УДК 81'32

COMBINED MORPHOLOGICAL AND SYNTACTIC DISAMBIGUATION FOR CROSS-LINGUAL DEPENDENCY PARSYNG

Ageeva Ekaterina Higher School of Economics Russia E-mail: yekaterina.ageeva@gmail.com
Tyers, Francis UiT Norgga arktalas universitehta Norway
E-mail: francis.tyers@uit.no

Abstract

This paper describes a system for combined morphological disambiguation and dependency parsing and applies it to cross-lingual parsing of two under-resourced Turkic languages, Crimean Tatar and Tuvan. The system is based on finite-state morphological analysis followed by greedy transition-based dependency parsing. We show that it is possible to parse a related Turkic language using only a Treebank designed with another Turkic language in mind.

Keywords: syntactic analysis, dependency grammar, machine learning, banks syntactic trees, cross-language analysis.

КОМБИНИРОВАННЫЕ МОРФОЛОГИЧЕСКИЕ И СИНТАКСИЧЕСКИЕ НЕОДНОЗНАЧНОСТИ ДЛЯ СИНТАКСИЧЕСКОГО АНАЛИЗА ЗАВИСИМОСТЕЙ КРОСС-ЯЗЫКОВ

Агеева Екатерина Высшая школа экономики, Россия,
E-mail: yekaterina.ageeva@gmail.com
Тайерс Френсис Норвегия, E-mail: francis.tyers@uit.no

Аннотация

В данной статье описана система для комбинированной морфологической неоднозначности и синтаксического анализа зависимостей и применяет его к кросс-язычной разборе двух стран с ограниченными ресурсами тюркских языков, крымскотатарских и тувинских. Система

основана на конечном состоянии морфологического анализа с последующими жадными в разборе перехода на основе зависимостей. Покажем, что можно разобрать, связанную с использованием тюркского языка только Treebank разработан в виду с другим тюркскими языками.

Ключевые слова: синтактический анализ, грамматика зависимостей, машинное обучение, банки синтактических деревьев, межъязыковой анализ.

1 Introduction

Morphological and syntactic analysis are the stepping stones for more complex language processing applications, such as machine translation, information retrieval, question-answering, and many others. We also explore the applicability of the joint method to cross-lingual parsing. Cross-lingual techniques are applied to different tasks, such as sentiment analysis (Wan, 2009), word sense disambiguation (Lefever et al., 2010), and others. The principal idea behind cross-lingual language processing is to apply the resources (e.g. corpora, treebanks, analysers) of one language to process a different language, which is usually under-resourced. Although cross-lingual dependency parsing has been performed before, by e.g. Xiao et al. (2014) and Tiedemann (2015), it did not benefit from combined morphosyntactic disambiguation. This work presents a free/open-source tool for combined morphological and syntactic parsing, and uses it to experiment with applying a parsing and disambiguation model trained on a Kazakh treebank to parse two related languages, Crimean Tatar and Tuvan.

	Treebank						Morphology	
	Train	Dev	Test	P	M	D	Coverage	Ambiguity
Kazakh	2,849	717	950	29	234	29	98.4	1.28
Tuvan	—	—	855	19	125	26	99.6	1.19
Crimean Tatar	—	—	639	19	103	26	99.5	1.77

Table 1: Statistics for the corpora, P is the set of part of speech tags, M is the set of unique morphological analyses and D is the number of dependency relations used.

2 Related work

Joint syntactic and morphological parsing (also called joint disambiguation) has emerged as an effort to improve parsing accuracy for languages with rich morphology, for which the standard parsing techniques perform poorly. One of the first experiments discussing the benefits of joint processing was carried out by Tsarfaty (2006). Tsarfaty explored the effects of joint morphological segmentation and part-of-speech tagging on parsing quality for Hebrew: the model that performed segmentation and tagging jointly had an advantage over the pipeline approach. Cohen et al. (2007) make the next step in joint parsing and include syntactic relations into their model. They trained two analysers, the first of which includes segmentation and part-of-speech tagging modules, and the second performs constituency parsing. The analysers are combined using the "product of experts" learning technique, which takes the product of independent probability functions to produce the final result. This work is also concerned with Modern Hebrew. Goldberg and Tsarfaty (2008) have developed the first model that incorporated morphology and syntax as a single classifier, as opposed to the two separate classifiers of Cohen et al. (2007) and achieve better results. Further experiments with joint morphological and syntactic disambiguation have explored its effects on parsing both morphologically rich languages and languages with high ambiguity of word forms, such as Chinese and English. Li et al. (2011) have developed several joint parsing models for Chinese, which incorporate different features and use various pruning strategies to reduce search space. These models have shown improvement over the pipeline models for Chinese. Similar work has been done by Bohnet and Nivre (2012), who also proposed to use the joint technique for dependency parsing, as opposed to constituency parsing in the works previous to this. Bohnet and Nivre develop a parser and experiment with Czech, German, English and Chinese, achieving state-of-the-art accuracy. Çetinoğlu et al. (2013) use this parser to process Turkish, and also report an improvement in parsing accuracy. Bohnet, Nivre, et al. (2013) further expand dependency parsing models by adding more sophisticated morphological information, and using word clusters to incorporate lexical information into the model. In addition, joint models may also deal with sub-word segments, and a number of works for Chinese word segmentation demonstrate improvement over the pipeline models (see e.g. Jiang et al. (2008), Kruengkrai et al. (2009), Sun (2011), and Zhang et al. (2008)).

3 Data sets and resources

Kazakh For training and testing we use the treebank developed by Tyers and Washington (2015). This consists of 402 sentences from different domains: learners' books, folk

Language	Gold tags	Pipeline	Oracle
Kazakh	71.7	61.4	61.8
Tuvan	71.9	50.4	51.0
Crimean Tatar	79.0	64.0	73.8

Table 2: Reference results (labelled attachment score, LAS) for the three languages in question. Gold tags means that the input to the parser was the part-of-speech tag and morphological information from the test corpus; Oracle is the result of parsing all the possible paths and selecting the one with highest LAS; Pipeline is the result of applying the statistical disambiguator described in Assylbekov et al. (2016) trained on the Kazakh treebank. Note that the Oracle may be lower than using the Gold tags as the morphological analyser may not cover all forms or return all valid analyses.

tales, legal texts and Wikipedia articles. The morphological analyser used was by Washington et al. (2014) and for pipeline disambiguation we used the hybrid tool developed by Assylbekov et al. (2016), consisting of approximately 150 hand-written rules in constraint grammar and a statistical model.

Tuvan For testing we used 115 grammar-book sentences from Anderson et al. (1999) annotated for universal dependencies (Nivre et al., 2016) distributed with the morphological analyser described in Tyers, Washington, et al. (2016).

Crimean Tatar For testing we use a treebank consisting of 150 grammar-book sentences from Kavitskaya (2010) annotated for universal dependencies. The morphological analyser is available from the `apertium-crh` package[1] under development at the Apertium project.

4 Reference system

In order to assess the capabilities of a combined parsing model, we have performed experiments with a non-combined reference parser. This parser is purely syntactic, and the part-of-speech and morphological information is pre-disambiguated. The non-combined parser is a basic greedy transition-based implementation as described by Kübler et al. (2009). We used the decision tree classifier and its internal tuning and cross-validation algorithms that have been implemented as part of the `scikit-learn` module for Python (Pedregosa et al., 2011)

In order to choose the best feature set for the model we performed experiments with different combinations of features, testing the performance of each set against the development set for Kazakh. The combination of features which resulted in the highest labelled-attachment score was chosen. This was part-of-speech, lemma and morphological information for the word at the top of the stack and the first four words at the front of the buffer.

Using the model described above, we conducted three experiments to determine the boundaries of what the combined model can achieve given our data. Each experiment has been run on the Kazakh corpus. In addition, we performed two cross-lingual experiments: the Crimean Tatar and Tuvan test corpora were using the model trained on Kazakh data. The experimental results are summarised in Table 2.

[1] `https://svn.code.sf.net/p/apertium/svn/incubator/apertium-crh/`

Gold The model was given the unambiguous input of the gold part-of-speech and morphological analysis from the testing section of the corpus.[2] This gives an upper bound on performance of the pipeline model. If our disambiguator had 100% accuracy with respect to the corpus, this is the parsing performance we could expect to achieve.

Pipeline This model can be considered to be the *state of the art* for each language. The output of the morphological analyser is first disambiguated (by a hybrid disambiguator, see Assylbekov et al. (2016)) and the best output of that disambiguation is given to the parsing model.

Oracle With this model, each path from the lattice output (see Figure 1) by the morphological analyser is expanded and parsed. The resulting output is scored with labelled-attachment score, and for each sentence, the best score is taken. This can be considered to be the upper bound of performance for the combined model.

4.1 Formats and metrics

As the Kazakh treebank takes advantage of the new tokenisation standards in the CoNLL-U format,[3] and the parser only supports CoNLL-X, certain transformations were needed to perform the experiments. The corpus was flattened with conjoined tokens receiving a dummy surface form. The converted data is available alongside the original.

Furthermore it was necessary to come up with a new format for expressing ambiguous analyses in a format similar to the CoNLL series of formats. The format is identical to CoNLL-U, but allows for each ID to be repeated with a different analysis.

ID	FORM	LEMMA	UPOS	XPOS	FEATS	HEAD	REL	DEPREL	MISC
1	Осында	осында	ADV	adv	_	_	_	_	_
1	Осында	осы	PRON	prn	_	_	_	_	_
2	орыс	орыс	NOUN	n	_	_	_	_	_
3	тілінде	тіл	NOUN	n	_	_	_	_	_
4	сөйлейтін	сөйле	VERB	n	_	_	_	_	_
4	сөйлейтін	сөйле	VERB	n	_	_	_	_	_
5	адам	адам	NOUN	n	_	_	_	_	_
6-8	бар ма	_	_	_	_	_	_	_	_
6	бар ма	бар	ADJ	n	_	_	_	_	_
7	бар ма	е	VERB	cop	_	_	_	_	_
8	бар ма	ма	PART	qst	_	_	_	_	_
9	?	?	PUNCT	sent	_	_	_	_	_

5 Combined model

5.1 Preprocessing

As both Crimean Tatar and Tuvan lack an annotated corpus with which to train a part-of-speech tagger or morphological disambiguator, so the input to the parser is a lattice (see Figure 1) rep-

[2] This is equivalent to taking the first six columns of CoNLL-U format and feeding them to the parser.

[3] http://universaldependencies.org/format.html

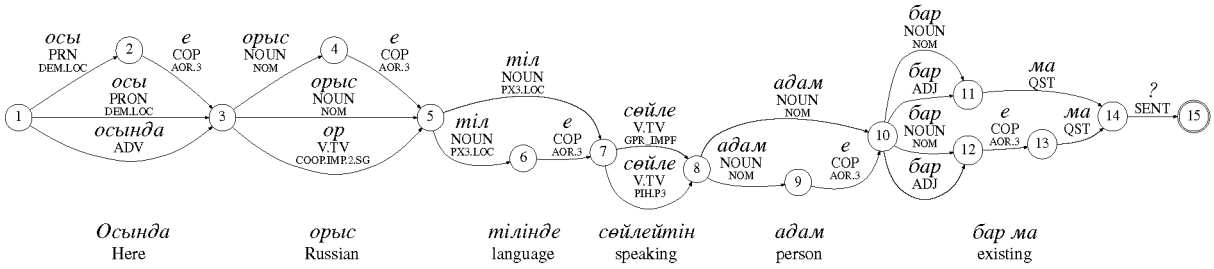

Figure 1: An example of ambiguous tokenisation for the sentence *Осында орыс тілінде сөйлейтін адам бар ма?* "Is there a person here who speaks Russian?" in Kazakh. The tokenisation path expressed in the treebank is in bold.

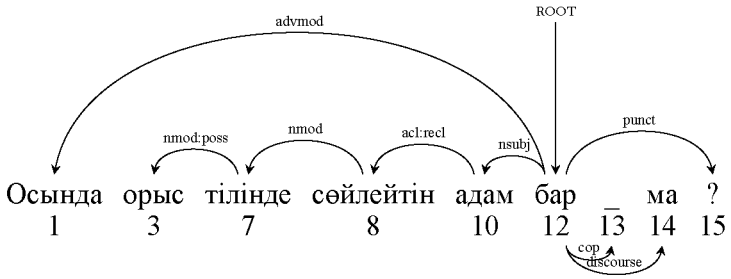

Figure 2: The dependency tree for the sentence given in Figure 1.

resenting the ambiguous output of the morphological analysers. The morphological analysers also perform tokenisation on the basis of a left-to-right longest match algorithm described in Garrido-Alenda et al. (2002). The mapping between space-separated 'surface forms' and syntactic tokens is non-trivial. In some cases a single surface token is equivalent to a single syntactic token (as in *сөйлейтін* 'speaking' in Figure 1), in other cases, multiple surface tokens may result in multiple syntactic tokens (as in *бар ма* 'is existing?' in Figure 1). There are a number of factors involved in determining if multiple surface tokens should be treated as a single token, including: does the token undergo any morphophonological processes? (e.g. the question suffix *ма* may also appear as *ме*, *ба* or *бе* depending on the ending of the previous token), and does an extra syntactic token (e.g. the zero-copula in the third person aorist) need to be introduced?

5.2 Morphological disambiguation

Having performed the baseline experiments, we have set out to develop the combined syntactic and morphological parser. We took our syntactic parser as a base and added the capability to do morphological disambiguation. We treat morphological disambiguation as a classification task, similar to determining the best next transition in the dependency parser. In this case, the items to classify also are configurations, and the label assigned to each is a concatenation of the part-of-speech and the morphological tags of the first word in the buffer. In reality, the entity classified is the first word in the buffer, but because we use the features from other parts of the configuration, technically, we classify configurations. We have chosen to perform disambiguation of the first word in the buffer. On one hand, it is best to disambiguate as late as possible, so that the syntactic parser can benefit from additional information for as long as

Language	Pipeline	Combined	Oracle
Kazakh	61.4	63.2	61.8
Tuvan	50.4	58.4	51.0
Crimean Tatar	64.0	62.4	73.8

Table 3: Results

possible. On the other hand, all transitions in the syntactic parser assume that the tokens are already disambiguated, and that the words that may participate in transitions are the first word in the buffer and the first word on the stack. Therefore, disambiguation happens just before the word can potentially participate in any transitions, but not earlier. We check if the buffer front needs disambiguation before predicting every following transition. We also accommodate for ambiguous tokenisation when the analyser may need to split a single 'surface form' into several structural tokens, which later form the dependency relations (for example, tokens *бар ма* in Figure 1). In this case, these tokens are *unwrapped*, and both the buffer and the underlying sentence shift to make place for the extra tokens.

The features we use for morphological disambiguation are the same as for dependency parsing, with a minor change. Because we only disambiguate the token when it reaches the buffer front, the features concerning other items in the buffer were modified to work with ambiguous tokens in the following way:

form: returns the form of the first analysis, or the unifying surface form for several syntactic tokens, if there are multiple;

part-of-speech: returns the ambiguity class of the token, i.e. a concatenation of all distinct part-of-speech tags seen in the analyses for this token. For the token *орыс* this would be NOUN|VERB.

morphological features: returns nothing if the token is ambiguous.

The dependency parser drives the process: it moves the state from one configuration to another by determining the next transition, until the buffer is empty. The classifier for morphological disambiguation works as a supplementary tool at each step, selecting the best analysis for the buffer front as described in the section above.

6 Cross-lingual parsing

It was necessary to make a number of small changes to the annotation scheme for the Crimean Tatar and Tuvan treebanks as the annotation conventions for a number of phenomena are not yet completely standardised. The universal dependency relation `iobj` 'second obligatory argument' (typically indirect object) is called `arg` in the Tuvan data, and `nmod:rcp` in the Crimean Tatar data. These were standardised to `iobj`. The Crimean data used `acl:relcl` for relative clauses, where the Tuvan and Kazakh data used `acl`. We standardised on `acl`. Finally, both Crimean Tatar and Tuvan distinguished clausal subjects, `csubj` from nominal subjects `nsubj`, a distinction which is currently not made in the Kazakh treebank, so we collapsed both of these labels into a single `subj` label.

There were also a number of idiosyncracies left in place, for example `nsubj:caus` for causative subjects if verbs in Crimean Tatar. There were no examples of this phenomenon in the Tuvan or Kazakh data.

7 Discussion

7.1 Error analysis

We analysed the errors made by the dependency parser and the morphological disambiguation component. This section reports the common error patterns we discovered. We have observed, although to a lesser extent than in the pipeline models, the error accumulation effect: if the morphological classifier selected an incorrect analysis for a given token, it will very likely enter an incorrect dependency relationship, which will at least partly affect the parse tree. The errors at this point may be incorrect tokenisation (cases when one surface form is analysed as several lemmas), incorrect part-of-speech label, or incorrect morphological analysis. Predictably, the parser also makes errors of its own, assigning incorrect head and/or dependency labels when the morphology has been determined correctly. We should note that the mistakes are not language-specific, and repeat across different corpora — which is not very surprising, provided we used the same model to parse them. The first, and perhaps the most expected category of errors deals with part of speech ambiguity. Words like *bu* / *бо* 'this' and o 'that' can be classified as determiners, demonstrative pronouns, or personal pronouns (*o* as 'that' vs 'he'); *bir* 'one' can be a numeral or an indefinite determiner, the distinctions not always correctly made by our model. It also tends to select the substantive interpretation over an attribute adjective, an error which has surfaced in the Tuvan and Kazakh corpora. Some of part of speech errors are common; others are made due to lack of training data. For example, the Kazakh word *көп* 'many' was misclassified once as an adjective as opposed to a determiner, and once correctly classified as an adverb, but it never occurred in the training corpus. In cases when part of speech has been determined correctly, the morphological information may have not been. The most common source of such errors is the distinction between verb forms. There are cases when passive transitive verbs have been classified as intransitive, and when the tag for a participle form has been assigned instead of (the correct) tag for verbal adverb form. These particular distinctions, however, have been up to debate in the annotation guidelines of the corpus, and can also depend on the interpretation. Other morphological errors are more straightforward and reveal that the parser may be rather ignorant about the surrounding words. For example, the verb *басталды* 'started' was classified twice as plural rather than singular, even though it had a correctly determined singular subject. In general, having more context may improve parsing accuracy, although at a cost of considering more possibilities at each step. A common error that speaks in favour of this is finding multiple subjects in rather simple sentences — a pattern that is infrequent in training data, and that could have been better learned. Consider a the Crimean Tatar sentence in Figure 3 in, where three words – my brother, every, and day – have been tagged as subjects.

We suspect that in such cases the parser (especially having made an incorrect part of speech decision) assigns the relation that is most likely given a local context. It does not consider the likelihood of a 6-word sentence having 3 subjects, knowledge of which would significantly improve its performance.

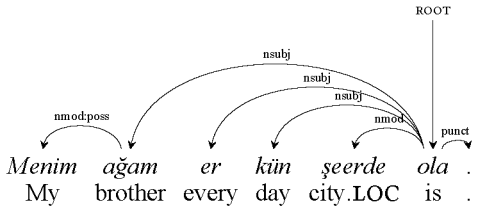

Figure 3: Multiple subjects in a simple sentence

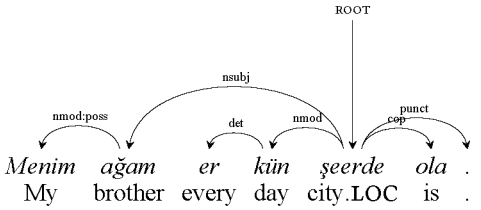

Figure 4: Treebank parse for Figure 3

Finally, a small fraction of errors come from rare categories, which have not been encountered often enough during training. For example, the only instance of a 'vocative' relation in the Kazakh testing corpus has not been correctly determined, but it only occurred twice in the training corpus. Another example is the dependency label 'parataxis', which signifies a relation between the main verb and the clause after a colon or a semicolon. This relation has no overt signs of coordination or subordination, and is therefore difficult to learn, especially given the 10 instances (0.3%) in the training corpus.

7.2 Future work

There are a number of avenues for future work. One aspect we would like to improve concerns the output of the morphological analyser. At the moment the lattice we give to the parser is unweighted, that is all of the analyses are considered equally probable. However, this is unlikely to be the case. A noun reading for the word *орыс* 'Russian' is far more likely than the cooperative imperative of the verb *ор* 'reap with me!'. It is possible to apply weights to a finite-state transducer either using corpora, or linguistic knowledge, and this is something we would like to incorporate into the model. On a similar vein, we would like to experiment with adding rule-based constraints. Given a small or non-existent treebank (in the case of cross-lingual parsing), is it possible to write simple rule-based constraints which could be incorporated into the model ? These constraints could be of the type "A clause may have at most one subject", "The copula verb cannot be the root of a sentence" or "A personal pronoun in nominative is never a nominal modifier". In considering the model, we would like to implement a real joint model, where we have a single classifier which predicts both the best transition and morphological disambiguation in a single step. Looking at adding word-embeddings (Mikolov et al., 2013), which can be calculated from inexpensive monolingual corpora would also be an interesting avenue for future work. Finally, we would like to apply this work to other Turkic languages, and possibly use the parser to bootstrap treebanks for other Kypchak languages.

8 Concluding remarks

This work has been concerned with cross-lingual dependency parsing enhanced by morphological disambiguation. We have developed a combined syntactic and morphological parser, which is transition-based and operates with two independent classifiers. We have shown that it is possible to use the classifiers trained on Kazakh data to parse corpora in Crimean Tatar and Tuvan. After adding morphological disambiguation to the dependency parsing process, we have improved the parsing quality for Kazakh and Tuvan over the baseline scores. All of the

code and data used in the experiments can be found on GitHub.[4]

Acknowledgements

Removed for review

References

Anderson, G. and Harrison, K. D. (1999). Tyvan. Lincom Europa.

Assylbekov, Z., Washintgon, J. N., Tyers, F. M., Nurkas, A., Sundetova, A., Karibayeva, A., Abduali, B., and Amirova, D. (2016). A free/open-source hybrid morphological disambiguation tool for Kazakh". 1st International Workshop on Turkic Computational Linguistics. In: *Proceedings of the 1st International Workshop on Turkic Computational Linguistics.*

Bohnet, B. and Nivre, J. (2012). A Transition-Based System for Joint Part-of-Speech Tagging and Labeled Non-Projective Dependency Parsing. In: *Proceedings of EMNLP.*

Bohnet, B., Nivre, J., Boguslavsky, I. M., Farkas, R., Ginter, F., and Hajič, J. (2013). Joint morphological and syntactic analysis for richly inflected languages. In: *Transactions of the Association for Computational Linguistics* 1, pp. 429–440.

Çetinoğlu, Ö. and Kuhn, J. (2013). Towards joint morphological analysis and dependency parsing of Turkish. In: *Proceedings of the Second International Conference on Dependency Linguistics (DepLing 2013)*, pp. 23–32.

Cohen, S. B. and Smith, N. A. (2007). Joint morphological and syntactic disambiguation. In: *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pp. 208–217.

Garrido-Alenda, A., Forcada, M. L., and Carrasco, R. C. (2002). Incremental construction and maintenance of morphological analysers based on augmented letter transducers. In: *Proceedings of the Conference on Theoretical and Methodological Issues in Machine Translation*, pp. 53–62.

Jiang, W., Huang, L., Liu, Q., and Lü, Y. (2008). A cascaded linear model for joint Chinese word segmentation and part-of-speech tagging. In: *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics.*

Kavitskaya, D. (2010). Crimean Tatar. Lincom Europa.

Kruengkrai, C., Uchimoto, K., Kazama, J., Wang, Y., Torisawa, K., and Isahara, H. (2009). An error-driven word-character hybrid model for joint Chinese word segmentation and POS tagging. In: *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP.* Vol. 1, pp. 513–521.

Kübler, S., MacDonald, R., and Nivre, J. (2009). Dependency Parsing. Morgan & Claypool.

Lefever, E. and Hoste, V. (2010). Semeval-2010 task 3: Cross-lingual word sense disambiguation. In: *Proceedings of the 5th International Workshop on Semantic Evaluation*, pp. 15–20.

[4] `http://github.com/Sereni/joint-parser`