Improving Low Resource Machine Translation using Morphological Glosses

Steven Shearing ssheari1@jhu.edu
Christo Kirov ckirov1@jhu.edu
Huda Khayrallah huda@jhu.edu
David Yarowsky yarowsky@jhu.edu
Johns Hopkins University

Abstract

Low-resource machine translation is a challenging problem, especially when the source language is morphologically complex. We describe a simple procedure for constructing glosses, or mappings between complex, inflected source-language words and equivalent multi-word English expressions. We demonstrate the utility of glosses, especially compared to entries in bilingual dictionaries, across several data-augmentation strategies designed to mitigate a lack of training data. In our experiments, we achieve improvements of up to 1 BLEU point in a Russian-English translation task and 2.4 BLEU points in a Spanish-English translation task over a strong baseline translation system.

1 Introduction

Low-resource machine translation, where only a small amount of parallel data is available between source and target languages, poses a significant challenge. Machine translation systems, especially those based on neural network models, tend to be data-hungry. Highly-inflected source languages further complicate the situation, presenting a significant sparsity problem in low-resource settings. Most possible inflected wordforms are likely to appear only once in the data or not at all.

In an effort to improve performance when limited parallel data is available for learning how to translate from a highly inflected source language into English, we experiment with two simple data augmentation strategies—appending and substitution. To alleviate sparsity, we experiment with appending entries from multilingual dictionaries directly to the bitext. We also leverage linguistic knowledge about the morphological grammar of the highly-inflected source language to generate multi-word English glosses. We show that these glosses, which better mimic in-situ translations, are more effective than dictionary entries when appended to the training data. We also employ glosses for a second strategy, directly substituting them in place of complex inflected forms in the source language. The overarching idea is to create a new version of the source, source$'$, that is more similar to the target language. In theory, this should improve performance by solving some portion of the translation problem before a final translation model is trained. We show that gloss substitution has a positive effect on BLEU scores compared to baseline systems.

We present experimental results translating from Russian and Spanish into English. While Russian and Spanish are not low-resource languages, we simulate extremely low-resource scenarios by relying only on representative language packs from DARPA's LORELEI (LOw REsource Languages for Emergent Incidents) program as our source of parallel training data.

These packs typically contain less than 50,000 bilingual sentence pairs in total, orders of magnitude below the amount used to train most state-of-the-art MT systems. We run our experiments using traditional phrase-based statistical machine translation models (PBMT). While neural machine translation offers state-of-the-art performance when training data is plentiful, PBMT remains competitive or superior in the low resource conditions we focus on (Koehn and Knowles, 2017).

2 Multilingual Dictionaries Versus Glosses

We define an entry in a multilingual dictionary as a mapping between a lemma form in the source language to one or more definitions in the target language.

бежать,VERB,to run, to be running

While useful, these types of entries have several notable drawbacks when used as bitext for a translation system. First, on the source side, the dictionary forms of words, or lemmas, are typically uninflected, and may not be in common usage. For example, the dictionary form of verbs in many languages is the infinitive, but in actual text tensed forms are much more common. Second, on the target side, dictionary definitions are not necessarily equivalent to in-situ translations of a word, and often contain additional descriptive text.

Glosses, as we define them, are intended to remedy these problems. A gloss is a mapping between an inflected form of a word, and an in-situ translation. In many cases, English uses syntactic constructions to express distinctions made by inflectional morphology in a source language. As a result, single source words are often glossed as multi-word expressions in English.

бегут, бежать,V;IPFV;PRS;3;PL,(they/NNS) are running; (they/NNS) run

Generating a gloss for an inflected word follows a general process outlined in Hewitt et al. (2016). In this work, however, we simplify many of the steps. Our implementation is fully described in the Experiments section below.

1. Apply morphological analysis to an input inflected word to recover its base lemma and morphological features, e.g.,

   comprábamos → *comprar*, V;1;PL;PST;IPFV

2. Using a separate lemma-to-lemma dictionary, recover a target lemma for the source word:

   comprar → *buy*

3. Specify a conversion from each vector of source morphological features to a target gloss template. For many language pairs, this can be done manually:

   V;1;PL;PST;IPFV → '(we) were VBG.'

   Here, VBG is a Penn Treebank tag[1] which indicates that the template can be filled with the gerund (*-ing*) form of an English verb.

4. Given a gloss template from (3), and a target lemma from (2), replace the PTB placeholder in the template, inflecting the target lemma as needed with a morphological generation tool or lookup table:

   '(we) were VBG' + *buy* → '(we) were *buying*'

   This completes the gloss generation process:

   comprábamos → '(we) were buying'

[1] https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html

3 Related Work

Prior work has both explored ways of generating morphological information, and incorporating such morphological information into Phrase Based Machine Translation. While there is work on generating rich morphology on the target side of translations (for example: (Toutanova et al., 2008; Huck et al., 2017), we focus on rich source side morphology in this work.

Hewitt et al. (2016) created glosses by re-purposing instructional prompts found in a special corpus designed to elicit inflectional paradigms from bilingual speakers (Sylak-Glassman et al., 2016). For example, the sample prompt '(The apple) has been eaten.' was designed to elicit third person present perfect verb forms from bilingual Spanish speakers. They heuristically interpolated between multiple prompts to generate new gloss templates for each possible feature vector and lemma. They experiment with both appending their synthetic translations to the parallel text as well as using an additional phrase table (and the combination of both), but did not find that one method was consistently superior.

Broadly, the goal of our substitution approach is to transform the source language into a form that is more similar to the target. A number of previous strategies used in MT have fallen under this umbrella. Compound splitting (Koehn and Knight, 2003; Macherey et al., 2011) of, for example, German source-side words increases similarity with English as English doesn't use nearly as many compounds as German. Fraser and Marcu (2005) use stemming to reduce Romanian source side vocabulary size to improve Romanian-English word alignment. Ding et al. (2016) compare supervised (ChipMunk (Cotterell et al., 2015)) and unsupervised (Morfessor (Virpioja et al., 2013), and Byte-Pair encoding (Sennrich et al., 2015)) morphological segmentation methods on the source side of the PBMT system for the WMT Turkish-English Translation Task.

4 Experiments

4.1 Model

We use Moses (Koehn et al., 2007) as the Phrase Based Machine Translation (PBMT) system to run all our translation experiments. Data is tokenized and truecased using standard Moses scripts. We use GIZA++ (Och and Ney, 2003) for alignment with the grow-diag-final-and setting. We set the maximum sentence length to 80 and the maximum phrase length to 5. For decoding, we use Cube Pruning (Huang and Chiang, 2007). We also weigh potential translations using a 5-gram KenLM (Heafield, 2011) language model.

4.2 Data

Bitext. For our base bitext, we use the Russian-English Corpus from the LORELEI Russian Representative Language Pack (LDC2016E95 V1.1), and the Spanish-English Corpus from the LORELEI Spanish Representative Language Pack (LDC2016E97). The corpora primarily consist of news and web forums. While they are included in the LORELEI corpora, we exclude Tweets from these experiments. We also remove sentences longer than 80 words.

For the Russian baselines, we randomly split the remaining data into train (46,746), tune (2,233), and test (2,462) sentence pairs. For the Spanish baselines, we randomly split the remaining data into train (22,311), tune (2,031), and test (2,032) sentence pairs. See table 1 for the number of sentences in the training corpus for each experiment condition.[2] The tuning and test sets remain the same for all experiments in a language.

[2] Since we perform the length filtering on the training set after gloss substitution in the corresponding condition, some sentences are removed as substitution makes them too long.

Dictionary. The LORELEI language packs from Russian and Spanish also contain dictionaries mapping source lemmas into target definitions in a custom XML format. Some entries include multiple definitions. In this case, each definition was split into its own line of bitext. Furthermore, some of the definitions include notes on gender (e.g. артистка: artist (female).) or topic (e.g. бить: chime (about clock)), or other comments (e.g. бензель: (rare) paintbrush). We remove any text within parenthesis, and then remove any entries with non-English words. After all post-processing was complete, we were left with 58,856 dictionary entries for Russian, and 64,450 dictionary entries for Spanish.

Glosses. Glosses for Russian and Spanish were created as follows: Lists of inflected wordforms were obtained via a union of the UniMorph database (`unimorph.githbub.io`), which provides a mapping from inflected forms to their lemmas and morphological feature vectors, and a tokenization of the monolingual corpus released by the LDC as part of LORELEI language packs for Russian and Spanish.

For each word in the list, additional morphological analyses were obtained. For Russian, we applied the PyMorphy2 package (Korobov, 2015)[3] to each word, while for Spanish we used the Freeling package (Padró and Stanilovsky, 2012).[4] For both Russian and Spanish, we also applied a custom sequence-2-sequence neural network analyzer trained on the raw data in the UniMorph database. The network used an architecture, training scheme, and hyperparameters identical to that used in (Kann and Schütze, 2016). It mapped sequences of characters representing an inflected word directly to a sequence representing the its underlying lemma and features (*c o m p r á b a m o s → c o m p r a r V 1 PL PST IPFV*) The feature vectors output by all analysis methods were manually mapped into the UniMorph feature schema standard (Sylak-Glassman et al., 2015). As the total set of of unique feature vectors remaining after this mapping was limited for both Russian (569 vectors) and Spanish (239 vectors), we were able to manually produce one or more gloss templates for each vector (e.g., V;1;PL;PST;IPFV → '(we) were VBG.').

Source lemmas in Russian and Spanish were converted to English lemmas via lookup in, preferably, Wiktionary-derived lemma translation data (Kirov et al., 2016), or PanLex (Baldwin et al., 2010). English lemmas were then inflected using the tools provided by Smedt and Daelemans (2012) and inserted into the corresponding gloss templates.

We further post-processed the glosses by removing anything in parenthesis or brackets, and then removed entries containing non-English words in the translation, after which we were left with 3,122,470 Russian glosses, and 589,188 Spanish glosses.

5 Conditions

We trained three baseline models and five additional experimental setups. Total sizes of training datasets for each condition, in number of paired sentences, are shown in Table 1.

For Baseline 1, in both Russian and Spanish, we simply trained a default Moses system on the base bitext in each LORELEI language pack. The language model and truecaser used during decoding were trained only on the target-side portion of the parallel training data.

For Baseline 2, we made use of the extensive monolingual data available for English. Following previous work, we trained a new truecaser and a much larger language model from the English side of the Russian-English parallel text plus text from the Associated Press Worldstream, English Service, a subset of the English Gigaword corpus (Parker et al., 2011) (a total of 54,287,116 sentences). Given the clear performance benefit, we continued to use this larger language model for all subsequent experiments. While the large language model was trained

[3] `https://github.com/kmike/pymorphy2`

[4] `http://nlp.lsi.upc.edu/freeling`

Condition	Russian	Spanish
Baseline 1 (Small LM)	46,460	22,311
Baseline 2 (Big LM)	46,460	22,311
Append Dictionary	105,316	86,861
Append Glosses	3,168,675	611,439
Append Dictionary + Glosses	3,227,531	675,989
Substitute Glosses	46,414	22,226
Substitute Glosses + Identity Alignment	95,603	40,724

Table 1: Number of train sentences for Russian-English and Spanish-English Translation.

source	Женщина была почти при смерти
reference	the woman had nearly died
substitution	woman were почти при to death

Table 2: An example of gloss substitution in the Russian-English training set.

using the target-side of the Russian-English data, this contributed a minuscule amount in proportion to the contribution of data from the Gigaword corpus. Thus, we use the same language model for both Russian and Spanish experiments.

For Baseline 3, in both Russian and Spanish, we train PBMT system on the glosses and dictionary (without any parallel sentences). We use the larger language model from Baseline 2.

Appending. Our next experimental conditions involved appending additional data to the base bitext for each language. We experimented with appending the processed dictionary entries or generated glosses, as well as appending both the dictionary and the glosses in one system. Each of these modifications increased the total size of the training data, as seen in Table 1.

Substitution. Finally, we substituted our glosses directly into the base bitext. Any inflected source word appearing in the list of glosses was a candidate for substitution. Many words had multiple glosses available. To decide which one to use for substitution, we considered the following confidence hierarchy. First, any gloss corresponding to a pre-existing entry in the UniMorph database was preferred. Next, we preferred entries corresponding to an off-the-shelf morphological analysis (derived from PyMorphy2 in Russian, and Freeling in Spanish). Glosses based on the custom-trained neural-network analyzer were used when a UniMorph entry was not available and both PyMorphy2 and Freeling failed to provide an analysis. An example of the substitution process is shown in table 2.

While substitution is intended to make the source language appear more like the target (in this case literally, since target language words are substituted directly into the source), the alignment algorithm in the PBMT system is not character-aware and therefore has no sense of identity between the source and target vocabularies. To get around this, we add a condition attempting to bias the aligner to notice identical source and target phrases. In particular, for each gloss substitution, we append a gloss-to-gloss identity mapping to the bitext.

6 Results & Discussion

Table 3 indicates the lowercased BLEU scores achieved by the model in each experimental condition in Russian-English and Spanish-English settings. Results were consistent across both language pairs. As expected, using a larger target-side language model (Baseline 2) provides a significant boost over the initial baseline (Baseline 1) with language model trained only on

Condition	Russian BLEU	%OOV(Type/Token)	Spanish BLEU	%OOV (Type/Token)
Baseline 1 (bitext + Small LM)	15.6	31.1/11.2	18.6	24.6/8.0
Baseline 2 (bitext + Big LM)	16.2	31.1/11.2	20.6	24.6/8.0
Baseline 3 (dictionary + Glosses + Big LM)	7.7	26.5/26.7	14.6	12.1/13.1
Append Dictionary	17.0	28.7/10.2	21.5	20.2/6.6
Append Glosses	17.8	15.2/5.5	23.8	6.8/2.3
Append Dictionary + Glosses	18.0	15.1/5.5	23.9	6.6/2.3
Substitute Glosses	17.7	26.3/5.0	22.2	19.5/4.0
Substitute Glosses + Identity Alignment	17.8	26.3/5.0	22.9	19.5/4.0

Table 3: Lowercased BLEU for Russian-English and Spanish-English Translation.

the available bitext. This was true even for Spanish, where the large language model was partially trained on the Russian-English target-side data, and was potentially out-of-domain. Every augmentation strategy provided some further improvement. Baseline 3 demonstrates that simply using the lexical resources and a strong language model can produce decent results in the absence of bitext (particularly in the Spanish experiments).

As we hypothesized, appending the glosses to the training data results in better performance than just appending dictionary pairs. This is likely because glosses are closer to actual translations. However, there is an additive effect of appending both dictionary items and glosses, suggesting that the two external data sources contain at least some complementary information.

The substitution trials did not fare as well as the appending trials. They did, however, still provide an improvement over Baseline 2. This was even without adding identity pairs to the training data in order to bias alignment (so the model was not aware which parts of the source sentences were actually English), or increasing the amount of bitext in any way. This suggests that using simple techniques like gloss substitution to transform the source into something closer to the target language makes learning a complex MT model after the fact more effective. In the Russian gloss substitution, 67% of tokens were replaced in the training set. In Spanish, 80% of tokens were replaced.

Table 3 includes the out-of-vocabulary rate (type and token) for each experiment. The low out-of-vocabulary rate for baseline 3 demonstrates the coverage of the the dictionary and glosses (particularity for Spanish). The glosses in particular have very broad coverage. They provide a dramatic drop in OOV's (dropping the rate by over 50%). In addition to the BLEU improvement, reducing the OOV rate can greatly improve the usability of low resource machine translation.

7 Conclusions & Future Work

We showed that glosses of morphologically complex source words are a useful resource for rapidly improving machine translation performance in extremely low-resource scenarios. As glosses mimic in-situ translations of inflected words, they are more informative than dictionary items, which map lemmas to definitions. Glosses are useful both for augmenting training data with additional bitext, or transforming source language data into a form that is more similar to the target language. Future work will explore different ways of generating glosses, and apply additional transformations to the language data to ease the amount a translation model needs to learn. This would include changing both the source language, *and* making reversible changes to the target. If a non-English, morphologically complex target is used, these might include target-side morphological segmentation.

Acknowledgments

This material is based upon work supported in part by DARPA under Contract No. HR0011-15-C-0113. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of DARPA.

References

Baldwin, T., Pool, J., and Colowick, S. M. (2010). PanLex and LEXTRACT: Translating all words of all languages of the world. In *Proceedings of the 23rd International Conference on Computational Linguistics: Demonstrations*, COLING '10, pages 37–40, Stroudsburg, PA, USA. Association for Computational Linguistics.

Cotterell, R., Müller, T., Fraser, A., and Schütze, H. (2015). Labeled morphological segmentation with Semi-Markov models. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning*, pages 164–174, Beijing, China. Association for Computational Linguistics.

Ding, S., Duh, K., Khayrallah, H., Koehn, P., and Post, M. (2016). The JHU machine translation systems for WMT 2016. In *Proceedings of the First Conference on Machine Translation*, pages 272–280, Berlin, Germany. Association for Computational Linguistics.

Fraser, A. and Marcu, D. (2005). ISI's participation in the Romanian-English alignment task. In *Proceedings of the ACL Workshop on Building and Using Parallel Texts*, ParaText '05, pages 91–94, Stroudsburg, PA, USA. Association for Computational Linguistics.

Heafield, K. (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.

Hewitt, J., Post, M., and Yarowsky, D. (2016). Automatic construction of morphologically motivated translation models for highly inflected, low-resource languages. *AMTA 2016, Vol.*, page 177.

Huang, L. and Chiang, D. (2007). Forest rescoring: Faster decoding with integrated language models. In *Annual Meeting-Association For Computational Linguistics*, volume 45, page 144.

Huck, M., Tamchyna, A., Bojar, O., and Fraser, A. (2017). Producing unseen morphological variants in statistical machine translation. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 369–375, Valencia, Spain. Association for Computational Linguistics.

Kann, K. and Schütze, H. (2016). Single-model encoder-decoder with explicit morphological representation for reinflection. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 555–560, Berlin, Germany. Association for Computational Linguistics.

Kirov, C., Sylak-Glassman, J., Que, R., and Yarowsky, D. (2016). Very-large scale parsing and normalization of Wiktionary morphological paradigms. In Calzolari, N., Choukri, K., Declerck, T., Goggi, S., Grobelnik, M., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France. European Language Resources Association (ELRA).

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. (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.

Koehn, P. and Knight, K. (2003). Empirical methods for compound splitting. In *Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics-Volume 1*, pages 187–193. Association for Computational Linguistics.

Koehn, P. and Knowles, R. (2017). Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39, Vancouver. Association for Computational Linguistics.

Korobov, M. (2015). Morphological analyzer and generator for Russian and Ukrainian languages. In Khachay, M. Y., Konstantinova, N., Panchenko, A., Ignatov, D. I., and Labunets, V. G., editors, *Analysis of Images, Social Networks and Texts*, volume 542 of *Communications in Computer and Information Science*, pages 320–332. Springer International Publishing.

Macherey, K., Dai, A. M., Talbot, D., Popat, A. C., and Och, F. (2011). Language-independent compound splitting with morphological operations. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 1395–1404, Stroudsburg, PA, USA. Association for Computational Linguistics.

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

Padró, L. and Stanilovsky, E. (2012). FreeLing 3.0: Towards wider multilinguality. In *Proceedings of the Language Resources and Evaluation Conference (LREC 2012)*, Istanbul, Turkey. ELRA.

Parker, R., Graff, D., Kong, J., Chen, K., and Maeda, K. (2011). English Gigaword Fifth Edition LDC2011T07. Philadelphia. Linguistic Data Consortium.

Sennrich, R., Haddow, B., and Birch, A. (2015). Neural machine translation of rare words with subword units. *CoRR*, abs/1508.07909.

Smedt, T. D. and Daelemans, W. (2012). Pattern for Python. *Journal of Machine Learning Research*, 13(Jun):2063–2067.

Sylak-Glassman, J., Kirov, C., Post, M., Que, R., and Yarowsky, D. (2015). *A Universal Feature Schema for Rich Morphological Annotation and Fine-Grained Cross-Lingual Part-of-Speech Tagging*, pages 72–93. Springer International Publishing, Cham.

Sylak-Glassman, J., Kirov, C., and Yarowsky, D. (2016). Remote elicitation of inflectional paradigms to seed morphological analysis in low-resource languages. In Chair), N. C. C., Choukri, K., Declerck, T., Goggi, S., Grobelnik, M., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France. European Language Resources Association (ELRA).

Toutanova, K., Suzuki, H., and Ruopp, A. (2008). Applying morphology generation models to machine translation. In *Proceedings of ACL-08: HLT*, pages 514–522, Columbus, Ohio. Association for Computational Linguistics.

Virpioja, S., Smit, P., Grönroos, S.-A., Kurimo, M., et al. (2013). Morfessor 2.0: Python implementation and extensions for Morfessor Baseline.