Fine-Grained Morpho-Syntactic Analysis for the Under-Resourced Language Chaghatay

Kenneth Steimel†, **Akbar Amat**‡, **Arienne Dwyer**‡, **Sandra Kübler**†
† Indiana University
‡ University of Kansas
ksteimel@iu.edu, akbar.amat@ku.edu, arienne@ku.edu, skuebler@indiana.edu

Abstract

We investigate part of speech (POS) tagging for Chaghatay, a historical language with a considerable amount of morphology but few available resources such as POS annotated corpora. In a situation where we have little training data but a large POS tagset, it is not obvious which method will be best to obtain an accurate POS tagger. We experiment with a conditional random field and a Recurrent Neural Network, augmenting the models with coarse grained POS tag information, and by utilizing additional data, either additional unannotated data used to train a language model or annotated data from a modern relative, Uyghur. Our results show that the combination of an RNN and pretraining with coarse grained POS tags reaches the highest accuracy of 76.17%.

1 Introduction

Part of Speech (POS) tagging has often been considered a solved problem. For languages with large annotated resources, POS tagging has reached accuracies in the high 90s: For English, the state of the art[1] has reached 97.85% (Akbik et al., 2018), and for French, 97.80% (Denis and Sagot, 2009). However, this is definitely not the case for many other languages with fewer resources, which often also exhibit considerable morphology. In such cases, the POS tags may go beyond pure word classes and may include a range of morphological information[2].

The current paper presents work on creating a POS tagger for Chaghatay (ISO-639 code: chg), using a manually created, annotated corpus[3]. However, in terms of modern POS annotated corpora, this linguistically annotated corpus is small, and the POS tagset is complex, including a considerable amount of morphological information. This is one of the most challenging settings for POS taggers. We investigate which of the approaches to POS tagging that are currently considered state of the art, using conditional random fields (CRF) or recurrent neural networks (RNN), can be successful in such a setting.

Given the complex tagset, we are also interested in determining whether a first analysis using a coarse grained POS tagset can be beneficial. Our assumption is that if we can determine the coarse word class reliably, this information can guide the full POS tagger by restricting the available choices for a given word in context. We finally investigate whether additional data, either additional unannotated Chaghatay data, or annotated data from modern Uyghur, one of the language's modern relatives[4], can be employed to improve accuracy.

Our main goal is creating a POS tagger that can, in the future, be integrated in the annotation process, to alleviate the burden on human annotators. This is especially important for languages such as Chaghatay, where highly specialized knowledge is required for every annotation step, including transcription of the manuscript.

[1] As documented at `https://aclweb.org/aclwiki/POS_Tagging_(State_of_the_art)`.

[2] For convenience, we will use the term POS tag even though the annotations are a combination of POS tags and morphological annotations.

[3] `https://uyghur.ittc.ku.edu/atmo.html`

[4] Another option would be to use data from Uzbek, the other modern relative, but we are not aware of any POS annotated corpus.

Our result show that for POS tagging without any modifications, the CRF reaches a higher accuracy than the RNN. However, adding coarse-grained POS information allows the RNN to surpass the CRF. Adding data from additional sources does not seem to be useful.

The remainder of the paper is structured as follows: Section 2 provides an overview of the language, the corpus, and the tagset. Section 3 explains our research questions in more detail. Section 4 describes the experimental setup. Section 5 discusses our findings, and section 6 concludes.

2 Chaghatay

2.1 Overview of Chaghatay

Chaghatay [trk:chg] was a koiné variety used by Central Asian Turks from the 14th to early 20th century as a prestige literary language from Bukhara to Kashghar. It amalgamated Eastern Turkic, Kwārazm Turkic, and an increasing amount of Persian. Today it is regarded as Classical Uzbek or Classical Uyghur. Since Chaghatay was the prestige form used primarily by élites as a literary and erudite lingua franca, it was fairly uniform, despite its use over a large territory. A late eastern variety of Chaghatay is under examination here.

2.2 The Chaghatay Corpus

The corpus consists of late 19th-early 20th century Chaghatay manuscripts collected in the Kashgar area of the southern Tarim Basin, in Eastern Turkestan (Xinjiang). Comprising the Jarring Collection at Lund University, they were collected by the philologist and diplomat Gunnar Jarring and predecessor Swedish missionaries in the southern Tarim. Metadata, and those manuscripts scanned by the Lund University Library, are available online[5]. Transcriptions (in the original Perso-Arabic script), transliterations (into a lossless Latin script), English translations, and POS annotation for selected manuscripts are also available online[6]. Medicine, healing, and networks were the topical foci.

2.3 The POS Tagset

The tagset is primarily of inflectional morphology, and is described by Dwyer (2018). The tagset is relatively large (about 500 items), given the rich morphology of Turkic languages, and given that the tagset was originally designed for manual part of speech annotation and Interlinear Glossing (ILG) of both Chaghatay and its descendant, Modern Uyghur.

The annotation scheme is primarily sentence-based (for linguists), and for text scholars, line and page breaks were later added. Each sentential unit has the following annotation tiers: a transcription of the original Perso-Arabic; a lossless Latin-script version of the former, and a segmented tier. Each segment was then annotated in two morphological tiers, an all-caps form/function "POS" tag from the tagset, and an interlinear glossing (ILG) tier, in which substantives are glossed in English, and grammatical categories are repeated from the POS tier with all-caps tags. A free translation of the sentence constitutes the next tier, and a final tier contains textual or linguistic comments.

We show an image of an original manuscript opening in Figure 1 and the sentence-based annotation tiers in Figure 2.

Textual scholars are likely to be interested in line and page breaks in the manuscript. Therefore, the annotation scheme also accounts for a line or page break within a sentence using the element <phr/> (phrase), as shown in Figure 4. In the unpunctuated example in Figure 3, we can see that the sentence `akr kmrshnynk / astyma bwlmaqy tn aġyr bwlmaqy āġzy tatlyġ bwlmaq` runs over two lines (here with a slash inserted to represent the line break).

[5] http://www.alvin-portal.org/alvin/resultList.jsf?dswid=2113

[6] https://uyghur.ittc.ku.edu/manuscripts/index.xhtml

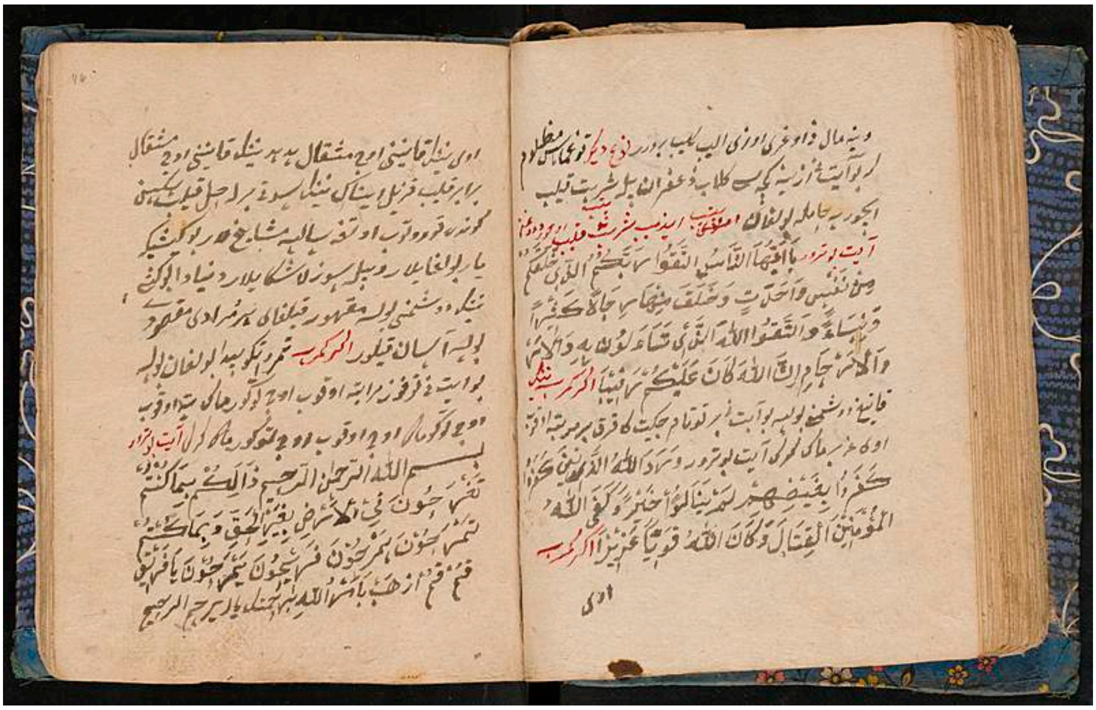
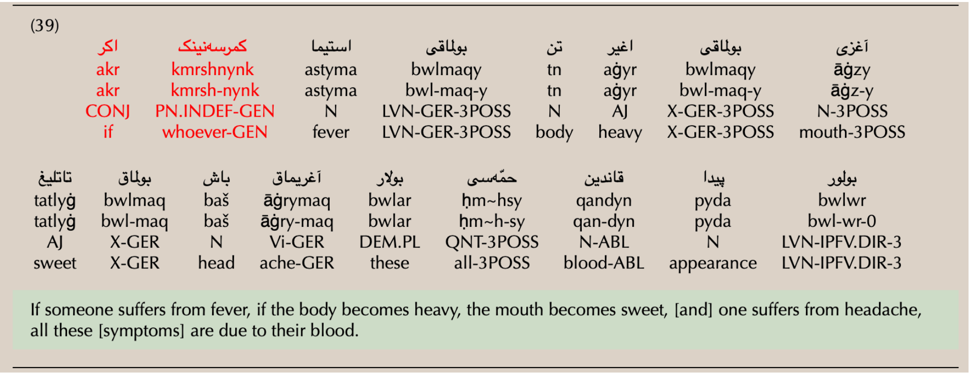


Figure 1: A page from the original manuscript.



(39)

اكر	كمرسەنينک	استيما	بولماقى	تن	اغير	بولماقى	آغزى
akr	kmrshnynk	astyma	bwlmaqy	tn	aġyr	bwlmaqy	āġzy
akr	kmrsh-nynk	astyma	bwl-maq-y	tn	aġyr	bwl-maq-y	āġz-y
CONJ	PN.INDEF-GEN	N	LVN-GER-3POSS	N	AJ	X-GER-3POSS	N-3POSS
if	whoever-GEN	fever	LVN-GER-3POSS	body	heavy	X-GER-3POSS	mouth-3POSS

تاتليغ	بولماق	باش	آغريماق	بولار	حمّەسى	قاندين	پيدا	بولور
tatlyġ	bwlmaq	baš	āġrymaq	bwlar	ḥm~hsy	qandyn	pyda	bwlwr
tatlyġ	bwl-maq	baš	āġry-maq	bwlar	ḥm~h-sy	qan-dyn	pyda	bwl-wr-0
AJ	X-GER	N	Vi-GER	DEM.PL	QNT-3POSS	N-ABL	N	LVN-IPFV.DIR-3
sweet	X-GER	head	ache-GER	these	all-3POSS	blood-ABL	appearance	LVN-IPFV.DIR-3

If someone suffers from fever, if the body becomes heavy, the mouth becomes sweet, [and] one suffers from headache, all these [symptoms] are due to their blood.

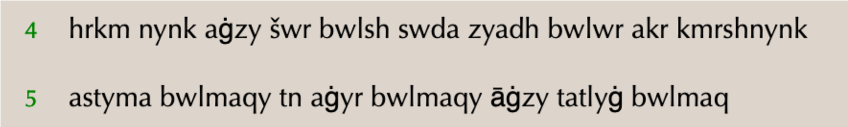
Figure 2: Example of the sentence-based tiers.

3 Research Questions

POS tagging for Chaghatay is one of the most challenging settings for POS tagging in general. The Chaghatay corpus is an ongoing project, thus little annotated data is available. Additionally, since Chaghatay is no longer spoken, there is only a limited amount of textual data available, restricting our ability to train a language model or use semi-supervised strategies. Finally, the POS tagset is large and includes a detailed analysis of morphological features. This leads us to consider the following questions:

3.1 Choice of Classifier

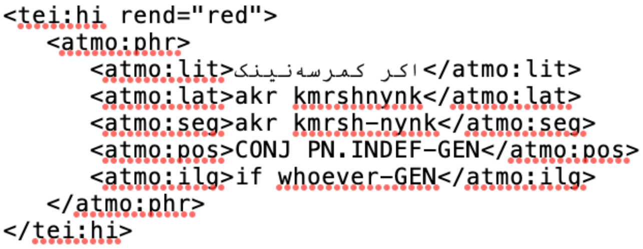
Given the combination of a small training set and a large POS tagset, the choice of classifier is not obvious. We decided to focus on two approaches that have been shown to be successful in POS tagging: Conditional Random Fields (CRF) (Gahbiche-Braham et al., 2012) and Recurrent Neural Networks (RNN) (Shao et al., 2017). RNNs are considered state of the art, but it is well known that they work best when they have access to large amounts of training data (Horsmann and Zesch, 2017).



4 hrkm nynk aġzy šwr bwlsh swda zyadh bwlwr akr kmrshnynk

5 astyma bwlmaqy tn aġyr bwlmaqy āġzy tatlyġ bwlmaq

Figure 3: Example of an unpunctuated example.



```
<tei:hi rend="red">
   <atmo:phr>
      <atmo:lit>اکر کمرسه‌نینک</atmo:lit>
      <atmo:lat>akr kmrshnynk</atmo:lat>
      <atmo:seg>akr kmrsh-nynk</atmo:seg>
      <atmo:pos>CONJ PN.INDEF-GEN</atmo:pos>
      <atmo:ilg>if whoever-GEN</atmo:ilg>
   </atmo:phr>
</tei:hi>
```

Figure 4: Example of a phrase element.

CRFs may be more amenable to small training data sets, but they may not scale up to a large label set (Horsmann and Zesch, 2017). Additionally, neural models can be pretrained on additional data from other domains and then optimized on our small training set.

3.2 Utilizing Coarse Grained POS Tagging as Preprocessing

The large tagset in this corpus is ideal for corpus-based analysis but provides challenges for statistical taggers. We investigate methods to overcome the challenges of a large tagset by using coarse POS tags as a first step in predicting the fine-grained tags. The most basic approach to POS tagging the data involves simply performing sequence tagging on the data using the fine grained POS tagset. However, given the combination of large tagset and small training set size, it is possible that the fined grained POS tagger could utilize information about the coarse grained category of a word. For example, knowing that a word is a noun will constrain the possible fine grained POS tags. Thus, we investigate for both types of models, CRFs and RNNs, whether utilizing coarse tags will improve the performance of the fine grained POS tagger.

Our approach involves separate coarse taggers, one for the CRF and one for the RNN, that are then leveraged by a more granular tagger. The CRF model uses coarse tags as additional features while the neural model uses transfer learning from a coarse tagger.

For the CRF model, we create a separate model trained on coarse tags. Then the coarse tagger is applied to a text, and the coarse tags predicted are included as features to the fine-grained CRF model. For this two-stage approach to be realistic, the coarse tagger needs to be trained using jackknifing (see Section 4.1.2 for details).

However, where the CRF tagger uses these coarse tags as features in its joint probability model, the neural approach does not use the coarse tag for making a decision about a specific word. Instead, it uses coarse grained tagging to provide a better initialization for the network. A standard method to obtain a better initialization would be to use off-the-shelf embeddings, which have been trained on a large data set of texts. For a low-resource language like Chaghatay, this is is not an option as such embeddings do not exist, and insufficient data is available to create traditional word embeddings like Word2Vec (Mikolov et al., 2013) and FastText (Bojanowski et al., 2017). Instead, we train a coarse-grained part of speech tagger and then transfer that model to fine-grained tagging by optimizing it on the more challenging task. This will provide a better weight initialization, similar to that provided by external embeddings.

3.3 Utilizing Training Data in Different Structure Formats

Since the corpus annotation process has evolved over time (see Section 2.2), we have manually annotated data in three different formats with regard to the marking of units: sentence segmented, phrase segmented, and line segmented. Since our task is POS tagging, we assign one label per word, but we also use information about sentence boundaries. Thus, the most relevant data are the sentence segmented documents. However, we only have very few of those, which raises the question whether we can use the other types of data to augment the training set. Does the additional data help guide the POS tagger, or is the missing information about sentence boundaries detrimental for the POS tagger? Does the difference in segmentation have any effect on POS tagging, or does the need for data override the need for sentence boundary information?

3.4 Pre-Training the RNN

For neural sequence tagging architectures, language model pre-training has been shown to be beneficial (Peters et al., 2018; Ortiz Suárez et al., 2020). In contrast to the CRF model, the transfer learning approach used for the neural network can be adapted from a variety of different initial tasks. We investigate whether this method can be used successfully in a setting where we have access to very little data in the target language. Since we do not have much additional data for Chaghatay, we experiment with two settings: 1) We use data from Chaghatay's modern relative, Uyghur, in the assumption that Uyghur is close enough to Chaghatay to provide a good starting point for the POS tagger. 2) We also experiment with pretraining the RNN using language modeling of Chaghatay as the task. For this pretraining step, we can leverage more training data since we can use all Chaghatay texts, including those that have been annotated for parts of speech yet.

4 Methodology

4.1 Data

As described in Section 2.2, the corpus has been annotated in different phases, with different underlying basic units of annotation, ranging from sentences, to lines in the manuscript and phrases. We use the term "structure" to refer to any of these units. The annotated Chaghatay data used for part of speech tagging contains 5 508 structures including 1 244 sentences, 1 348 lines, and 2 916 phrases. In total there are 30 666 words and 8 767 unique tokens.

Data from all available segmentation formats is combined during model training and evaluation for most models. However, in cases where the performance of models trained on different structure formats are compared, sentence data is used for the test set, and a combination of all structure formats are used for training data.

4.1.1 Data Splits for the Chaghatay Corpus

With the minimal amounts of data available, dedicated training and test datasets would provide a narrow view of the performance of our taggers. To make our results more robust, the available tagged data was randomly divided into 5 parts, and 5-fold cross validation was performed. These parts for cross validation are independent, non-stratified random samples across all three structure formats described in Section 2.3.

4.1.2 Creating Coarse Grained POS Tags

We extract coarse-grained POS tags by breaking a complex morphological tag into a series of smaller tags and looking up each of these smaller tags in a table to identify the appropriate coarse tag of the complex tag. First, a complex tag like VT-ANT.DIR-3=CZR for the word *swrdy0ky* would be split on markers for morpheme boundaries and clitic boundaries ('=' and '-') giving the following smaller tags: 'VT', 'ANT.DIR.3', and 'CZR'. These tags are then each looked up in a table of the correspondences between fine-grained parts of speech and coarse ones. This table was created during the creation of the annotation guidelines. A separate list of inflectional tags (like ANT.DIR.3) is also maintained. We then choose the coarse tag corresponding to the first fine-grained segment included in the table. If none of the tag segments is in the correspondence table, and all are listed in the list of inflectional tags, the coarse tag is INFL. If none of the segments are in the fine to coarse correspondence table, and not all the tag segments are listed as inflectional tags, an unknown coarse tag (XXXX) is assigned. This only affects 78 of the 27 782 words in the corpus.

4.1.3 Data for Language Modeling

As Chaghatay is a historical language, the standard method for collecting data for language models, i.e., scraping text from websites, is impossible. However, language model pretraining can still be useful for part of speech tagging in Chaghatay.

Because the overall annotation process in the Chaghatay corpus is quite time consuming, a considerable number of texts have been transliterated but not linguistically annotated yet. 9 518 structures have been transliterated but, as discussed in Section 4.1, roughly half this number of structures have annotations. These 9 518 structures are used to train the simple language models discussed in Section 4.2.4.

4.1.4 The Modern Uyghur Corpus

For the modern Uyghur data, we use the Uyghur Treebank (Eli et al., 2016), which is part of the Universal Dependencies (UD) project (McDonald et al., 2013). This treebank uses Universal POS tags, conforming to the UD annotation standards. The Universal POS tagset is a very coarse tagset consisting of 17 POS tags. The Uyghur Dependency treebank uses only 16 of those.

The Uyghur treebank is substantially larger than the Chaghatay data we are working with. In total, there are 3 459 sentences and 40 236 words. The data is divided into train, development, and test portions by the treebank creators. Only the training portion is used for pretraining our Chaghatay model with modern Uyghur data.

4.2 Models

We compare two different types of common sequence models: Conditional Random Fields (CRF) and Recurrent Neural Network (RNN) taggers. For both models, no special accommodations were made for out-of-vocabulary (OOV) tokens during training. Instead, feature representations derived from the characters in a word were leveraged.

4.2.1 Conditional Random Field Tagger

A Conditional Random Field model (CRF) (Lafferty et al., 2001) is similar to a Hidden Markov Model (the traditional approach for POS tagging), but with a flexible feature model and a discriminative probability model. CRFs have been shown to be well suited for sequence tagging tasks (Gahbiche-Braham et al., 2012; Sun, 2014). We use the CRF implementation by Okazaki (2007).

For our part of speech tagging task, we model the structure of the sentence as a sequence of words. For each word in an input sentence, we extract the following features: 1) The lowercase word, 2) the identity of the first 10 characters of the word as separate features, 3) the identity of the last 10 characters of the word as separate features, 4) the previous word in the sentence, and 5) the next word in the sentence.

All fine-grained CRF taggers were trained using the averaged perceptron training algorithm. For the coarse grained model, the LBFGS training algorithm was used as training time for the coarse tagset was quite short, and the LBFGS algorithm produced slightly better results.

4.2.2 Neural Tagger

Neural networks have been shown to work well for mono-lingual as well as multi-lingual POS tagging (Huang et al., 2015; Shao et al., 2017). Our neural tagger is a relatively simple Gated-Recurrent-Unit (GRU) network. This network consists of a word embeddings layer, a character embedding layer (the final state of a GRU over the characters in a word), a bidirectional GRU with varying numbers of layers, and a final softmax layer. For all experimental settings, the embeddings are updated during the training process; freezing of layers is not performed. In

Classifier	# hidden layers	Accuracy
CRF	n/a	**74.57**
RNN	1	*73.48*
	2	71.57
	3	68.02

Table 1: Accuracy of CRF and RNN using the fine-grained POS tagset.

the default setting, the character embeddings and word embeddings are randomly initialized. In the various transfer learning settings, the entire network, including the character and word embeddings are intialized using the weights learned from the previous task.

4.2.3 Using Coarse Grained POS Tagging

For creating the coarse CRF tagger, we apply jackknifing: We use 5-fold cross validation on the training data, such that a model is trained on 4 of the folds and predicts coarse tags on the remaining fold. This means we have the full dataset automatically POS tagged for coarse parts of speech.

The coarse RNN model is trained for 50 epochs while the fine-grained model is trained for an additional 75 epochs. To transfer from the coarse part of speech tagging model to the fine-grained model, the top softmax layer of the network is removed and replaced with a new softmax layer containing the relevant number of classes (where each potential tag is a class). The new softmax layer is randomly initialized.

4.2.4 Pretraining the RNN

The neural model allows for us to pretrain using a variety of different tasks. In this case, we pretrain the RNN model on language modeling and part of speech tagging for Uyghur. For these additional experiments, we only use the RNN architecture.

The architecture used for language modeling is very similar to the architecture of the part of speech tagger: A word embeddings layer is concatenated with a GRU-based character embeddings layer, this then passes through some number of GRU layers. The only difference is that the language model calculates a softmax over all possible words for both the forward and backward directions where the part of speech tagger had one softmax over the possible part of speech tags. As with pretraining on coarse part of speech tagging, the top part of the network is removed, and the final linear layers of the part of speech tagger are added on and randomly initialized.

For pretraining on modern Uyghur part of speech tagged data, the same design described in Section 4.2.3 is used.

5 Results

5.1 Choice of Classifier

We first look into the performance of the two classifiers, the CRF and the RNN, when performing fine-grained tagging. For the RNN, we experimented with 1, 2, and 3 hidden layers. These results are shown in Table 1, averaged over 5-fold cross-validation. We reach the best results of 74.57% using the CRF model. The best results for the RNN are 1 point lower, at 73.48%.

For the neural network models, the best results are reached with a single hidden layer in the main GRU. This indicates that larger networks are somewhat over-parameterized given the relatively small size of the training corpus. Reducing the number of hidden units for the single layer RNN shows a decrease in performance.

5.2 Utilizing Coarse Tags

For the second experiment we use coarse part of speech tags, either as a first tagging step for the CRF or as pretraining for the RNN (1 hidden layer). The results are shown in Table 2.

Classifier	Setting	Accuracy	In vocab. acc.	OOV acc.
CRF	fine-grained	74.57	83.15	42.91
	plus coarse-grained	74.68	83.05	43.93
RNN	fine-grained	73.48	81.75	41.78
	plus coarse-grained	**76.17**	**83.37**	**48.65**

Table 2: Comparison between the fine-grained only setting and the setting adding coarse-grained POS tagging, reporting overall accuracy, in-vocabulary accuracy and out-of-vocabulary accuracy.

CRF+coarse			CRF		
gold	tagger	#	gold	tagger	#
AJ	N	363	AJ	N	306
N	AJ	216	N	AJ	229
FOR	N	155	FOR	N	139
DEM	PN.DEM	97	Npr	N	93
Npr	N	85	DEM	PN.DEM	92
N	FOR	71	N	FOR	82
N	Npr	75	N	Npr	71
Npr	FOR	62	FOR	Npr	69
PN.DEM	DEM	65	PN.DEM	DEM	67
AJ	AV	44	AJ	AV	54

RNN+coarse			RNN		
gold	tagger	#	gold	tagger	#
AJ	N	222	AJ	N	290
N	AJ	215	N	AJ	276
Npr	N	128	FOR	N	145
DEM	PN.DEM	107	DEM	PN.DEM	115
N	Npr	92	N	FOR	99
FOR	N	87	Npr	N	96
PN.DEM	DEM	83	N+ACC	N-ACC	88
N+ACC	N-ACC	82	N	Npr	80
N	FOR	77	FOR	Npr	76
FOR	Npr	76	AV	AJ	72

Table 3: The 10 most frequent confusions per setting.

This setup provides a negligible increase in performance for the CRF model, from 74.57% to 74.68%. However, for the neural model, pretraining on coarse tagging is very beneficial. This setup increases accuracy from 73.48% to 76.17%, thus also improving over the CRF model.

When we evaluate in-vocabulary and out-of-vocabulary words separately, we see the same trend, the CRF sees a negligible decrease for known words, and it gains about 1% absolute for out-of-vocabulary words. In comparison, the RNN starts with a low accuracy on known words (81.57% versus 83.30 for the CRF) but gains 1.5% on known words and almost 7% on out-of-vocabulary words.

We also had a look at the confusion matrix for these four settings. The 10 most frequent confusions per setting with their frequencies are shown in Table 3. These confusions show that all models have the tendency to label words as noun (N, Npr, N-ACC): 5-6 of the 10 most frequent confusions involve this label. This is likely due to the prevalence of this tag: Over 23% of all words are tagged as nouns, and thus the model has strong tendencies to confuse other tags for nouns.

All models have difficulty distinguishing proper nouns (Npr) from conventional nouns (N),

Segmentation	Size training data (# structures)	Accuracy
Sentence data	800	67.69
Phrase data	800	52.14
Line data	800	53.34

Table 4: CRF model accuracy with training data from single structure types.

	Size training data	Accuracy	
Type of segmentation	(# structures)	CRF	RNN
Sentences	844	67.24	67.69
Sentences+phrases	844	65.31	64.26
Sentences+lines	844	65.73	61.81
Sentences+phrases	2 192	69.35	70.10
Sentences+lines	3 735	70.50	68.96
Sentences+phrases+lines	5 108	**71.90**	71.22

Table 5: Effectiveness of additional data sources.

likely due to the similar syntactic contexts both can be found in. Demonstrative pronouns (PN.DEM) and demonstratives (DEM) are also frequently mistagged by all four model types. The taggers seem to have difficulty identifying Arabic words. FOR, by definition, denotes an unanalyzed string surrounded by whitespace, usually a code-switch into an Arabic phrase.

When we compare the condition using coarse grained POS information to the base conditions directly performing fine grained POS tagging, we see that most of the error types are the same. It is interesting to see that the CRF+coarse model has higher numbers on the most frequent confusions than the base CRF, which implies that the CRF+coarse has fewer error categories overall while for the base CRF, the errors are distributed over more categories.

5.3 Different Structure Formats

Here, we investigate whether it is more important to have sentence segmented data, or if we need more data even if it is segmented differently. Given this question, we restrict our initial training set and the test set to sentence segmented data. We use 400 sentences from the 1 244 sentences as test data, the rest serves as initial training set.

We first look at the quality of the different segmentation styles. I.e., we carry out an experiment in which we train on the one dataset, using a single segmentation. We use the CRF model for this experiment since it showed a higher performance in the setting without coarse grained POS tags. Note that these results cannot be directly compared to the results in Table 2 since we do not use cross-validation here.

Table 4 shows that the quality of the line and phrase structures is not sufficient to substitute the sentence data: Both types of data result in a decrease of accuracy around 15% absolute even though the training set size is 2.5-6 times larger than for the sentence structures. This shows very clearly that the end of sentences marking is important for the POS tagging task.

Next, we look at the effectiveness of adding data from line and phrase structures to the sentence structures for annotating the sentence level test data. I.e., we start with the sentence segmented training set, and then add the line segmented and phrase segmented data. Note that this means that the size of the training set changes across settings. We also created balanced training sets so that the final training set size is the same as that of the sentence data, 844.

The results of adding training data in different segmentations are shown in Table 5. The results show that the additional data sources are beneficial to both the CRF and RNN models when added to the sentence segmented data. The accuracy increases from 67.24% to 71.90% for the CRF and from 67.69% to 71.22% for the RNN. When we compare the setting of the balanced training sets in the upper part of the table to the setting with all additional data, we see that the

	Fine-grained	Transfer from		
# hidden layers	tagger	coarse POS	Chaghatay lg. model	Uyghur
1	73.48	**76.17**	74.73	68.32
2	71.57	71.74	72.55	63.94
3	68.02	70.02	71.09	56.81

Table 6: Accuracy of different neural taggers

balanced cases lead to lower accuracies than using only sentences. The RNN is more susceptible to the difference in segmentation, reaching 64.26% for sentences plus phrases and 61.81% for sentences plus lines, as opposed to 67.69% for sentences only. Both architectures profit from the additional data, and the CRF reaches 71.90% when all available training data are combined. This shows that while the differences in segmentation do influence the POS taggers, having more training data outweighs these differences. We also see that initially, both architectures show a very similar performance, but the CRF model reaches a higher accuracy on the largest dataset, thus showing that it is better suited to using variable training data successfully.

5.4 Pretraining the RNN

In the final question, we investigate whether we can use pretraining via a Chaghatay language model or with modern Uyghur data to alleviate the data sparsity problem. The flexibility of neural networks allows us to use other tasks for network pretraining.

The results of this experiment are shown in Table 6. For ease of comparison, we repeat the results for the initial setting in the fine-grained setting, where we simply train on the fine-grained training set, and for optimized RNN initially trained on coarse-grained Chaghatay POS tags. The results show that the RNN profits from pretraining using a language modeling task with Chaghatay data. For the model with 1 hidden layer, accuracy increases from 73.48% to 74.73%. Pretraining on the Modern Uyghur data, in contrast, results in a considerable drop in performance by more than 5% absolute. This may illustrate the significant lexical, syntactic, and certainly orthographic differences between the two languages. Another reason for the decrease in accuracy may be the differences in the POS tagsets. This seems unlikely since pretraining on coarse-grained POS tags from the Chaghatay corpus has a beneficial effect, resulting in the highest results in our experiments.

Some spelling differences between Modern Uyghur and Chaghatay that likely lead to errors include the following: In Chaghatay, with the exception of (long) alef, vowels are unspecified or represented with consonants; in Modern Uyghur, each vowel has its own glyph. Further, Chaghatay typically lacks punctuation and (as seen above re: FRAG), scribes may insert a line break in the middle of a word or even morpheme. Finally, due to phonological changes such as vowel raising, modern Uyghur spelling often deviates from that of cognate forms in Chaghatay. This suggests that it may be more useful to use Uzbek data (which has less of these kinds of phonological changes) instead of Modern Standard Uyghur[7].

We have also experimented with different numbers of hidden layers, but the same pattern holds regardless of pretraining: The best results are reached with a single hidden layer.

6 Conclusion and Future Work

We have investigated POS tagging for Chaghatay, in a situation where we have a small training set but a large POS tagset. Our results show that without additional pretraining, the Conditional Random Fields tagger performs better than its neural counterpart. By using pretraining on coarse grained POS tags, the neural models are able to surpass the CRF model's performance. Using additional data from a language model or from modern Uyghur did not improve results.

[7] But we are not aware of any POS annotated corpus for Uzbek.

For the future, we are planning to investigate how well the different POS tagging architectures support manual postcorrection. I.e., does a higher overall accuracy also translate into higher manual annotation rates, or are the types of errors, which we have shown to differ between the architectures, are the determining factor? We will also start annotating the corpus for Universal Dependencies (McDonald et al., 2013).

References

Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018. Contextual string embeddings for sequence labeling. In *Proceedings of the 27th International Conference on Computational Linguistics (COLING)*, pages 1638–1649, Santa Fe, NM.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Pascal Denis and Benoit Sagot. 2009. Coupling an annotated corpus and a morphosyntactic lexicon for state-of-the-art POS tagging with less human effort. In *Proceedings of the Pacific Asia Conference on Language, Information and Computation (PACLIC 23)*, Hong Kong, China.

Arienne M Dwyer. 2018. Morphological annotation in the ATMO project. Technical report, University of Kansas. `http://uyghur.ittc.ku.edu/manuals/MorphologicalAnnotation.xhtml`.

Marhaba Eli, Weinila Mushajiang, Tuergen Yibulayin, Kahaerjiang Abiderexiti, and Yan Liu. 2016. Universal dependencies for Uyghur. In *Proceedings of the Third International Workshop on Worldwide Language Service Infrastructure and Second Workshop on Open Infrastructures and Analysis Frameworks for Human Language Technologies (WLSI/OIAF4HLT2016)*, pages 44–50, Osaka, Japan.

Souhir Gahbiche-Braham, Hélène Bonneau-Maynard, Thomas Lavergne, and François Yvon. 2012. Joint segmentation and POS tagging for Arabic using a CRF-based classifier. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC)*, pages 2107–2113, Istanbul, Turkey.

Tobias Horsmann and Torsten Zesch. 2017. Do LSTMs really work so well for PoS tagging? – A replication study. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 727–736, Copenhagen, Denmark.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional LSTM-CRF models for sequence tagging. Technical Report arXiv:1508.01991 arXiv:1508.01991, arXiv.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning (ICML)*, pages 282–289, San Francisco, CA.

Ryan McDonald, Joakim Nivre, Yvonne Quirmbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith Hall, Slav Petrov, Hao Zhang, Oscar Täckström, Claudia Bedini, Núria Bertomeu Castelló, and Jungmee Lee. 2013. Universal dependency annotation for multilingual parsing. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 92–97, Sofia, Bulgaria.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *Proceedings of the International Conference on Learning Representations (ICLR)*, Scottsdale, AZ.

Naoaki Okazaki. 2007. CRFsuite: a fast implementation of Conditional Random Fields (CRFs). `http://www.chokkan.org/software/crfsuite/`.

Pedro Javier Ortiz Suárez, Laurent Romary, and Benoît Sagot. 2020. A monolingual approach to contextualized word embeddings for mid-resource languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1703–1714, Online.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2227–2237, New Orleans, LA.

Yan Shao, Christian Hardmeier, Jörg Tiedemann, and Joakim Nivre. 2017. Character-based joint segmentation and POS tagging for Chinese using bidirectional RNN-CRF. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (IJCNLP)*, pages 173–183, Taipei, Taiwan.

Xu Sun. 2014. Structure regularization for structured prediction. In *Proceedings of Neural Information Processing Systems (NIPS)*, pages 2402–2410, Montreal, Canada.