Part-of-speech Tagging with Rich Language Description

Daniil Anastasyev, Andrew Andrianov, Eugene Indenbom

ABBYY, Moscow

Dialogue
23rd International Conference on Computational Linguistics
Moscow, RSUH, 1st June, 2017

Outline

1. Problem overview
2. Method description
   - Morphological model
   - Features
   - Target classes
   - Neural network
3. Evaluation
   - Train data
   - Evaluation results
   - Example of work
   - Errors analysis
4. Summary

- The morphological analysis is key step in many NLP pipelines.

- The morphological analysis is key step in many NLP pipelines.
- The results of morphological analysis are used in syntactic and semantic parsing in ABBYY Compreno.

- The morphological analysis is key step in many NLP pipelines.
- The results of morphological analysis are used in syntactic and semantic parsing in ABBYY Compreno.
- Accurate morphological analyser can highly increase speed of the syntactic parsing by reducing the number of obtained hypotheses.

Broad outline

The main features of our model:

1. Vast morphological description.
   - Aims to reduce the set of possible grammatical values.

Broad outline

The main features of our model:

1. Vast morphological description.
   - Aims to reduce the set of possible grammatical values.
2. LSTM neural network.
   - Aims to construct the best grammatical values.

Russian morphological description

- Words are combined into paradigms.

Russian morphological description

- Words are combined into paradigms.
- Each paradigm provides information about grammatical value.

Possible grammatical values of the word «мыла»

1. *мыть* **Par1** VERB Tense=Past Number=Sing Gender=Fem Voice=Act VerbForm=Fin Mood=Ind (freq=$2.03 \cdot 10^{-7}$)
2. *мыло* **Par2** NOUN Animacy=Inan Case=Gen Gender=Neut Number=Sing (freq=$7.57 \cdot 10^{-7}$)
3. *мыло* **Par2** NOUN Animacy=Inan Case=Nom Gender=Neut Number=Plur (freq=$3.81 \cdot 10^{-7}$)
4. *мыло* **Par2** NOUN Animacy=Inan Case=Acc Gender=Neut Number=Plur (freq=$3.48 \cdot 10^{-7}$)

Russian morphological description

- Words are combined into paradigms.
- Each paradigm provides information about grammatical value.
- The provided analysis is ambiguous.

Possible grammatical values of the word «мыла»

1. *мыть* **Par1** VERB Tense=Past Number=Sing Gender=Fem Voice=Act VerbForm=Fin Mood=Ind (freq=$2.03 \cdot 10^{-7}$)
2. *мыло* **Par2** NOUN Animacy=Inan Case=Gen Gender=Neut Number=Sing (freq=$7.57 \cdot 10^{-7}$)
3. *мыло* **Par2** NOUN Animacy=Inan Case=Nom Gender=Neut Number=Plur (freq=$3.81 \cdot 10^{-7}$)
4. *мыло* **Par2** NOUN Animacy=Inan Case=Acc Gender=Neut Number=Plur (freq=$3.48 \cdot 10^{-7}$)

Russian morphological description

- Words are combined into paradigms.
- Each paradigm provides information about grammatical value.
- The provided analysis is ambiguous.
- The description's tagset is taken from ABBYY Compreno.
  - It differs considerably from Universal Dependencies.

Possible grammatical values of the word «мыла»

1. *мыть* **Par1** VERB Tense=Past Number=Sing Gender=Fem Voice=Act VerbForm=Fin Mood=Ind (freq=$2.03 \cdot 10^{-7}$)
2. *мыло* **Par2** NOUN Animacy=Inan Case=Gen Gender=Neut Number=Sing (freq=$7.57 \cdot 10^{-7}$)
3. *мыло* **Par2** NOUN Animacy=Inan Case=Nom Gender=Neut Number=Plur (freq=$3.81 \cdot 10^{-7}$)
4. *мыло* **Par2** NOUN Animacy=Inan Case=Acc Gender=Neut Number=Plur (freq=$3.48 \cdot 10^{-7}$)

Unknown words processing

- Texts usually contains large number of out-of-vocabulary words:
  rare words («Анастасьев»), neologisms («заддосить»), misspellings («мыпа»).

Unknown words processing

- Texts usually contains large number of out-of-vocabulary words:
  rare words («Анастасьев»), neologisms («заддосить»), misspellings («мыпа»).
- The analysis of unknown words is performed as follows:
  1. Possible endings are determined:
     - заддосить (empty, like «резюме»)
     - заддосить (like «февраль»)
     - заддосить (like «закосить»)

Unknown words processing

- Texts usually contains large number of out-of-vocabulary words:
  rare words («Анастасьев»), neologisms («заддосить»), misspellings («мыпа»).
- The analysis of unknown words is performed as follows:
  1. Possible endings are determined:
     - заддосить (empty, like «резюме»)
     - заддосить (like «февраль»)
     - заддосить (like «закосить»)
  2. Paradigms occurred with such endings are collected.

Unknown words processing

- Texts usually contains large number of out-of-vocabulary words:
  rare words («Анастасьев»), neologisms («заддосить»), misspellings («мыпа»).
- The analysis of unknown words is performed as follows:
  1. Possible endings are determined:
     - заддосить (empty, like «резюме»)
     - заддосить (like «февраль»)
     - заддосить (like «закосить»)
  2. Paradigms occurred with such endings are collected.
  3. Obtained analyses are sorted by statistics of suffixes of known words in the paradigms.
     - Unknown word has suffix similar to some suffix of known word $\rightarrow$ it's likely that they share same paradigm.
     - E.g., «ддосить» and «гундосить».
     - $Q(form) = \mathbf{P}(paradigm(form), suffix(form))$.

Unknown words processing

- Texts usually contains large number of out-of-vocabulary words:
  rare words («Анастасьев»), neologisms («заддосить»), misspellings («мыпа»).
- The analysis of unknown words is performed as follows:
  1. Possible endings are determined:
     - заддосить (empty, like «резюме»)
     - заддосить (like «февраль»)
     - заддосить (like «закосить»)
  2. Paradigms occurred with such endings are collected.
  3. Obtained analyses are sorted by statistics of suffixes of known words in the paradigms.
     - Unknown word has suffix similar to some suffix of known word $\rightarrow$ it's likely that they share same paradigm.
     - E.g., «ддосить» and «гундосить».
     - $Q(form) = \mathbf{P}(paradigm(form), suffix(form))$.
- The found score is treated in the same way as probability of known word.

Choice of classifier

Have to take into account context of the analysed word:

1. Usage of context features.
    - Features from 2-3 words in left and right contexts of the analysed word

Choice of classifier

Have to take into account context of the analysed word:

1. Usage of context features.
   - Features from 2-3 words in left and right contexts of the analysed word
   - Linear classifiers (Logistic regression or Linear SVM) or gradient boosting performs good.

Choice of classifier

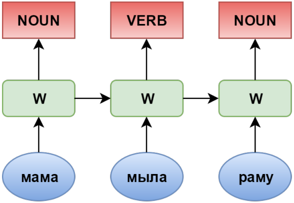
Have to take into account context of the analysed word:

1. Usage of context features.
   - Features from 2-3 words in left and right contexts of the analysed word
   - Linear classifiers (Logistic regression or Linear SVM) or gradient boosting performs good.
   - Sometimes such context is not sufficient.

Choice of classifier

Have to take into account context of the analysed word:

1. Usage of context features.
   - Features from 2-3 words in left and right contexts of the analysed word
   - Linear classifiers (Logistic regression or Linear SVM) or gradient boosting performs good.
   - Sometimes such context is not sufficient.
2. Usage of recurrent neural networks.



Choice of classifier

Have to take into account context of the analysed word:

1. Usage of context features.
   - Features from 2-3 words in left and right contexts of the analysed word
   - Linear classifiers (Logistic regression or Linear SVM) or gradient boosting performs good.
   - Sometimes such context is not sufficient.
2. Usage of recurrent neural networks.
   - Words are fed one by one.

Choice of classifier

Have to take into account context of the analysed word:

1. Usage of context features.
    - Features from 2-3 words in left and right contexts of the analysed word
    - Linear classifiers (Logistic regression or Linear SVM) or gradient boosting performs good.
    - Sometimes such context is not sufficient.
2. Usage of recurrent neural networks.
    - Words are fed one by one.
    - Bidirectional recurrent layer gives information from both left and right contexts.

Choice of classifier

Have to take into account context of the analysed word:

1. Usage of context features.
    - Features from 2-3 words in left and right contexts of the analysed word
    - Linear classifiers (Logistic regression or Linear SVM) or gradient boosting performs good.
    - Sometimes such context is not sufficient.
2. Usage of recurrent neural networks.
    - Words are fed one by one.
    - Bidirectional recurrent layer gives information from both left and right contexts.
    - Usually LSTM or GRU are used.

Choice of classifier

Have to take into account context of the analysed word:

1. Usage of context features.
   - Features from 2-3 words in left and right contexts of the analysed word
   - Linear classifiers (Logistic regression or Linear SVM) or gradient boosting performs good.
   - Sometimes such context is not sufficient.
2. Usage of recurrent neural networks.
   - Words are fed one by one.
   - Bidirectional recurrent layer gives information from both left and right contexts.
   - Usually LSTM or GRU are used.
   - Can handle longer dependencies.

Used features

- Grammatical value.

Feature description

- Probabilities of each possible grammeme by dictionary.
- Мыла:
  - Freq of Genitive form – $7.57 \cdot 10^{-7}$.
  - Freq of Nominative form – $3.81 \cdot 10^{-7}$.
  - Freq of Accusative form – $3.48 \cdot 10^{-7}$.
- $\mathbf{P}(\text{Case}=\text{Acc}) = \frac{3.48 \cdot 10^{-7}}{(7.57+3.81+3.48) \cdot 10^{-7}} \approx 0.234$.

Used features

- Grammatical value.
- Ambiguity classes' probabilities.

Feature description

- Probabilities of each possible grammatical values.

Used features

- Grammatical value.
- Ambiguity classes' probabilities.
- Punctuation.

Feature description

- Presence of particular punctuation mark at the left or right.
- Binary feature.

Used features

- Grammatical value.
- Ambiguity classes' probabilities.
- Punctuation.
- Word's case type.

Feature description

- Proper, lower, UPPER or FiXed capitalisation.
- Binary feature.

Used features

- Grammatical value.
- Ambiguity classes' probabilities.
- Punctuation.
- Word's case type.
- Suffixes.

Feature description

- Last 1-3 letters of the word:
  - -ыла, -ла, -а.
- Used separate embedding layer for each suffix length.

Used features

- Grammatical value.
- Ambiguity classes' probabilities.
- Punctuation.
- Word's case type.
- Suffixes.
- Word embeddings.

Feature description

Can be initialised:

- Uniformly, for the first ≤ 20000 most frequent words.
- By pretrained embeddings.
  - Led to larger model and did not give any significant improvement.

Target classes

Two ways to design the classification task:

1. Multiclass classification between all possible grammatical values.
   - E.g., «NOUN Animacy=Inan Case=Nom Gender=Neut Number=Sing» refers to one of the classes.

Target classes

Two ways to design the classification task:

1. Multiclass classification between all possible grammatical values.
    - E.g., «NOUN Animacy=Inan Case=Nom Gender=Neut Number=Sing» refers to one of the classes.
    - Leads to large number of classes.

Target classes

Two ways to design the classification task:

1. Multiclass classification between all possible grammatical values.
    - E.g., «NOUN Animacy=Inan Case=Nom Gender=Neut Number=Sing» refers to one of the classes.
    - Leads to large number of classes.
    - Some classes are underrepresented in the train set.

Target classes

Two ways to design the classification task:

1. Multiclass classification between all possible grammatical values.
   - E.g., «NOUN Animacy=Inan Case=Nom Gender=Neut Number=Sing» refers to one of the classes.
   - Leads to large number of classes.
   - Some classes are underrepresented in the train set.
2. Multiclass classification inside all of the grammatical categories.
   - E.g., classification between three possible values of category «Number»: Singular, Plural, Not defined.

Target classes

Two ways to design the classification task:

1. Multiclass classification between all possible grammatical values.
   - E.g., «NOUN Animacy=Inan Case=Nom Gender=Neut Number=Sing» refers to one of the classes.
   - Leads to large number of classes.
   - Some classes are underrepresented in the train set.
2. Multiclass classification inside all of the grammatical categories.
   - E.g., classification between three possible values of category «Number»: Singular, Plural, Not defined.
   - Leads to fewer number of classes.

Target classes

Two ways to design the classification task:

1. Multiclass classification between all possible grammatical values.
   - E.g., «NOUN Animacy=Inan Case=Nom Gender=Neut Number=Sing» refers to one of the classes.
   - Leads to large number of classes.
   - Some classes are underrepresented in the train set.
2. Multiclass classification inside all of the grammatical categories.
   - E.g., classification between three possible values of category «Number»: Singular, Plural, Not defined.
   - Leads to fewer number of classes.
   - Did not enhance quality of the model.

Proposed neural network architecture

- Input layers:

Proposed neural network architecture

- Input layers:
  - Grammatical value representation with punctuation and orthographic features.

Proposed neural network architecture

- Input layers:
  - Grammatical value representation with punctuation and orthographic features.
  - Word embeddings input.

Proposed neural network architecture

- Input layers:
  - Grammatical value representation with punctuation and orthographic features.
  - Word embeddings input.
  - Suffixes embeddings inputs.

Proposed neural network architecture

- Input layers:
  - Grammatical value representation with punctuation and orthographic features.
  - Word embeddings input.
  - Suffixes embeddings inputs.
- One or two Bidirectional LSTM layers.

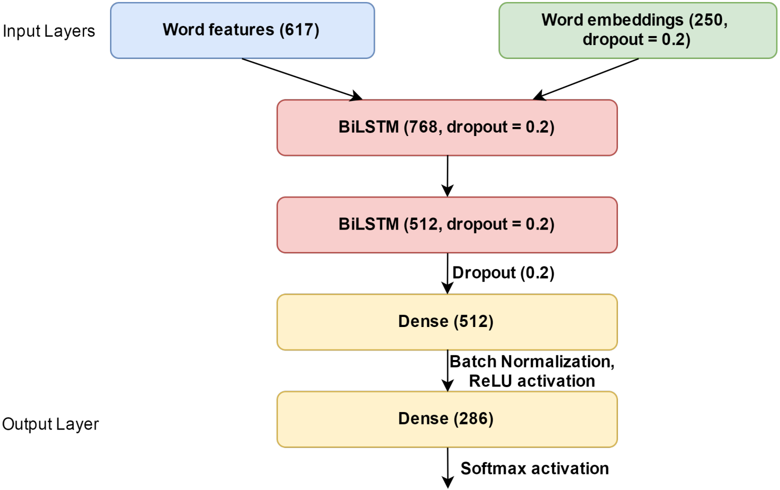
Proposed neural network architecture

- Input layers:
  - Grammatical value representation with punctuation and orthographic features.
  - Word embeddings input.
  - Suffixes embeddings inputs.
- One or two Bidirectional LSTM layers.
- Dense layer with Dropout and ReLU activation.

Proposed neural network architecture

- Input layers:
  - Grammatical value representation with punctuation and orthographic features.
  - Word embeddings input.
  - Suffixes embeddings inputs.
- One or two Bidirectional LSTM layers.
- Dense layer with Dropout and ReLU activation.
- Dense output layer with Softmax activation.

Structure of the neural network



Train data

- GICR subcorpus.
  - Contains about 1 million words.
  - Uses Universal Dependencies tagset and conventions that correspond to test set.
- Wikipedia subcorpus.
  - Contains more than 3 million words.
  - Uses Compreno tagset.
- Corpus of novels.
  - Contains about 30 million words.
  - Uses Compreno tagset.

Tagsets conversion

- The annotation format of Wikipedia and Novels corpora was automatically converted.
- The achieved annotation had number of differences from the competition format.
- This differences were smoothed out by proper training of the neural network.

Fine tuning of the model

- Model gains from usage of large corpus on pretrain stage.
- Model requires training on corpus with more appropriate annotation.

Performance of the model on validation set

Train corpus	**Accuracy on validation set**
GICR	96.41%
Novels	95.36%
Pretrained on Novels, Trained on GICR	97.78%

Evaluation results

- Model achieved following results on the test set.
- The last model was pretrained on Novels corpus, others were trained on GICR subcorpus only.

Results with different parameters

Model	Fiction	News	Social media
1BiLSTM(768) + 0.2 Dropout	94.95% / 69.54%	97.01% / 75.70%	94.30% / 71.30%
1BiLSTM(768) + Dense(768) + 0.2 Dropout	95.35% / 71.83%	97.20% / 76.82%	94.66% / 73.94%
1BiLSTM(768) + 2BiLSTM(512) + Dense(768) + 0.2 Dropout	95.57% / 73.10%	97.37% / 78.77%	95.13% / 74.47%
1BiLSTM(768) + 2BiLSTM(512) + Dense(768) + 0.5 Dropout	95.30% / 73.35%	97.54% / 79.89%	95.15% / 75.00%
Third model pretrained on large corpus (final results)	97.45% / 81.98%	97.37% / 87.71%	96.52% / 81.34%

Example of work on out-of-vocabulary words

- Analysis of sentence that fully consists of unknown words.

Глокая *глокий* ADJ Case=Nom Gender=Fem Number=Sing Degree=Pos		**куздра** *куздра* NOUN Case=Nom Gender=Fem Number=Sing	**штеко** *штеко* ADV Degree=Pos	**будланула** *будлануть* VERB Gender=Fem Number=Sing Tense=Past Voice=Act	**бокра** *бокра* NOUN Case=Nom Gender=Fem Number=Sing
Глокая *глокать* VERB Voice=Act Tense=Notpast VerbForm=Conv	,	**куздра** *куздра* NOUN Case=Nom Gender=Fem Number=Sing	**штеко** *штеко* ADV Degree=Pos	**будланула** *будлануть* VERB Gender=Fem Number=Sing Tense=Past Voice=Act	**бокра** *бокра* NOUN Case=Nom Gender=Fem Number=Sing

Errors summary

- Ambiguity between nominative and accusative cases – $\sim 30\%$ of all mistakes.
- Ambiguity in the numbers of nouns – $\sim 11\%$ of all mistakes.

The most common mistakes

Correct tag	**Number of occurrences**	**Predicted tag**	**Number of errors**
Nominative	2650	Accusative	60
Accusative	1644	Nominative	37
Plural	2777	Singular	28
Nominative	2650	Genitive	19
DET	656	PRON	14
PRON	1133	DER	11

Examples of errors

- Some errors in identification of nominative and accusative cases and numbers of noun.

Минуту спустя	**кровать** *кровать* NOUN Case=Acc Gender=Fem Number=Sing	его принималась скрипеть...
	Минуты *минута* NOUN Case=Gen Gender=Fem Number=Sing	, проведенные дедом...

Examples of errors

- Some errors in test data annotation.

…эти слова в ласку и	**нежность** *нежность* NOUN Animacy=Inan Case=Nom Gender=Fem Number=Sing	, их жестокости не скроешь…
при некоторых видах	**плетения** *плетение* NOUN Animacy=Inan Case=Nom Gender=Neut Number=Plur	когда вращение купола…

Summary

- The proposed model shows fine accuracy.

Summary

- The proposed model shows fine accuracy.
- It gains from:

Summary

- The proposed model shows fine accuracy.
- It gains from:
    1. Comprehensive language description.

Summary

- The proposed model shows fine accuracy.
- It gains from:
  1. Comprehensive language description.
  2. Analysis of unknown words.

Summary

- The proposed model shows fine accuracy.
- It gains from:
    1. Comprehensive language description.
    2. Analysis of unknown words.
    3. Usage of LSTM neural network.

Summary

- The proposed model shows fine accuracy.
- It gains from:
  1. Comprehensive language description.
  2. Analysis of unknown words.
  3. Usage of LSTM neural network.
  4. Pretraining on large corpus.