A Neural Network Approach to Morphological Disambiguation Based on the LSTM Architecture in the National Corpus of the Tatar Language

Rinat Gilmullin[1], Bulat Khakimov[1,2], and Ramil Gataullin[1]

[1] Institute of Applied Semiotics of the Tatarstan Academy of Sciences, Kazan, Russia
[2] Kazan Federal University, Kazan, Russia
rinatgilmullin@gmail.com

Abstract. This paper presents the results of experiments on morphological disambiguation in the National corpus of the Tatar language "Tugan tel". The experiments were conducted using the LSTM based neural network model. The tagged socio-political sub-corpus of the National corpus of the Tatar language "Tugan tel" with a volume of 2,4 million words was used as training data. Experiments have shown that LSTM models are language-independent and can be applied to the Tatar language too. The results for Tatar are on a comparable level with those for other agglutinative languages, such as Hungarian and Turkish.

Keywords. Morphological disambiguation, Tatar language, Tatar National Corpus, corpus data, morphological tagging, LSTM, neural architectures

1 Introduction

Morphological disambiguation is one of the main tasks of automatic natural language processing. Its results can be used to improve accuracy and quality of the methods used in such tasks as text classification and clustering, machine translation, and information retrieval.

The complexity and peculiarities of morphological disambiguation vary for each particular language. For example, for English with its poor morphology and rigid word order in the sentence, the morphological disambiguation, as a rule, is reduced to the task of POS tagging and is based on rather simple methods. In Russian, morphological ambiguity is not so salient as in English, but, nevertheless, it is inherent. Free word order in Russian adds complexity to the task. In the Tatar language, as in other agglutinative languages of the Turkic group, morphemes are the most important meaningful language units that carry both semantic and syntactic information. With a theoretically unlimited number of morphemes attached to the stem, morphological ambiguity takes on various forms, which greatly complicates the disambiguation.

Up to now, a basic paradigm of methods for disambiguation has been formed [1]. This includes the rule-based methods [2,3], machine learning methods based on the probabilistic models [4,5], and hybrid methods [6,7,8]. Developing the National corpus of the Tatar language "Tugan tel" (http://tugantel.tatar/) and the socio-political sub-corpus with manual morphological disambiguation made it possible to study this problem using statistical methods based on machine learning [3,8].

Analysis of open source codes developed for this task over the past few years has shown that one of the most effective tools is PurePos 2.0 [6] which implements a hybrid model based on hidden Markov models, as well as a neural network model based on recurrent neural networks with long short-term memory LSTM [5]. Hidden Markov model is a process model in which a process is considered a Markov process, and it is not known what state the system is in (its states are hidden), but each state can produce, with some probability, an event that can be observed. In other words, the Markov process with unknown parameters is studied, and the task is to recognize these unknown parameters basing on observables. The results of recognizing POS tags of Tatar words showed an accuracy of 97% [8].

Another approach that rather successfully solves the problem of morphological ambiguity is based on a recurrent neural network with a long short-term memory (LSTM) [9,10]. In [5], the results (see Table 1) of applying this approach to Turkish, Russian and Arabic are given.

Table 1. Results of experiments using the LSTM neural network architecture for morphological disambiguation.

Language	**Turkish**		**Russian**		**Arabic**	
	% from ambiguous words	% from all tokens	% from ambiguous words	% from all tokens	% from ambiguous words	% from all tokens
Without context (baseline)	88.65	95.45	64.97	88.58	72.22	78.06
Local context	89.18	95.67	71.56	90.72	80.10	84.29
Whole sentence (surface form)	91.03	96.41	69.49	90.05	86.45	88.95
Left-to-Right	90.50	96.19	68.55	89.75	89.30	91.27
CRF	90.24	96.09	72.78	91.13	-	-

The analysis of the used context size in [5] deserves a special attention. The authors compared different sizes and types of contexts and experimentally revealed the most appropriate type for each language. It turned out that for the Turkish language it is sufficient to construct vectors based on surface word forms without explicitly defining their morphological features, but using all the words in the sentence. Whereas for Russian, agreement in gender, number and case is important, which in turn requires not only surface word forms, but also their morphological features in the context. At the same time, optimization based on the conditional random field method (CRF) helps to achieve better results (disambiguation accuracy 91.13%). The situation is similar with the Arabic language, when surface word forms are not enough for full disambiguation. This can be explained by the fact that in Arabic the level of ambiguity is higher than Turkish. If, for example, in Turkish, on average, there are 2.81 parsing options per word, and in Russian 5.81, then in Arabic there are 11.31. Therefore, for correct model training, a completely disambiguated tagged context is required.

This article describes the results of applying the neural network model based on the LSTM architecture to morphological disambiguation in the National corpus of the Tatar language.

2 The Tatar Language

The Tatar language belongs to the Turkic group that forms a subfamily of Altaic languages. It is spoken in West-central Russia (in the Volga region) and in the southern parts of Siberia. The number of Tatars in Russia in 2010 was 5,31 million people [9]. In 2013, the existing language classifications [12, 13] described Tatar as an under-resourced language.

3 LSTM model for morphological disambiguation

Model training requires tagged disambiguated texts. The method supposes that each parse of an ambiguous word and its context is juxtaposed with vectors. In the first case, the vector is based on its lemma and morphological features, and in the second case, on the surface forms of the surrounding words; in addition, the vector can be expanded by morphological features. Here, the context is not limited to several words of the immediate vicinity of words and can reach the size of the entire sentence. After that, on the basis of the resulting pair of vectors, the distribution of conditional probabilities is constructed; from these the most probable parse is selected as the correct one.

According to [5], the LSTM model is designed to build a vector representation of an ambiguous word (vectors are constructed on the basis of the lemma and morphological features of each of the alternatives, then they are united into R matrix and the surrounding context (indicated by h vector). After using the *softmax* function on the product of R matrix and h vector, the distribution of probabilities of each parsing option in this particular context is constructed, on the basis of which morphological ambiguity is resolved in favor of the most likely alternative:

$$p(y_t = a|x) = softmax(R_{x_t}\, vo\, h_t)$$

3.1 Vector representation of the ambiguous word and its context

Let us take an ambiguous word with the following morphological parsing:

$$stem_i + tag_{i,1} + tag_{i,2} + \cdots + tag_{i,L}$$

where $stem_i = (stem_{i,1}, stem_{i,2}, \ldots, stem_{i,K})$, a lemma K symbols long of the ith parsing option; each $tag_{i,j}$ is the jth tag (morphological feature) of the ith parsing option (which contains L of such tags). To construct the vector of the lemma, a bidirectional LSTM is used on top of each symbol of the lemma; for the vector of morphological features, we use a bidirectional LSTM over the tags. First, the bidirectional LSTM creates g_x representation of the input vector $x = (x_1, x_2, \dots, x_T)$ by computing the direct $\vec{g}$ and the inverse $\overleftarrow{g}$ sequence, and combines the two sequences using the Rectified Linear Unit (ReLU).

$$\vec{g}_t = f(x_t, \vec{g}_{t-1})$$
$$\overleftarrow{g}_t = f(x_t, \overleftarrow{g}_{t+1})$$
$$g = ReLU(\vec{g}_T, \overleftarrow{g}_0),$$

where f(x, y) is a LSTM function with input values x and y.

Thus, the corresponding vector representations are constructed separately for the lemma and for the tag sequence (morphological features). Next, the resulting vectors are combined using the hyperbolic tangent:

$$r_i = \tanh(g_{stem_i} + g_{tag_i})$$

Next, $r_i$vectors are combined into R matrix, where each row belongs to a particular parse.

$$R = [r_1, r_2, \dots, r_N]$$

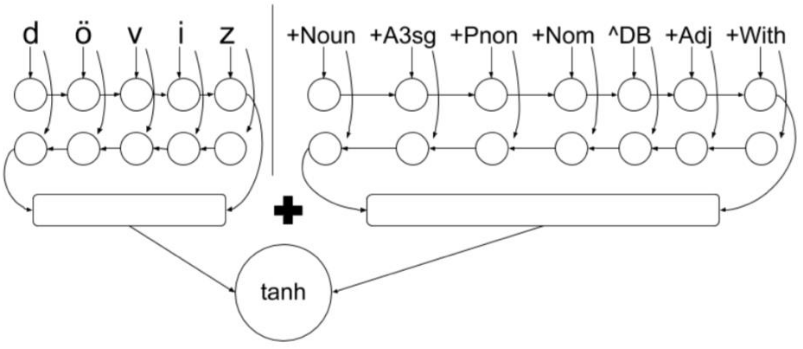

Fig. 1. LSTM neural network architecture for obtaining a vector representation of the morphological parse.

One of the methods for constructing the context vector described in [5] is to use only the surface forms of the surrounding words (without morphological features). For this, the bidirectional LSTM model is used over each x_i word, constructing a separate vector for each word. Then for the left context, the vectors are assembled from right to left, and for the right context – from left to right (see Fig. 2.). After that, the vectors are combined using the hyperbolic tangent:

$$\vec{c}_t = f(x_t, \vec{c}_{t-1})$$
$$\overleftarrow{c}_t = f(x_t, \overleftarrow{c}_{t+1})$$
$$h_t = \tanh(\vec{c}_t, + \overleftarrow{c}_t]$$

Next, in order to perform the morphological disambiguation, the distribution of alternative probabilities is constructed – for this, *softmax* function from the product of $\mathrm{h_t}$ vector and R matrix is taken, and the most probable parse is selected as the correct one (according to the same formula as described in the previous section):

$$p(y_t \; = \; a|x) \; = \; softmax(R_{x_t}\, vo\; h_t)$$

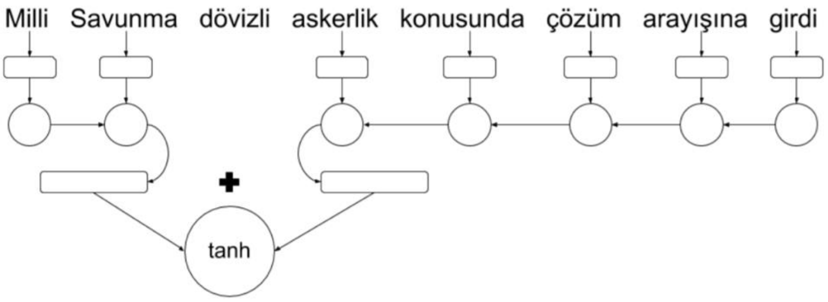

Fig. 2. Neural network architecture for obtaining a context vector.

Sometimes, surface forms of the surrounding words in the context are not enough for morphological disambiguation. Apart from these, it is necessary that all ambiguities in the surrounding words are resolved, i.e. data on the lemma and on all morphological features corresponding to the given context are needed. In such cases, the remedy is sequential disambiguation, when information about the allowed option is transmitted further, and the next case of ambiguity is resolved on its basis (in [5], this approach is defined as Left-to-Right).

In such cases, the LSTM model builds a vector based on the lemma and morphological features of the word from the context (if they are ambiguous, then the one in favor of which the disambiguation was made is selected) and thus m_t vector is calculated and then the disambiguation is performed:

$$m_t = f(r_i^t, m_{t-1})$$

where r_i^t is a vector fromR_{x_t}, the parsing option selected at the previous disambiguation stage.

$$r_i^{t+1} = \tanh\left(g_{stem_i}^{t+1} + g_{tag_i}^{t+1} + m_t\right)$$
$$R_{x_{t+1}} = [r_1^{t+1}, r_2^{t+1}, \dots, r_N^{t+1}]$$

$$p(y_t = a|x, y_1, y_2, \dots, y_{t-1}) = softmax(R_{x_t} \times h_t)$$
$$\hat{y} = argmax_{\tilde{y} \in Y_X} \prod_{i=1}^{T} p(\tilde{y}_t|x, \tilde{y}_1, \tilde{y}_{2\dots\dots}\tilde{y}_{t-1})$$

4 Data preparation

At the initial stage of work, statistical data on the frequency of word forms with multivariate parses, presented in Table 2, were obtained from the text base of the National corpus of the Tatar language “Tugan tel” [3]. The morphological module implemented on the basis of the HFST toolkit is used for the morphological tagging of the corpus. [14].

Table 2. Distribution of morphological parsing options.

Parsing options	Number	Share in the corpus
Total number of word forms with multivariate parses	5.650.820	25,75%
2 parses	4.282.108	19,51%
3 pares	1.045.392	4,76%
4 parses	296.547	1,35%
5 and more parses	26.773	0,12%
Total in the sample	21.940.452	100%

The total volume of the corpus at this stage was 21.940.452 tokens; the share of tokens with multivariate parses was 25.75%.

At the same time, the maximum length of the word form presented in the corpus consists of the stem and twelve grammatical affixes.

To carry out experiments with model training, it was necessary to have a morphologically disambiguated corpus. The part of socio-political sub-corpus of the National corpus of the Tatar language “Tugan tel” was used as training data. The sub-corpus statistics are given in Table 3.

Table 3. Statistics of the training and test samples on the socio-political corpus.

	Training sample	Test sample
Number of contexts (sentences)	54.580	944
Number of tokens (including punctuation)	600.480	11.655
Number of multivariate parses	125.480 (21%)	2.527 (21%)
Number of unique word forms	29.953	2.788
Number of unique lemmas	7.117	1.226

Number of unique morphological forms	1.898	346

Manual morphological disambiguation of the socio-political sub-corpus was carried out by experts using a Web-based toolkit for morphological disambiguation in the corpus of the Tatar language [15].

Manual morphological disambiguation was organized in several stages.

At stage 1 selected texts from socio-political sub-corpus were automatically tagged using the morphological analyze. Then certain types of ambiguity were automatically disambiguated where possible, as well as redundant and incorrect parses were removed.

At stage 2 annotators performed manual disambiguation using web-toolkit for morphological disambiguation in dialog mode. They selected the right parsing option based on the context.

At stage 3 main experts performed total manual review of the tagged texts disambiguated at stage 2. This double-checking helped us make sure that the tagging and disambiguation of the training data is correct.

As a result, for our experiments 56.524 morphologically disambiguated sentences were prepared.

5 Experiments and Evaluation

As one can see from Table 2, the tagged data sample was divided into a training sample and a test sample. LSTM models were trained only using the training set, and the test sample was used just for testing. Based on approach described in [5], we considered each sentence to be a minibatch for training. The objective function used for training was the total cross-entropy loss between the selected parse and the correct parse for every token in the sentence. Stochastic gradient descent and backpropagation were used to adjust the parameters for our model. All LSTMs in our models were trained with a single hidden layer. We used a hidden dimension size of 100 for the tag, stem, and surface form LSTMs and 200 for the context and previous parse LSTMs.

Tables 4, 5 provide an estimate of the accuracy of several indicators: lemma recognition, morpheme sequence recognition and disambiguation.

Table 4. Indicators of accuracy of recognition of lemmas and morpheme sequences.

Indicators	**LSTM NN**
Lemma Recognition Accuracy	11299 / 11655 = 96.94%
Morpheme Sequence Recognition Accuracy	11127 / 11655 = 95.46%

Table 5 shows how the algorithm processes the different types of ambiguity according to the number of parsing options. As expected, the best result is for words with only two parsing options: 84.61%, when overall accuracy is 79.10%. In one hand, more variants increase complexity, in another hand, such words (see Table 2) do not have enough examples, so as a result, model lacks accuracy with them.

Table 5. The number of morphological parsing options and accuracy of disambiguation.

Number of options	**LSTM NN**
n=2	1545 / 1826 = 84.61 %
n=3	268 / 424 = 63.21 %
n=4	141 / 192 = 73.44 %
n=5	7 / 9 = 77.78 %
n=6	37 / 72 = 51.39 %
n=7	0 / 2 = 0.00 %
n=8	0 / 1 = 0.00 %
Total	1999 / 2527 = 79.10%

The results of LSTM are virtually close to those of other disambiguation methods. The main benefit of the proposed method is that the model can be trained taking into account the size and peculiarities of the context. So the highest accuracy rate of the morphological disambiguation in the corpus of the Tatar language was achieved with the construction of vectors taking into account all the words in the sentence as the surrounding context. In addition, the vector of the surrounding context was expanded using morphological features.

6 Conclusions

This paper presents the results of work on morphological disambiguation of the Tatar language using the neural network model based on the LSTM architecture. Given the limited set of corpus data for training, the results of experiments showed a fairly good level of accuracy for morphological disambiguation, 79.10%. We believe that the lower accuracy of the neural network model is primarily related to the amount of training data, since systems with neural networks are not sufficiently effective when training on a limited set of data.

At the same time, the obtained results can be effectively used in creating a morphologically disambiguated “golden” sub-corpus, significantly reducing the number of multivariate parses requiring manual morphological disambiguation.

References

1. Gataullin, R.R.: Analiticheskij obzor metodov razresheniya morfologicheskoj mnogoznachnosti [Analytical review of the methods of morphological disambiguation]. Rossijskij nauchnyj elektronnyj zhurnal (Elektronnye biblioteki) 19(2), 98-114 (2016).
2. Zin'kina, Yu.V., Pyatkin, N.V., Nevzorova, O.A.: Razreshenie funkcional'noj omonimii v russkom yazyke na osnove kontekstnyh pravil [Context rule based functional disambigua-

tion in Russian]. In: Proceedings of the International Conference “Dialog-2005”, pp. 198–202. Nauka, Moscow (2005).

3. Gataullin, R., Khakimov, B., Suleymanov, D., Gilmullin, R.: Context-Based Rules for Grammatical Disambiguation in the Tatar Language. In: Nguen, N.T. et al. (eds). ICCCI 2017, Part II, LNAI, vol.10449, pp. 529-537 (2017).
4. Sak, H., Gongur, T., Saraclar. M.: Morphological disambiguation of Turkish text with perceptron algorithm. In: Computational Linguistics and Intelligent Text Processing, 8th International Conference CICLing, Mexico City, Mexico, February 2007, pp. 107–118. Mexico (2007).
5. Qinlan Shen, Daniel Clothiaux, Emily Tagtow, Patrick Littell, Chris Dyer.: The Role of Context in Neural Morphological Disambiguation. In: Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, pp. 181–191. Osaka (2016) http://aclweb.org/anthology/C16-1018, last accessed 2018/10/25.
6. Orosz, G. and Novák, A.: PurePos 2.0: a hybrid tool for morphological disambiguation. In: Proceedings of Recent Advances in Natural Language Processing, pp. 539–545. (2013) http://aclweb.org/anthology//R/R13/R13-1071.pdf, last accessed 2018/10/25.
7. Yuret, D., Ture, F.: Learning morphological disambiguation rules for Turkish. In: Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL, pp. 328–334. New York (2006).
8. Gilmullin, R.A., Gataullin, R.R.: Razreshenie morfologicheskoj mnogoznachnosti tekstov na tatarskom yazyke na osnove instrumentariya PurePos [Morphological disambiguation of Tatar texts using PurePos]. In: Proceedings of the V International Conference on Turkic Languages Processing «TurkLang 2017», pp. 30-37. Fen, Kazan (2017).
9. Berment, V.: Me´thodes pour informatiser des langues et des groups de langues peu dote´es, Ph.D. Thesis, J. Fourier University, Grenoble (2004).
10. Krauwer, S.: The basic language resource kit (BLARK) as the first milestone for the language resources roadmap. In: Proceedings of International Workshop on Speech and Computers - SPEECOM, pp. 8–15. Moscow (2003).
11. Hochreiter, S.: Long short-term memory. Neural Computation 9(8), 1735–1780 (1997).
12. Gers, F.A.: Learning to Forget: Continual Prediction with LSTM. Neural Computation 12(10), 2451–2471 (2000).
13. Lewis, M. Paul, Gary F. Simons, and Charles D. Fennig (eds.): Ethnologue: Languages of the World, http://www.ethnologue.com last accessed 2018/10/25.
14. Gilmullin, R., Gataullin, R.: Morphological Analysis System of the Tatar Language. In: Nguen, N.T. et al. (eds). ICCCI 2017, Part II, LNAI, vol.10449, pp. 519-528 (2017).
15. Gataullin, R.R.: Web-instrumentarij dlya snyatiya morfologicheskoj mnogoznachnosti v tekstovom korpuse tatarskogo yazyka [Web-based toolkit for morphological disambiguation in the corpus of Tatar texts]. In: Proceedings of the V International Conference on Preservation and Development of Native Languages in a Multinational State, pp. 71-73. Otechestvo, Kazan (2014).