Initial Experiments In Data-Driven Cross-Lingual Morphological Analysis Using Morpheme Segmentation

Vladislav Mikhailov
National Research University
Higher School of Economics
School of Linguistics
Moscow
vnmikhaylov@edu.hse.ru

Lorenzo Tosi
National Research University
Higher School of Economics
School of Linguistics
Moscow
ltosi@edu.hse.ru

Anastasia Khorosheva
National Research University
Higher School of Economics
School of Linguistics
Moscow
aakhorosheva@edu.hse.ru

Oleg Serikov
National Research University
Higher School of Economics
School of Linguistics
Moscow
oaserikov@edu.hse.ru

Abstract

The paper describes initial experiments in data-driven cross-lingual morphological analysis of open-category words using a combination of unsupervised morpheme segmentation, annotation projection and an LSTM encoder-decoder model with attention. Our algorithm provides lemmatisation and morphological analysis generation for previously unseen low-resource language surface forms with only annotated data on the related languages given. Despite the inherently lossy annotation projection, we achieved the best lemmatisation F1-score in the VarDial 2019 Shared Task on Cross-Lingual Morphological Analysis for both Karachay-Balkar (Turkic languages, agglutinative morphology) and Sardinian (Romance languages, fusional morphology).

1 Introduction

This paper describes our submission to the VarDial 2019 Shared Task on Cross-Lingual Morphological Analysis (Zampieri et al., 2019). It is the task of producing lemma, part-of-speech tag and morphosyntactic annotations for previously unseen surface forms based on annotated data in related languages. Since surface forms are likely to be ambiguous, morphological analysis systems are supposed to produce a complete list of all possible and only valid analyses. These may represent not only multiple sets of morphosyntactic features, but also distinct lemmas and part-of-speech tags. For example, given a Turkish word *girdi* ‘to enter, entry’, the morphological analyzer should generate a full set of morphological readings the word can attain. (see Table 1).

In this paper we explore the task of data-driven cross-lingual morphological analysis. We apply this to two relatively low-resource languages: Karachay-Balkar and Sardinian; the former is Turkic with agglutinative morphology, while the latter is Romance and has fusional morphology.

We believe that it is possible to transfer morphology across related languages by exploiting cross-lingual inflection patterns despite language-specific morphological features inventories. For example, we can observe orthography specific common substring in *NOUN* surface forms in multiple related languages which stores the same tag values set *Case = Loc, Number = Plur, Number[psor] = {Sing, Plur}, Person[psor] = 3* (see Table 2).

Our method is inspired by previous approaches to neural-network based morphological analysis using inflection patterns for Polish (Jędrzejowicz and Strychowski, 2005), to cross-lingual morphological tagging for low-resource

Proceedings of VarDial, pages 144–152
Minneapolis, MN, June 7, 2019 ©2019 Association for Computational Linguistics

iso	word form	lemma	POS	MSD
tur	girdi	gir	VERB	Aspect = Perf, Number = Plur, Person = 3, Tense = Past, Valency = 1, VerbForm = Fin
tur	girdi	girdi	NOUN	Case = Nom
tur	girdi	gir	VERB	Aspect = Perf, Number = Plur, Person = 3, Tense = Past, Valency = 2, VerbForm = Fin

Table 1: A complete set of morphological annotations for the Turkish word form *girdi*, meaning *gir*- '(to) enter', *girdi*- 'entry'.

iso	word form	lemma	POS	MSD
tur	kenarlarında	kenar	NOUN	Case = Loc, Number = Plur, Number[psor] = {Sing, Plur}, Person[psor] = 3
kir	клеткаларында	клетка	NOUN	Case = Loc, Number = Plur, Number[psor] = {Sing, Plur}, Person[psor] = 3
tat	мәктәпләрендә	мәктәп	NOUN	Case = Loc, Number = Plur, Number[psor] = {Sing, Plur}, Person[psor] = 3
bak	далаларында	дала	NOUN	Case = Loc, Number = Plur, Number[psor] = {Sing, Plur}, Person[psor] = 3
kaz	аңғарларында	аңғар	NOUN	Case = Loc, Number = Plur, Number[psor] = {Sing, Plur}, Person[psor] = 3

Table 2: An example of morphological grapheme level pattern for a set of NOUN tag values: *-ләрендә* (tat), *-larında* (tur) and *-ларында* (kir, bak, kaz).

languages (Buys and Botha, 2016) and to data-driven morphological analysis for Finnish (Silfverberg and Hulden, 2018). In contrast to these approaches, our algorithm[1] produces full morphological analyses for previously unseen surface forms: it provides both morphological tags and lemmas as output and it can return multiple alternative analyses for one input word form.

We now give a brief description of our algorithm. First, we orthographically normalize and automatically transliterate both source and target language data into a joint orthographic representation using lookup tables. We model morpheme-to-tag inventory for each language family employing unsupervised morpheme segmentation with Morfessor (Virpioja, 2013). After this, we cluster all the target surface forms by making predictions over only part-of-speech tag with Morphnet (Silfverberg and Tyers, 2019), an LSTM encoder-decoder model with attention implemented using the OpenNMT neural machine translation toolkit (Klein et al., 2017). Within each cluster, we apply a greedy annotation algorithm using the cross-lingual morpheme-to-tag inventory. The next step is the annotation projection based on string intersections between source language data and target language data. Finally, we transliterate the analyzed target language data back to its non-normalized format.

2 Related works

Our method is similar to alignment-based distant supervision approach, where the aim is to train a morphological tagger in the low-resource language through annotations projected across aligned bilingual text corpora with a high-resource language. (Buys and Botha, 2016) propose an embedding-based model using Wsabie, a shallow neural network that makes predictions at each token independently. To project annotations onto the target language, one uses type and token constraints across parallel text corpora. However, we transfer morphology for target surface forms only through the same language family without manual constraint implementation, and cover lemmatisation with ambiguous annotations produced.

Another trend in cross-lingual morphological analysis is transfer learning. The key idea is to employ multi-task learning,

[1] Code available at `https://github.com/NIS-2018-CROSS-M/vardial-cma`

treating each individual language as a single task and train a joint model for all the tasks. All learned representations are jointly embedded into a shared vector space to transfer morphological knowledge in a language-to-language manner. (Cotterell and Heigold, 2017) propose a character-level recurrent neural morphological tagger to learn language specific features by forcing character embeddings for both high-resource language and low-resource language to share the same vector space. In contrast to the projection-based approach, this model requires a minimal amount of annotated data in the low-resource target language. However, we do not use the target language annotated data and morphological tagging datasets provided by the Universal Dependencies (UD) treebanks; and our algorithm generates lemmas and multiple sets of morphosyntactic annotations.

Our work is most closely related to the Morphnet model in (Silfverberg and Tyers, 2019). The essential idea is to analyze previously unseen surface forms using a corpus of morphologically annotated data in the related language. This can represent the solution to low coverage inherent in rule-based morphological analyzers. They don't require constant updating to keep working, but need to be updated to cover new surface forms. In our algorithm, we use Morphnet to cluster target words by predicting over part-of-speech tags and then to generate a full set of morphological readings for words not being analyzed with the greedy annotation algorithm. However, we do not use the Universal Dependencies (UD) treebanks and learn the algorithm to analyze four open class words: nouns, verbs, adjectives and adverbs.

3 Data

The data used in the experiments consisted of tab separated files with five columns: language code, surface form, lemma, part-of-speech tag and morphosyntactic description (MSD). We used unannotated data for two Turkic target languages (Crimean Tatar, Karachay-Balkar) and for two Romance target languages (Asturian, Sardinian). We also used annotated data for five Turkic source languages (Bashkir, Kazakh, Kyrgyz, Tatar, Turkish) and for five Romance source languages (Catalan, French, Italian, Portuguese, Spanish). We compiled a corpus of segmented target language surface forms with Morfessor.

3.1 Normalization

We discarded all the diacritics in the Romance languages set, e.g. a Portuguese word *seqüência* 'sequence' becomes *sequencia*. In the Turkic languages set, the Turkish data was transliterated from Latin into Cyrillic, e.g. *gelen* 'coming' becomes *гелен*. In the Karachay-Balkar data, we discarded grapheme '*ъ*' , e.g. *чыкъгъанды* 'appeared' becomes *чыкганды*. We also discarded all the diacritics in the Bashkir, Kazakh, Kyrgyz and Tatar data, e.g. Kazakh word *шыгармалардың* 'complete works' becomes *шыгармалардын*.

3.2 Morpheme segmentation

We employed default recursive training of the Morfessor model. In recursive training, the current split for the processed surface form is removed from the model and its morphemes are updated accordingly. After this, all possible splits are tried by choosing one split and running the algorithm recursively on the created morphemes. The best split is selected and the training continues with the next surface form. We did not tune the Morfessor model with the average morpheme length and the approximate number of desired morpheme types because we want to use the algorithm unsupervised. To train the model, we split the data into 80% train data and 20% test data. Consider an example of the output for the following input Karachay-Balkar surface forms, where the + sign implies the morpheme boundary:

нюзюрлеринде	*нюзюр+леринде*
экземпляр	*эк+з+е+м+п+ля+р*
политикасында	*политика+сында*

4 Methodology

In this section we describe our approach to data-driven cross-lingual morphological analysis implemented specifically for the Turkic languages. We refer to the source language annotated data as *Source*, to the target language unannotated data as *Target* and to the corpus of segmented target surface forms as *Morf*. *POS* is always one of *NOUN*, *VERB*, *ADJ*, or *ADV*.

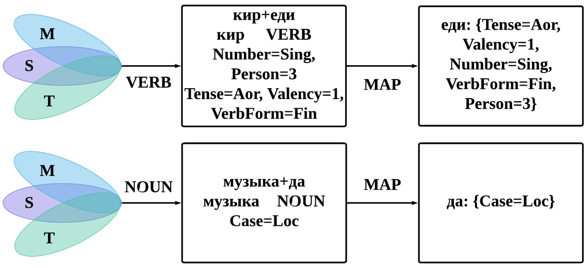

Figure 1: A graphical structure for modeling cross-lingual morpheme-to-tag inventory for Turkic languages. Source, Target and Morf are represented as S, T and M, respectively.

4.1 Morpheme-to-tag inventory

The key idea of modeling cross-lingual morpheme-to-tag inventory is automatic revealing of cross-lingual inflection patterns using unsupervised morpheme segmentation and string intersections between *Source*, *Target* and *Morf*.

We assume the morpheme-to-tag inventory to be specific to each part-of-speech tag and we do not merge the inventories in the experiments. We also do not compile the inventory for *ADJ* and *ADV* surface forms, since the latter do not store any morphosyntactic description in *Source*, e.g. *борын* ‘langsyne’ (word form), *борын* (lemma), *ADV* (POS), ’_’ (MSD). 116 out of 4146 (2%) *ADJ* surface forms in *Source* store only one tag value *Degree=Comp* which we consider to be statistically insignificant for model performance.

In agglutinative languages (Turkic family) the stem is invariant across different word forms. We generate distinct morphosyntactic features with a single root-word and map morphemes with morphological features, e.g. morpheme *лар* stores the feature *Number=Plur* for nouns and verbs. We represent each word as a grapheme level sequence $stem_i + m_{1i} + ... + m_{ni}$, so that $stem_i$ is a $lemma_i$ for $word\ form_i$. For example, the stem of a word *нюзюрлеринде* is *нюзюр* ‘promise’ and the lemma is *нюзюр*. We also refer to the first morpheme m_1 in each morphologically segmented word form $m_1, ..., m_n$ in $Morf$ as the Morfessor lemma. Consider the segmented word form *нюзюр+леринде* with the Morfessor lemma *нюзюр* and m_2 *леринде*.

The overall scheme for modeling *NOUN* and *VERB* cross-lingual morpheme-to-tag inventories is outlined in Figure 1, where the respective string intersections between *Source*, *Target* and *Morf* are found. Here, we first compute the word form intersection between *Source* and $Target$, and the lemma intersection between *Source* and $Morf$. If the word form in $Target$ can be found in *Source* and if the respective lemma in $Morf$ can be found in *Source*, we generate the following unit sequence: *Target word form*, *segmented Target word form*, *Morfessor lemma*, *Source word form*, *Source lemma*, *Source POS* and *Source MSD*. Within each unit sequence in the intersection, we project the *Source MSD* of the word form onto the second morpheme m_2 in the $Target$ segmented word form. Finally, we create the respective morpheme-to-tag pair.

For example, we have the following analysis in *Source*: *музыкада* (word form), *музыка* ‘music’ (lemma), *NOUN* (POS), *Case=Loc* (MSD). We can also find the same word form *музыкада* in $Target$ and the respective segmented word form *музыка+да* in $Morf$ (the Morfessor lemma is *музыка* and the segmented morpheme is *да*). On the basis of the word form intersection and the lemma intersection, we project the MSD *Case=Loc* onto the morpheme *да*. Thus, we create the morpheme-to-tag pair *да : Case=Loc*.

Since a single morpheme-to-tag pair can represent a concatenated string of distinct morphemes mapped with a set of morphological tag values, we additionally retrieve morpheme-to-tag pairs within each cross-lingual inventory. It is achieved by computing the difference between the morpheme strings and the difference between the tag values sets. For example, *NOUN* morpheme-to-tag inventory stores the following pairs:

ларын	*Case=Acc, Number=Plur, Number[psor]={Sing, Plur}, Person[psor]=3*
ын	*Case=Acc, Number[psor]={Sing, Plur}, Person[psor]=3*

In this case, we compute the difference between the two morpheme-to-tag pairs, i.e.

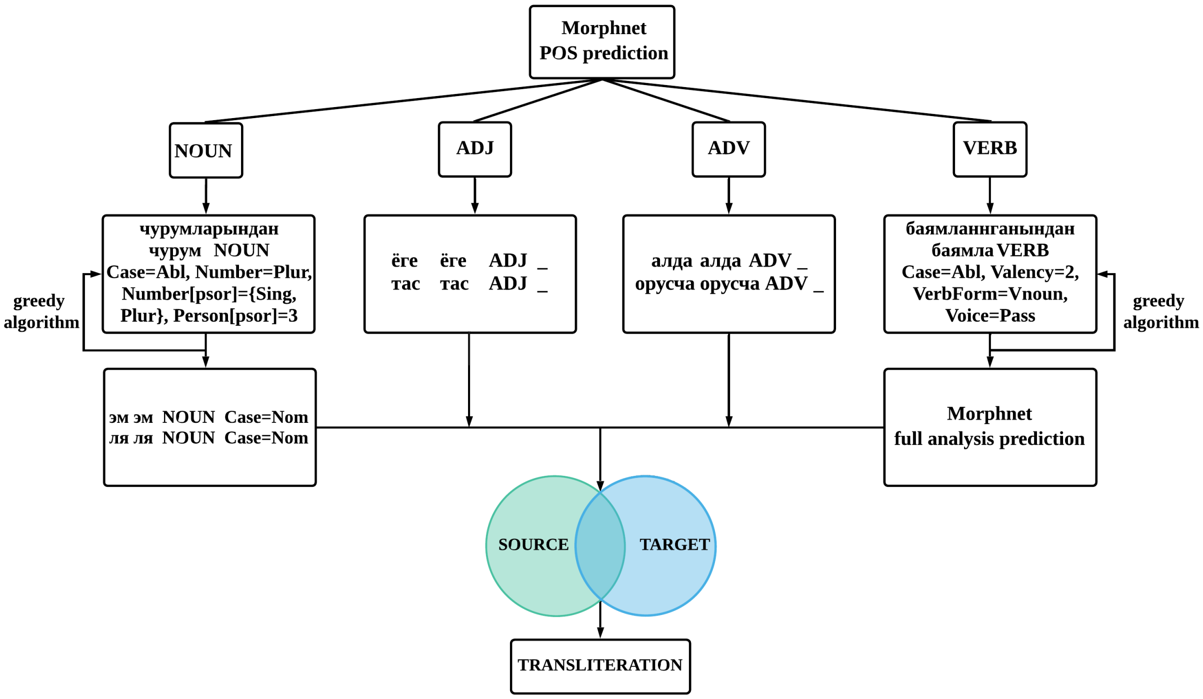

Figure 2: A graphical structure for morphological analysis of Karachay-Balkar surface forms.

$$\textit{ларын} - \textit{ын} \rightarrow \left\{\begin{array}{r} Case = Acc, \\ Number = Plur, \\ Number[psor] = \\ \{Sing, Plur\}, \\ Person[psor] = 3 \end{array}\right\} \setminus \left\{\begin{array}{l} Case = Acc, \\ Number[psor] = \\ \{Sing, Plur\}, \\ Person[psor] = 3 \end{array}\right\}$$

As a result, we map previously unretrieved morpheme *лар* with the tag value *Number=Plur*.

4.2 A greedy annotation algorithm

We cluster all the target surface forms by making predictions over only part-of-speech tags with Morphnet trained through *Source*. Each target surface form is processed in the following manner (see Figure 2). Due to the reasons described in Section 4, we consider $word\ form_i$ to be $lemma_i$ and MSD_i to be '_' for all the surface forms in ADJ and ADV clusters.

In $NOUN$ and $VERB$ clusters, we apply a greedy annotation algorithm to inflect a lemma and a morphosyntactic description for each surface form, which we now describe.

All morpheme-to-tag pairs in $NOUN$ and $VERB$ cross-lingual inventories are sorted by the morpheme length in the descending order. First, the longest cross-lingual morpheme is matched with a substring of the processed surface form starting from its end. If match, the respective inflection pattern is projected onto the surface form. If it fails to match, the next surface form is processed. After this, we deinflect a lemma by computing the difference between the surface form string and the matched morpheme string.

For example, one of the longest morphemes *ларындан (Case=Abl, Number=Plur, Number[psor]={Sing, Plur}, Person[psor]=3)* is matched with a target surface form *ызларындан*. The respective inflection pattern is projected onto the surface form. The inflected lemma is the difference *ызларындан* – *ларындан* → *ыз* 'trace'. Finally, we get a full analysis set: *ызларындан* (word form), *ыз* (lemma), $NOUN$ (POS), *{Case=Abl, Number=Plur, Number[psor]={Sing, Plur}, Person[psor]=3}* (MSD).

Out-of-vocabulary cross-lingual morpheme-to-tag pairs and ambiguous target surface forms are the potential weak points of the greedy algorithm. If the analysis with the greedy algorithm fails:

- we consider non-analyzed $NOUN$ surface forms to have the following analysis. Since a zero affix stores *'Case=Nom'* tag value, we assume $wordform$ string to be the respective $lemma$ and MSD to be *'Case=Nom'*. For example, *юг* 'South' (word form), *юг* (lemma), $NOUN$ (POS), '_' (MSD).

- we give non-analyzed $VERB$ surface forms to Morphnet as the input. The output is a full morphological analysis with ambiguous annotations, merged with the previously analyzed surface forms.

To correct the analyses acquired with the greedy algorithm and Morphnet predictions, we project annotations from $Source$ across $Target$ on the basis of string intersection (one-to-one orthographic match). We suppose that the intersection keeps loan words and cognates, which share the same set of morphological annotations. Finally, we employ transliteration of the analyzed target data back to its non-normalized format.

5 Experiments and results

We present results for six experiments and compare the performance of our algorithm on the VarDial 2019 CMA Shared Task with the baseline system. Since analyzed surface forms can have multiple morphological analyses, the results are evaluated on precision, recall and F1-score (Silfverberg and Tyers, 2019).

5.1 Experiment on Test Data

We performed four experiments in the orthographically non-normalized and normalized data scenarios on each language family.

Experiment 1. Turkic languages, normalized data
$Source$: Bashkir (bak), Kazakh (kaz), Kyrgyz (kir), Tatar (tat), Turkish (tur).
$Target$: Crimean Tatar (crh).

Experiment 2. Turkic languages, non-normalized data
$Source$: Bashkir (bak), Kazakh (kaz), Kyrgyz (kir), Tatar (tat), Turkish (tur).
$Target$: Crimean Tatar (crh).

Experiment 3. Romance languages, normalized data
$Source$: Catalan (cat), French (fra), Italian (ita), Portuguese (por), Spanish (spa).
$Target$: Asturian (ast).

Experiment 4. Romance languages, non-normalized data
$Source$: Catalan (cat), French (fra), Italian (ita), Portuguese (por), Spanish (spa).
$Target$: Asturian (ast).

5.2 Experiment on Surprise Language

In these experiments the target languages were unknown before the data release. We performed two experiments only in the orthographically normalized data scenario since the normalization improved the performance on the test data.

Experiment 5. Turkic languages, normalized data
$Source$: Bashkir (bak), Kazakh (kaz), Kyrgyz (kir), Tatar (tat), Turkish (tur).
$Target$: Karachay-Balkar (krc).

Experiment 6. Romance languages, normalized data
$Source$: Catalan (cat), French (fra), Italian (ita), Portuguese (por), Spanish (spa).
$Target$: Sardinian (srd).

Tables 3, 4, 5, 6 show the results on complete analyses including lemma, part-of-speech tag and morphosyntactic description. Our algorithm delivers the best F1-score on lemma prediction for Karachay-Balkar and Sardinian languages.

6 Discussion

Our approach of representing a surface form as a grapheme level sequence of stem and morphemes, along with retrieving cross-lingual inflection patterns improves performance on lemmatisation comparing to the baseline system. Despite the fact that this approach naively appears suitable only for agglutinative morphology, we yet achieve the best results for Sardinian (fusional morphology) in the VarDial 2019 Shared Task on CMA.

We looked at the analyses for the Karachay-Balkar language and classified the errors into nine categories: (1) Out-of-vocabulary morpheme-to-tag pairs; (2) Boundary between stem and morphemes; (3) Part-of–speech tag prediction; (4) Analysis overgeneration; (5) Insufficient analysis set; (6) Statistical assumption based error; (7) Back transliteration; (8) Substandard forms; (9) Derivational morphemes.

A common source of first-category errors is found in lemmatisation and morphosyntactic description of the surface forms storing out-of-vocabulary morphemes. For example, the morpheme _ла_ was not retrieved when modeling the cross-lingual inventory. As a result, the algorithm produced the lemma _*эмиратла_ and the MSD _*Case=Acc_ of the _NOUN эмиратланы_ rather than _эмират_ 'emirate' and _Case=Acc, Number=Plur_.

The algorithm also generated both correct and incorrect annotations for the same input form; this can be considered as the second error category. For example, we get one correct analysis for the word _башланды_ with the lemma _башлан_ 'beginning' and the incorrect one with the lemma _*башла_. It can be explained as the overenthusiastic greedy annotation when the cross-lingual morpheme being a substring of the lemma string.

For the third error category consider the _VERB юлеширге_ 'to divide' analyzed with Morphnet as _*NOUN_. Consequently, the algorithm produced the lemma _*юлешир_ and the incorrect MSD _*Case=Dat_ instead of _юлеш_ and _Case=Dat, Tense=Aor, Valency=2, VerbForm=Vnoun_.

The fourth error category includes superfluous analysis generation, e.g. we get two analyses for the _VERB эта_ (a form of the auxiliary verb 'to be') with the correct morphosyntactic annotation _Aspect=Imp, Valency=1, VerbForm=Conv_ and the redundant one _*Aspect=Imp, Valency=2, VerbForm=Conv_. This error can also occur when one surface form is predicted with two different part-of-speech tags, e.g. the word _къарачай-малкъар_ 'Karachay-Balkar' is analyzed as both _NOUN_ and _*VERB_.

For errors of the fifth type we have the _NOUN джолларын_ (lemma _джол_, meaning 'road') given only one correct annotation _Case=Acc, Number=Plur, Number[psor]={Sing, Plur}, Person[psor]=3_ instead of two possible. Moreover, there are ambiguous morphemes which store more than one tag value set. Consider the compound _NOUN премьер-министрни_ with the lemma _премьер-министр_, 'prime-minister' having two correct MSDs _Case=Acc_ and _Case=Gen_. In contrast, our algorithm provided only one correct MSD _Case=Acc_.

Errors of the seventh category can be

model	Recall	Precision	F1
experiment 1	66.91	33.32	44.49
experiment 2	25.07	9.88	14.17
experiment 3	62.09	31.82	42.07
experiment 4	67.70	13.83	22.97
baseline crh	36.43	44.74	40.16
baseline ast	66.64	70.73	68.62
experiment 5	43.01	35.59	38.95
experiment 6	74.58	37.15	49.60
baseline krc	39.59	50.94	44.55
baseline srd	66.42	67.28	66.85

Table 3: Results for morphosyntactic description prediction.

model	Recall	Precision	F1
experiment 1	76.75	34.85	47.94
experiment 2	32.43	13.15	18.72
experiment 3	35.34	21.02	26.36
experiment 4	58.53	13.43	21.85
baseline crh	56.87	59.66	58.23
baseline ast	62.28	59.90	61.07
experiment 5	63.30	51.82	56.99
experiment 6	48.07	32.55	38.82
baseline krc	54.90	56.91	55.89
baseline srd	35.73	35.59	35.66

Table 4: Results for lemma prediction.

model	Recall	Precision	F1
experiment 1	87.72	67.47	76.27
experiment 2	80.29	33.94	47.71
experiment 3	75.66	61.09	67.60
experiment 4	73.71	23.16	35.25
baseline crh	77.37	79.38	78.36
baseline ast	75.40	73.53	74.45
experiment 5	87.87	67.61	76.42
experiment 6	87.29	62.28	72.69
baseline krc	77.38	79.13	78.25
baseline srd	68.12	68.60	68.36

Table 5: Results for POS prediction.

model	Recall	Precision	F1
experiment 1	58.39	25.53	35.53
experiment 2	24.93	9.13	13.36
experiment 3	26.15	12.76	17.15
experiment 4	49.22	9.54	15.99
baseline crh	29.29	36.04	32.32
baseline ast	44.56	44.26	44.41
experiment 5	39.57	32.38	35.61
experiment 6	36.54	17.08	23.28
baseline krc	34.77	44.94	39.21
baseline srd	26.85	26.10	26.47

Table 6: Results for full analysis prediction: lemma + POS + MSD.

found in the lemmas containing mismatched graphemes *къ* and *гъ*. For example, the *NOUN* *де-факто* ‘de facto’ receives the lemma **де-факъто* instead of *де-факто*.

The eighth error category is represented by substandard attested word forms, e.g. the generated lemma **энтта* for the *ADV* *энтта* is confused with the correct lemma *энтда* ‘again’.

Errors of the ninth category are considered to be the incorrect lemmas for the word forms containing the derivational morphemes *ду*, *ды*, and *ди*. These can be found specifically in *ADJ* and do not store any morphosyntactic description. The algorithm produced the lemma **джокъду* for the *ADJ* *джокъду* ‘no’ derived from the underlying stem and the actual lemma *джокъ* ‘nothingness’.

We suggest that the rate of errors in the first and the second categories can probably be reduced by applying GBUSS algorithm (Shalonova et al., 2009) which proved to perform better than Morfessor. Another approach is to include the morphological information on lemma and suffixes into the character-based word representations learned by the bi-LSTMs (Özateş et al., 2018). The errors of the third, fourth and fifth categories might be partially resolved with the Morphnet hyperparameters tuning, e.g. increasing a probability threshold and adjusting the maximal number of output candidates specifically for each POS cluster. Back transliteration errors can be reduced by employing byte-pair encoding (BPE) which allows to eliminate the orthographic normalization. Finally, semi-supervised learning and retrieving *ADJ* cross-lingual morpheme-to-tag pairs might solve the sixth, eighth and ninth error categories.

Future Work

In future work we are planning to experiment with Slavic languages (fusional morphology).

Hard attention models for morphological analysis are the object of our further exploration since these proved to deliver a better performance in the low-resource language scenario (Cotterell et al., 2018).

Another direction is to make use of cognate identification as it might improve morphology transferring across the single language family. Cognates tend to share the same language knowledge, e.g. English *tooth* and German *Zahn* have the same semantic meaning and morphosyntactic features. This can be achieved with applying phoneme level Siamese convolutional networks (Rama, 2016) or generating multilingual cognate tables by clustering surface forms from existing lexical resources (Wu and Yarowsky, 2018).

Conclusion

In this paper we proposed an approach for data-driven cross-lingual morphological analysis in a low-resource language setting based on a combination of unsupervised morpheme segmentation, annotation projection and an LSTM encoder-decoder model with attention. Despite the morphological differences between agglutinative and fusional languages, our algorithm obtains the best performance on lemmatisation for Karachay-Balkar and Sardinian languages in the VarDial 2019 Shared Task on CMA.

Acknowledgements

We wish to thank the anonymous reviewers for helpful comments and suggestions.

References

Jan Buys and Jan A Botha. 2016. Cross-lingual morphological tagging for low-resource languages. *arXiv preprint arXiv:1606.04279*.

Ryan Cotterell and Georg Heigold. 2017. Cross-lingual, character-level neural morphological tagging. *arXiv preprint arXiv:1708.09157*.

Ryan Cotterell, Christo Kirov, John Sylak-Glassman, Géraldine Walther, Ekaterina Vylomova, Arya D McCarthy, Katharina Kann, Sebastian Mielke, Garrett Nicolai, Miikka Silfverberg, et al. 2018. The conll–sigmorphon 2018 shared task: Universal morphological reinflection. *arXiv preprint arXiv:1810.07125*.

Piotr Jędrzejowicz and Jakub Strychowski. 2005. A neural network based morphological analyser of the natural language. In *Intelligent Information Processing and Web Mining*, pages 199–208. Springer.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander M. Rush. 2017. OpenNMT: Open-source toolkit for neural machine translation. In *Proc. ACL*.

Şaziye Betül Özateş, Arzucan Özgür, Tunga Gungor, and Balkız Öztürk. 2018. A morphology-based representation model for lstm-based dependency parsing of agglutinative languages. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 238–247.

Taraka Rama. 2016. Siamese convolutional networks for cognate identification. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1018–1027.

Ksenia Shalonova, Bruno Golénia, and Peter Flach. 2009. Towards learning morphology for under-resourced fusional and agglutinating languages. *IEEE Transactions on Audio, Speech, and Language Processing*, 17(5):956–965.

Miikka Silfverberg and Mans Hulden. 2018. Initial experiments in data-driven morphological analysis for finnish. In *Proceedings of the Fourth International Workshop on Computational Linguistics of Uralic Languages*, pages 98–105.

Miikka Silfverberg and Francis Tyers. 2019. Data-driven morphological analysis for uralic languages. In *Proceedings of the Fifth International Workshop on Computational Linguistics for Uralic Languages*, pages 1–14.

Peter; Grönroos Stig-Arne; Kurimo Mikko Virpioja, Sami; Smit. 2013. Morfessor 2.0: Python implementation and extensions for morfessor baseline. D4 julkaistu kehittämis- tai tutkimusraportti tai -selvitys.

Winston Wu and David Yarowsky. 2018. Creating large-scale multilingual cognate tables. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*.

Marcos Zampieri, Shervin Malmasi, Yves Scherrer, Tanja Samardžić, Francis Tyers, Miikka Silfverberg, Natalia Klyueva, Tung-Le Pan, Chu-Ren Huang, Radu Tudor Ionescu, Andrei 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 (VarDial)*. Association for Computational Linguistics.