ONLINE: ISSN 2075-7182

Компьютерная лингвистика и интеллектуальные технологии

По материалам ежегодной международной конференции «Диалог» (2023)

Выпуск 22

Computational Linguistics and Intellectual Technologies

Papers from the Annual International Conference “Dialogue” (2023)

Issue 22

Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной международной конференции «Диалог». Вып. 22. 2023. С. I–602.

Сборник включает 54 доклада международной конференции по компьютерной лингвистике и интеллектуальным технологиям «Диалог 2023», представляющих широкий спектр теоретических и прикладных исследований в области описания естественного языка, моделирования языковых процессов, создания практически применимых компьютерных лингвистических технологий.

Для специалистов в области теоретической и прикладной лингвистики и интеллектуальных технологий.

Предисловие

22-й выпуск ежегодника «Компьютерная лингвистика и интеллектуальные технологии» содержит избранные материалы 29-й международной онлайн-конференции «Диалог». В 2023 году для публикации в основном томе сборника редколлегией были отобраны 54 доклада из 120, поданных на конференцию. Работы, представленные в сборнике, отражают те направления исследований в области компьютерного моделирования и анализа естественного языка, которые по традиции представляются на Диалоге:

- **Интеллектуальный анализ документов (Intelligent Document Processing):** классификация, Name Entity & Relation Extraction, саммаризация, генерация, анализ тональности, Argumentation Mining, Propaganda & Fake News Detection, etc., мультимодальные подходы (совместное использование моделей NLP и Computer Vision);
- **Глубокое обучение в компьютерной лингвистике:** методики применения нейронных сетей в исследованиях, содержательная интерпретация;
- **Компьютерные лингвистические ресурсы:** новые датасеты и новые сценарии и типы разметки, Evaluation Benchmarks;
- Компьютерный анализ Social Media;
- **Корпусная лингвистика и корпусометрия:** методики создания, использования и оценки корпусов;
- **Компьютерная семантика:** аналитические и дистрибуционные модели, связь между ними;
- Лингвистические онтологии и автоматическое извлечение знаний;
- **Мультимодальная коммуникация:** аналитические и нейронные модели речевого акта;
- Модели общения и диалоговые агенты;
- Лингвистический анализ текста: морфология, синтаксис, семантика (модели анализа);
- Компьютерная лексикография;
- **Полевая компьютерная лингвистика:** применение методов NLP для малоресурсных языков.

В соответствии с традициями «Диалога», конференции по компьютерной лингвистике с почти полувековой историей, отбор работ основывается на представлении о важности соединения новых методов и технологий анализа языковых данных с полноценным лингвистическим анализом. Диалог является де-факто крупнейшим форумом по проблемам создания современных компьютерных ресурсов, моделей и технологий для русского языка, поэтому ключевым событием «Диалога» является подведение итогов технологических соревнований между разработчиками систем лингвистического анализа русскоязычных текстов — *Dialogue Evaluation*. В этом году состоялись 4 соревнования:

- **RuCoCo:** Соревнование по разрешению кореференции;
- **RuSentNE:** Соревнование по анализу тональности к именованным сущностям в новостных текстах;
- **RECEIPT-AVQA:** Соревнование по генерации ответов на вопросы к изображениям;
- **SEMarkup:** Соревнование по автоматической семантической разметке.

Статьи в сборнике публикуются на русском и английском языках. При выборе языка публикации действует следующее правило:

- доклады по компьютерной лингвистике подаются на английском языке. Это расширяет их аудиторию и позволяет привлекать к рецензированию международных экспертов;
- доклады, посвященные лингвистическому анализу русского языка, предполагающие знание этого языка у читателя, подаются на русском языке (с обязательной аннотацией на английском).

Несмотря на традиционную широту тематики представленных на конференции и отобранных в сборник докладов, они не могут дать полной картины направлений «Диалога». Ее можно получить с помощью сайта конференции www.dialog-21.ru, на котором представлены обширные электронные архивы «Диалогов» последних лет и все результаты проведенных тестирований Dialogue Evaluation.

Мы обращаем внимание авторов и читателей сборника, что с 2018 года Редсовет отказался от печати сборника на бумаге. Все сборники размещаются на сайте конференции. С 2014 года основной том индексируются Scopus.

Программный комитет конференции «Диалог»
Редколлегия сборника «Компьютерная лингвистика и интеллектуальные технологии»

Рецензенты

Азарова Ирина Владимировна
Андрианов Андрей Иванович
Антонова Александра Александровна
Баранов Анатолий Николаевич
Беликов Владимир Иванович
Богданова-Бегларян Наталья Викторовна
Богуславский Игорь Михайлович
Бурцев Михаил Сергеевич
Васильев Виталий Геннадьевич
Гусев Илья Олегович
Добров Борис Викторович
Добровольский Владимир Андреевич
Добровольский Дмитрий Олегович
Жарков Андрей Алехандрович
Зализняк Анна Андреевна
Захаров Леонид Михайлович
Золотухин Денис Денисович
Иванов Владимир Владимирович
Ивойлова Александра Михайловна
Ильвовский Дмитрий Алексеевич
Инденбом Евгений Михайлович
Инькова Ольга Юрьевна
Иомдин Леонид Лейбович
Киосе Мария Ивановна
Клышинский Эдуард Станиславович
Клячко Елена Леонидовна
Князев Сергей Владимирович
Кобозева Ирина Михайловна
Козеренко Елена Борисовна
Копотев Михаил Вячеславович
Коротаев Николай Алексеевич
Котельников Евгений Вячеславович
Котов Артемий Александрович
Куратов Юрий Михайлович
Кутузов Андрей Борисович
Лапошина Антонина Николаевна
Левонтина Ирина Борисовна
Лобанов Борис Мефодьевич
Логинов Василий Васильевич
Лукашевич Наталья Валентиновна
Малафеев Алексей Юрьевич
Митрофанова Ольга Александровна
Мичурина Мария Александровна
Недолужко Анна
Никишина Ирина Юрьевна
Орлов Евгений Анатольевич
Пазельская Анна Германовна
Переверзева Светлана Игоревна
Петрова Мария Владимировна
Подлесская Вера Исааковна
Рыгаев Иван Петрович
Селегей Владимир Павлович
Слюсарь Наталия Анатольевна
Смирнов Иван Валентинович
Смуров Иван Михайлович
Татевосов Сергей Георгиевич
Урысон Елена Владимировна
Федорова Ольга Викторовна
Феногенова Алена Сергеевна
Хохлова Мария Владимировна
Циммерлинг Антон Владимирович
Шаврина Татьяна Олеговна
Шамардина Татьяна Вячеславовна
Шаров Сергей Александрович
Янко Татьяна Евгеньевна

Contents[1]

[*] The reports of each section are ordered by the surname of the first author in compliance with the English alphabet.

Disambiguation in context in the Russian National Corpus: 20 yeas later

Olga Lyashevskaya
HSE University
Vinogradov Russian Language Institute RAS
Moscow, Russia
olesar@yandex.ru

Ilia Afanasev
HSE University
MTS AI
Moscow, Russia
szrnamerg@gmail.com

Stefan Rebrikov
HSE University
Kurchatov Institute
Moscow, Russia
robstef85@gmail.com

Yana Shishkina
HSE University
Moscow Institute of Physics and Technology
Moscow, Russia
yanaalekseevna2000@mail.ru

Elena Suleymanova **Igor Trofimov** **Natalia Vlasova**
A. K. Ailamazyan Program Systems Institute of RAS
Pereslavl-Zalessky, Russia
yes2helen@gmail.com itrofimov@gmail.com nathalie.vlassova@gmail.com

Abstract

An updated annotation of the Main, Media, and some other corpora of the Russian National Corpus (RNC) features the part-of-speech and other morphological information, lemmas, dependency structures, and constituency types. Transformer-based architectures are used to resolve the homonymy in context according to a schema based on the manually disambiguated subcorpus of the Main corpus (morphology and lexicon) and UD-SynTagRus (syntax). The paper discusses the challenges in applying the models to texts of different registers, orthographies, and time periods, on the one hand, and making the new version convenient for users accustomed to the old search practices, on the other. The re-annotated corpus data form the basis for the enhancement of the RNC tools such as word and n-gram frequency lists, collocations, corpus comparison, and Word at a glance.

Keywords: morphological tagging; dependency parsing; lemmatization; disambiguation; NLP evaluation; Russian National Corpus; Russian

DOI: 10.28995/2075-7182-2023-22-307-318

Разрешение неоднозначности в контексте для Национального корпуса русского языка: 20 лет спустя

О. Н. Ляшевская[1,2], **И. А. Афанасьев**[1,3], **С. А. Ребриков**[1,4], **Я. А. Шишкина**[1,5], **Е. А. Сулейманова**[6], **И. В. Трофимов**[6], **Н. А. Власова**[6]

[1]Национальный исследовательский университет «Высшая школа экономики»
[2]Институт русского языка им. В. В. Виноградова РАН
[3]МТС ИИ
[4]НИЦ «Курчатовский институт»
[5]МФТИ
Москва, Россия
[6]Институт программных систем им. А. К. Айламазяна РАН
г. Переславль-Залесский, Ярославская обл., Россия
olesar@yandex.ru, {szrnamerg, robstef85}@gmail.com, yanaalekseevna2000@mail.ru,
{yes2helen, itrofimov, nathalie.vlassova}@gmail.com

Аннотация

Обновление разметки Основного, Газетного и ряда других корпусов Национального корпуса русского языка (НКРЯ) касается информации о части речи, других морфологических признаках, леммах (словарных формах слов), структурах зависимостей предложения и типах составляющих. Для разрешения лингвистической неоднозначности в контексте используются нейросетевые архитектуры на основе трансформеров. Разметка воспроизводит схему, применяемую в подкорпусе Основного корпуса со снятой вручную грамматической омонимией (морфология и леммы) и UD-SynTagRus (синтаксис). В статье рассматриваются проблемы применения моделей к текстам, написанным в различных функциональных стилях, орфографиях и в разные периоды времени. Поскольку в ряде случаев текстовому фрагменту в заданном контексте можно сопоставить более одного теоретически возможного лингвистического разбора, необходимо принимать во внимание поддержку множественных разборов. Кроме того, обсуждаются вопросы совместимости старой и новой разметки в плане адаптации пользователей к новому поисковому функционалу корпуса. Автоматически дизамбигуированные данные больших корпусов позволили улучшить существующие и разработать новые сервисы поисковой платформы НКРЯ, такие как частотные списки слов и n-грамм, коллокации, сравнение корпусов и портрет слова.

Ключевые слова: автоматическое разрешение лексико-грамматической неоднозначности, морфологическая разметка, синтаксическая разметка, русский язык, Национальный корпус русского языка

1 Introduction

For almost 20 years, the lexico-grammatical annotation of the Russian National Corpus (RNC) existed in three formats. (1) In the Syntactic corpus (SynTagRus, 1.4 MW), each word was provided with one and only one morphological and lemma analysis appropriate in context, and each sentence was analysed as one syntactic dependency tree. (2) In the the manually disambiguated subcorpus of the Main corpus ("Snyatnik", 6 MW) and in the Educational corpus (0,6 MW), only morphology and lemmas were analysed based on a somewhat different tagset and grammatical dictionary compared to SynTagRus. The majority of historical RNC corpora were annotated generally in the same way and oriented on their own markup schemas, tagsets, and dictionaries. (3) Finally, there were no disambiguation in the largest part of the modern Russian texts (more than 1 billion words) and Church Slavonic texts (5,3 MW): each word corresponded to as many analyses as the grammatical dictionary stores, regardless of the context. If the word form of a modern language is not attested in the dictionary, the MyStem hypothesis module assigns a few of the most probable annotations to it (Segalovich, 2003; Zobnin and Nosyrev, 2015).

One of the objectives of the Corpus 2.0 project (2020-2022) was to add syntactic annotations and resolve lexical and morphological ambiguity in modern Russian texts. Firstly, this allows users to constraint the search window by defining syntactic relations between elements or setting up a certain type of clause or phrase within which the elements should occur. Secondly, this makes it possible to significantly reduce the number of irrelevant examples in the search output. Thirdly, other search facilities such as lexical groups-based search, frequency lists, collocations, associated words, etc. definitely benefit from the less noisy annotation input. Fourthly, the use of syntactic n-grams based on dependency parses (Goldberg and Orwant, 2013) in addition to ordinary sequential n-gram opens the way to a new kind of high-quality tools for researchers. All these changes also involve technical improvements in the infrastructure of the corpus search engine such as reducing the size of the search indices and the time spent performing the calculations, extending the amount of annotated data and information conveyed to the user.

2 Related Work

The approaches to the three grammar tasks that form the basic NLP pipeline, namely, part-of-speech/morphological tagging, lemmatisation, and dependency parsing, rapidly developed for the last half a century (Hann, 1974) (Spyns, 1996) (Aduriz et al., 1996) (Branco and Silva, 2003)] (Qi et al., 2020) (Kumar et al., 2022). Currently pipeline models that combine part-of-speech/morphological tagging, lemmatisation , and parsing, dominate the landscape (Straka and Straková, 2017) (Kondratyuk, 2019) (Kanerva et al., 2021). However, despite this pursuit to develop the language-independent tagger for benchmark datasets (Toleu et al., 2022) that provide satisfying for all the included languages, yet moderate for each of them results, there is a growing concern that low-resourced language NLP, and

probably NLP in general, is going to suffer from the trend (Alonso-Alonso et al., 2022). Frw works clearly state the intention to make a universal tagger, which is based upon the multi-lingual training and switching parameters to fine-tune for a single language (Üstün et al., 2020). The models, trained for the particular task-language pair, still seem to deserve attention, as (Dyer, 2022) states for the case of Wolof language.

Automatic morphological tagging systems currently employ the pair of dominating approaches, the single-language rule-based one (Gambäck, 2012), and the machine learning-based one, which can assume both monolingual (Berdičevskis et al., 2016) (Qi et al., 2018) (Qi et al., 2020) (Scherrer, 2021) and multi-lingual (Straka and Straková, 2017) forms. Instead of targeting the multi-lingual level, now morphological tagging shifts into the multi-lect one to be able to deal with the very close (Obeid et al., 2022), yet significantly different lects, as is the cast with Arabic (Inoue et al., 2022) (Fashwan and Alansary, 2022). This also provokes a lot of discussion for morphological tagging of low-resouced languages (Blum, 2022) (Wiemerslage et al., 2022). The discussion about data quality takes place within the common morphology tagging discourse (Muradoglu and Hulden, 2022). New methods are being developed, for instance, graph-based part-of-speech tagging (ImaniGooghari et al., 2022), or using compressed FastText models (Nevěřilová, 2022). Specifically concerning Russian, joined morphological analysis and morpheme segmentation models were proposed recently (Bolshakova and Sapin, 2022).

Lemmatisation follows the same patterns that morphological tagging does. Currently, there is a division between the universal lemmatisation tools (Straka and Straková, 2017) (Bergmanis and Goldwater, 2018) (Kanerva et al., 2021), and language, or domain-specific (Fernández, 2020) The sequence-to-sequence architecture (Sutskever et al., 2014) (Cho et al., 2014) prevails now, and within it the encoder-decoder transformers dominate (Lewis et al., 2020) The ensemble models that enhance lemmatisation efficiency with external resources (Milintsevich and Sirts, 2021) are gaining popularity (de Graaf et al., 2022)

Dependency parsing is probably the most dynamically developing area of the three, as it still presents the highest challenge of the three for the automated corpus tools. New methods are constantly being implemented: the last three years witnessed a combination of the second-order graph-based and headed-span-based projective dependency parsing (Yang and Tu, 2022), the domain adaptation (Li et al., 2022) and the dependency parsing being treated as machine reading comprehension (MRC)-based span-span prediction (Gan et al., 2022) and using structure preserving embeddings for dependency parsing (Kádár et al., 2021) The state-of-the-art method, biaffine parsing, is modified (Xu et al., 2022). The previously under-utilised concepts, such as *nuclei* (semantically independent units consisting of a content word together with its grammatical markers, regardless of whether the latter are realised in dependent words or not (Basirat and Nivre, 2021)), are introduced to the frameworks. The data augmentation techniques are implemented to enhance the performance of the models (Goodwin et al., 2022). (Eggleston and O'Connor, 2022) and (Langedijk et al., 2022) introduce cross-lect dependency parsing, getting in line with papers that consider low-resourced languages (Tian et al., 2022) and zero-shot (de Lhoneux et al., 2022) (Shi et al., 2022) dependency parsing. The issues of the dataset construction that affect evaluation are discussed in (Krasner et al., 2022) Artificial performance inflation is a problem that should be addressed across the pipeline of morphological tagging, lemmatisation and part-of-speech tagging (Goldman et al., 2022).

3 Data for Training and Evaluation

We conducted experiments involving a diverse panel of text samples. A variety of genres, types, domains, time periods of creation, and orthographies were presented in the following datasets for modern Russian (1700-2020s):

- `SynTagRus UD 2.8` - 1,1 M tokens (contemporary fiction, popular science, newspaper and journal articles dated between 1960 and 2016, texts of online news etc.). This portion of the RNC Syntactic Corpus converted to the Universal Dependencies (UD) format was the main training dataset used in the GramEval-2020 shared task.
- `SynTagRus UD 2015` - 400k tokens. An addition to the RNC Syntactic Corpus annotated in 2015-

GramEval-2020 (Taiga)	dev	test	New RNC datasets	dev	test
fiction	1.0k	1.0k	prose-XX	10.4k	20.0 k
news	1.0k	1.0k	newspapers-XXI	7.8k	14.4k
poetry	1.0k	1.0k	prose-XIX	41.7k	80.7k
social	1.0k	1.0k	poetry-XIX	1.4k	1.4k
wiki	1.0k	1.0k	old-orthography	14.8k	14.8k
			old-orthography-XVIII	6.1k	6.1k
			Middle Russian: LEG	16.5k	39.0k
			bezobrazov		519.0k

Table 1: Size of the validation and test sets, tokens.

2020; converted and added to UD v.2.9. New genres: wikipedia.

- Taiga - 200 k tokens. Modern text samples extracted from Taiga Corpus, MorphoRuEval-2017 and GramEval-2020 shared tasks collections. Genres include electronic communication (VK, Twitter and other social media, YouTube comments, questions & answers from otvet.mail.ru, reviews from reviews.yandex.ru); poetry from stihi.ru (naïve poetry) and RNC Corpus of Russian poetry; fiction; news (lenta.ru etc.); wiki (Russian wikipedia). Taiga includes, among others, development and test data of the GramEval-2020 shared task (modern Russian), which was subdivided into the following subsets: fiction, news, poetry, social, wiki.
- newspapers-XXI - 34 k tokens. Samples extracted from the RNC National media and Regional and international media corpora.
- prose-XX - 423 k tokens. Texts of the 20th c. and the beginning of the 21th c. in modern orthography (RNC Main corpus). Fiction includes stories by V. M. Shukshin, I. V. Evdokimov, and M. K. Pervukhin, non-fiction - diaries and memories, journalism covers general news, finance, church news, recipes and tips.
- prose-XIX - 108 k tokens. Texts of the 19th c. in modern orthography (RNC Main corpus). The dataset includes drama by A. V. Sukhovo-Kobylin, A. Pisemsky, M. Gorky, etc., fiction by N. V. Gogol, S. T. Aksakov, E. A. Salias etc., non-fiction on history, hygiene, memories and essays.
- poetry-XIX - 50 k tokens. Samples from the RNC Russian Poetry Corpus written before 1917 and provided in modern orthography.
- old-orthography - 108 k tokens. Texts of the 19th - early 20th cc. in pre-revolutionary orthography (S. T. Aksakov, P. A. Kulish, M. Pogodin, A. Spaso-Kukotsky, N. I. Grech)
- old-orthography-XVIII - 6 k tokens. 18th century texts in old orthography (by Peter the Great, S. Pufendorf, P. I. Pogoretsky, F. A. Emin)

As for historical Russian data (1400-1700s), we used official legal and business writing texts, as the other RNC Middle Russian collections, like vernacular gramotki, were distinctly different in the occurrences of old grammatical forms and constructions, in phonetic features reflected in orthography, and in genre-specific lexical distributions. We split the taken texts into two datasets:

- LEG(acy) texts written in 15th – 17th cc. (ca. 1.1 M tokens), and
- Bezobrazov - recently added to the RNC texts of the latter half of the 17th c. from Bezobrazov's archive (500 k tokens).

Table 1 summarises the size of the development and test data used in experiments. In the experiments reported below, the models were trained on a joined modern Russian training dataset (1700-2020s) or historical Russian data (1400-1700s).

All data are presented in the CONLL-U format and annotated according to the Russian UD-Ext scheme (Lyashevskaya, 2019). This scheme assumes the use of a standard inventory of the UD-Russian dependency relations and common RNC and UD policy for lemmatisation. Enhanced dependency relations are not provided. To make morphological annotations of the RNC Main corpus and Russian UD compatible,

the following features are added to the GramEval2020 and SynTagRus data and used in all new datasets:

- parts of speech: PRED for predicatives (eg. *можно, холодно, жаль*), ADVPRO for pronominal adverbs (eg. *тут*), PREDPRO for pronominal predicatives (eg. *некого*), PARENTH for parentheticals (eg. *конечно*), ANUM for ordinal numerals (eg. *второй*).
- grammatical features: Transit={Tran,Intr} for transitivity, Case={Acc2,Loc2} for secondary cases, Degree=Cmp2 for comparatives with the prefix *по-*, Anom=Yes for anomalous forms.

PoS-tags that are absent from the UD format were added by automatic replacement with the use of wordlists. Some PoS-tags were added manually, e.g. ANUM for numerals written with numbers, PRED for ambiguous words. PoS-tag disambiguation (e.g. *холодно* - ADV vs. ADJ vs. PRED; мало NUM vs. ADVPRO vs. PRED) and corresponding correction of dependency relations were performed manually. Necessary grammatical features were corrected or added using the wordlists and lists of tokens with manual correction. The transitivity feature was manually checked in context with the dependency relations correction.

4 Rubic: a Model for Tagging and Parsing

The study is divided into the following parts. In the first one we examine the previous results of the GramEval-2020 shared task. From this data, we form our expectations for the next suitable model to achieve in morphological tagging, lemmatisation, and dependency parsing. The second stage of the research is the description of the new model, and its results on the GramEval data. In some tasks, the model is challenged by the other models, specifically trained for this task on the particular dataset, to explore the possible enhancements. The third part of the study is dedicated to the analysis of the key errata that the proposed model makes, and whether the other models struggle with the same issues.

The model that we are starting with, our baseline, is the one that has been previously used for the annotation of the RNC corpus data, qbic (Anastasyev, 2020), a winner of the GramEval-2020 shared task. Qbic is a RuBERT encoder accompanied by three classifier decoders performing the part-of-speech classification, lemmatisation, and dependency parsing, respectively. Lemmatisation is conducted in two stages, with the classifier assigning the particular rule to a token, after which the rules themselves are applied. Each lemmatisation rule specifies the number of characters to be cut and a combination of characters to be added, thus comprising a total of 1000 to 2000 rules, depending on the amount of training data (cf. also “less than 1,000 classes of rules in total” in (Michurina et al., 2021)). The rules form in the following manner:

- Training set yields sequences of transformations that are required to transform a token into its lemma (delete postfix/suffix of a certain length > add some sequence of characters to the end > capitalise/decapitalise)
- We take the sequences of transformations that are met more than 3 times (to exclude noise)
- The remaining sequences become rules

Table 2 shows the performance of qbic on the re-annotated GramEval-2020 datasets. A standard CONLL18 script was used to calculate accuracy scores for parts of speech (PoS), morphological features, lemmas, and labeled attachment score for syntactic dependencies (LAS, basic relation inventory, ie. `nummod` and `nummod:gov` are considered the same). The model performed in a satisfactory way in most of the aspects. However, its performance on dependency parsing was below expectations. Non-standard patterns in poetry, social media texts, and wiki presented an especially hard challenge for it. Additionally, qbic was not robust in full morphological tagging and lemmatisation in the case of social media, poetry, diaries, and encyclopedic texts, which contain abbreviations, non-standard punctuation, transcript notes, rare named entities, and especially in the case of the RNC subcorpus of older orthographies (ca. 13M tokens).

To meet this challenge, we present Rubic, a model that utilises the same architecture as qbic, with enhancements, see Figure 1. For an encoder, we use sberbank-ai/ruBert pretrained on 30 GB data. In our model, the lemmatisation module receives additional information from the part-of-speech tagging classifier. Rubic checks lemma candidates against a supplementary dictionary compiled manually. The dictionary is a pair of lemma and part of speech, split by tab, e.g. *автоматизм NOUN*. Besides that,

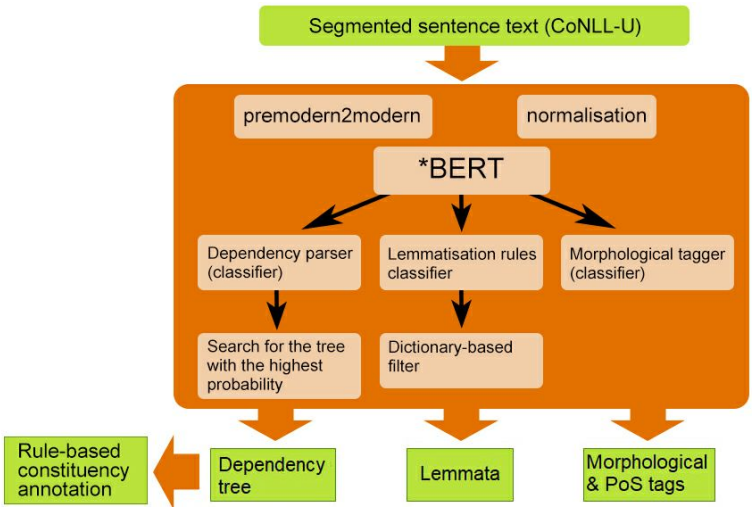

Figure 1: Key principles and architecture of Rubic.

Dataset	fiction	news	poetry	social	wiki
PoS	98.0	96.6	96.9	94.7	92.7
Morph.features	98.7	96.1	96.7	94.7	94.4
Lemmatisation	98.0	98.2	95.3	96.0	93.6
LAS	89.6	91.2	81.4	80.7	78.1

Table 2: Accuracy score of qbic on GramEval-2020 dataset, %

the symbol sequences unlikely to occur in Russian texts are preprocessed. We specifically set up Rubic to process data with non-standard orthography by implementing a graphic premodern2modern heuristic, and mapping the tokens in older orthography to tokens in modern orthography.

We perform data augmentation when training Rubic. We use the calculation of "the lexical usefulness weight" that prioritise the use of rare tokens for the further pipeline of data augmentation. If a sentence contains two, and exactly two quotation marks, we add another sentence to the dataset, that contains guillemets instead (we add 450 sentences via this heuristic). We use the heuristic of jo-fication, transforming e into $\ddot{e}$, in words, where it is possible (we add more than 800 sentences via this heuristic). We use the capitalisation heuristic, when the tokens are randomly capitalised for the purposes of better recognition (we acquire nearly 2000 additional sentences via this heuristic. %; we take only 20% of the sentences, generated by the previous heuristic).

With all these enhancements, the results of the model expectedly grow. We provide the difference between accuracy scores in Table 3. Rubic improves in parsing, and some improvements can be seen in tagging and lemmatisation. It underperforms on the fiction dataset, and wiki morphology presents it with some challenges. All this may also signal about overfitting, so we use the other datasets of the modern Russian language: CONLL18, and IWPT21. The results are presented in Table 4.

We also evaluated Rubic on the RNC test sets prepared specifically for the task of full corpus re-annotation. The results are shown in Table 5. In all datasets, Rubic performs well on major and most frequent part of speech categories such as verbs, nouns, proper nouns, prepositions, and coordinate conjunctions. Noun case accuracy is above 98% in all datasets except poetry and old orthography-XVIII. Mixing adjectives vs. participles, adjectives vs. adverbs is higher in the latter datasets and Taiga. Annotation of predicatives and corresponding syntactic structures is problematic in poetry, fiction and non-fiction written in the 20th c. and earlier, in which a wider variety of constructions and lexical fillers is available. Expectedly, parsing quality drops on longer sentences, and non-standard symbols, non-

Dataset	fiction	news	poetry	social	wiki
PoS	+0.1	+1.4	+1.7	+1.0	+0.5
Morph.features	-0.1	+0.3	+0.1	+0.6	-0.4
Lemmatisation	-0.3	+0.0	+0.2	+0.6	+0.5
LAS	+0.5	+0.8	+1.3	+0.3	+2.8

Table 3: Change in accuracy score for Rubic compared to qbic, %, GramEval-2020 datasets

Dataset	CONLL18	IWPT21
PoS	99.23	99.14
Morph.features	98.27	98.19
Lemmatisation	97.49	97.83
LAS	95.51	95.47

Table 4: Accuracy score of Rubic on standard modern Russian datasets, %

standard place of punctuation marks and other non-letters, and out-of-vocabulary abbreviations misleads the model.

5 Lemmatisation: Further Experiments

Rubic, thus, does not overfit for GramEval-2020 datasets. However, we wanted to see if there is a possibility to enhance its performance. To test this, we picked the lemmatisation task and trained two BART-large-based lemmatiser models (Lewis et al., 2020). This is a sequence-to-sequence state-of-the-art multilingual method that can help to reveal critical points in which Rubic needs enhancement.

The comparison is based on the following data: modern RNC datasets, historical LEG and Bezobrazov datasets. Both Rubic and BART-large were separately fine-tuned for modern and historical data. The results of comparison between BART-large and Rubic are in Table 6.

The news dataset witnesses a better performance of Rubic, by 0.1 per cent: the Rubic heuristics adapt the model for the specific language variety. However, it seems that the texts of the Middle Russian period require much more intricate heuristics, which leads to the striking 12 to 20, depending on data quality, per cent difference between BART-large and Rubic accuracy in favour of the former. Overall, BART-large beats Rubic by a significant margin of 0.4 to 3 per cent. The main challenges are non-standard orthography and syntactic structures of XIX century poetry, which encourage a more generalising approach of BART-large.

The Rubic model, despite implemented heuristics, is challenged by two main classes of words: non-productive verb models (*скорбатъ* instead of *скорбетъ* 'mourn'), and proper names (*Любовя* instead of *Любовъ* 'Lyubov'). The non-standard modern orthography also takes its toll: *наср@ла* is returned instead of *насратъ* 'do not give a damn about smth' likely due to the special symbol that was not normalised. Sometimes model generates empty lemmata, due to the rule-based nature of its lemmatiser module.

BART-large sequence-to-sequence architecture helps to deal with the aforementioned problems. It still overgeneralises, creating the syntagmae, similar to *-жо-* in verbs (*ожотъся* instead of *ожечься* 'get fired by'), or choosing the more general ending, completely confusing the word class, cf. *Стоцка* instead of *Стоцкая* 'Stotskaja'. Generalisation also leads into the model being unable to deal with orthography issues (odd *с* in *естесственный* 'natural'; odd *о* in *-пр-*, cf. *предупорежденне* instead of *предупреждение* 'warning'). Probably, the same factor leads to the appearance of hyphens in lemmas for the words that were transitioned from string to string somewhere in the data, sometimes with character replacing, for instance, in *пеп-льница* instead of *пепельница* 'ashpot'. Compound pronouns, such as *ни о чём* 'about nothing', often lose their negative particle (*ни*) part. The words that contain similar

Dataset	Taiga	newspa-pers-XXI	prose -XX	prose -XIX	poetry -XIX	old ortho-graphy	old ortho-graphy-XVIII
PoS	97.8	99.0	98.9	99.2	97.4	98.9	95.8
Morph.features	94.6	97.3	97.2	97.7	94.2	95.9	90.1
Lemmatisation	97.6	99.1	98.3	98.9	95.9	97.5	93.7
LAS	85.7	95.1	94.1	94.6	85.6	94.0	83.7

Table 5: The accuracy score of Rubic on RNC datasets, %

Dataset	Rubic, accuracy, %	BART-large, accuracy, %
Taiga	97.6	**98.0**
newspapers-XXI	**99.1**	99.0
prose-XX	98.3	**98.7**
prose-XIX	98.9	**99.3**
poetry-XIX	95.9	**98.9**
old orthography	97.4	**98.7**
old orthography-XVIII	93.7	**93.8**
LEG(al) test, 1400-1700	85.4	**98.0**
Bezobrazov	73.8	**85.0 (92.6 with normalisation)**

Table 6: Lemmatisation accuracy scores for Rubic and BART-large models on RNC datasets. The best results are highlighted in bold.

syllables, such as *царица* 'empress', are often reduced to a single syllable, in this case, *ца*: probably, the original BART-large dataset was trained to eliminate reduplication. The model clearly lacks knowledge of how the lemmas in particular language should look, which leads to generating adjective lemmas that after the adjectival affix *-ск-* have *-ив-* instead of *-ий-*. The model often does not pay attention to the morphology tagging (generated verb lemmas with *Aspect=Perf* tag often contain *-ывать*, which is a strong marker of continuous aspect in Russian verbs; prefix *по-* for *Degree=Cmp2* adjectives generated lemmas).

BART-large experiments show that sequence-to-sequence is not a necessarily ideal solution. It appears to be slow when annotating large amount of texts. However, this method reveals room for improvement of models like Rubic, particularly when it concerns the dataset construction, non-standard orthography, and low-productive paradigms, such as proper names and some verb classes. We are going to dedicate further research to these particular issues.

6 Corpus annotation and future development

At the moment, Main corpus, Regional Media, and Educational corpora are annotated by Rubic. In order to make it easier for users to switch from the old version to the new one, two lemma layers – annotations provided by Mystem and Rubic – are searchable. By default, the search is conducted on the layer automatically disambiguated by Rubic only.

We decided to apply three techniques to improve the Rubic outcome. Firstly, although the neural model is set up to produce only one analysis per token, in the case of theoretically plausible equivalent linguistic interpretations (eg. adjective vs. participle, see the practice of the manually disambiguated RNC subcorpus) additional morphological and lexical analyses were provided by rules. Secondly, lemmas that occur 30 times and more in the corpus and are not found in the Mystem dictionary, were checked and corrected manually. Thirdly, a number of heuristics were applied to the dependency annotations to provide search by constituency types and unlabeled tree configurations (eg. search within subordinate clauses; within participial phrases; search words that do not have dependents).

In the future, based on the results of the users' feedback, more disambiguated RNC corpora will be made available, with necessary adjustments in the annotation methods. RNC services such as frequency lists, graphs by year, lemma-based corpus portraits and comparison, collocation tools, Word at a glance sketch tool, and search by lexico-semantic features, depend critically on the quality of data lemmatisation. More work should be done in terms of finding new text classes on which the models underperform and adding relevant excepts to training; balancing the training collection by text types; balancing learning rate for different task. Decoding of abbreviated words is likely to be formulated as a separate since the distribution of such forms in large corpora cannot be modeled in the same way as lemmatisation rules.

The project's repository containing supplementary materials is available at: `https://github.com/olesar/RNC2.0`.

Acknowledgements

This work was carried out within the framework of the grant from the Ministry of Science and Higher Education of the Russian Federation within Agreement No. 075-15-2020-793: "Next-generation computational linguistics platform for the Russian language digital recording: infrastructure, resources, research".

References

Itziar Aduriz, Izaskun Aldezabal, Iñaki Alegria, Xabier Artola, Nerea Ezeiza, and Ruben Urizar. 1996. Euslem: A lemmatiser/tagger for basque. // Martin Gellerstam, Jerker Järborg, Sven-Göran Malmgren, Kerstin Norén, Lena Rogström, and Catalina Röjder Papmehl, *Proceedings of the 7th EURALEX International Congress*, P 27–35, Göteborg, Sweden, aug. Novum Grafiska AB.

Iago Alonso-Alonso, David Vilares, and Carlos Gómez-Rodríguez. 2022. The fragility of multi-treebank parsing evaluation. // *Proceedings of the 29th International Conference on Computational Linguistics*, P 5345–5359, Gyeongju, Republic of Korea, October. International Committee on Computational Linguistics.

Daniil Anastasyev. 2020. Exploring pretrained models for joint morphosyntactic parsing of Russian. // *Computational Linguistics and Intellectual Technologies: Papers from the Annual Conference "Dialogue"*, volume 19, P 1–12.

Ali Basirat and Joakim Nivre. 2021. Syntactic nuclei in dependency parsing – a multilingual exploration. // *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, P 1376–1387, Online, April. Association for Computational Linguistics.

Aleksandrs Berdičevskis, Hanna Eckhoff, and Tatiana Gavrilova. 2016. The beginning of a beautiful friendship: rule-based and statistical analysis of Middle Russian. // *Komp'yuternaya lingvistika i intellektual'nye tekhnologii. Trudy mezhdunarodnoj konferencii «Dialog»*, P 99–111, Moscow, Russia. RSSU.

Toms Bergmanis and Sharon Goldwater. 2018. Context sensitive neural lemmatization with Lematus. // *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, P 1391–1400, New Orleans, Louisiana, June. Association for Computational Linguistics.

Frederic Blum. 2022. Evaluating zero-shot transfers and multilingual models for dependency parsing and POS tagging within the low-resource language family tupían. // *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, P 1–9, Dublin, Ireland, May. Association for Computational Linguistics.

Elena I Bolshakova and Alexander S Sapin. 2022. Building a combined morphological model for Russian word forms. // *Analysis of Images, Social Networks and Texts: 10th International Conference, AIST 2021, Tbilisi, Georgia, December 16–18, 2021, Revised Selected Papers*, P 45–55. Springer.

Antònio Branco and Jõao Silva. 2003. Portuguese specific issues in the rapid development of state of the art taggers. // *Proceedings of the Workshop on Language Technology for Normalisation of Less-Resourced Languages (SALTMIL8/AfLaT2012)*, P 7–9, Paris. European Language Resources Association.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder–decoder for statistical machine translation. // *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, P 1724–1734, Doha, Qatar, October. Association for Computational Linguistics.

Evelien de Graaf, Silvia Stopponi, Jasper K. Bos, Saskia Peels-Matthey, and Malvina Nissim. 2022. AGILe: The first lemmatizer for Ancient Greek inscriptions. // *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, P 5334–5344, Marseille, France, June. European Language Resources Association.

Miryam de Lhoneux, Sheng Zhang, and Anders Søgaard. 2022. Zero-shot dependency parsing with worst-case aware automated curriculum learning. // *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, P 578–587, Dublin, Ireland, May. Association for Computational Linguistics.

Bill Dyer. 2022. New syntactic insights for automated Wolof Universal Dependency parsing. // *Proceedings of the Fifth Workshop on the Use of Computational Methods in the Study of Endangered Languages*, P 5–12, Dublin, Ireland, May. Association for Computational Linguistics.

Chloe Eggleston and Brendan O'Connor. 2022. Cross-dialect social media dependency parsing for social scientific entity attribute analysis. // *Proceedings of the Eighth Workshop on Noisy User-generated Text (W-NUT 2022)*, P 38–50, Gyeongju, Republic of Korea, October. Association for Computational Linguistics.

Amany Fashwan and Sameh Alansary. 2022. Developing a tag-set and extracting the morphological lexicons to build a morphological analyzer for Egyptian Arabic. // *Proceedings of the The Seventh Arabic Natural Language Processing Workshop (WANLP)*, P 142–160, Abu Dhabi, United Arab Emirates (Hybrid), December. Association for Computational Linguistics.

Laura García Fernández. 2020. A contribution to old english lexicography. *NOWELE / North-Western European Language Evolution*, 73(2):236–251.

Björn Gambäck. 2012. Tagging and verifying an amharic news corpus. // *Proceedings of the Workshop on Language Technology for Normalisation of Less-Resourced Languages (SALTMIL8/AfLaT2012)*, P 79–84, Paris. European Language Resources Association.

Leilei Gan, Yuxian Meng, Kun Kuang, Xiaofei Sun, Chun Fan, Fei Wu, and Jiwei Li. 2022. Dependency parsing as MRC-based span-span prediction. // *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, P 2427–2437, Dublin, Ireland, May. Association for Computational Linguistics.

Yoav Goldberg and Jon Orwant. 2013. A dataset of syntactic-ngrams over time from a very large corpus of english books. // *Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity*, P 241–247.

Omer Goldman, David Guriel, and Reut Tsarfaty. 2022. (un)solving morphological inflection: Lemma overlap artificially inflates models' performance. // *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, P 864–870, Dublin, Ireland, May. Association for Computational Linguistics.

Emily Goodwin, Siva Reddy, Timothy O'Donnell, and Dzmitry Bahdanau. 2022. Compositional generalization in dependency parsing. // *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, P 6482–6493, Dublin, Ireland, May. Association for Computational Linguistics.

Michael Hann. 1974. Principles of automatic lemmatisation. *ITL Review of Applied Linguistics*, 23(1):3–22.

Ayyoob ImaniGooghari, Silvia Severini, Masoud Jalili Sabet, François Yvon, and Hinrich Schütze. 2022. Graph-based multilingual label propagation for low-resource part-of-speech tagging. // *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, P 1577–1589, Abu Dhabi, United Arab Emirates, December. Association for Computational Linguistics.

Go Inoue, Salam Khalifa, and Nizar Habash. 2022. Morphosyntactic tagging with pre-trained language models for arabic and its dialects. // *Proceedings of the Findings of the Association for Computational Linguistics: ACL2022*, Dublin, Ireland, May. Association for Computational Linguistics.

Ákos Kádár, Lan Xiao, Mete Kemertas, Federico Fancellu, Allan Jepson, and Afsaneh Fazly. 2021. Dependency parsing with structure preserving embeddings. // *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, P 1684–1697, Online, April. Association for Computational Linguistics.

Jenna Kanerva, Filip Ginter, and Tapio Salakoski. 2021. Universal lemmatizer: A sequence-to-sequence model for lemmatizing universal dependencies treebanks. *Natural Language Engineering*, 27(5):545–574.

Dan Kondratyuk. 2019. Cross-lingual lemmatization and morphology tagging with two-stage multilingual BERT fine-tuning. // *Proceedings of the 16th Workshop on Computational Research in Phonetics, Phonology, and Morphology*, P 12–18, Florence, Italy, August. Association for Computational Linguistics.

Nathaniel Krasner, Miriam Wanner, and Antonios Anastasopoulos. 2022. Revisiting the effects of leakage on dependency parsing. // *Findings of the Association for Computational Linguistics: ACL 2022*, P 2925–2934, Dublin, Ireland, May. Association for Computational Linguistics.

C S Ayush Kumar, Advaith Maharana, Srinath Murali, Premjith B, and Soman Kp. 2022. BERT-based sequence labelling approach for dependency parsing in Tamil. // *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*, P 1–8, Dublin, Ireland, May. Association for Computational Linguistics.

Anna Langedijk, Verna Dankers, Phillip Lippe, Sander Bos, Bryan Cardenas Guevara, Helen Yannakoudakis, and Ekaterina Shutova. 2022. Meta-learning for fast cross-lingual adaptation in dependency parsing. // *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, P 8503–8520, Dublin, Ireland, May. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. // *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, P 7871–7880, Online, July. Association for Computational Linguistics.

Ying Li, Shuaike Li, and Min Zhang. 2022. Semi-supervised domain adaptation for dependency parsing with dynamic matching network. // *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, P 1035–1045, Dublin, Ireland, May. Association for Computational Linguistics.

Olga Lyashevskaya. 2019. A reusable tagset for the morphologically rich language in change: A case of Middle Russian. // *Komp'juternaja Lingvistika i Intellektual'nye Tehnologii*, P 422–434.

Mariia Michurina, Alexandra Ivoylova, Nikolay Kopylov, and Daniil Selegey. 2021. Morphological annotation of social media corpora with reference to its reliability for linguistic research. // *Komp'juternaja Lingvistika i Intellektual'nye Tehnologii*, P 492–504.

Kirill Milintsevich and Kairit Sirts. 2021. Enhancing sequence-to-sequence neural lemmatization with external resources. // *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, P 3112–3122, Online, April. Association for Computational Linguistics.

Saliha Muradoglu and Mans Hulden. 2022. Eeny, meeny, miny, moe. how to choose data for morphological inflection. // *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, P 7294–7303, Abu Dhabi, United Arab Emirates, December. Association for Computational Linguistics.

Zuzana Nevěřilová. 2022. Compressed FastText Models for Czech Tagger. // *Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2022*, P 79–87, Tribun EU. European Language Resources Association.

Ossama Obeid, Go Inoue, and Nizar Habash. 2022. Camelira: An Arabic multi-dialect morphological disambiguator. // *Proceedings of the The 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, P 319–326, Abu Dhabi, UAE, December. Association for Computational Linguistics.

Peng Qi, Timothy Dozat, Yuhao Zhang, and Christopher D. Manning. 2018. Universal Dependency parsing from scratch. // *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, P 160–170, Brussels, Belgium, October. Association for Computational Linguistics.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. // *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, P 101–108, Online, July. Association for Computational Linguistics.

Yves Scherrer, 2021. *Adaptation of Morphosyntactic Taggers*, P 138–166. Studies in Natural Language Processing. Cambridge University Press.

Ilya Segalovich. 2003. A fast morphological algorithm with unknown word guessing induced by a dictionary for a web search engine. // *MLMTA*, P 273–280, 01.

Freda Shi, Kevin Gimpel, and Karen Livescu. 2022. Substructure distribution projection for zero-shot cross-lingual dependency parsing. // *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, P 6547–6563, Dublin, Ireland, May. Association for Computational Linguistics.

Peter Spyns. 1996. A tagger/lemmatiser for Dutch medical language. // *Proceedings of the 16th Conference on Computational Linguistics - Volume 2*, COLING '96, P 1147–1150, USA. Association for Computational Linguistics.

Milan Straka and Jana Straková. 2017. Tokenizing, POS tagging, lemmatizing and parsing UD 2.0 with UDPipe. // *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, P 88–99, Vancouver, Canada, August. Association for Computational Linguistics.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. // *Advances in neural information processing systems*, P 3104–3112.

Yuanhe Tian, Yan Song, and Fei Xia. 2022. Enhancing structure-aware encoder with extremely limited data for graph-based dependency parsing. // *Proceedings of the 29th International Conference on Computational Linguistics*, P 5438–5449, Gyeongju, Republic of Korea, October. International Committee on Computational Linguistics.

Alymzhan Toleu, Gulmira Tolegen, and Rustam Mussabayev. 2022. Language-independent approach for morphological disambiguation. // *Proceedings of the 29th International Conference on Computational Linguistics*, P 5288–5297, Gyeongju, Republic of Korea, October. International Committee on Computational Linguistics.

Ahmet Üstün, Arianna Bisazza, Gosse Bouma, and Gertjan van Noord. 2020. UDapter: Language adaptation for truly Universal Dependency parsing. // *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, P 2302–2315, Online, November. Association for Computational Linguistics.

Adam Wiemerslage, Miikka Silfverberg, Changbing Yang, Arya D. McCarthy, Garrett Nicolai, Eliana Colunga, and Katharina Kann. 2022. Morphological processing of low-resource languages: Where we are and what's next. // *Findings of the Association for Computational Linguistics: ACL 2022*, P 988–1007. Association for Computational Linguistics.

Ziyao Xu, Houfeng Wang, and Bingdong Wang. 2022. Multi-layer pseudo-Siamese biaffine model for dependency parsing. // *Proceedings of the 29th International Conference on Computational Linguistics*, P 5476–5487, Gyeongju, Republic of Korea, October. International Committee on Computational Linguistics.

Songlin Yang and Kewei Tu. 2022. Combining (second-order) graph-based and headed-span-based projective dependency parsing. // *Findings of the Association for Computational Linguistics: ACL 2022*, P 1428–1434, Dublin, Ireland, May. Association for Computational Linguistics.

AI Zobnin and GV Nosyrev. 2015. Morfologicheskij analizator MyStem 3.0. *Trudy Instituta russkogo yazyka im. VV Vinogradova*, 6:300–310.

Abstracts

RECEIPT-AVQA-2023 CHALLENGE

Begaev A., **Orlov E.**, Budapest, Hungary

In this work, we introduce a new challenging Document VQA dataset, named Receipt AVQA, and present the results of the associated RECEIPT-AVQA-2023 shared task. Receipt AVQA is comprised of 21,835 questions in English over 1,957 receipt images. The receipts contain a lot of numbers, which means discrete reasoning capability is required to answer the questions. The associated shared task has attracted 4 teams that have managed to beat an extractive VQA baseline in the final phase of the competition. We hope that the published dataset and promising results of the contestants will inspire further research on understanding documents in scenarios that require discrete reasoning.

CONSTRUCTING A SEMANTIC CORPUS FOR RUSSIAN: SEMONTOCOR

Boguslavsky I. M.[1,2], **Dikonov V. G.**[1], **Inshakova E. S.**[1], **Iomdin L. L.**[1], **Lazursky A. V.**[1], **Rygaev I. P.**[1], **Timoshenko S. P.**[1], **Frolova T. I.**[1], [1]A. A. Kharkevich Institute for Information Transmission Problems, Moscow, Russia; [2]Universidad Politécnica de Madrid, Madrid, Spain

The SemOntoCor project focuses on creating a semantic corpus of Russian based on linguistic and ontological resources. It is a satellite project with regard to a semantic parser (SemETAP) being developed, the latter aiming at producing semantic structures and drawing various types of inferences. SemETAP is used to annotate SemOntoCor in a semi-automatic mode, whereupon SemOntoCor, when reaching sufficient maturity, will help create new parsers and other semantic applications. SemOntoCor can be viewed as a further step in the development of SynTagRus with its several layers of annotation. SemOntoCor builds on top of the morpho-syntactic annotation of SynTagRus and assigns each sentence a Basic Semantic Structure (BSemS). BSemS represents the direct layer of meaning of the sentence in terms of ontological concepts and semantic relations between them. It abstracts away from lexico-syntactic variation and in many cases decomposes lexical meanings into smaller elements. The first phase of SemOntoCor consists in annotating a Russian translation of the novel "The Little Prince" by Antoine de Saint-Exupery (1532 sentences, 13120 tokens).

PSEUDO-LABELLING FOR AUTOREGRESSIVE STRUCTURED PREDICTION IN COREFERENCE RESOLUTION

Bolshakov V.[1,2], **Mikhaylovskiy N.**[1,3], [1]NTR Labs; [2]BMSTU, Moscow, Russia; [3]Higher IT School, Tomsk State University, Tomsk, Russia

Coreference resolution is an important task in natural language processing, since it can be applied to such vital tasks as information retrieval, text summarization, question answering, sentiment analysis and machine translation. In this paper, we present a study on the effectiveness of several approaches to coreference resolution, focusing on the RuCoCo dataset as well as results of participation in the Dialogue Evaluation 2023. We explore ways to increase the dataset size by using pseudo-labelling and data translated from another language. Using such technics we managed to triple the size of dataset, make it more diverse and improve performance of autoregressive structured prediction (ASP) on coreference resolution task. This approach allowed us to achieve the best results on RuCoCo private test with increase of F1-score by 1.8, Precision by 0.5 and Recall by 3.0 points compared to the second-best leaderboard score. Our results demonstrate the potential of the ASP model and the importance of utilizing diverse training data for coreference resolution.

LIGHT COREFERENCE RESOLUTION FOR RUSSIAN WITH HIERARCHICAL DISCOURSE FEATURES

Chistova E. V., **Smirnov I. V.**, FRC CSC RAS, Moscow, Russia

Coreference resolution is the task of identifying and grouping mentions referring to the same real-world entity. Previous neural models have mainly focused on learning span representations and pairwise scores for coreference decisions. However, current methods do not explicitly capture the referential choice in the hierarchical discourse, an important factor in coreference resolution. In this study, we propose a new approach that incorporates rhetorical information into neural coreference resolution models. We collect rhetorical features from automated discourse parses and examine their impact. As a base model, we implement an end-to-end span-based coreference resolver using a partially fine-tuned multilingual entity-aware language model LUKE. We evaluate our method on the RuCoCo-23 Shared Task for coreference resolution in Russian. Our best model employing rhetorical distance between mentions has ranked 1st on the development set (74.6% F1) and 2nd on the test set (73.3% F1) of the Shared Task. We hope that our work will inspire further research on incorporating discourse information in neural coreference resolution models.

PARTITIVE GENITIVE IN RUSSIAN: DICTIONARY AND CORPUS DATA

Chuikova O. Iu., Herzen State Pedagogical University of Russia, St. Petersburg, Russia

The paper aims at comprehensive analysis of the verbs compatible with the partitive genitive object. Based on the Dictionary of Russian Language, the list of perfective verbal lexemes that are able to take the genitive object is compiled and semantic features that unite these verbs are revealed. The features are divided into two groups: aspectually relevant features and aspectually irrelevant features. The corpus-based analysis of the use of the verbs that take both genitive and accusative objects makes it possible to identify features that increase the likelihood of certain object case-marking.

BIMODAL SENTIMENT AND EMOTION CLASSIFICATION WITH MULTI-HEAD ATTENTION FUSION OF ACOUSTIC AND LINGUISTIC INFORMATION

Dvoynikova A. A., **Karpov A. A.**, St. Petersburg Federal Research Center of the Russian Academy of Sciences, Saint-Petersburg, Russia

This article describes solutions to couple of problems: CMU-MOSEI database preprocessing to improve data quality and bimodal multitask classification of emotions and sentiments. With the help of experimental studies, representative features for acoustic and linguistic information are identified among pretrained neural networks with Transformer architecture. The most representative

features for the analysis of emotions and sentiments are EmotionHuBERT and RoBERTa for audio and text modalities respectively. The article establishes a baseline for bimodal multitask recognition of sentiments and emotions – 63.2% and 61.3%, respectively, measured with macro F-score. Experiments were conducted with different approaches to combining modalities – concatenation and multi-head attention. The most effective architecture of neural network with early concatenation of audio and text modality and late multi-head attention for emotions and sentiments recognition is proposed. The proposed neural network is combined with logistic regression, which achieves 63.5% and 61.4% macro F-score by bimodal (audio and text) multi-tasking recognition of 3 sentiment classes and 6 emotion binary classes.

INTRODUCTION MODEL IN RUSSIAN «PEAR REPORTAGES»: THE ROLE OF COMMON GROUND

Fedorova O. V. , Lomonosov Moscow State University, Moscow, Russia

In this study, the peculiarities of the character introduction in the genre of live reportage were studied. The participants were 25 students oh the Lomonosov Moscow State University. Speech production was elicited by means of the “Pears Film” by W. Chafe. Different types of the collective common ground were considered. It turned out that, unlike narratives of other genres, the chronological scale is more important for the introduction than the status scale. It was also shown that the collected reportages from the point of view of the introduction peculiarities are more similar to classical retellings than to the sports reportages.

FOREGROUND AND BACKGROUND IN RUSSIAN SIGN LANGUAGE NARRATIVES: THE ROLE OF ASPECT AND ACTIONALITY

Filimonova E. V., Russian State University for the Humanities; Institute of linguistics, Russian Academy of Sciences, Moscow, Russia

The paper explores the role of aspect and actionality in foregrounding and backgrounding of clauses in Russian Sign Language narratives. Corpus study shows similarities to functions of aspectual markers and actionality in spoken languages. Besides grammatical markers and predicate types, non-manual marking and prosodic features of verbal sign can contribute to clause foregrounding and backgrounding.

MULTIMODAL DISCOURSE TREES IN FORENSIC LINGUISTICS

Galitsky B. A.[1], **Ilvovsky D. A.**[2], **Goncharova E. F.**[2,3], [1]Knowledge Trail Inc., San Jose, CA, USA; [2]NRU HSE; [3]AIRI, Moscow, Russia

We extend the concept of a discourse tree (DT) in the discourse representation of text towards data of various forms and natures. The communicative DT to include speech act theory, extended DT to ascend to the level of multiple documents, entity DT to track how discourse covers various entities were defined previously in computational linguistics, we now proceed to the next level of abstraction and formalize discourse of not only text and textual documents but also various kinds of accompanying data. We call such discourse representation Multimodal Discourse Trees (MMDTs). The rational for that is that the same rhetorical relations that hold between text fragments also hold between data values, sets and records, such as Reason, Cause, Enablement, Contrast, Temporal sequence. MMDTs are evaluated with respect to the accuracy of recognition of criminal cases when both text and data records are available. MMDTs are shown to contribute significantly to the recognition accuracy in cases where just keywords and syntactic signals are insufficient for classification and discourse-level information needs to be involved.

INCREMENTAL TOPIC MODELING FOR SCIENTIFIC TREND TOPICS EXTRACTION

Gerasimenko N.[1,2], **Chernyavskiy A.**[3], **Nikiforova M.**[1], **Ianina A.**[4], **Vorontsov K.**[2,4], [1]Sberbank, [2]MSU Institute for Artificial Intelligence, [3]National Research University Higher School of Economics, [4]Moscow Institute of Physics and Techonology (MIPT)

Rapid growth of scientific publications and intensive emergence of new directions and approaches poses a challenge to the scientific community to identify trends in a timely and automatic manner. We denote trend as a semantically homogeneous theme that is characterized by a lexical kernel steadily evolving in time and a sharp, often exponential, increase in the number of publications. In this paper, we investigate recent topic modeling approaches to accurately extract trending topics at an early stage. In particular, we customize the standard ARTM-based approach and propose a novel incremental training technique which helps the model to operate on data in real-time. We further create the Artificial Intelligence Trends Dataset (AITD) that contains a collection of early-stage articles and a set of key collocations for each trend. The conducted experiments demonstrate that the suggested ARTM-based approach outperforms the classic PLSA, LDA models and a neural approach based on BERT representations. Our models and dataset are open for research purposes.

FINE-TUNING TEXT CLASSIFICATION MODELS FOR NAMED ENTITY ORIENTED SENTIMENT ANALYSIS OF RUSSIAN TEXTS

Glazkova A., University of Tyumen, Tyumen, Russia

The paper presents an approach to named entity oriented sentiment analysis of Russian news texts proposed during the RuSentNE evaluation. The approach is based on RuRoBERTa-large, a pre-trained RoBERTa model for Russian. We compared several types of entity representation in the input text, and evaluated strategies for handling class imbalance and resampling entity tags in the training set. We demonstrated that some strategies improve the results of pre-trained models obtained on the dataset presented by the organizers of the evaluation.

ASPECT-BASED ARGUMENT GENERATION IN RUSSIAN

Goloviznina V. S., **Fishcheva I. N.**, **Peskisheva T. A.**, **Kotelnikov E. V.**, Vyatka State University, Kirov, Russia

The paper explores the argument generation in Russian based on given aspects. An aspect refers to one of the sides or property of the target object. Five aspects were considered: "Safety", "Impact on health", "Reliability", "Money", "Convenience and comfort". Various approaches were used for aspect-based generation: fine-tuning, prompt-tuning and few-shot learning. The ruGPT-3Large model was used for experiments. The results show that traditionally trained model (with fine-tuning) generates 51.6% of the arguments on given aspects, with the prompt-tuning approach – 33.9%, and with few-shot learning – 10.6%. The model also demonstrated the ability to generate arguments on new, previously unknown aspects.

RUSENTNE-2023: EVALUATING ENTITY-ORIENTED SENTIMENT ANALYSIS ON RUSSIAN NEWS TEXTS

Golubev A. A.[1], **Rusnachenko N. L.**[2], **Loukachevitch N. V.**[1], [1]Lomonosov Moscow State University, [2]Bauman Moscow State Technical University, Moscow, Russia

The paper describes the RuSentNE-2023 evaluation devoted to targeted sentiment analysis in Russian news texts. The task is to predict sentiment towards a named entity in a single sentence. The dataset for RuSentNE-2023 evaluation is based on the Russian news corpus RuSentNE having rich sentiment-related annotation. The corpus is annotated with named entities and sentiments towards these entities, along with related effects and emotional states. The evaluation was organized using the CodaLab competition framework. The main evaluation measure was macro-averaged measure of positive and negative classes. The best results achieved were of 66% Macro Fmeasure (Positive+Negative classes). We also tested ChatGPT on the test set from our evaluation and found that the zero-shot answers provided by ChatGPT reached 60% of the F-measure, which corresponds to 4th place in the evaluation. ChatGPT also provided detailed explanations of its conclusion. This can be considered as quite high for zero-shot application.

FREQUENCY DYNAMICS AS A CRITERION FOR DIFFERENTIATING INFLECTION AND WORD FORMATION (IN RELATION TO RUSSIAN ASPECTUAL PAIRS)

Gorbova E. V., independent researcher, **Chuikova O. Iu.**, Herzen State Pedagogical University of Russia

The paper reports the results of the critical evaluation of the quantitative approach to the distinction between inflection and word formation through the analysis of the trends in the frequency of word forms. The possibility of such analysis is provided by voluminous corpus data and tools for visualizing these trends. Both theoretical foundations of the proposed approach and the results of the pilot study of its applying to Russian aspectual triplets were considered. These cast doubt on the validity of distinguishing between inflection and word formation based on the trends in the frequency of word forms as a reliable tool used to reveal the unity or difference of lexical semantics and thus to define textual units as belonging to the same or different language units.

COMPUTER-ASSISTED DETECTION OF TYPOLOGICALLY RELEVANT SEMANTIC SHIFTS IN WORLD LANGUAGES

Gruntov I., Institute of Linguistics, Moscow, Russia, **Rykov E.**, HSE University, Moscow, Russia

The paper contains the description of a semi-authomatic method for the detection of typologically relevant semantic shifts in the world's languages. The algorithm extracts colexified pairs of meanings from polysemous words in digitised bilingual dictionaries. A machine learning classifier helps to separate those semantic shifts that are relevant to the lexical typology. Clustering is applied to group similar pairs of meanings into semantic shifts.

VAGUE REFERENCE IN EXPOSITORY DISCOURSE: MULTIMODAL REGULARITIES OF SPEECH AND GESTURE

Iriskhanova O.[1, 2], **Kiose M.**[1, 2], **Leonteva A.**[1, 2], **Agafonova O.**[1], [1]Moscow State Linguistic University; [2]Institute of Linguistics RAS, Moscow, Russia

The paper looks into the vague reference expressed in speech and gesture distribution in expository discourse. The research data are the monologues of 19 participants with total length of 2 hours 38 minutes. In these monologues, the use of vague reference (expressed in placeholders and approximators, with total amount of 2528) and functional gesture types (deictic, representational, pragmatic and adaptors, with total amount of 2309) was explored, with the aim of identifying the regular patterns of speech and gesture distribution and co-occurrence. The multimodal regularities include 1) the proportional frequency of four gesture types use equal to 6.8 / 14.4 / 28.7 / 50.1, which manifests overall distribution of co-speech gesture in expository discourse, 2) the significant difference in co-speech gesture use with placeholders and approximators which manifests itself in the use of three gesture types, adaptors, representational and pragmatic gestures, 3) the individually maintained significant difference in co-speech gesture use with placeholders and approximators which manifests itself in adaptors. These regularities can serve as predictors for identifying the specifics of vague reference in multimodal expository discourse.

A NEW DATASET FOR SENTENCE-LEVEL COMPLEXITY IN RUSSIAN

Ivanov V.[1, 2], **Elbayoumi M. G.**[2], [1]Kazan Federal University, Kazan, Russia; [2]Innopolis University, Innopolis, Russia

Text complexity prediction is a well-studied task. Predicting complexity sentence-level has attracted less research interest in Russian. One possible application of sentence-level complexity prediction is more precise and fine-grained modeling of text complexity. In the paper we present a novel dataset with sentence-level annotation of complexity. The dataset is open and contains 1,200 Russian sentences extracted from SynTagRus treebank. Annotations were collected via Yandex Toloka platform using 7-point scale. The paper presents various linguistic features that can contribute to sentence complexity as well as a baseline linear model.

THE PROBLEM OF LINGUISTIC MARKUP CONVERSION: THE TRANSFORMATION OF THE COMPRENO MARKUP INTO THE UD FORMAT

Ivoylova A. M.[1], **Dyachkova D. S.**[1], **Petrova M. A.**[2], **Michurina M. A.**[1], [1]RSUH; [2]A4 Technology, Moscow, Russia

The linguistic markup is an important NLP task. Currently, there are several popular formats of the markup (Universal Dependencies, Prague Dependencies, and so on), which are mostly focused on morphology and syntax. Full semantic markup can be found in the ABBYY Compreno model. However, the structure of the format differs significantly from the models mentioned above. In the given work, we convert the Compreno markup into the UD format, which is rather popular among NLP researchers, and enrich it with the semantical pattern.

Compreno and UD present morphology and syntax differently as far as tokenization, POS-tagging, ellipsis, coordination, and some other things are concerned, which makes the conversion of one format into another more complicated. Nevertheless, the conversion allowed us to create the UD-markup containing not only morpho-syntactic information but also the semantic one.

KNOWLEDGE TRANSFER BETWEEN TASKS AND LANGUAGES IN THE MULTI-TASK ENCODER-AGNOSTIC TRANSFORMER-BASED MODELS

Karpov D., **Konovalov V.**, MIPT, Dolgoprudny, Russia

We explore the knowledge transfer in the simple multi-task encoder-agnostic transformer-based models on five dialog tasks: emotion classification, sentiment classification, toxicity classification, intent classification, and topic classification. We show that these models' accuracy differs from the analogous single-task models by ~0.9%. These results hold for the multiple transformer backbones. At the same time, these models have the same backbone for all tasks, which allows them to have about 0.1% more parameters than any analogous single-task model and to support multiple tasks simultaneously. We also found that if we decrease the dataset size to a certain extent, multi-task models outperform singletask ones, especially on the smallest datasets. We also show that while training multilingual models on the Russian data, adding the English data from the same task to the training sample can improve model performance for the multi-task and single-task settings. The improvement can reach 4–5% if the Russian data are scarce enough. We have integrated these models to the DeepPavlov library and to the DREAM dialogue platform.

ATTENTION-BASED ESTIMATION OF TOPIC MODEL QUALITY

Kataeva V., **Khodorchenko M.**, ITMO University, St Petersburg, Russia

Topic modeling is an essential instrument for exploring and uncovering latent patterns in unstructured textual data, that allows researchers and analysts to extract valuable understanding of a particular domain. Nonetheless, topic modeling lacks consensus on the matter of its evaluation. The estimation of obtained insightful topics is complicated by several obstacles, the majority of which are summarized by the absence of a unified system of metrics, the one-sidedness of evaluation, and the lack of generalization. Despite various approaches proposed in the literature, there is still no consensus on the aspects of effective examination of topic quality. In this research paper, we address this problem and propose a novel framework for evaluating topic modeling results based on the notion of attention mechanism and Layer-wise Relevance Propagation as tools for discovering the dependencies between text tokens. One of our proposed metrics achieved a 0.71 Pearson correlation and 0.74 $_{K}$ correlation with human assessment. Additionally, our score variant outperforms other metrics on the challenging Amazon Fine Food Reviews dataset, suggesting its ability to capture contextual information in shorter texts.

FOREGROUNDING AND ACCESSIBILITY EFFECTS IN THE GAZE BEHAVIOR OF THE READERS WITH DIFFERENT COGNITIVE STYLE

Kiose M.[1,2], **Rzheshevskaya A.**[1], **Izmalkova A.**[3,1], **Makeev S.**[4], [1]Moscow State Linguistic University; [2]Institute of Linguistics RAS; [3]Higher School of Economics; [4]Lomonosov Moscow State University, Moscow, Russia

This paper explores accessibility effects in the gaze behavior of readers with different cognitive style, impulsive and reflective, as mediated by graphological and linguistic foregrounding in the discursive acts in 126 areas of interest (AOIs). The study exploits 1890 gaze behavior probes available at open access Multimodal corpus of oculographic reactions MultiCORText. We identified that while graphological foregrounding makes initial or final components of discursive act more accessible for the impulsive readers, reflective readers also observe the components within the act. Linguistic foregrounding produces higher access with impulsive readers in case the linguistic form is visually focalized (phonological foregrounding and parallel structures); meanwhile, with reflective readers this is the information density appearing in elliptical and one-component sentences which maintains higher access.

TOWARDS A RUSSIAN MULTIMEDIA POLITENESS CORPUS

Klokova K.[1], **Krongauz M.**[2], **Shulginov V.**[1,2], **Yudina T.**[1], [1]MIPT, [2]HSE, Moscow, Russia

Communication involves an exchange of information as well as the use of linguistic means to begin, sustain, and end conversations. Politeness is seen as one of the major language tools that facilitate smooth communication. In English, politeness has been an area of great interest in pragmatics, with various theories and corpus annotation approaches used to understand the relationship between politeness and social categories like power and gender, and to build Natural Language Processing applications. In Russian linguistics, politeness research has largely focused on lexical markers and speech strategies. This paper introduces the ongoing work on the development of the Russian Multimedia Politeness Corpus and discusses an annotation framework for oral communicative interaction, with an emphasis on adapting politeness theories for discourse annotation. The proposed approach lies in the identification of frames that encompass contextual information and the selection of relevant spatial, social, and relational features for the markup. The frames are then used to describe standard situations, which are marked by typical intentions and politeness formulae and paraverbal markers.

AN EXPERIMENTAL STUDY OF ARGUMENT EXTRACTION FROM PRESUPPOSITIONAL CLAUSES IN RUSSIAN

Knyazev M., Institute for Linguistic Studies, Russian Academy of Sciences, Saint Petersburg, Russia; HSE University, Saint Petersburg, Russia; Lomonosov Moscow State University, Moscow, Russia

The paper discusses two acceptability rating studies testing wh-interrogative and relative extractions of arguments from *čto*-clauses of presuppositional predicates like *žalet'* 'regret', as contrasted with nonpresuppositional predicates like *nadejat'sja* 'hope' and nominalized (*to čto*) clauses. The results show a difference in extraction between bare and nominalized clauses but no difference between presuppositional and nonpresuppositional clauses, raising potential doubts about the analysis of presuppositional clauses as DPs with a silent D.

COLLABORATIVE CONSTRUCTIONS IN RUSSIAN CONVERSATIONS: A MULTICHANNEL PERSPECTIVE

Korotaev N. A., Institute of Linguistics RAS; Russian State University for the Humanities, Moscow, Russia

The talk provides a multichannel description of how interlocutors co-construct utterances in conversation. Using data from the "Russian Pears Chats & Stories", I propose for a tripartite sequential scheme of collaborative constructions. When the scheme is fully realized, its first step not only includes the initial component of the construction, but also presupposes that the first participant makes a request for a co-operative action; the final component of the construction is provided by the second participant during the

second step; while the third step consists of the first participant's reaction. On each step, the participants combine vocal and non-vocal resources to achieve their goals. In some cases, non-vocal phenomena provide an essential clue to what is actually happening during co-construction, including whether the participants act in a truly co-operative manner. I distinguish between three types of communicative patterns that may take place during co-construction: "Requested Cooperation", "Unplanned Cooperation", and "Non-realized Interaction". The data suggest that these types can be influenced by the way the knowledge of the discussed events is distributed among the participants.

FACT-CHECKING BENCHMARK FOR THE RUSSIAN LARGE LANGUAGE MODELS

Kozlova A., **Shevelev D.**, **Fenogenova A.**, SberDevices, Moscow, Russia

Modern text-generative language models are rapidly developing. They produce text of high quality and are used in many real-world applications. However, they still have several limitations, for instance, the length of the context, degeneration processes, lack of logical structure, and facts consistency. In this work, we focus on the fact-checking problem applied to the output of the generative models on classical downstream tasks, such as paraphrasing, summarization, text style transfer, etc. We define the task of internal fact-checking, set the criteria for factual consistency, and present the novel dataset for this task for the Russian language. The benchmark for internal fact-checking and several baselines are also provided. We research data augmentation approaches to extend the training set and compare classification methods on different augmented data sets.

TEXT COMPLEXITY AS A NON-DISCRETE VALUE: RUSSIAN L2 TEXT COMPLEXITY DATASET ANNOTATION BASED ON ELO RATING SYSTEM

Laposhina A. N., Pushkin State Russian Language Institute, Moscow, Russia

The task of assessing text complexity for L2 learners can be approached as either a classification or regression problem, depending on the chosen scale. The primary bottleneck in such research lies in the limited availability of appropriate data samples. This study presents a combined approach to create a dataset of Russian texts for L2 learners, placed on a continuous scale of complexity, involving expert pairwise comparisons and the Elo rating system. For this pilot dataset, 104 texts from Russian L2 textbooks, TORFL tests, and authentic sources were selected and annotated. The resulting data is useful for evaluation of the automated models for assessing text complexity.

WHOSE WORD? PROBLEMS OF LEXICOGRAPHIC REPRESENTATION OF IDEOLOGICALLY MARKED WORDS (THE LEXICON OF THE RUSSIAN-UKRAINIAN CONFLICT)

Levontina I. B., **Shmeleva E. Ya.**, Vinogradov Russian Language Institute of the Russian Academy of Sciences, Moscow, Russia

The article deals with the problems of presenting ideologically marked words in the dictionary. It is based on the analysis of the words that appeared in the Russian language or received new meanings during the Russian-Ukrainian conflict. The difficulty of the lexicographic representation of such words is that their evaluative potential is mobile, for example, offensive nicknames can be assimilated by "offended" ones and become neutral words. Ideologically marked words can either exist in the lexicon for a long time or be quickly replaced by other lexical units. Therefore, in the interpretation of ideologically marked words, it is advisable to indicate the approximate time of their existence. In addition to temporary indicators, in the dictionary entry of such words, it is necessary to indicate whose word it is, that is, on whose behalf an assessment is given to a person or event. Since we believe that explanatory dictionaries should contain not only common names, but also proper names, the article also discusses geographical names.

PARAMETER-EFFICIENT TUNING OF TRANSFORMER MODELS FOR ANGLICISM DETECTION AND SUBSTITUTION IN RUSSIAN

Lukichev D.[1 2], **Kryanina D.**[1], **Bystrova A.**[1], **Fenogenova A.**[3], **Tikhonova M.**[1, 3], [1]HSE University; [2]Sber; [3]SberDevices, Moscow, Russia

This article is devoted to the problem of Anglicisms in texts in Russian: the tasks of detection and automatic rewriting of the text with the substitution of Anglicisms by their Russian-language equivalents. Within the framework of the study, we present a parallel corpus of Anglicisms and models that identify Anglicisms in the text and replace them with the Russian equivalent, preserving the stylistics of the original text.

DISAMBIGUATION IN CONTEXT IN THE RUSSIAN NATIONAL CORPUS: 20 YEAS LATER

Lyashevskaya O. N.[1, 2], **Afanasev I. A.**[1, 3], **Rebrikov S. A.**[1, 4], **Shishkina Y. A.**[1, 5], **Suleymanova E. A.**[6], **Trofimov I. V.**[6], **Vlasova N. A.**[6], [1]HSE University; [2]Vinogradov Russian Language Institute RAS; [3]MTS AI; [4]Kurchatov Institute; [5]Moscow Institute of Physics and Technology, Moscow, Russia; [6]A. K. Ailamazyan Program Systems Institute of RAS, Pereslavl-Zalessky, Russia

An updated annotation of the Main, Media, and some other corpora of the Russian National Corpus (RNC) features the part-of-speech and other morphological information, lemmas, dependency structures, and constituency types. Transformer-based architectures are used to resolve the homonymy in context according to a schema based on the manually disambiguated subcorpus of the Main corpus (morphology and lexicon) and UD-SynTagRus (syntax). The paper discusses the challenges in applying the models to texts of different registers, orthographies, and time periods, on the one hand, and making the new version convenient for users accustomed to the old search practices, on the other. The reannotated corpus data form the basis for the enhancement of the RNC tools such as word and n-gram frequency lists, collocations, corpus comparison, and Word at a glance.

MULTIMODAL HEDGES FOR COMPANION ROBOTS: A POLITENESS STRATEGY OR AN EMOTIONAL EXPRESSION?

Malkina M. P.[1], **Zinina A. A.**[2, 3, 1], **Arinkin N. A.**[2, 3], **Kotov A. A.**[2, 3], [1]MSLU; [2]Kurchatov Institute, [3]RSUH, Moscow, Russia

We examine the use of multimodal hedges (a politeness strategy, like saying *A kind of!*) by companion robots in two symmetric situations: (a) user makes a mistake and the robot affects user's social face by indicating this mistake, (b) robot makes a mistake, loses its social face and may compensate it with a hedge. Within our first hypothesis we test the politeness theory, applied to

robots: the robot with hedges should be perceived as more polite, threat to its social face should be reduced. Within our second hypothesis we test the assumption that multimodal hedges, as the expression (or simulation) of internal confusion, may make the robot more emotional and attractive. In our first experiment two robots assisted users in language learning and indicated their mistakes by saying *Incorrect!* The first robot used hedges in speech and gestures, while the second robot used gestures, supporting the negation. In our second experiment two robots answered university exam questions and made minor mistakes. The first robot used hedges, while the second robot used addressive strategy in speech and gestures, e. g. moved its hand to the user and said *That's it!* We have discovered that the use of hedges as the politeness strategy in both situations makes the robot *comfortable to communicate with*. But robot with hedges looks more *polite* only in the experiment, where it affects user's social face, and not when the robot makes mistakes. However, the usage of hedges as an emotional cue works in both cases: the robot with hedges seems to be *cute* and *sympathy provoking* both when it attacks user's social face or loses its own social face. This spectrum of hedge usage can demonstrate its transition from an expressive cue of a negative emotion (nervousness) to a marker of speaker's friendliness and competence.

AUGMENTATION METHODS FOR SPELLING CORRUPTIONS

Martynov N., **Baushenko M.**, **Abramov A.**, **Fenogenova A.**, SberDevices, Moscow, Russia

The problem of automatic spelling correction is vital to applications such as search engines, chatbots, spellchecking in browsers and text editors. The investigation of spell-checking problems can be divided into several parts: error detection, emulation of the error distribution on the new data for model training, and automatic spelling correction. As the data augmentation technique, the adversarial training via error distribution emulation increases a model's generalization capabilities; it can address many other challenges: from overcoming a limited amount of training data to regularizing the training objectives of the models. In this work, we propose a novel multi-domain dataset for spelling correction. On this basis, we provide a comparative study of augmentation methods that can be used to emulate the automatic error distribution. We also compare the distribution of the single-domain dataset with the errors from the multi-domain and present a tool that can emulate human misspellings.

AUTOCORRELATIONS DECAY IN TEXTS AND APPLICABILITY LIMITS OF LANGUAGE MODELS

Mikhaylovskiy N.[1,2], **Churilov I.**[2], [1]Higher IT School, Tomsk State University, Tomsk, Russia; [2]NTR Labs, Moscow, Russia

We show that the laws of autocorrelations decay in texts are closely related to applicability limits of language models. Using distributional semantics we empirically demonstrate that autocorrelations of words in texts decay according to a power law. We show that distributional semantics provides coherent autocorrelations decay exponents for texts translated to multiple languages. The autocorrelations decay in generated texts is quantitatively and often qualitatively different from the literary texts. We conclude that language models exhibiting Markovian behavior, including large autoregressive language models, may have limitations when applied to long texts, whether analysis or generation.

NAMED ENTITY-ORIENTED SENTIMENT ANALYSIS WITH TEXT2TEXT GENERATION APPROACH

Moloshnikov I.[1], **Skorokhodov M.**[1], **Naumov A.**[1], **Rybka R.**[1,2], **Sboev A.**[3,1], [1]NRC "Kurchatov Institute"; [2]Russian Technological University "MIREA"; [3]National Research Nuclear University "MEPhI", Moscow, Russia

This paper describes methods for sentiment analysis targeted toward named entities in Russian news texts. These methods are proposed as a solution for the Dialogue Evaluation 2023 competition in the RuSentNE shared task. This article presents two types of neural network models for multi-class classification. The first model is a recurrent neural network model with an attention mechanism and word vector representation extracted from language models. The second model is a neural network model for text2text generation. High accuracy is demonstrated by the generative model fine-tuned on the competition dataset and CABSAR open dataset. The proposed solution achieves 59.33 over two sentiment classes and 68.71 for three-class classification by f1-macro.

"PEARS ARE BIG GREEN": GESTURES WITH CONCRETE OBJECTS

Nikolaeva Y. V., Lomonosov Moscow State University, Interdisciplinary Scientific and Educational School "Preservation of the World Cultural and Historical Heritage", Moscow, Russia

The paper examines hand gestures when referring to inanimate referents. The aim of the study was to explore which factors determine the features of a gesture within the framework of modes of representation. Four main types of modes of representation were considered: drawing or shaping the form of the referent, acting, pointing, and presentation (PUOH); in addition, a new category of beat gestures was added.

As a result, it was shown that communicative dynamism or other referent characteristics such as control of the object or its inferability from the previous context do not fully determine the use of gestures with the referent. As an alternative hipothesis, we propose a notion of gesture information hierarchy, where discursive factors, such as previous mentions of the referent and the introduction or change of the protagonist along with the way an object is used determines the form of the gesture.

RUSSIAN CONSTRUCTICON 2.0: NEW FEATURES AND NEW PERSPECTIVES OF THE BIGGEST CONSTRUCTICON EVER BUILT

Orlov A. V.[1], **Butenko Z. A.**[1,2], **Demidova D. A.**[2], **Starchenko V. M.**[1], **Rakhilina E. V.**[1,3], **Lyashevskaya O. N.**[1,3], [1]HSE University, Moscow, Russia; [2]UiT The Arctic University of Norway, Tromsø, Norway; [3]Vinogradov Institute for Russian language RAS, Moscow, Russia

Russian constructicon is an open-access linguistic database containing detailed descriptions of over 3,800 Russian grammatical constructions. In this paper we present a new, enlarged and updated version of Russian Constructicon (RusCxn) as well as new trajectories of development which were opened for the resource after the update. Since its first release, RusCxn, has undergone many significant changes. Our team has expanded the number of constructions present in the database 1,5 times, introduced new meta-information features such as glosses, significantly reworked the architecture and the design of Russian Constructicon's website, and improved the search facilities. The above-mentioned changes not only make RusCxn more attractive and convenient-to-use, but they can also greatly facilitate typological research in the field of Construction Grammar and improve the mapping between constructicography-orinented resources for different languages.

LINGUISTIC ANNOTATION GENERATION WITH CHATGPT: A SYNTHETIC DATASET OF SPEECH FUNCTIONS FOR DISCOURSE ANNOTATION OF CASUAL CONVERSATIONS

Ostyakova L.[1,2], **Petukhova K.**[2], **Smilga V.**[2], **Zharikova D.**[2], [1]HSE University; [2]Moscow Institute of Physics and Technology, Moscow, Russia

This paper is devoted to examining the hierarchical and multilayered taxonomy of Speech Functions, encompassing pragmatics, turn-taking, feedback, and topic switching in open-domain conversations. To evaluate the distinctiveness of closely related pragmatic classes, we conducted comparative analyses involving both expert annotators and crowdsourcing workers. We then carried out classification experiments on a manually annotated dataset and a synthetic dataset generated using ChatGPT. We looked into the viability of using ChatGPT to produce data for such complex topics as discourse. Our findings contribute to the field of prompt engineering techniques for linguistic annotation in large language models, offering valuable insights for the development of more sophisticated dialogue systems.

POLY-PREDICATION IN INFORMAL MONOLOGICAL DISCOURSE (ACCORDING TO «WHAT I SAW» CORPUS)

Panysheva D., Russian State University for the Humanities, Moscow, Russia

The article discusses the relationship between the mode of discourse and quantitative metrics of poly-predication. Based on the material of the corpus "What I Saw", oral and written versions of stories are compared according to the relative frequency of polypredicative constructions and the representation of certain types of polypredication, the features of semantics and grammatical labeling of such structures are described. Using the nonparametric Wilcoxon criterion, the absence of statistical significance between the density of poly-predication in the oral and written parts of the corpus is proved.

RUSSIAN ADDITIVE MARKERS *TAKŽE* AND *TOŽE*: A SYNCHRONIC AND DIACHRONIC PERSPECTIVE

Pekelis O. E., Russian State University for the Humanities/Moscow, Russia; HSE University/Moscow, Russia

It is well known that Russian additive markers *takže* and *tože* differ in terms of information structure: the scope of *takže* is focus, while the scope of *tože* is topic. Based on data of several corpora of Russian, this paper shows that in modern Russian, *takže* and tože are opposed on other language levels as well, namely syntactically (in terms of word order), lexically (a variant of *takže* that is synonymous with *tože* including at the level of the information structure, is going out of use), stylistically and as far as their involvement in grammaticalization processes is concerned (*takže* but not *tože* developed into a coordinate conjunction and a discourse marker). However, as evidenced by Russian National Corpus data, most of these contrasts were absent or less pronounced in the Russian language of the 18th-19th centuries. Thus, in the last two centuries *takže* and *tože* evolved toward their consistent differentiation.

THE COBALD ANNOTATION PROJECT: THE CREATION AND APPLICATION OF THE FULL MORPHO-SYNTACTIC AND SEMANTIC MARKUP STANDARD

Petrova M. A.[1], **Ivoylova A. M.**[2], **Bayuk I. S.**[1], **Dyachkova D. S.**[2], **Michurina M. A.**[2], [1]A4 Technology; [2]RSUH, Moscow, Russia

The current paper is devoted to the Compreno-Based Linguistic Data (CoBaLD) Annotation Project aimed at creating text corpora annotated with full morphological, syntactic and semantic markup. The first task of the project is to suggest a standard for the full universal markup which would include both morphosyntactic and semantic patterns. To solve this problem, one needs the markup model, which includes all necessary markup levels and presents the markup in a format convenient for users. The latter implies not only the fullness of the markup, but also its structural simplicity and homogeneity. As a base for the markup, we have chosen the simplified version of the Compreno model, and as data presentation format, we have taken Universal Dependencies.

At the second stage of the project, the Russian corpus with 400 thousand tokens (CoBaLD-Rus) has been created, which is annotated according to the given standard. The third stage is devoted to the testing of the new format. For this purpose, we have held the SEMarkup Shared Task aimed at creating parsers which would produce full morpho-syntactic and semantic markup. Within this task, we have elaborated neural network-based parser trained on our dataset, which allows one to annotate new texts with the CoBaLD-standard. Our further plans are to create fully annotated corpora for other languages and to carry out the experiments on language transfers of the current markup to other languages.

HALF-MASKED MODEL FOR NAMED ENTITY SENTIMENT ANALYSIS

Podberezko P., **Kaznacheev A.**, **Abdullayeva S.**, **Kabaev A.**, MTS AI, Moscow, Russia

Named Entity Sentiment analysis (NESA) is one of the most actively developing application domains in Natural Language Processing (NLP). Social media NESA is a significant field of opinion analysis since detecting and tracking sentiment trends in the news flow is crucial for building various analytical systems and monitoring the media image of specific people or companies.

In this paper, we study different transformers-based solutions NESA in RuSentNE-23 evaluation. Despite the effectiveness of the BERT-like models, they can still struggle with certain challenges, such as overfitting, which appeared to be the main obstacle in achieving high accuracy on the RuSentNE-23 data. We present several approaches to overcome this problem, among which there is a novel technique of additional pass over given data with masked entity before making the final prediction so that we can combine logits from the model when it knows the exact entity it predicts sentiment for and when it does not. Utilizing this technique, we ensemble multiple BERTlike models trained on different subsets of data to improve overall performance. Our proposed model achieves the best result on RuSentNE-23 evaluation data and demonstrates improved consistency in entity-level sentiment analysis.

PROSODIC PORTRAIT OF THE RUSSIAN CONNECTOR PRICHOM IN THE MIRROR OF THE MULTIMEDIA CORPUS

Podlesskaya V. I., Institute of linguistics, Russian Academy of Sciences; Russian State University for the Humanities, Moscow, Russia

Based on data from the multimedia subcorpus of the Russian National Corpus, the paper addresses prosodic features of discourse fragments introduced by the connector *prichom* 'and besides'. The data of instrumental and perceptual analysis show that the fragment with *prichom* has communicative-prosodic autonomy: firstly, it has an internal thematic structure with an obligatory rheme and an optional theme; and secondly, there is a prosodic break before this fragment. The autonomy of the fragment introduced by *prichom* is preserved in

a variety of contexts: (i) both in cases where this fragment is a complete clause and when it is a fragmented clause; (ii) both in those cases when the previous fragment is prosodically realized as final (projecting no continuation), and when it is realized as non-final (projecting continuation); (iii) both in those cases when the fragment introduced by *prichom* is an element of the main narrative chain, and when it is inserted parenthetically inside another fragment. In addition to the above, a fragment with *prichom* can form a separate turn in the conversation. Thus, the detected prosodic features of the fragment with *prichom* make it possible to objectify the idea earlier expressed in the literature (Kiselyova 1971, Vinogradov 1984, Inkova 2018, inter alia): that structures with *prichom* are built in two "communicative steps", or that they are used to express "concomitance established at the level of speech acts". Clauses connected by the relationship of syntactic subordination quite often lose their prosodic autonomy (Podlesskaya 2014 a, b), and vice versa, clauses in coordinated constructions tend to retain prosodic autonomy. Therefore, the prosodic autonomy of the components of the construction with *prichom*, retained in various contexts, speaks in favor of its coordinated status, while a number of syntactic tests proper speak of the opposite.

HWR200: NEW OPEN ACCESS DATASET OF HANDWRITTEN TEXTS IMAGES IN RUSSIAN

Potyashin I.[1], **Kaprielova M.**[1,2], **Chekhovich Y.**[1,2], **Kildyakov A.**, **Seil T.**[1], **Finogeev E.**[1], **Grabovoy A.**[1,2], [1]Antiplagiat; [2]FRC CSC RAS, Moscow, Russia

Handwritten text image datasets are highly useful for solving many problems using machine learning. Such problems include recognition of handwritten characters and handwriting, visual question answering, near-duplicate detection, search for text reuse in handwriting and many auxiliary tasks: highlighting lines, words, other objects in the text. The paper presents new dataset of handwritten texts images in Russian created by 200 writers with different handwriting and photographed in different environment. We described the procedure for creating this dataset and the requirements that were set for the texts and photos. The experiments with the baseline solution on fraud search and text reuse search problems showed results of results of 60% and 83% recall respectively and 5% and 2% false positive rate respectively on the dataset.

SIMPLE YET EFFECTIVE NAMED ENTITY ORIENTED SENTIMENT ANALYSIS

Sanochkin L.[1,2], **Bolshina A.**[1], **Cheloshkina K.**[1,2], **Galimzianova D.**[1,2], **Malafeev A.**[1,2], [1]MTS AI, [2]HSE, Moscow, Russia

Sentiment analysis, i.e. the automatic evaluation of the emotional tone of a text, is a common task in natural language processing. Entity-Oriented Sentiment Analysis (EOSA) predicts the sentiment of entities mentioned in a given text. In this paper, we focus on the EOSA task for the Russian news. We propose a text classification pipeline to solve this task and show its potential in such tasks. Moreover, in general, EOSA implies labeling both named entities and their sentiment, which can require a lot of annotator labour and time and, thus, presents a major obstacle to the development of a production-ready EOSA system. To help alleviate this, we analyse the potential of applying an Active learning approach to EOSA tasks. We demonstrate that by actively selecting instances for labeling in EOSA the annotation effort required for training machine learning models can be significantly reduced.

IS IT POSSIBLE TO MAKE THE RUSSIAN PUNCTUATION RULES MORE EXPLICIT?

Shmelev A., Vinogradov Russian Language Institute of the Russian Academy of Sciences, Moscow, Russia

This paper deals with some issues related to the Russian punctuation rules and their account in computer checkers and correctors (both “analytic” and “synthetic”). It also discusses variation of punctuation. The paper offers a critical assessment of reference books devoted to punctuation and makes special reference to certain verbs of propositional attitude and their parenthetical use (in particular, *dumat’* ‘to think,’ *videt’* ‘to see,’ and *slyshat’* ‘to hear). It claims that the inherent characteristics of the verbs under consideration influence the punctuation, and therefore every verb deserves a detailed description (lexicographic portrait). In particular, *videt’* and *slyshat’* behave quite differently when used as parenthetical verbs. A step towards making the punctuation rules more explicit may consist in providing an index of words mentioned in the rules together with a subject index.

THE ROLE OF INDICATORS IN ARGUMENTATIVE RELATION PREDICTION

Sidorova E., **Akhmadeeva I.**, **Kononenko I.**, **Chagina P.**, A.P. Ershov Institute of Informatics Systems, Siberian Branch, Russian Academy of Sciences, Novosibirsk, Russia

The article presents a comparative study of methods for argumentative relation prediction based on a neural network approach. The distinctive feature of the study is the use of argumentative indicators in the preparation of the training sample. The indicators are generated based on the discourse marker dictionary. The experiments were carried out using an annotated corpus of scientific and popular science texts, including 162 articles available on the ArgNet-Bank Studio web platform. A set of all argumentative relations is described by internal connections of arguments and include the conclusion and the premise. In the first stage of training set construction, fragments of text that included two consecutive sentences were examined. In the second stage, indicators were retrieved from the corpus texts and, for each indicator, statements presumably corresponding to the premise and conclusion of the argument were extracted. In total, 4.2 thousand indicator-based training contexts and 13.6 thousand pairs of sentences were obtained from the corpus with annotation of the presence of an argumentative relation. Based on this training sample, four classifiers were built: without indicators, with marking indicators in sentences using tags, taking into account segmentation of text based on indicators, with segmentation and tags. The results of the experiments on argumentative relation prediction are presented.

TEXT VQA WITH TOKEN CLASSIFICATION OF RECOGNIZED TEXT AND RULE-BASED NUMERICAL REASONING

Surkov V. O., **Evseev D. A.**, Moscow Institute of Physics and Technology, Dolgoprudny, Russia

In this paper, we describe a question answering system on document images which is capable of numerical reasoning over extracted structured data. The system performs optical character recognition, detection of key attributes in text, generation of a numerical reasoning program, and its execution with the values of key attributes as operands. OCR includes the steps of bounding boxes detection and recognition of text from bounding boxes. The extraction of key attributes, such as quantity and price of goods, total etc., is based on the BERT token classification model. For expression generation we investigated the rule-based approach and the T5-base model and found that T5 is capable of generalization to expression types unseen in the training set. The proposed architecture of the question answering system utilizes the structure of independent blocks, each of which can be enhanced or replaced while keeping other components unchanged. The proposed model was evaluated in the Receipt-AVQA competition and on FUNSD dataset.

SCALAR STRUCTURE FOR *POLU- HALF*

Tatevosov S. G., Lomonosov Moscow State University Interdisciplinary School "Preservation of the World Cultural and Historical Heritage", Moscow, Russia, **Kisseleva X. L.**, Vinogradov Russian Language Institute of the Russian Academy of Sciences, Moscow, Russia

This paper explores restrictions on the distribution of *polu-* 'half' in combination with adjectival stems in Russian. Relying on the literature on degree semantics, we analyze *polu-* as a degree modifier that specifies the degree to which the adjective maps an individual as ½ of the maximal degree. This correctly predicts that *polu-* can only combine with upper closed scales. We argue that unlike *half* in English, *polu-* does not require a scale be lower closed.

TEXT SIMPLIFICATION AS A CONTROLLED TEXT STYLE TRANSFER TASK

Tikhonova M., HSE University, SberDevices, Moscow, Russia, **Fenogenova A.**, SberDevices, Moscow, Russia

The task of text simplification is to reduce the complexity of the given piece of text while preserving its original meaning to improve readability and understanding. In this paper, we consider the simplification task as a subfield of the general text style transfer problem and apply methods of controllable text style to rewrite texts in a simpler manner preserving their meaning. Namely, we use a paraphrase model guided by another style-conditional language model. In our work, we perform a series of experiments and compare this approach with the standard fine-tuning of an autoregressive model.

AN ATTEMPT TO DETERMINE A PREPOSITION AND DELIMIT THE CLASS OF DERIVED PREPOSITIONS IN RUSSIAN

Uryson E., Russian Language Institute RAS, Moscow, Russia

The object of the paper are Russian words traditionally described as derived prepositions. The problem is that there is no formal definition of preposition in theoretical or applied linguistics. Non-derivative, or primitive prepositions are given in grammar by the closed list, so strictly speaking there is no need to define this class of words. However. we must have criteria for determining derived prepositions. I suggest a set of necessary conditions that a preposition must satisfy. I demonstrate that so called adverbial prepositions in Russian do not satisfy them and should be described as adverbs. Similarly, some Russian verbal prepositions, and some Russian denominative prepositions should not be described as prepositions.

ESTIMATING COGNITIVE TEXT COMPLEXITY WITH AGGREGATION OF QUANTILE-BASED MODELS

Veselov A. S., Lomonosov Moscow State University, Moscow, Russia, **Eremeev M. A.**, New York University, New York, USA, **Vorontsov K. V.**, Moscow Institute of Physics and Technology, Moscow, Russia

In this paper, we introduce a novel approach to estimating the cognitive complexity of a text at different levels of language: phonetic, morphemic, lexical, and syntactic. The proposed method detects tokens with an abnormal frequency of complexity scores. The frequencies are taken from the empirical distributions calculated over the reference corpus of texts. We use the Russian Wikipedia for this purpose. Ensemble models are combined from individual models from different language levels. We created datasets of pairs of text fragments taken from social studies textbooks of different grades to train the ensembles. Empirical evidence shows that the proposed approach outperforms existing methods, such as readability indices, in estimating text complexity in terms of accuracy. The purpose of this study is to create one of the important components of the system of recommendation of scientific and educational content.

MAXPROB: CONTROLLABLE STORY GENERATION FROM STORYLINE

Vychegzhanin S. V., **Kotelnikova A. V.**, **Sergeev A. V.**, **Kotelnikov E. V.**, Vyatka State University, Kirov, Russia

Controllable story generation towards keywords or key phrases is one of the purposes of using language models. Recent work has shown that various decoding strategies prove to be effective in achieving a high level of language control. Such strategies require less computational resources compared to approaches based on fine-tuning pre-trained language models. The paper proposes and investigates the method *MaxProb* of controllable story generation in Russian, which works at the decoding stage in the process of text generation. The method uses a generative language model to estimate the probability of its tokens in order to shift the content of the text towards the guide phrase. The idea of the method is to generate a set of different small sequences of tokens from the language model vocabulary, estimate the probability of following the guide phrase after each sequence, and choose the most probable sequence. The method allows evaluating the consistency of the token sequence for the transition from the prompt to the guide phrase. The study was carried out using the Russian-language corpus of stories with extracted events that make up the plot of the story. Experiments have shown the effectiveness of the proposed method for automatically creating stories from a set of plot phrases.

THE PROSODY OF THE RUSSIAN QUESTION

Yanko T. E., Institute of Linguistics, Russian Academy of Sciences, Moscow, Russia

The analysis of Russian interrogative prosody is based on a model of a question as consisting of the two components: the illocutionary proper component and the illocutionary improper component. The illocutionary improper component includes the data for information retrieval. The illocutionary proper component can be formed both by segmental means of expression (by an interrogative word or a particle) or solely by prosody (as in Russian yes-no questions). The prosody of Russian questions having the interrogative words or the interrogative particle li is highly variable, whereas the prosody of Russian yes-no questions expressed by prosody is stable. The latter is the Russian rising accent, which has a rise on the tonic syllable of the accent-bearer followed by a fall on the post-tonics if any. The illocutionary improper component can be located sentence initially and carry a specific falling accent (namely, a late fall). A specific type of a question with the interrogative proper component omitted is recognized. Such questions carry a late fall, or a falling-rising accent on the accent-bearer. The analysis is exemplified by the frequency tracings of the sound sentences taken from the Russian National Corpus and other open sources. As the instrument for verifying the acoustic data, we used the computer system Praat. The paper is illustrated throughout with pitch contours of sound records.

PARALLEL CORPUS AS A TOOL FOR SEMANTIC ANALYSIS: THE RUSSIAN DISCOURSE MARKER STALO BYT' (*CONSEQUENTLY*)

Zalizniak Anna A. , Institute of Linguistics of the RAS, Moscow, Russia, Dobrovol'ski jD. O. , Russian Language Institute of the RAS; Institute of Linguistics of the RAS, Moscow, Russia

The article examines the semantics of the Russian discourse marker *stalo byt',* using the data obtained by analyzing translational correspondences extracted from parallel corpora of the Russian National Corpus (RNC). Typically, this discourse marker is an indicator of inferential evidentiality, by which the speaker marks the fact that the given statement is a conclusion made by the speaker on the basis of the information they received and accepted as true by default. In addition, *stalo byt'* has two secondary types of usage—"rhetorical" and "narrative"—where the basic semantics of this discourse marker is subject to certain modifications. One of the key points of analysis is the reconstruction of semantic mechanisms providing the actual semantics of *stalo byt'*.

RUSSIAN PREDICATIVES AND FREQUENCY METRICS

Zimmerling A. V., Pushkin State Russian Language Institute; Institute of Linguistics, Russian Academy of Science, Moscow, Russia

This paper introduces five metrics for measuring the frequencies of dative predicatives in Russian. A dative predicative is a word or multiword expression licensing the dative-predicative-structure, where the semantic subject of the non-agreeing non-verbal predicate is marked by the dative case. I measure the frequencies of the predicatives in the contact position <−1;1> with the same-clause dative subject pronouns in 1Sg (*m*-metrics) and 3Sg (*e*-metrics). The *m*-metrics is applied for retrieving a list of dative predicatives from a corpus. I argue that for each large text collection there is a minimal *m*-value confirming that an item belongs to the core of the dative-predicative structure. The *m/e* score makes up the third metrics that shows whether an element is oriented towards the use in the 1st person or not. Basing on the *m*-metrics, I retrieved 3 lists of predicatives in the subcorpus of 2000–2021 texts included in the Russian National Corpus. The A list includes 87 items with $m \geq 10$, the B list includes 44 items with $m \geq 50$, the C list includes 24 items with $m \geq 100$. 72–79% of items in each list have an *m/e* value ≥ 1,25. A linguistic interpretation of this result is that for each list of dative predicatives it is true that the majority of its elements are autoreferential expressions oriented towards the use in the 1st person present indicative tense in the direct speech. The fourth metrics shows the total number of occurrences of a word or multiword expression in the corpus (*N*). I argue that the *N* score must be measured before POS tagging, and lemmatization. The fifth and the last metrics is the *m/N* score. The RNC data suggest an inverse correlation between the score of an item in the context specific for dative-predicative structures (*m*) and its overall frequency in the corpus (*N*). This effect is explained by the regular homonymy of high frequent predicatives with high frequent adverbials and parenthetical expressions.

Авторский указатель

Author Index

Научное издание

Компьютерная лингвистика и интеллектуальные технологии

По материалам ежегодной
международной конференции «Диалог»

Выпуск 22, 2023

Ответственный за выпуск **А. В. Ульянова**
Вёрстка **К. А. Климентовский**