Cognate Projection for Low-Resource Inflection Generation

Bradley Hauer, Amir A. Habibi, Yixing Luan, Rashed Rubby Riyadh, Grzegorz Kondrak
Department of Computing Science
University of Alberta, Edmonton, Canada
{bmhauer,amirahmad,yixing1,riyadh,gkondrak}@ualberta.ca

Abstract

We propose cognate projection as a method of crosslingual transfer for inflection generation in the context of the SIGMORPHON 2019 Shared Task. The results on four language pairs show the method is effective when no low-resource training data is available.

1 Introduction

In this description of the University of Alberta systems, we discuss our approach to Crosslingual Transfer for Inflection Generation (Task 1) in the SIGMORPHON 2019 Shared Task on Crosslinguality and Context in Morphology (McCarthy et al., 2019). The task of inflection generation is to produce an inflected word-form given a lemma and a sequence of abstract morphological tags. For example, the Latin citation form *fucō* with the tag `V;IND;FUT;3;SG` should yield the form *fucābit*.[1] The goal is to examine how best to do this in a cross-lingual setting.

We focus on depth over breadth, performing experiments on only four language pairs which represent a range of diachronic relationships. Kashubian is so closely related to Polish that it is sometimes viewed as a dialect. Occitan and Spanish are less closely related, but share many morphological features. Romanian evolved from Latin over the course of 1500 years. Hindi and Bengali are also related, but written in distinct scripts.

In order to alleviate the training data sparsity in the low-resource setting, we attempt to leverage external text corpora, from which we extract target language word lists for both inflection generation and cognate projection. The results show that this strategy improves the overall results for some of the tested language pairs.

[1]For an unknown reason, only the inflected Latin forms in the data include vowel length diacritics.

As our principal contribution, we propose and test the idea of performing cognate projection to leverage high-resource training data for low-resource inflection generation. The results demonstrate that an implementation of this concept can perform better than the baselines in the scenario when no low-resource inflection data is available.

2 Prior Work

Our methods build upon the prior work of the University of Alberta teams for three previous SIGMORPHON shared tasks on type-level morphological generation (Cotterell et al., 2016, 2017, 2018). We view inflection as a string transduction task. Our discriminative transduction models stem from the DIRECTL+ transducer of Jiampojamarn et al. (2008), which was originally designed for grapheme-to-phoneme conversion.

Nicolai et al. (2016) apply discriminative string transduction to morphological reinflection. They show that the approach of Nicolai et al. (2015) performs well on typologically diverse languages. They also discuss language-specific heuristics and errors.

Nicolai et al. (2017) combine a discriminative transduction system with neural models. The results on five languages show that the approach works well in the low-resource setting. Additionally, they propose adaptations designed to handle small training sets, such as tag re-ordering and particle processing.

Najafi et al. (2018a) make further progress on the combination of neural and non-neural models for low-resource reinflection. Their best system obtains the highest accuracy on 34 out of 103 languages. They achieve additional improvements in accuracy by leveraging unannotated text corpora using the non-standard approaches of Nicolai et al. (2018) and Najafi et al. (2019).

Proceedings of the 16th Workshop on Computational Research in Phonetics, Phonology, and Morphology, pages 6–11
Florence, Italy. August 2, 2019

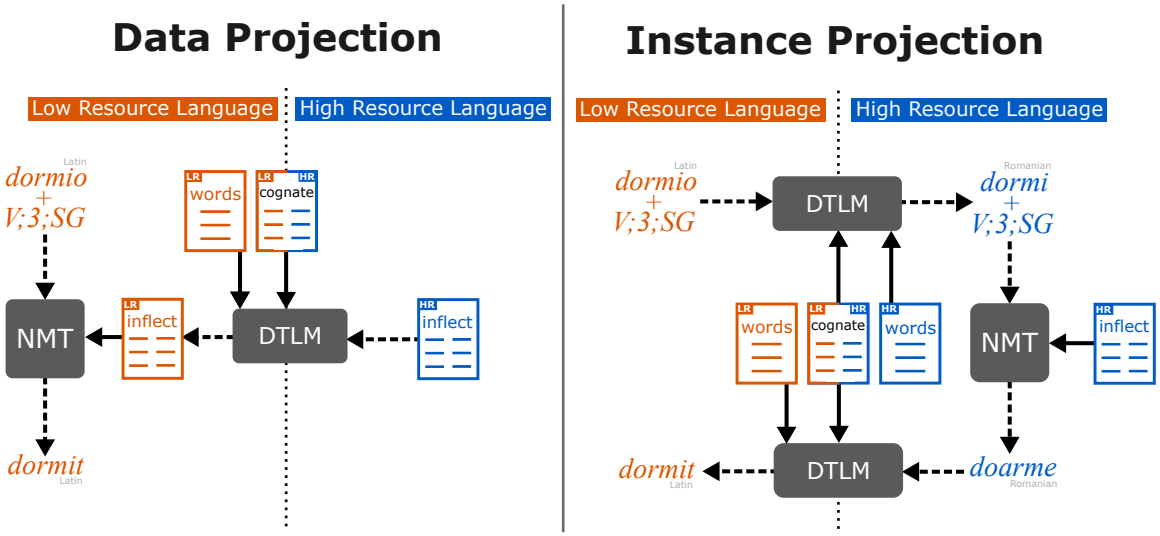

Figure 1: Two approaches to applying cognate projection to inflection generation. DTLM and NMT denote projection and inflection models, respectively. Dashed arrows show transduction. Solid arrows indicate training data. The LR and HR components are shown in orange and blue.

3 Tools

In this section, we describe our two principal tools: DTLM for cognate projection and low-resource inflection generation, and OpenNMT for high-resource inflection generation.

3.1 DTLM

DTLM (Nicolai et al., 2018) combines discriminative transduction with character and word language models derived from large unannotated corpora, with the language-model features integrated into the transducer. DTLM employs a many-to-many alignment method, which is referred to as precision alignment.

Nicolai et al. (2018) demonstrate that DTLM achieves superior results in low-data scenarios on several transduction tasks, including inflection generation, transliteration, phoneme-to-grapheme conversion, and cognate projection. In the CoNLL–SIGMORPHON 2018 Shared Task on Universal Morphological Reinflection (Cotterell et al., 2018), DTLM was our best performing individual system. It was also successfully used in the NEWS 2018 shared task on transliteration (Najafi et al., 2018b).

3.2 OpenNMT

OpenNMT (Klein et al., 2017) is an open-source neural machine translation tool based on sequence to sequence model with attention mechanism. Klein et al. (2017) demonstrates that OpenNMT generally performs better quality of machine translation than other existing open-source machine translation systems and is fairly efficient in terms of training and test speed.

Machine translation models have been successfully applied to other transduction tasks (Kann and Schütze, 2016). We employ OpenNMT as a vanilla HR morphological inflection tool, by simply concatenating the lemma and the tags to form the input sequence. Each individual tag is encoded as a single input token. No target wordlists are used.

4 Cognate Projection Methods

Each dataset in this shared task pairs a low-resource (LR) language with a related high-resource (HR) language. Genetically related languages share *cognates*, words with a common linguistic origin (St Arnaud et al., 2017). For example, the Latin word *oculus* 'eye' is cognate with the Romanian word *ochi*. Cognate pairs exhibit phonetic and semantic similarity (Kondrak, 2013). The correspondences between substrings in cognates tend to follow regular patterns (Kondrak, 2009).

Cognate projection, also referred to as cognate production (Beinborn et al., 2013; Ciobanu, 2016), is the task of predicting the spelling of a hypothetical cognate in another language. For example, the projection of *oculus* from Latin to Romanian should generate *ochi*. Even if a cognate

Language	Source	UniMorph	Words
Kashubian	Wikipedia	509	60286
Occitan	Wikipedia	8316	318706
Latin	UniMorph	509182	357951
Bengali	UniMorph	4443	2752

Table 1: The size of the UniMorph datasets and our target word lists.

Pair	k	t	Train	Dev	Test
pol↔csb	7500	0.4	6500	500	500
spa↔oci	5300	0.4	4500	500	300
ron↔lat	4612	0.4	4000	300	312
hin↔ben	1816	0.5	1456	180	180

Table 2: Our cognate projection datasets.

word does not exist, cognate projection should produce a target form that incorporates the interlingual sound correspondences and the phonotactic constraints of the target language. We hypothesize that the projected forms exhibit some of the morpho-phonetic properties of the actual words. For example, the projection of the Spanish verbal form *tomaré* ('I will take') into a (non-existent) Latin word *tomābō* could provide useful information for inflecting actual Latin verbs.

We propose two projection-based approaches for inflection generation which are based on the above hypothesis (Figure 1). We refer to those approaches as *Data Projection* and *Instance Projection* respectively. Both approaches aim at taking advantage of the HR inflection training data to perform LR inflection. Morphological tags are left unchanged. For cognate projection, we train transduction models (Section 3.1) on lists of cognate pairs extracted from small bitexts. The projection models are strengthened by target wordlists extracted from freely-available monolingual corpora.

The *Data Projection* approach simply projects the entire HR training data, which consists of lemmas and the corresponding inflected forms, into the LR language. For example, the Romanian training pair *"dormi+V;3;SG = doarme"* projects into Latin *"dormio+V;3;SG = dormit"*. This produces a relatively large, synthetic LR training set from which an LR inflection model can be derived (Section 3.2). The underlying idea is that the HR inflection patterns may be reflected in the corresponding LR inflection patterns, especially if the languages are closely related.

The *Instance Projection* approach is more complex, consisting of three transduction steps: (1) project an individual LR test instance into the HR language; (2) inflect the resulting form using a model trained on the HR training data, and (3) project the result back into the LR language. For example, Latin *"dormio+V;3;SG"* would first be projected into Romanian *"dormi+V;3;SG"*, then inflected using the Romanian model into *doarme*, and finally projected back into Latin as *dormit*. Unlike in Data Projection, inflection is performed entirely in the HR language. We aim to determine whether the higher HR inflection accuracy can offset the errors introduced at either of the projection steps.

5 Development

In this section, we describe our external resources and development results.

5.1 External Resources

For low-resource tasks, in both inflection generation and cognate projection, it makes obvious sense to leverage additional resources, which are freely available for many under-resourced languages. We extract the target word lists for DTLM from UniMorph[2] (Kirov et al., 2018). and Wikipedia[3], as summarized in Table 1.[4]

For cognate projection, we need training sets composed of cognate pairs, Finding good parallel bitexts for low-resource languages is quite challenging. Small bitexts exist in special domains, such as technical documentation or Bible translations. For Polish-Kashubian and Spanish-Occitan, we use software documentation from OPUS[5] (Tiedemann, 2012). For Hindi-Bengali, we use the OpenSubtitles (v2018) data, also from OPUS. For Romanian-Latin, we use a parallel corpus which contains a verse-by-verse alignment of the Bible translations in 100 languages (Christodouloupoulos and Steedman, 2015).

[2] https://unimorph.github.io

[3] https://dumps.wikimedia.org

[4] We are aware that the test data for the shared task may come from UniMorph. We use UniMorph solely for deriving the target word language model, without taking advantage of the morphological annotations. All our submissions that use external data are declared as non-standard.

[5] http://opus.nlpl.eu/

Data	System	ID	Word Accuracy				Levenshtein Distance			
			csb	oci	lat	ben	csb	oci	lat	ben
None	Copy Baseline	5	12.0	1.0	2.4	1.0	1.90	3.01	3.83	3.56
LR only	DTLM (standard)	1	60.0	**64.0**	14.3	55.0	0.58	**0.93**	2.56	0.86
	DTLM + wordlists	2	58.0	63.0	**34.0**	**64.0**	**0.56**	0.97	**1.85**	**0.72**
HR only	No Projection	-	16.0	7.0	0.3	n/a	1.72	2.91	5.13	n/a
	Instance Projection	4	28.0	5.0	0.0	0.0	2.02	2.99	6.58	6.51
	Data Projection	3	30.0	3.0	0.8	0.0	1.54	3.20	4.85	6.63
LR + HR	Data Projection	6	**66.0**	23.0	1.4	17.0	0.78	2.22	4.25	3.33
	1-MONO	-	62.0	60.0	5.1	31.0	0.60	1.11	3.34	1.91

Table 3: Inflection results on test sets of the shared task.

5.2 Inflection Generation

We perform inflection generation with DTLM in the low-resource setting, and OpenNMT in the high-resource setting. For DTLM, we apply the tag splitting and particle handling techniques described in Nicolai et al. (2017). In particular, we split tag sequences into component tags, and append them at both the beginning and end of the lemma, treating each of them as an atomic symbol. We tune the hyper-parameters of both the aligner and transducer using grid search for each language. For OpenNMT, we split tag sequences, and append them to the lemma. All parameters are set to default values.

The task of leveraging HR training data for LR inflection generation is complicated by two types of inconsistencies. First, there are unavoidable typological differences, especially between less similar languages. For example, Latin nominal inflection paradigms include six cases, most of which do not exist in Romanian, which instead distinguishes between definite and indefinite forms. Second, the order of the tags in the data may differ. For example, the person tag follows the tense tag in the Spanish data, while the order is reversed in the Occitan data. We do not perform any tag re-ordering in the current shared task, but see Nicolai et al. (2017) for a principled solution to this problem.

5.3 Cognate Projection

We train our cognate models on lists of HR-LR word pairs acquired from the bitexts. The bitexts are aligned with FAST_ALIGN (Dyer et al., 2013). We extract all aligned word pairs, and sort them by the alignment frequency. For Hindi and Bengali, which are written in different scripts, we compute the inter-lingual orthographic similarity after romanizing all words using *uroman* (Hermjakob et al., 2018). We discard all pairs with orthographic similarity below a threshold t, which is manually tuned for each language pair. The similarity is computed as $1 - D/L$, where D is the Levenshtein distance, and L is the length of the longer of the two strings. Furthermore, we discard pairs which involve any words that are English, are shorter than 4 characters, or include digits. We take the top k HR-LR pairs, and randomly divide them into training, development, and test sets, as summarized in Table 2.

For each language pair, we train a DTLM model in each direction on the training set, using the development set to prevent over-fitting, as well as a target-language word list (Section 5.1). The results of the intrinsic evaluation of the projection models on the in-domain test sets are shown in Table 4. The accuracy of the Romanian-Latin is relatively low, which may be due to the Bible domain.

6 Results and Discussion

We test several systems, as listed in Table 3. (Submission IDs are given here in parentheses.) A naive copy baseline (5) simply outputs the unchanged input lemmas. DTLM models with and without target wordlists (2 and 1) make no use of HR data (the latter is our only standard submission, which uses no external resources). The next three systems make use of only the HR training sets provided as part of the shared task. This emulates a scenario[6] where no LR inflection data is available. *Data Projection* (3) and *Instance Projection* (4) implement the two methods illustrated in Figure 1, while *No Projection* simply applies an inflection model trained on HR data to LR in-

[6]We note the similarity to the setup in the shared task on Cross-lingual Morphological Analysis of VarDial 2019 (Zampieri et al., 2019).

Pair	WA (LD)	Pair	WA (LD)
pol→csb	28.6 (1.97)	csb→pol	49.8 (1.32)
spa→oci	47.3 (1.76)	oci→spa	46.7 (2.15)
ron→lat	5.5 (2.88)	lat→ron	17.9 (2.26)
hin→ben	22.2 (2.92)	ben→hin	29.8 (2.62)

Table 4: Intrinsic evaluation of cognate projection.

stances. The last system (6) combines the projected HR inflection data with LR data, which probably comes closest to the spirit of this shared task. 1-MONO is the first-order monotonic hard attention system of Wu and Cotterell (2019).

The test results are shown in Table 3. The best result on each language is shown in bold. When only LR data is used, the results confirm the finding of (Nicolai et al., 2018) that leveraging target wordlists from monolingual corpora can improve inflection accuracy for less-closely related languages. With the exception of Polish-Kashubian, the standard DTLM model is better than the competitive baselines. However, the Polish-Kashubian results demonstrate that cognate projection can outperform the *Copy* and *No Projection* baselines when only HR data is used. Finally, augmenting the LR training data with the projected HR data does not improve the inflection accuracy in most cases.

7 Conclusion

We described the details of the systems that we tested on four language pairs in the SIGMORPHON 2019 Shared Task. In particular, we successfully experimented with leveraging cognate projection for inflection generation. We view our Polish-Kashubian results as a proof of concept that should motivate further research on this new idea.

Acknowledgments

We thank Garrett Nicolai for the assistance with DTLM. We thank the shared task organizers for their effort.

This research was supported by the Natural Sciences and Engineering Research Council of Canada.

References

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

Christos Christodouloupoulos and Mark Steedman. 2015. A massively parallel corpus: the Bible in 100 languages. *Language Resources and Evaluation*, 49(2):375–395.

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

Ryan Cotterell, Christo Kirov, John Sylak-Glassman, Géraldine Walther, Ekaterina Vylomova, Arya D. McCarthy, Katharina Kann, Sebastian Mielke, Garrett Nicolai, Miikka Silfverberg, David Yarowsky, Jason Eisner, and Mans Hulden. 2018. The CoNLL–SIGMORPHON 2018 shared task: Universal morphological reinflection. In *Proceedings of the CoNLL–SIGMORPHON 2018 Shared Task: Universal Morphological Reinflection*, pages 1–27, Brussels. Association for Computational Linguistics.

Ryan Cotterell, Christo Kirov, John Sylak-Glassman, Géraldine Walther, Ekaterina Vylomova, Patrick Xia, Manaal Faruqui, Sandra Kübler, David Yarowsky, Jason Eisner, and Mans Hulden. 2017. The CoNLL-SIGMORPHON 2017 shared task: Universal morphological reinflection in 52 languages. In *Proceedings of the CoNLL-SIGMORPHON 2017 Shared Task: Universal Morphological Reinflection*, Vancouver, Canada. Association for Computational Linguistics.

Ryan Cotterell, Christo Kirov, John Sylak-Glassman, David Yarowsky, Jason Eisner, and Mans Hulden. 2016. The SIGMORPHON 2016 shared task—morphological reinflection. In *Proceedings of the 2016 Meeting of SIGMORPHON*, Berlin, Germany. Association for Computational Linguistics.

Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. A simple, fast, and effective reparameterization of IBM model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–648, Atlanta, Georgia. Association for Computational Linguistics.

Ulf Hermjakob, Jonathan May, and Kevin Knight. 2018. Out-of-the-box universal Romanization tool uroman. In *Proceedings of ACL 2018, System Demonstrations*, pages 13–18, Melbourne, Australia. Association for Computational Linguistics.

Sittichai Jiampojamarn, Colin Cherry, and Grzegorz Kondrak. 2008. Joint processing and discriminative training for letter-to-phoneme conversion. In *Proceedings of ACL-08: HLT*, pages 905–913, Columbus, Ohio. Association for Computational Linguistics.

Katharina Kann and Hinrich Schütze. 2016. MED: The LMU system for the SIGMORPHON 2016 shared task on morphological reinflection. In *Proceedings of the 14th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 62–70, Berlin, Germany. Association for Computational Linguistics.

Christo Kirov, Ryan Cotterell, John Sylak-Glassman, Géraldine Walther, Ekaterina Vylomova, Patrick Xia, Manaal Faruqui, Sebastian J Mielke, Arya McCarthy, Sandra Kübler, et al. 2018. Unimorph 2.0: Universal morphology. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander M. Rush. 2017. OpenNMT: Open-source toolkit for neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics - System Demonstrations*, pages 67–72. Association for Computational Linguistics.

Grzegorz Kondrak. 2009. Identification of cognates and recurrent sound correspondences in word lists. *Traitement automatique des langues et langues anciennes (TAL)*, 50(2):201–235.

Grzegorz Kondrak. 2013. Word similarity, cognation, and translational equivalence. In Lars Borin and Anju Saxena, editors, *Approaches to Measuring Linguistic Differences*, volume 265 of *Trends in Linguistics*, pages 375–386. De Gruyter Mouton.

Arya D. McCarthy, Ekaterina Vylomova, Shijie Wu, Chaitanya Malaviya, Lawrence Wolf-Sonkin, Garrett Nicolai, Christo Kirov, Miikka Silfverberg, Sebastian Mielke, Jeffrey Heinz, Ryan Cotterell, and Mans Hulden. 2019. The SIGMORPHON 2019 shared task: Crosslinguality and context in morphology. In *Proceedings of the 16th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, Florence, Italy. Association for Computational Linguistics.

Saeed Najafi, Colin Cherry, and Grzegorz Kondrak. 2019. Efficient sequence labeling with actor-critic training. In *Canadian Conference on Artificial Intelligence*, pages 466–471. Springer.

Saeed Najafi, Bradley Hauer, Rashed Rubby Riyadh, Leyuan Yu, and Grzegorz Kondrak. 2018a. Combining neural and non-neural methods for low-resource morphological reinflection. In *Proceedings of the CoNLL–SIGMORPHON 2018 Shared Task: Universal Morphological Reinflection*, pages 116–120, Brussels. Association for Computational Linguistics.

Saeed Najafi, Bradley Hauer, Rashed Rubby Riyadh, Leyuan Yu, and Grzegorz Kondrak. 2018b. Comparison of assorted models for transliteration. In *Proceedings of the Seventh Named Entities Workshop*, pages 84–88, Melbourne, Australia. Association for Computational Linguistics.

Garrett Nicolai, Colin Cherry, and Grzegorz Kondrak. 2015. Inflection generation as discriminative string transduction. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 922–931. Association for Computational Linguistics.

Garrett Nicolai, Bradley Hauer, Mohammad Motallebi, Saeed Najafi, and Grzegorz Kondrak. 2017. If you can't beat them, join them: the University of Alberta system description. *Proceedings of the CoNLL SIGMORPHON 2017 Shared Task: Universal Morphological Reinflection*, pages 79–84.

Garrett Nicolai, Bradley Hauer, Adam St Arnaud, and Grzegorz Kondrak. 2016. Morphological reinflection via discriminative string transduction. In *Proceedings of the 14th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 31–35, Berlin, Germany. Association for Computational Linguistics.

Garrett Nicolai, Saeed Najafi, and Grzegorz Kondrak. 2018. String transduction with target language models and insertion handling. In *Proceedings of the Fifteenth Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 43–53.

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

Jrg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

Shijie Wu and Ryan Cotterell. 2019. Exact hard monotonic attention for character-level transduction. *arXiv preprint arXiv:1905.06319*.

Marcos Zampieri, Shervin Malmasi, Yves Scherrer, Tanja Samardžić, Francis Tyers, Miikka Silfverberg, Natalia Klyueva, Tung-Le Pan, Chu-Ren Huang, Radu Tudor Ionescu, Andrei M. Butnaru, and Tommi Jauhiainen. 2019. A report on the third VarDial evaluation campaign. In *Proceedings of the Sixth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 1–16. Association for Computational Linguistics.