Olga Lyashevskaya, Tatiana Shavrina, Igor Trofimov, Natalia Vlasova

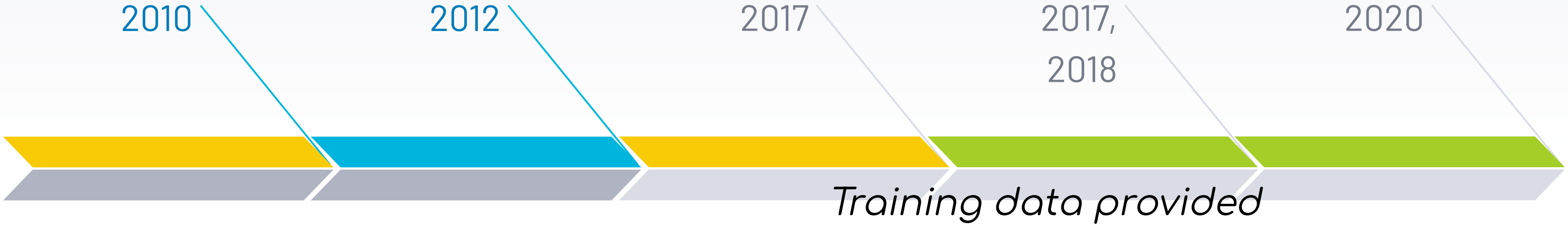
Dialogue Conference 2020

GramEval 2020

Russian Full Morphology and Universal Dependencies Parsing Shared Task

Russian morphology & syntax

in NLP shared tasks



2010

2012

2017

2017, 2018

2020

Training data provided

RU-EVAL

Lemma
PoS
Morphology
Rare Words
Disambiguation:
 Lemma, PoS

RU-EVAL

Syntax:
Dependencies
 General Texts
 News

MorphoRuEval

Lemma
PoS
Morphology
 Closed Track
 Open Track
Social
News
Fiction

Russian in CONLL UD

Token
Lemma
PoS
Morphology
Syntax:
Dependencies

GramEval

Lemma
PoS
Morphology
Syntax:
Dependencies
 6 genres

Motivation

We welcome systems that perform equally well on Russian texts of different genres and registers, genre- and time-specific words and constructions.

A new annotation benchmark for 6 genres:

News - Social Media - Wiki - Fiction - Poetry - Historical texts (17 century)

Existing pipelines "tokenization - morphology - lemmatization - syntax" accumulate errors at each stage. Should multi-level language structures be labelled in a more complex way?

Do contextual embedding architectures solve the problem?

From textbook to real life: heterogeneity in available source data as a yet another quest.

Objective

Building systems that implement full morphological and syntactic annotation and lemmatization according to Universal Dependencies (UD v2) format.

A cumulative evaluation score is computed on all tokens taking into account:

- POS (part of speech) accuracy
- morphological features accuracy
- LAS accuracy (labeled attachment score for dependency relations)
- lemmatization accuracy

Full Annotation Benchmark

Existing pipelines “tokenization - morphology - lemmatization - syntax” accumulate errors at each stage.
We believe that multi-level language structures need to be labelled together, otherwise errors in one tag level will lead to errors in the following.

During the competition, participants aim to build systems that define:

- Morphological characteristics of the word (part-of-speech and full tags).
- Lemma of the word.
- Syntactic relations (dependencies).

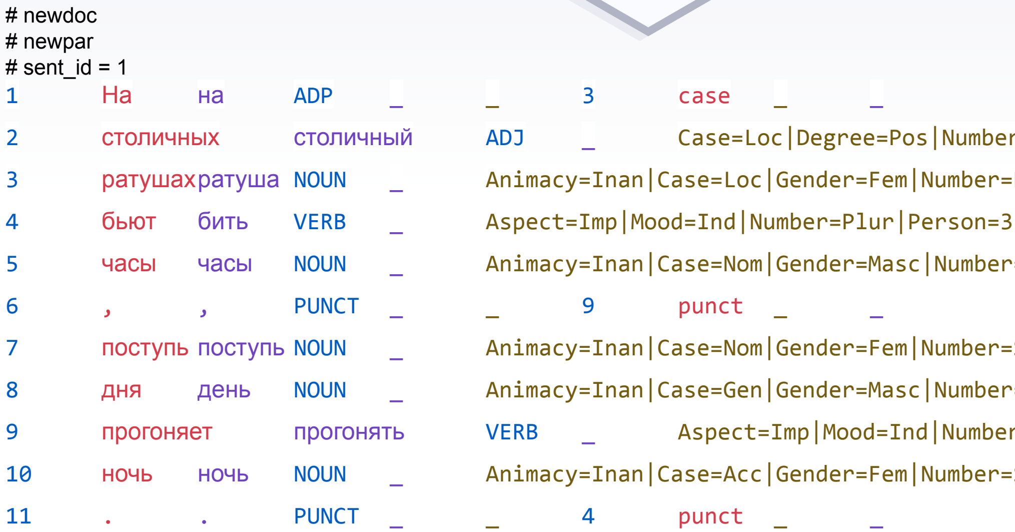
Genres: News - Fiction - Wiki - Social Media - Historical texts (17 century)

Data

Data

Data Format: UD CONLLU

- **training data** with full annotation - the resulting work of our team of annotators and existing UD treebanks
- **additional data** with automatic ("dirty") annotation
- **additional materials** such as frequency lists and models based on the third-party resources
- **development sets** (open test data) for preliminary evaluation of the model



```
# newdoc
# newpar
# sent_id = 1
1       На      на      ADP     _       _       3       case    _       _
2       столичных       столичный       ADJ     _       Case=Loc|Degree=Pos|Number
3       ратушахратуша   NOUN    _       Animacy=Inan|Case=Loc|Gender=Fem|Number=
4       бьют    бить    VERB    _       Aspect=Imp|Mood=Ind|Number=Plur|Person=3
5       часы    часы    NOUN    _       Animacy=Inan|Case=Nom|Gender=Masc|Number
6       ,       ,       PUNCT   _       _       9       punct   _       _
7       поступь поступь NOUN    _       Animacy=Inan|Case=Nom|Gender=Fem|Number=
8       дня     день    NOUN    _       Animacy=Inan|Case=Gen|Gender=Masc|Number
9       прогоняет       прогонять       VERB    _       Aspect=Imp|Mood=Ind|Number
10      ночь    ночь    NOUN    _       Animacy=Inan|Case=Acc|Gender=Fem|Number=
11      .       .       PUNCT   _       _       4       punct   _       _
```

Training Data

- **training data** with full annotation - the resulting work of our team of annotators and existing UD treebanks
- **additional data** with automatic ("dirty") annotation
- **additional materials** such as frequency lists and models based on the third-party resources
- **development sets** (open test data) for preliminary evaluation of the model

Clean Train Data

SynTagRus-UD
a harmonized version with semi-manual corrections
Russian data from the SynTagRus corpus

UD Russian GSD (wiki)
Russian Universal Dependencies Treebank annotated and converted by Google (96K tokens, wiki).

UD Russian Taiga
social UD_Taiga 26K, poetry UD_Taiga 13K,
news UD_Taiga 0.3K, Manual annotation

MorphoRuEval 2017
news UD_RuEval2017 (Lenta.ru, 5K)
fiction UD_RuEval2017 (magazines.gorky.media, 7K)
social UD_RuEval2017 (VK, 5K)

historical UD_OldRussian-RNC
A subcorpus of the Middle Russian corpus, texts of the 17th century, hybrid automatic with partial manual post-correction

Additional Data

- MorphoRuEval 2017
  Russian Corpus Data with manual verification, including SynTagRus, OpenCorpora, GICR, RNC.
- Twitter - Corpus of Russian tweets with sentiment annotation from study.mokoron.com
- Wikipedia -dump of Russian Wikipedia, first 100k articles, UDPipe pipeline
- Youtube - Comments from Russian Youtube Trends, UDPipe pipeline
- Lenta Ru news, up to 2018, UDPipe pipeline
- Stihi ru (Taiga)
- Proza ru (Taiga)
- Fiction Magazines (Taiga)

Dev and Test

- news UD_MorphoRuEval2017 1K dev + 1K test
- social networks UD_MorphoRuEval2017 1K dev + 1K test
- wiki UD_GSD 1K dev + 1K test
- fiction UD_SynTagRus 1K dev + 1K test
- poetry UD_Taiga 1K dev + 1K test
- 17th century UD_MidRussian-RNC 1K dev + 1K test

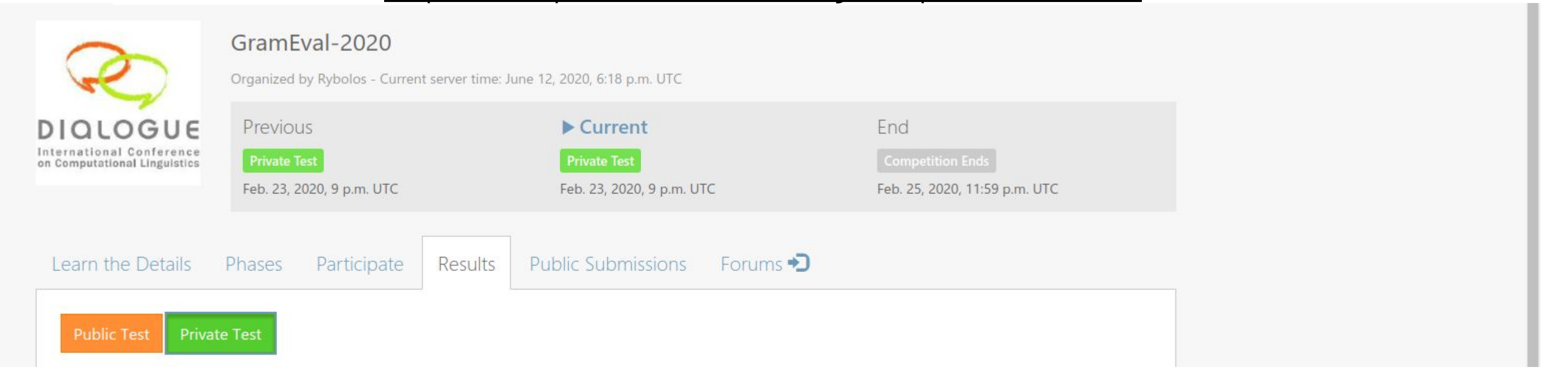
During test phase, all test sentences were mixed into additional vertical texts from various sources

Phases

Phase 1: Dev (Public test)
Participants can make dev submissions on Codalab
There are answers available for the dev set
The sources of the texts are known
Up to 1000 submissions per participant

Phase 2: Test (Private test)
No answers available
No sources of the texts provided
Up to 100 submissions per participant

Phase 2 never ends https://competitions.codalab.org/competitions/22902



GramEval-2020

Organized by Rybolos - Current server time: June 12, 2020, 6:18 p.m. UTC

Previous
Private Test
Feb. 23, 2020, 9 p.m. UTC

▶ Current
Private Test
Feb. 23, 2020, 9 p.m. UTC

End
Competition Ends
Feb. 25, 2020, 11:59 p.m. UTC

Metrics

The main metric of the competition is accuracy, averaged over all categories

Score by segment:

Score = Mean (POS_accuracy, Feature_accuracy, Lemma_accuracy, LAS)

Overall score:

Overall score = Mean (news_score, wiki_score, social_score, fiction_score, poetry_score, 17-c._score)

Baseline: UDPipe + RNNMorph = Overall score 80.3%

Additional metrics:

F1 metrics for pos, features, and dependency relations, lemmatization,
UAS, MLAS, BLEX metrics according to the CoNLL method

Token alignment
% of corrupted tokenization
During the private test phase, all systems had their alignment score of 100%.

Shared Task Results

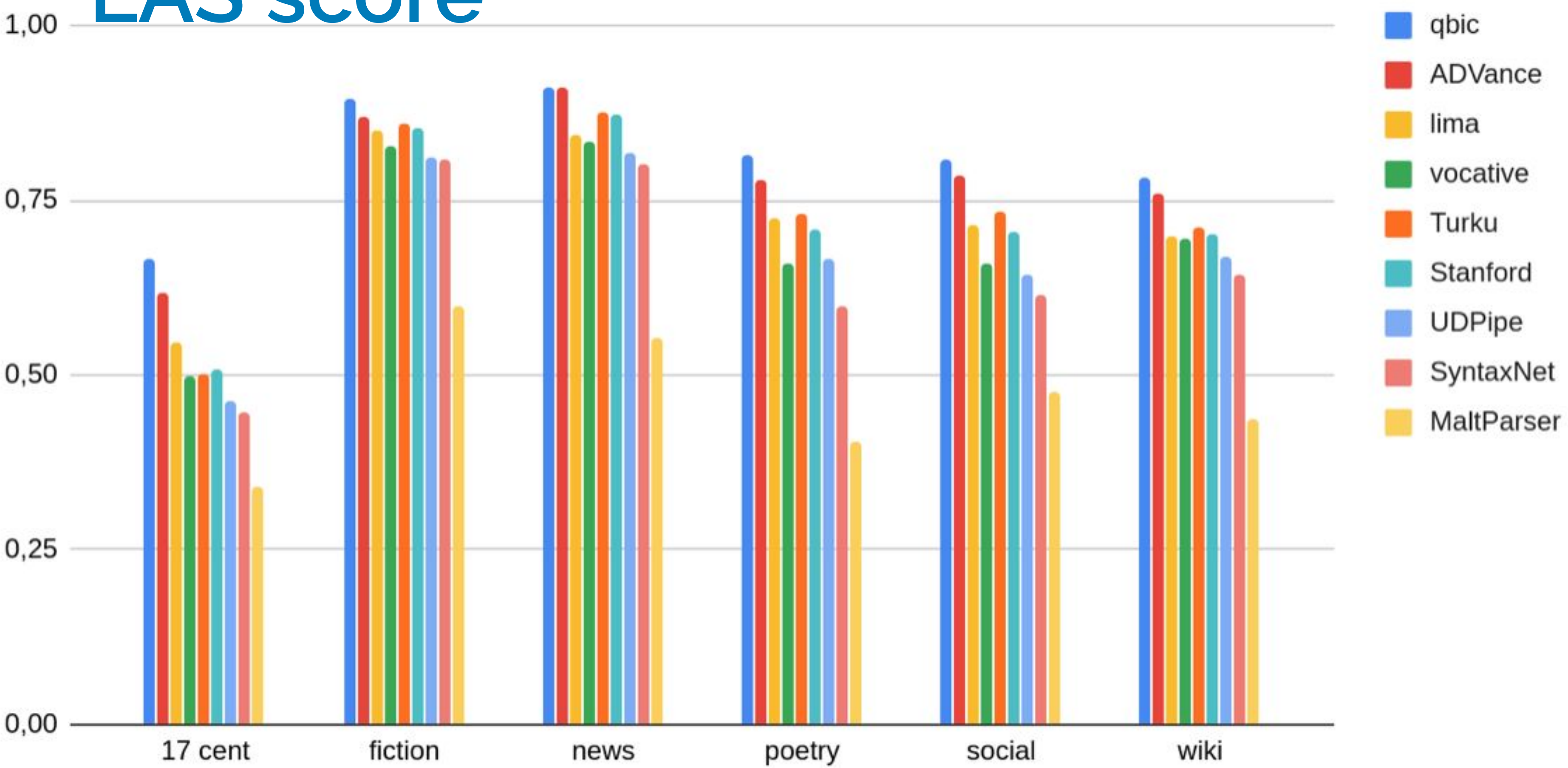
Main Results

4 new open source resources
More than +10% to the baseline!

Teams Overall score

- qbic - **0.91609** https://github.com/DanAnastasyev/GramEval2020
- ADVance - **0.90762** https://github.com/AlexeySorokin/GramEval2020
- lima - **0.87870** https://github.com/aymara/lima/tree/grameval-2020
- vocative - **0.85198** https://github.com/Koziev/rupostagger2 https://github.com/Koziev/rulemma
- baseline - **0.80377** https://github.com/dialogue-evaluation/GramEval2020/tree/master/baseline

LAS score


1,00
0,75
0,50
0,25
0,00
17 cent
fiction
news
poetry
social
wiki
qbic
ADVance
lima
vocative
Turku
Stanford
UDPipe
SyntaxNet
MaltParser

Team Results

Team	Data	Architecture	Embeddings
qbic	All GramEval data except SynTagRus	End-to-End parser: features, lemmas, and dependencies are predicted by joint BERT model with independent modules. Encoder is a single-layer LSTM, decoders are simple feedforward models for predicting lemmas and features , as well as a biaffine attention model for dependencies and their labels	Pretrained RuBert
Advance	All GramEval data + poetry Taiga corpus for embedding training	Classifier of 4 main data sources - normative fiction, 17 c., poetry, social. + Morphotagger and parser on BERT, pretrained on SynTagRus 2.5 + 17 c. lemmer on rules	4 separately trained BERTs on GramEval data

Team Results

Team	Data	Architecture	Embeddings
lima	All GramEval data	Original implementation of Dozat & Manning: embedding layer + LSTM layer + feedforward layer. Differs from the original models in that morphology and syntax are trained simultaneously in multitask learning mode	Pretrained FastText
Vocative	GramEval2020 data with rule-based parser validation for extracting good training samples for pos-tagging and parsing + own treebank data for pos tagging training	Ensemble model: 1) dictionary-based lemmatizer 2) LSTM-CRF pos tagger, considering the context and features + pure CRF pos tagger for sentences longer than 30 words + Russian UDpipe for pos and features 3) parser: UDPipe trained on GramEval data 4) rule-based correction for 17 c. data	Pretrained word2vec wordchar2vector

Analysis of submitted annotations

General notes

The competing systems make similar mistakes in morphological analysis.

Most errors are associated with:

- uppercase uses;
- non-standard spellings.

Beginning of the sentence or of the line in poetry, proper names sharing ambiguity with common nouns, words with spelling errors, author spelling, hashtags,abbreviations and acronyms

Special attention to 17 c. dataset!

Lemmatization and pos-tagging errors

- words with rare inflectional model, pluralia tantum, plural homonyms;
- homonymy in pos-tagging;
- participles vs. verbal adjectives;
- words such as **нельзя, надо, пора** tagged as VERB vs. ADV vs. NOUN

Low quality markup in training sets?

Inconsistency of markup in different training data?

Errors in morphological features

- animacy in adjectives, pronouns and numerals;
- gender, case and degree errors;
- aspect in biaspectual verbs;
- features of the verb **БЫТЬ**;
- voice in verbs.

In general all the systems do well with the paradigm syncretism.

Errors in dependency relation labelling

N	gold	predicted	N	gold	predicted
74	punct	discourse	19	obj	nsubj
49	parataxis	appos	18	amod	nummod
42	iobj	obl	18	punct	parataxis
40	list	parataxis	17	obj	obl
38	parataxis	conj	16	conj	parataxis
32	discourse	parataxis	15	appos	nmod
24	amod	appos	15	discourse	advmod
24	nmod	appos	15	mark	advmod
23	obl	nmod	15	xcomp	obl
20	nsubj	obj	14	appos	parataxis

1

Open Questions

1. Data Change

   Have replaced 1 segment of training data – deleted SynTagRus data

   Major modern morphological parsers have already been trained

2. Changed submitting procedure from dev phase to test phase:

   - 1st phase, public test – 6 vertical test files (genre is known)
   - 2nd phase, private test – 6 vertical test files in 1

     (all genres mixed + additional data added)

3. Data Annotation Problems

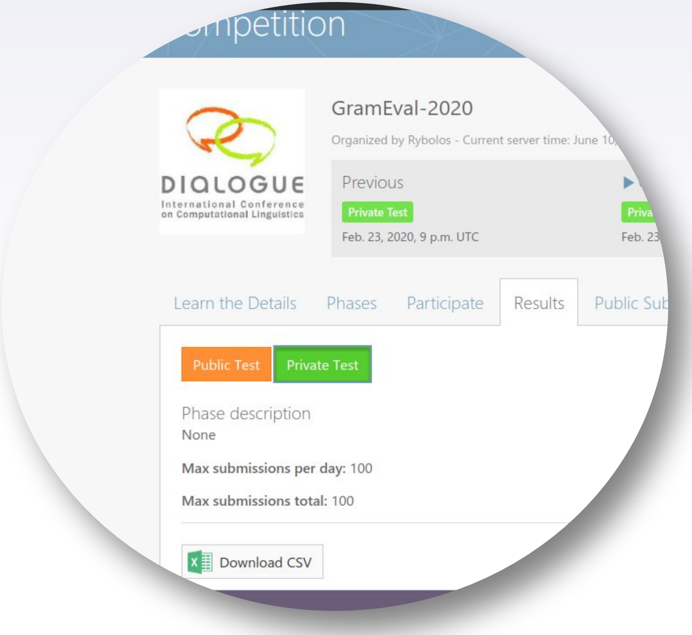
   - 17th century data lacks lemmatization
   - Automatic annotation on 17th century data
   - Orphograpy inconsistency (ѣ-problem)

Thank you!

Find all materials here

CodaLab: https://competitions.codalab.org/competitions/22902

GitHub: https://github.com/dialogue-evaluation/GramEval2020



GramEval-2020

Organized by Rybolos - Current server time: June 10

Previous

Private Test

Feb. 23, 2020, 9 p.m. UTC

Priva

Feb. 23

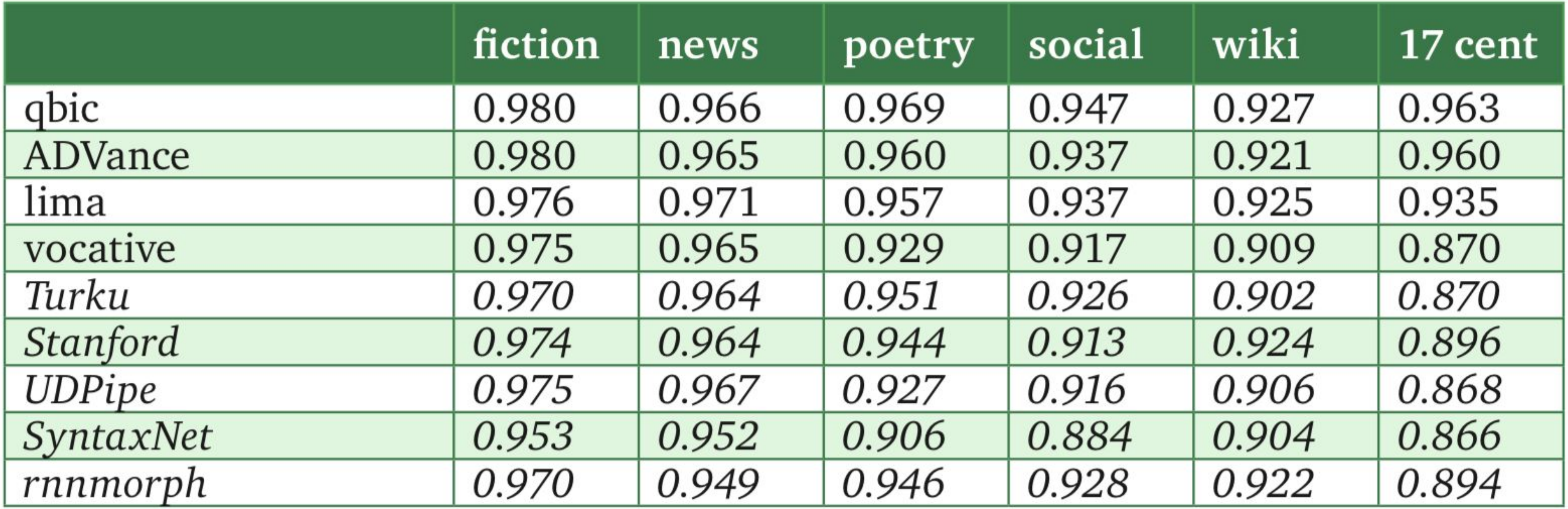
Phase description

None

Max submissions per day: 100

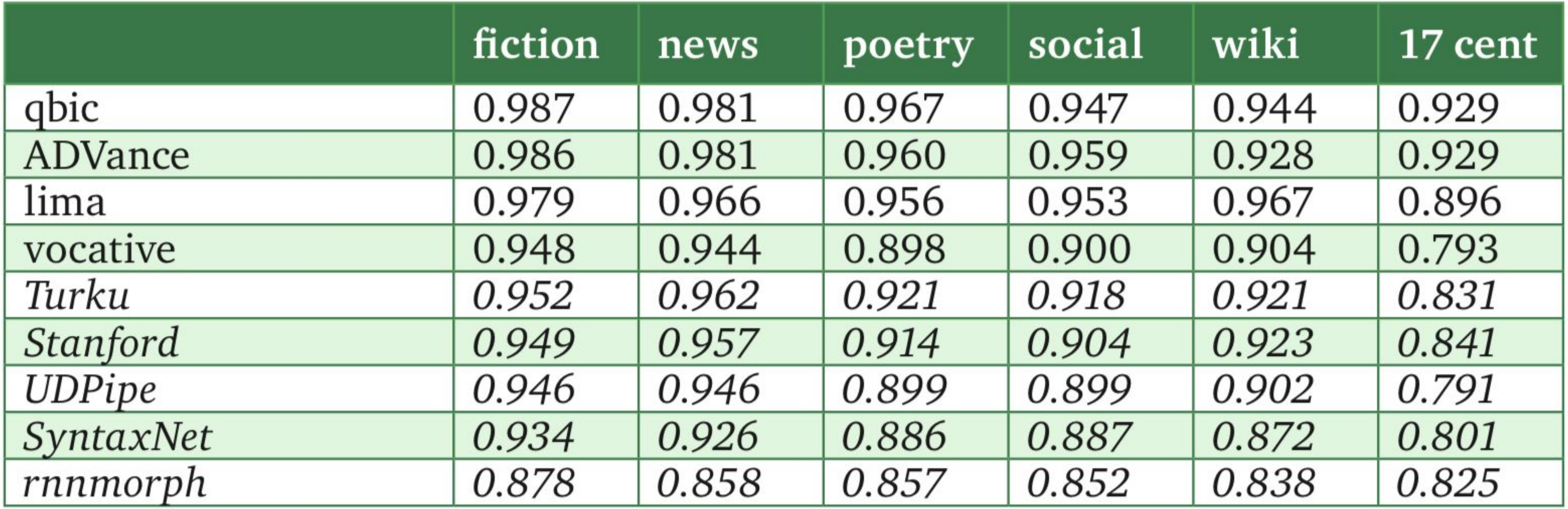
Max submissions total: 100

Parts of speech score



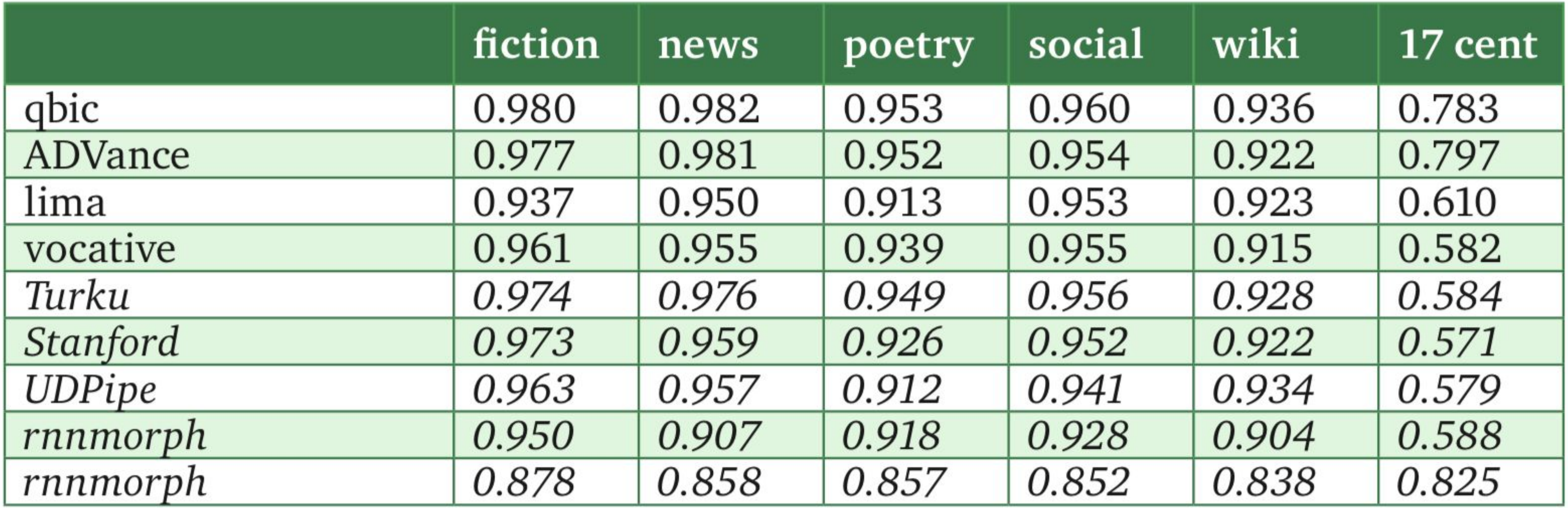
	fiction	news	poetry	social	wiki	17 cent
qbic	0.980	0.966	0.969	0.947	0.927	0.963
ADVance	0.980	0.965	0.960	0.937	0.921	0.960
lima	0.976	0.971	0.957	0.937	0.925	0.935
vocative	0.975	0.965	0.929	0.917	0.909	0.870
Turku	*0.970*	*0.964*	*0.951*	*0.926*	*0.902*	*0.870*
Stanford	*0.974*	*0.964*	*0.944*	*0.913*	*0.924*	*0.896*
UDPipe	*0.975*	*0.967*	*0.927*	*0.916*	*0.906*	*0.868*
SyntaxNet	*0.953*	*0.952*	*0.906*	*0.884*	*0.904*	*0.866*
rnnmorph	*0.970*	*0.949*	*0.946*	*0.928*	*0.922*	*0.894*

Morphological features score



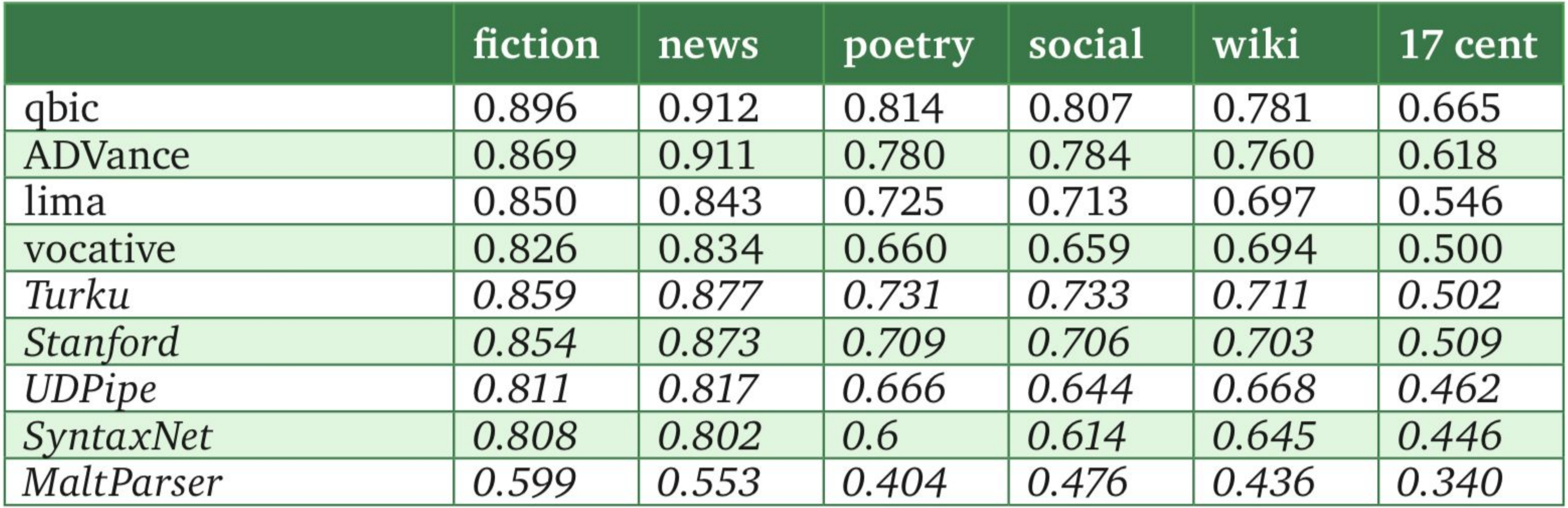
	fiction	news	poetry	social	wiki	17 cent
qbic	0.987	0.981	0.967	0.947	0.944	0.929
ADVance	0.986	0.981	0.960	0.959	0.928	0.929
lima	0.979	0.966	0.956	0.953	0.967	0.896
vocative	0.948	0.944	0.898	0.900	0.904	0.793
Turku	*0.952*	*0.962*	*0.921*	*0.918*	*0.921*	*0.831*
Stanford	*0.949*	*0.957*	*0.914*	*0.904*	*0.923*	*0.841*
UDPipe	*0.946*	*0.946*	*0.899*	*0.899*	*0.902*	*0.791*
SyntaxNet	*0.934*	*0.926*	*0.886*	*0.887*	*0.872*	*0.801*
rnnmorph	*0.878*	*0.858*	*0.857*	*0.852*	*0.838*	*0.825*

Lemma score

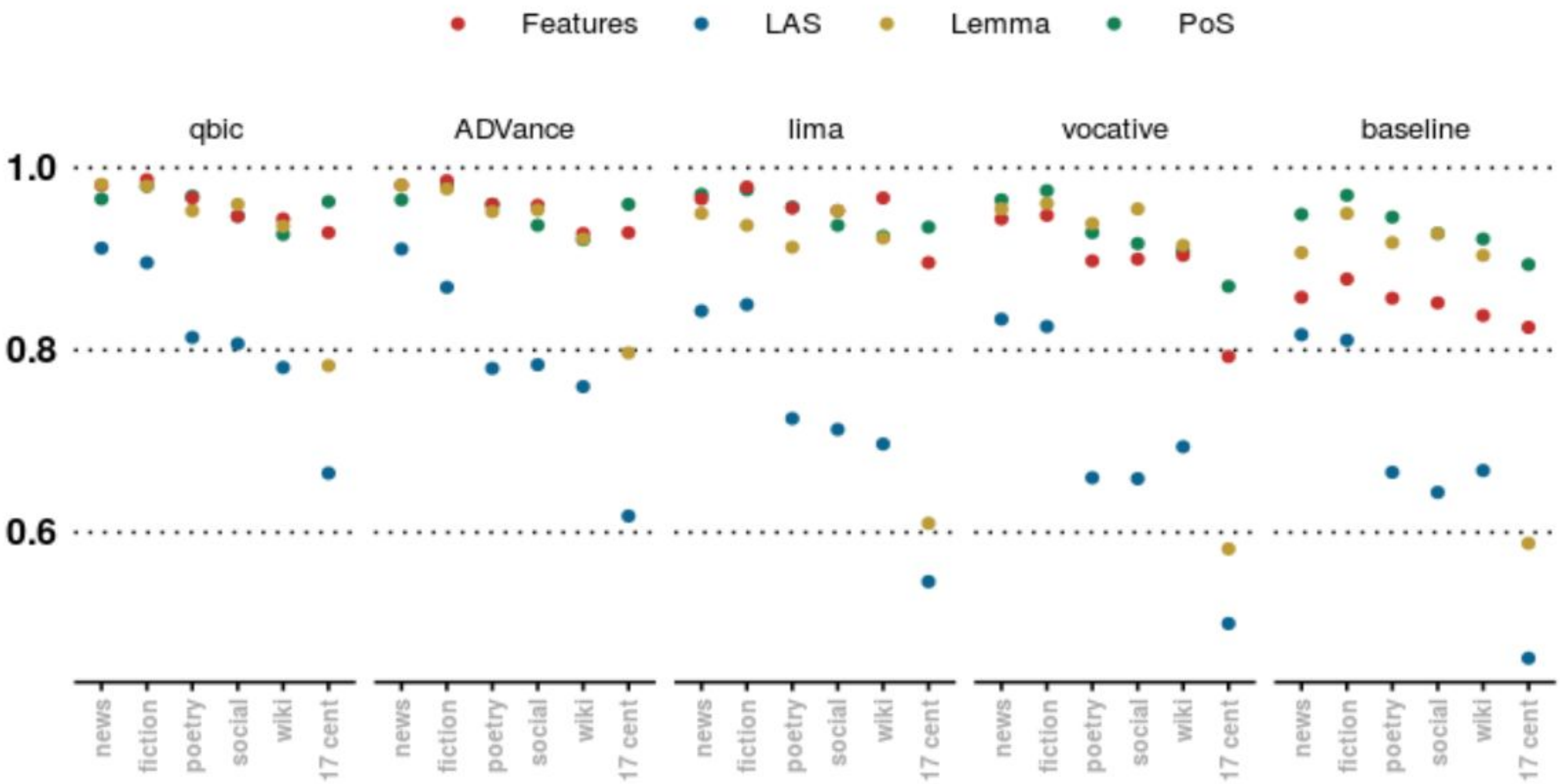


	fiction	news	poetry	social	wiki	17 cent
qbic	0.980	0.982	0.953	0.960	0.936	0.783
ADVance	0.977	0.981	0.952	0.954	0.922	0.797
lima	0.937	0.950	0.913	0.953	0.923	0.610
vocative	0.961	0.955	0.939	0.955	0.915	0.582
Turku	*0.974*	*0.976*	*0.949*	*0.956*	*0.928*	*0.584*
Stanford	*0.973*	*0.959*	*0.926*	*0.952*	*0.922*	*0.571*
UDPipe	*0.963*	*0.957*	*0.912*	*0.941*	*0.934*	*0.579*
rnnmorph	*0.950*	*0.907*	*0.918*	*0.928*	*0.904*	*0.588*
rnnmorph	*0.878*	*0.858*	*0.857*	*0.852*	*0.838*	*0.825*

LAS score



	fiction	news	poetry	social	wiki	17 cent
qbic	0.896	0.912	0.814	0.807	0.781	0.665
ADVance	0.869	0.911	0.780	0.784	0.760	0.618
lima	0.850	0.843	0.725	0.713	0.697	0.546
vocative	0.826	0.834	0.660	0.659	0.694	0.500
Turku	*0.859*	*0.877*	*0.731*	*0.733*	*0.711*	*0.502*
Stanford	*0.854*	*0.873*	*0.709*	*0.706*	*0.703*	*0.509*
UDPipe	*0.811*	*0.817*	*0.666*	*0.644*	*0.668*	*0.462*
SyntaxNet	*0.808*	*0.802*	*0.6*	*0.614*	*0.645*	*0.446*
MaltParser	*0.599*	*0.553*	*0.404*	*0.476*	*0.436*	*0.340*


● Features
● LAS
● Lemma
● PoS
qbic
ADVance
lima
vocative
baseline
1.0
0.8
0.6
news
fiction
poetry
social
wiki
17 cent