Using Related Languages to Enhance Statistical Language Models

Anna Currey, Alina Karakanta
Department of Computational Linguistics, Saarland University, Saarbrücken, Germany
amscurrey@gmail.com, alinak@coli.uni-saarland.de

Abstract

The success of many language modeling methods and applications relies heavily on the amount of data available. This problem is further exacerbated in statistical machine translation, where parallel data in the source and target languages is required. However, large amounts of data are only available for a small number of languages; as a result, many language modeling techniques are inadequate for the vast majority of languages. In this paper, we attempt to lessen the problem of a lack of training data for low-resource languages by adding data from related high-resource languages in three experiments. First, we interpolate language models trained on the target language and on the related language. In our second experiment, we select the sentences most similar to the target language and add them to our training corpus. Finally, we integrate data from the related language into a translation model for a statistical machine translation application. Although we do not see many significant improvements over baselines trained on a small amount of data in the target language, we discuss some further experiments that could be attempted in order to augment language models and translation models with data from related languages.

1 Introduction

Statistical language modeling methods are an essential part of many language processing applications, including automatic speech recognition (Stolcke, 2002), machine translation (Kirchhoff and Yang, 2005), and information retrieval (Liu and Croft, 2005). However, their success is heavily dependent on the availability of suitably large text resources for training (Chen and Goodman, 1996). Such data can be hard to obtain, especially for low-resource languages. This problem is especially acute when language modeling is used in statistical machine translation, where a lack of parallel resources for a language pair can be a significant detriment to quality.

Our goal is to exploit a high-resource language to improve modeling of a related low-resource language, which is applicable to cases where the target language is closely related to a language with a large amount of text data available. For example, languages that are not represented in the European Parliament, such as Catalan, can be aided by related languages that are, such as Spanish. The data available from the related high-resource language can be adapted in order to add to the translation model or the language model of the target language. This paper is an initial attempt at using minimally transformed data from a related language to enhance language models and increase parallel data for SMT.

2 Background and Previous Work

2.1 Domain Adaptation

This problem can be seen as a special case of domain adaptation, with the in-domain data being the data in the target language and the out-of-domain data being the data in the related language (Nakov and Ng, 2012). Domain adaptation is often used to leverage resources for a specific domain, such as biomedical text, from more general domains like newswire data (Dahlmeier and Ng, 2010). This idea can be applied to SMT, where data from the related lan-

Proceedings of NAACL-HLT 2016, pages 116–123,
San Diego, California, June 12-17, 2016. ©2016 Association for Computational Linguistics

guage can be adapted to look like data from the low-resource language. It has been shown that training on a large amount of adapted text significantly improves results compared to training on a small in-domain corpus or training on unadapted data (Wang et al., 2012). In this paper, we apply two particular domain adaptation approaches. First, we interpolate language models from in-domain and out-of-domain data, following Koehn and Schroeder (2007). We also attempt to select the best out-of-domain data using perplexity, similar to what was done in Gao et al. (2002).

2.2 Machine Translation

In contrast to transfer-based and word-based machine translation, for statistical machine translation, quality is heavily dependent on the amount of parallel resources. Given the difficulty of obtaining sufficient parallel resources, this can be a problem for many language pairs. For those cases, a third language can be used as a pivot. The process of using a third language as a bridge instead of directly translating is called triangulation (Singla et al., 2014). Character-level translation combined with word-level translation has also been shown to be an improvement over phrase-based approaches for closely related languages (Nakov and Tiedemann, 2012). Similarly, transliteration methods using cognate extraction (Nakov and Ng, 2012) and bilingual dictionaries (Kirschenbaum and Wintner, 2010) can be used to aid the low-resource language.

3 Experimental Framework

3.1 Choice of Languages

For the purpose of our experiments, we treat Spanish as if it were a low-resource language and test Spanish language models and English-Spanish translations. We use Italian and Portuguese as the closely-related languages. Using these languages for our experiments allows us to compare the results to the language models and machine translations that can be created using large corpora.

Spanish, Portuguese, and Italian all belong to the Romance family of Indo-European languages. Spanish has strong lexical similarity with both Portuguese (89%) and Italian (82%) (Lewis, 2015). Among major Romance languages, Spanish and Portuguese have been found to be the closest pair in automatic corpus comparisons (Ciobanu and Dinu, 2014) and in comprehension studies (Voigt and Gooskens, 2014), followed by Spanish and Italian.

3.2 Data

We used the Europarl corpus (Koehn, 2005) for training and testing. In order to use the data in our experiments, we tokenized[1] the corpus, converted all words to lowercase, and collapsed all numerical symbols into one special symbol. Finally, we transliterated the Italian and Portuguese corpora to make them more Spanish-like; this process is described in section 3.3.

The data that was used to train, test and develop is split as follows: 10% of the Spanish data (196,221 sentences) was used for testing, 10% for development, and the remaining 80% (1,569,771 sentences) for training. The Italian and Portuguese corpora were split similarly and training sizes for the models varied between 30K and 1,523,304 and 1,566,015 sentences for Italian and Portuguese, respectively.

3.3 Transliteration

In order to use Italian and Portuguese data to model Spanish, we first transliterated the Italian and Portuguese training corpora using a naive rule-based transliteration method consisting of word-level string transformations and a small bilingual dictionary. For the bilingual dictionary, the 200 most common words were extracted from the Italian and the Portuguese training corpora and manually given Spanish translations. In translating to Spanish, an effort was made to keep cognates where possible, and to use the most likely or common meanings.

Table 1 gives translations used for the ten most common Italian words in the data. Even in this small sample, there is a problematic translation. The Italian preposition *per* can be translated to *por* or *para*. In keeping with the desire to use a small amount of data, we briefly read the Italian texts to find the translation we felt was more likely (*para*), and chose that as the translation for all instances of *per* in the training set. We also verified that *para* was more likely in the Spanish training text overall than *por*.

[1] We used the Tok-tok tokenizer by Jon Dehdari: https://github.com/jonsafari/tok-tok

Italian	Spanish	Gloss
di	de	of
e	y	and
che	que	that
la	la	the (f. sg.)
in	en	in
il	el	the (m. sg.)
per	para	for
a	a	to
è	es	is
un	un	a (m. sg.)

Table 1: Sample Italian-Spanish translations.

The rule-based component of the transliteration consisted of handwritten word-initial, word-final, and general transformation rules. We applied approximately fifty such rules per language to the data. In order to come up with the rules, we examined the pan-Romance vocabulary list compiled by EuroComRom (Klein, 2002); however, such rules could be derived by an expert with knowledge of the relevant languages with relatively little effort. Character clusters that were impossible in Spanish were converted to their most common correspondence in Spanish (in the word list). We also identified certain strings that had consistent correspondences in Spanish and replaced them appropriately. These rules were applied to all words in the Italian and Portuguese training data except for those that were in the bilingual dictionary. See table 2 for examples of string transformation rules used for the Italian case.

Type	Original	Translit.	Example
initial	sp	esp	Spagna
initial	qua	cua	qualità
initial	st	est	stare
final	ssioni	siones	impressioni
final	are	ar	stare
final	tà	dad	qualità
general	gn	ñ	Spagna
general	vv	v	improvviso
general	ò	o	però

Table 2: Sample Italian-Spanish transliterations.

Italian text
La difficoltà di conciliare questi obiettivi risiede nel fatto che le logiche di questi settori sono contraddittorie.
Transliteration into Spanish
La dificoldad de conciliar estos obietivos risiede en el hecho que las logique de estos setores son contraditorie.

Table 3: Example of transliterated text using our approach.

4 Experiments

4.1 Experiment 1: Language Model Interpolation

Our first experiment attempted to use language models trained on the transliterated data to increase the coverage of a language model based on Spanish data; this was modeled after Koehn and Schroeder (2007). The language models in this experiment were trigram models with Good-Turing smoothing built using SRILM (Stolcke, 2002).

As baselines, we trained Spanish (*es*) LMs on a small amount (30K sentences) and a large amount (1.5M sentences) of data. We also trained language models based on 30K transliterated and standard Italian (*it*) and Portuguese (*pt*) sentences. All were tested on the Spanish test set. Table 4 shows the perplexity for each of the baselines. As expected, more Spanish training data led to a lower perplexity. However, the transliterated Italian and Portuguese baselines yielded better perplexity with less data. Note also the strong effect of transliteration.

Language	Train Size	PP
es	30K	93.49
es	1.5M	55.84
it	30K	1683.31
it translit.	30K	96.21
it translit.	1.5M	207.60
pt	30K	1877.23
pt translit.	30K	151.06
pt translit.	1.5M	251.53

Table 4: Baseline results for experiment 1.

In the experiment, we interpolated LMs trained on different amounts of transliterated data with the LM trained on 30K Spanish sentences. We used

SRILM's compute-best-mix tool to determine the interpolation weights of the models. This parameter was trained on the Spanish development set.

Table 5 shows the results for the interpolation of the Spanish LM with Italian and Portuguese, both separately and simultaneously. The lambda values are the weights given to each of the language models. None of the interpolated combinations improves on the perplexity of the smallest Spanish baseline. The best results for interpolated language models are achieved when combining the 30K-sentence Spanish model with the 1.5M-sentence Portuguese model, which almost reaches the perplexity level of the Spanish-only model. As a comparison, we also interpolated two separate language models, each trained on 30K Spanish sentences; the weight for these models was close to 0.5.

In the best-performing language model mix that used all three languages, Portuguese was weighted with a lambda of about 0.17, whereas Italian was only weighted with 0.016. That shows that Portuguese, in this setup, is a better model of Spanish.

An open question has to do with the performance of the Portuguese language model in the experiment compared to the baselines. In table 4, we see that the language model does significantly worse when trained on more Portuguese data. However, the interpolation of the Spanish and Portuguese language models yields a lower perplexity when trained on a large amount of Portuguese data. Since the data was identical in the baselines and experiments, further exploration is needed to understand this behavior.

4.2 Experiment 2: Corpus Selection

For our second experiment, our goal was to select the most "Spanish-like" data from our Italian and Portuguese corpora. We concatenated this data with the Spanish sentences in order to increase the amount of training data for the language model. This is similar to what was done by Gao et al. (2002).

First, we trained a language model on our small Spanish corpus. This language model was then queried on a concatenation of the transliterated Italian and Portuguese data. The sentences in this corpus were ranked according to their perplexity in the Spanish LM. We selected the best 30K and 5K sentences, which were then concatenated with the Spanish data to form a larger corpus. Finally, we used KenLM (Heafield, 2011) to create a trigram language model with Kneser-Ney smoothing (Kneser and Ney, 1995) on that data. We also ran the same experiment on Italian and Portuguese separately.

Table 6 gives the results from these experiments. This table shows that the mixed-language models for each language performed better when they had a lower amount of non-Spanish data. This indicates that it is better to simply use a small amount of data in the low-resource language, rather than trying to augment it with the transliterated data from related languages. Using a smaller amount of the Spanish data, having a different strategy for selecting the non-Spanish data, using a different transliteration method, or using Italian and Portuguese data that was not a direct translation of the Spanish data may have all led to improvements. It is also interesting to note that the language models based on the corpus containing only Portuguese performed almost as well as those based on the corpus containing Portuguese and Italian. This indicates that the Portuguese data likely had more Spanish-like sentences than the Italian data. As mentioned in section 3.1, Portuguese is more similar to Spanish, so this makes intuitive sense. However, it is surprising given the results in table 4, which shows that the Italian-only language models performed better on Spanish data than the Portuguese-only language models.

4.3 Experiment 3: Statistical Machine Translation

Lastly, we experimented with translation models in order to see if our approach yielded similar results. For our baseline, we used a small parallel corpus of 30K English-Spanish (*en-es*) sentences from the Europarl corpus (Koehn, 2005). The data was preprocessed as described in section 3.2. Since SMT systems are often trained on large amounts of data, we expected poor coverage with this dataset. However, this size would be representative of the amount of data available for low-resource languages.

We used Moses (Koehn et al., 2007) to train our phrase-based SMT system on the above mentioned parallel corpus (*en-es*). We also trained a language model of 5M words of Spanish data from the same source, making sure that this data was strictly distinct from our parallel data. The language model was trained using KenLM (Heafield, 2011). The

Languages	Sentences	PP	Lambda es	Lambda it	Lambda pt
es + es	30K + 30K	86.59	0.502		
es + it	30K + 30K	95.19	0.9818	0.0182	
es + it	30K + 100K	96.08	0.9716	0.0284	
es + it	30K + 200K	96.49	0.9648	0.0352	
es + it	30K + 1.5M	96.91	0.9493	0.0507	
es + pt	30K + 30K	95.51	0.9340		0.0660
es + pt	30K + 100K	95.93	0.8939		0.1061
es + pt	30k + 200K	95.71	0.8709		0.1291
es + pt	30k + 1.5M	93.52	0.8170		0.1830
es + it + pt	30K + 30K + 30K	95.52	0.9298	0.0093	0.0608
es + it + pt	30K + 100K + 100K	95.94	0.8882	0.0126	0.0991
es + it + pt	30K + 200K + 200K	95.72	0.8655	0.0137	0.1207
es + it + pt	30K + 1.5M + 1.5M	93.53	0.8106	0.0161	0.1731

Table 5: Results of interpolated language models and optimal lambda values.

Languages	Sentences	PP
es	30K	84.57
es + it	30K + 5K	85.78
es + it	30K + 30K	94.10
es + pt	30K + 5K	85.11
es + pt	30K + 30K	90.31
es + it/pt	30K + 5K	85.13
es + it/pt	30K + 30K	90.24

Table 6: Results for the corpus selection experiment.

weights were set by optimizing BLEU using MERT on a separate development set of 2,000 sentences (English-Spanish). After decoding, we detokenized and evaluated the output. For the evaluation, we used a clean Spanish test set of 2,000 sentences from the same source. As an automatic evaluation measure, we used BLEU (Papineni et al., 2002) for quantitative evaluation.

For our experiments, we used Italian and Portuguese as auxiliary languages. We created two corpora of 30K sentences each from the Europarl corpus, *en-it* and *en-pt*. We first tokenized and transliterated the training corpus of the related language as described in section 3.3. Then, we concatenated the resulting corpora with our baseline corpus and trained our model. This is similar to what was done by Nakov and Ng (2012), although we attempt to translate into the low-resource language. We first experimented with each auxiliary language independently and then with both languages. In total we conducted the following experiments:

- English-Spanish (*en-es*) + English-Italian transliterated (*en-es*$_{it}$)
- English-Spanish (*en-es*) + English-Portuguese transliterated (*en-es*$_{pt}$)
- English-Spanish (*en-es*) + English-Italian transliterated (*en-es*$_{it}$) + English-Portuguese transliterated (*en-es*$_{pt}$)

In this experiment, we expected to observe some improvements compared to the language modeling experiments, as the mistakes in the transliterated output could be filtered out by the language model containing clean Spanish data. Moreover, we examined whether it is possible to have gains from using multiple related languages simultaneously.

Languages	Sentences	BLEU	p-value
en-es (Baseline)	30K	0.3360	
en-es + en-es$_{it}$	30K + 30K	0.3357	0.22
en-es + en-es$_{pt}$	30K + 30K	0.3349	0.08
en-es + en-es$_{it}$ + en-es$_{pt}$	30K + 30K + 30K	0.3384	0.041

Table 7: BLEU scores obtained for the different training sets and their sizes.

Table 7 shows the BLEU scores for the experiments. To determine whether our results were significant we used the bootstrap resampling method

(Koehn, 2004), which is part of Moses. There were no significant improvements in BLEU score when only one auxiliary language was used. Nonetheless, we observed a significant improvement when data from both Italian and Portuguese is used. This may be an indication that more out-of domain data, when used in the translation model and sufficiently transformed, can actually improve performance.

One open question at this point is whether the improvement was caused by the contribution of more than one language or simply by the increase in training data. It is possible that a similar improvent could be achieved by increasing the data of one language to 60K. However, in order to support our conjecture, it will be necessary to conduct experiments with different sizes and combinations of data from the related languages.

5 Discussion

We observed that a closely-related language cannot be used to aid in modeling a low-resource language without being properly transformed. Although our naive rule-based transliteration method strongly improved over the non-transliterated closely-related language data, it performed worse than even a small amount of target language data. In addition, adding more data from the related language caused the models to do worse; this may be because there were more words in the data that were not translated using the 200-word dictionary, so there was more noise from the rule-based transliterations in the data. Thus, we were not successful in using data from a related language to improve language modeling for a low-resource language.

For statistical machine translation, our results show gains from augmenting the translation models of a low-resource language with transliterated related-language data. We expect that by taking advantage of more sophisticated transliteration and interpolation methods as well as larger amounts of data from the closely-related language(s), larger improvements in BLEU can be achieved.

6 Future Work

We plan on experimenting with more sophisticated ways of transforming related language data, including unsupervised and semi-supervised transliteration methods. We would particularly like to experiment with neural network machine transliteration using a character-based LSTM network. This could be developed based on small parallel texts or lists of bilingual cognates of varying sizes. We could also use existing transliteration modules integrated in the SMT system (Durrani et al., 2014). In addition, we hope to explore using bilingual dictionaries without transliteration, as well as using phonological transcription as an intermediary between the two related languages. Finally, it would be beneficial to examine the contribution of each of the rules in our rule-based system separately.

A relatively simple modification to our experiments would be to use more data in creating the translation model (in experiment 3). While we found that using more of the high-resource language data in the language models yielded higher perplexity, the same did not carry over to BLEU scores, especially since we saw a slight improvement in BLEU score when using both Portuguese and Italian data. A similar option would be to select the best Italian and Portuguese data (as was done in experiment 2) for use in the translation model, instead of selecting random sentences.

In statistical machine translation, it would be interesting to explore methods of using data from related languages while preserving the reliable information from the low-resource language. One idea could be methods for interpolating phrase tables for the transliterated corpora as well as setting optimal weights for each of them, similar to the approach of Sennrich (2012). We would also like to improve the translation model coverage by filling up the phrase table for a low-resource language with data from a related language while keeping the useful data from the low-resource language (Bisazza et al., 2011) or by using the related languages as a back-off (Yang and Kirchhoff, 2006).

Finally, a weakness of our language modeling experiments was that we used almost parallel data between the related and the target languages. Hence, the related language was not likely to increase the vocabulary coverage of the models; instead, it just added misspellings of the target language words. In the future, we would like to run experiments with data from the related languages that is strictly distinct from the data of the low-resource language.

References

Arianna Bisazza, Nick Ruiz, and Marcello Federico. 2011. Fill-up versus interpolation methods for phrase-based SMT adaptation. In *IWSLT*, pages 136–143.

Stanley F Chen and Joshua Goodman. 1996. An empirical study of smoothing techniques for language modeling. In *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics*, pages 310–318. Association for Computational Linguistics.

Alina Maria Ciobanu and Liviu P Dinu. 2014. On the Romance languages mutual intelligibility. In *Proceedings of the 9th International Conference on Language Resources and Evaluation, LREC*, pages 3313–3318.

Daniel Dahlmeier and Hwee Tou Ng. 2010. Domain adaptation for semantic role labeling in the biomedical domain. *Bioinformatics*, 26(8):1098–1104.

Nadir Durrani, Hieu Hoang, Philipp Koehn, and Hassan Sajjad. 2014. Integrating an unsupervised transliteration model into statistical machine translation. *EACL 2014*, page 148.

Jianfeng Gao, Joshua Goodman, Mingjing Li, and Kai-Fu Lee. 2002. Toward a unified approach to statistical language modeling for Chinese. *ACM Transactions on Asian Language Information Processing (TALIP)*, 1(1):3–33.

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

Katrin Kirchhoff and Mei Yang. 2005. Improved language modeling for statistical machine translation. In *Proceedings of the ACL Workshop on Building and Using Parallel Texts*, pages 125–128. Association for Computational Linguistics.

Amit Kirschenbaum and Shuly Wintner. 2010. A general method for creating a bilingual transliteration dictionary. In *LREC*.

Horst G Klein. 2002. Eurocom-Rezeptive Mehrsprachigkeit und Neue Medien.

Reinhard Kneser and Hermann Ney. 1995. Improved backing-off for m-gram language modeling. In *Acoustics, Speech, and Signal Processing, 1995. ICASSP-95., 1995 International Conference on*, volume 1, pages 181–184. IEEE.

Philipp Koehn and Josh Schroeder. 2007. Experiments in domain adaptation for statistical machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 224–227. Association for Computational Linguistics.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions*, pages 177–180. Association for Computational Linguistics.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In Dekang Lin and Dekai Wu, editors, *Proceedings of EMNLP 2004*, pages 388–395, Barcelona, Spain, July. Association for Computational Linguistics.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT Summit*, volume 5, pages 79–86. Citeseer.

M. Paul Lewis, editor. 2015. *Ethnologue: Languages of the World*. SIL International, Dallas, TX, USA, eighteenth edition.

Xiaoyong Liu and W Bruce Croft. 2005. Statistical language modeling for information retrieval. *Annual Review of Information Science and Technology*, 39(1):1–31.

Preslav Nakov and Hwee Tou Ng. 2012. Improving statistical machine translation for a resource-poor language using related resource-rich languages. *Journal of Artificial Intelligence Research*, pages 179–222.

Preslav Nakov and Jörg Tiedemann. 2012. Combining word-level and character-level models for machine translation between closely-related languages. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2*, pages 301–305. 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. Association for Computational Linguistics.

Rico Sennrich. 2012. Perplexity minimization for translation model domain adaptation in statistical machine translation. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 539–549. Association for Computational Linguistics.

Karan Singla, Nishkarsh Shastri, Megha Jhunjhunwala, Anupam Singh, Srinivas Bangalore, and Dipti Misra Sharma. 2014. Exploring system combination approaches for Indo-Aryan MT systems. *LT4CloseLang 2014*, page 85.

Andreas Stolcke. 2002. SRILM - an extensible language modeling toolkit. In *Proceedings of the 7th International Conference on Spoken Language Processing (ICSLP 2002)*.

Stefanie Voigt and Charlotte Gooskens. 2014. Mutual intelligibility of closely related languages within the

Romance language family. *Language Contact: The State of the Art*, page 103.

Pidong Wang, Preslav Nakov, and Hwee Tou Ng. 2012. Source language adaptation for resource-poor machine translation. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 286–296. Association for Computational Linguistics.

Mei Yang and Katrin Kirchhoff. 2006. Phrase-based backoff models for machine translation of highly inflected languages. In *EACL*, pages 3–7.