common words, such as *'reading'* are not present in the lexicon, but *'to_read_aloud'* is. Additionally, there are frequently entries such as *'way_over_there'* and *'masculine_given_name'* that are challenging to process. As an attempt to mitigate this issue, we segmented such English entries, creating multiple Na–English entries for each. However, results in Table 3 show that this failed to show improvements. More sophisticated processing of the lexicon is required.

Lexicon size There are about 2,115 Na entries in the lexicon and 2,947 Na–English entries, which makes the lexicon especially small in comparison to the PanLex lexicon used in the previous experiments. Duong et al. (2016) report large reductions in performance of CLWEs on some tasks when lexicon size is scaled down to 10k.

To better understand how limited lexicon size could be affecting language model performance, we performed an ablation experiment where random entries in the PanLex English–German lexicon were removed in order to restrict its size. Figure 5 shows the performance of English language modeling when training data is restricted to 2k sentences (to emulate the Na case) and the size of the lexicon afforded to the CLWE training is adjusted. This can only serve as a rough comparison, since PanLex is large and so a 1k entry subset may contain many obscure terms and few useful ones. Nevertheless, results suggest that a critical point occurs somewhere in the order of 10k entries. However, since improvements are demonstrated even with smaller dictionaries, this is further evidence that more sophisticated preprocessing of the Na lexicon is required.

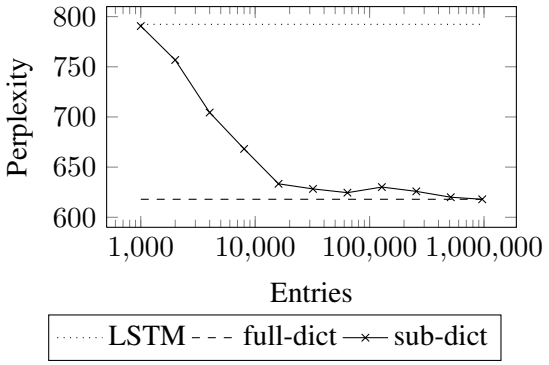

Figure 5: Perplexities of an English–German CLWE-pretrained language model trained on 2k English sentences as the dictionary size available in CLWE training increases to its full size (*sub-dict*). As points of comparison, *LSTM* is a long short-term memory language model with no pretraining and *full-dict* is a CLWE-pretrained language model with the full dictionary available.

Domain Another difference that may contribute to the results is that the domain of the text is significantly different. The Na corpus is a collection of spoken narratives transcribed, while the Wikipedia articles are encyclopaedic entries, which makes the registers very different.

5.3 Future Work on Na Language Modeling

Though the technique doesn't work out of the box, this sets a difficult and compelling challenge of harnessing the available Na data more effectively.

The lexicon is a rich source of other information, including part-of-speech tags, example sentences and multilingual translations. In addition to better preprocessing of the lexical information we have already used, harnessing this additional information is an important next step to improving Na language modeling. The corpus includes translations into French, Chinese and English, as well

as glosses. Some CLWE methods can additionally utilize such parallel data (Coulmance et al., 2015; Ammar et al., 2016) and we leave to future work incorporation of this information as well.

The tonal system is well described (Michaud, 2008), and so further Na-specific work should allow differences between surface form tones and tones in the lexicon to be bridged.

Our work corroborates the observation that MKN performs well on rare words (Chen et al., 2015). Interpolation is an effective means to harness this strength when training data is sparse. Furthermore, hybrid count-based and NNLMs (Neubig and Dyer, 2016) promise the best of both worlds for language modeling for low-resource languages.

6 Conclusion

In this paper we have demonstrated that CLWEs can remain resilient when training data in the target language is scaled down drastically. Such CLWEs continue to perform well on the WordSim353 task, as well as demonstrating downstream efficacy across a number of languages through initialization of NNLMs. This work supports CLWEs as a method of transfer of information to resource-poor languages by harnessing distributional information in a large source language. We can expect parameter initialization with CLWEs trained on such asymmetric data conditions to aid in other NLP tasks too, though this should be empirically assessed.

Acknowledgements

This work was conducted during Oliver Adams' internship at IBM Research Australia. We are grateful for support from NSF Award 1464553 and the DARPA/I2O, Contract Nos. HR0011-15-C-0114 and HR0011-15-C-0115.

References

Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. 2013. Polyglot: Distributed word representations for multilingual NLP. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 183–192.

Waleed Ammar, George Mulcaire, Yulia Tsvetkov, Guillaume Lample, Chris Dyer, and Noah A. Smith. 2016. Massively multilingual word embeddings. *arXiv:1602.01925*.

Jerome R. Bellegarda. 2004. Statistical language model adaptation: Review and perspectives. *Speech Communication*, 42(1):93–108.

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Janvin. 2003. A neural probabilistic language model. *The Journal of Machine Learning Research*, 3:1137–1155.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. *Proceedings of the 26th annual international conference on machine learning*, pages 41–48.

Parminder Bhatia, Robert Guthrie, and Jacob Eisenstein. 2016. Morphological priors for probabilistic neural word embeddings. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 490–500.

Steven Bird. 2011. Bootstrapping the language archive: New prospects for natural language processing in preserving linguistic heritage. *Linguistic Issues in Language Technology*, 6:1–16.

Sarath Chandar A P, Stanislas Lauly, Hugo Larochelle, Mm Mitesh Khapra, Balaraman Ravindran, Vikas C Raykar, and Amrita Saha. 2014. An autoencoder approach to learning bilingual word representations. In *Advances in Neural Information Processing Systems 27*, pages 1853–1861.

Stanley F Chen and Joshua Goodman. 1999. An empirical study of smoothing techniques for language modeling. *Computer Speech & Language*, 13(4):359–394.

Yanqing Chen, Bryan Perozzi, R Al-Rfou, and Steven Skiena. 2013. The expressive power of word embeddings. *arXiv:1301.3226*.

Welin Chen, David Grangier, and Michael Auli. 2015. Strategies for training large vocabulary neural language models. *arXiv:1512.04906*.

Ronan Collobert and Jason Weston. 2008. A unified architecture for natural language processing: Deep neural networks with multitask learning. *Proceedings of the 25th international conference on Machine learning*, pages 160–167.

Jocelyn Coulmance, Jean-Marc Marty, Guillaume Wenzek, and Amine Benhalloum. 2015. Trans-gram, fast cross-lingual word-embeddings. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1109–1113.

Thi-Ngoc-Diep Do, Alexis Michaud, and Eric Castelli. 2014. Towards the automatic processing of Yongning Na (Sino-Tibetan): developing a 'light' acoustic model of the target language and testing 'heavyweight' models from five national languages. In *4th International Workshop on Spoken Language Technologies for Under-resourced Languages*, pages 153–160.

Long Duong, Hiroshi Kanayama, Tengfei Ma, Steven Bird, and Trevor Cohn. 2016. Learning crosslingual word embeddings without bilingual corpora. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1285–1295.

Manaal Faruqui and Chris Dyer. 2014. Improving vector space word representations using multilingual correlation. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 462–471.

Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. 2001. Placing search in context: The concept revisited. In *Proceedings of the 10th International Conference on World Wide Web*, WWW '01, pages 406–414.

Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. 2013. Devise: A deep visual-semantic embedding model. In *Advances in Neural Information Processing Systems 26*, pages 2121–2129.

Ankur Gandhe, Florian Metze, and Ian Lane. 2014. Neural network language models for low resource languages. In *INTERSPEECH-2014*, pages 2615–2619.

Joshua Goodman. 2001. A Bit of Progress in Language Modeling. *Technical Report*.

Stephan Gouws and Anders Søgaard. 2015. Simple task-specific bilingual word embeddings. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1386–1390.

Alex Graves. 2013. Generating sequences with recurrent neural networks. *arXiv:1308.0850*.

Kenneth Heafield. 2011. Kenlm: Faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 187–197.

Karl Moritz Hermann and Phil Blunsom. 2013. A simple model for learning multilingual compositional semantics. *arXiv:1312.6173*.

Geoffrey Hinton, Li Deng, Dong Yu, George E Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N Sainath, and Others. 2012. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *Signal Processing Magazine, IEEE*, 29(6):82–97.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Arnar Jensson, Koji Iwano, and Sadaoki Furui. 2008. Development of a speech recognition system for icelandic using machine translated text. In *The first International Workshop on Spoken Languages Technologies for Under-resourced Languages*, pages 18–21.

Arnar Jensson, Tasuku Oonishi, Koji Iwano, and Sadaoki Furui. 2009. Development of a WFST based speech recognition system for a resource deficient language using machine translation. *Proceedings of Asia-Pacific Signal and Information Processing Association*, pages 50–56.

David Kamholz, Jonathan Pool, and Susan Colowick. 2014. PanLex: Building a resource for panlingual lexical translation. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation*, pages 3145–3150.

Alexandre Klementiev, Ivan Titov, and Binod Bhattarai. 2012. Inducing crosslingual distributed representations of words. In *Proceedings of COLING 2012*, pages 1459–1474.

Reinhard Kneser and Hermann Ney. 1995. Improved backing-off for m-gram language modeling. In *1995 International Conference on Acoustics, Speech, and Signal Processing*, pages 181–184.

Tomáš Kočiský, Karl Moritz Hermann, and Phil Blunsom. 2014. Learning bilingual word representations by marginalizing alignments. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 224–229.

Mikko Kurimo, Seppo Enarvi, Ottokar Tilk, Matti Varjokallio, André Mansikkaniemi, and Tanel Alumäe. 2016. Modeling under-resourced languages for speech recognition. *Language Resources and Evaluation*, pages 1–27.

Jey Han Lau and Timothy Baldwin. 2016. An empirical evaluation of doc2vec with practical insights into document embedding generation. In *1st Workshop on Representation Learning for NLP*, pages 78–86.

Antonio Valerio Miceli Barone. 2016. Towards cross-lingual distributed representations without parallel text trained with adversarial autoencoders. In *Proceedings of the 1st Workshop on Representation Learning for NLP*, pages 121–126.

Alexis Michaud. 2008. Phonemic and tonal analysis of Yongning Na*. *Cahiers de Linguistique Asie Orientale*, 37(2):159–196.

Alexis Michaud. 2016. Online Na-English-Chinese Dictionary. *https://halshs.archives-ouvertes.fr/halshs-01204638*. This is version 1.1 of the dictionary.

Tomas Mikolov, Martin Karafiát, Lukas Burget, Jan Cernockỳ, and Sanjeev Khudanpur. 2010. Recurrent neural network based language model. In *INTERSPEECH-2010*, pages 1045–1048.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representation in vector space. *arXiv:1301.3781*.

Tomas Mikolov, Quoc V Le, and Ilya Sutskever. 2013b. Exploiting similarities among languages for machine translation. *arXiv:1309.4168*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013c. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Andriy Mnih, Zhang Yuecheng, and Geoffrey Hinton. 2009. Improving a statistical language model through non-linear prediction. *Neurocomputing*, 72(7-9):1414–1418.

Graham Neubig and Chris Dyer. 2016. Generalizing and hybridizing count-based and neural language models. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1163–1172.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 1532–1543.

Radim Řehůřek and Petr Sojka. 2010. Software framework for topic modelling with large corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50.

Noam Shazeer, Ryan Doherty, Colin Evans, and Chris Waterson. 2016. Swivel: Improving embeddings by noticing what's missing. *arXiv:1602.02215*.

Ilya Sutskever, James Martens, and Geoffrey E Hinton. 2011. Generating text with recurrent neural networks. In *Proceedings of the 28th International Conference on Machine Learning*, pages 1017–1024.

Rui Wang, Hai Zhao, Sabine Ploux, Bao-Liang Lu, and Masao Utiyama. 2016. A novel bilingual word embedding method for lexical translation using bilingual sense clique. *arXiv:1607.08692*.

Min Xiao and Yuhong Guo. 2014. Distributed word representation learning for cross-lingual dependency parsing. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*, pages 119–129.

Ping Xu and Pascale Fung. 2013. Cross-lingual language modeling for low-resource speech recognition. *IEEE Transactions on Audio, Speech and Language Processing*, 21(6):1134–1144.

Wojciech Zaremba, Ilya Sutskever, and Oriol Vinyals. 2014. Recurrent neural network regularization. *arXiv:1409.2329*.

Jiajun Zhang, Shujie Liu, Mu Li, Ming Zhou, and Chengqing Zong. 2014. Bilingually-constrained phrase embeddings for machine translation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 111–121.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer learning for low-resource neural machine translation. *arXiv:1604.02201*.

Will Y. Zou, Richard Socher, Daniel Cer, and Christopher D. Manning. 2013. Bilingual word embeddings for phrase-based machine translation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1393–1398.