Совместный семинар Российской ассоциации Искусственного интеллекта и Федерального исследовательского центра «Информатика и управление» РАН

ИНТЕЛЛЕКТУАЛЬНЫЙ АНАЛИЗ ТЕКСТОВ НА ОСНОВЕ МЕТОДОВ РАЗНОУРОВНЕВОЙ ОБРАБОТКИ ЕСТЕСТВЕННОГО ЯЗЫКА

Смирнов Иван Валентинович

д.т.н, доцент

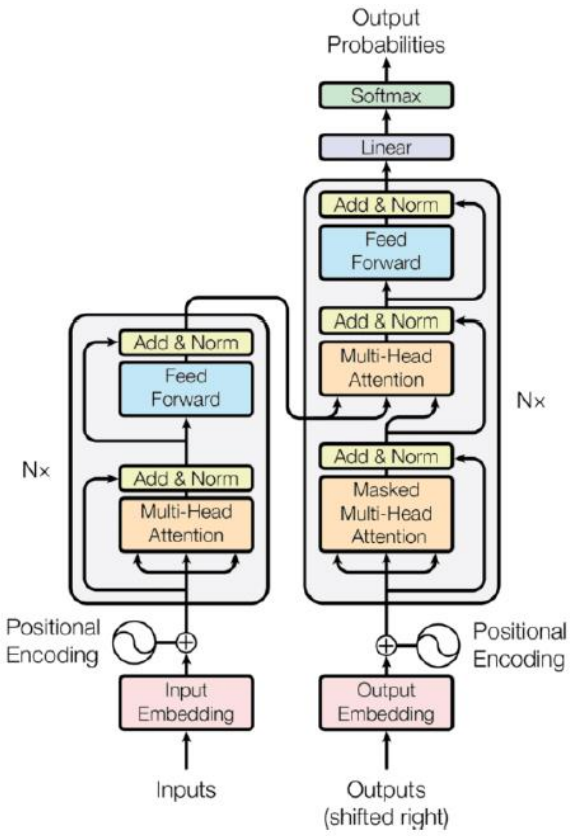
заведующий отделом «Интеллектуальный анализ информации»

ФИЦ ИУ РАН

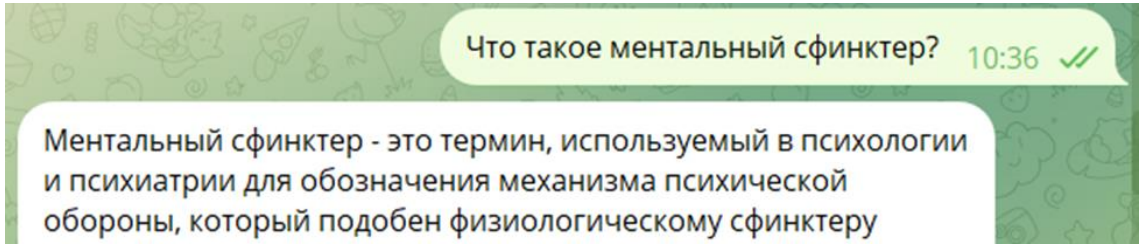
Интеллектуальный анализ текстов

- Тексты на естественном языке остаются одним из основных способов хранения и передачи информации
- Интеллектуальный анализ текстов направлен на получение из текстов полезной информации для поддержки принятия решений в различных областях
- Решаемые задачи:
  - Информационный и вопросно-ответный поиск
  - Извлечение информации из текстов
  - Классификация и кластеризация текстов, атрибуция текстов
  - Резюмирование текстов
  - Психолингвистический анализ текстов, анализ тональности
  - И другие
- Используются методы обработки информации и искусственного интеллекта
- Базовые задачи
  - Классификация фрагментов текста (символов, слов, словосочетаний, предложений, текстов…)
  - Классификация пар фрагментов текста
  - Генерация текста (по заданному фрагменту текста, по базе знаний\данных…)

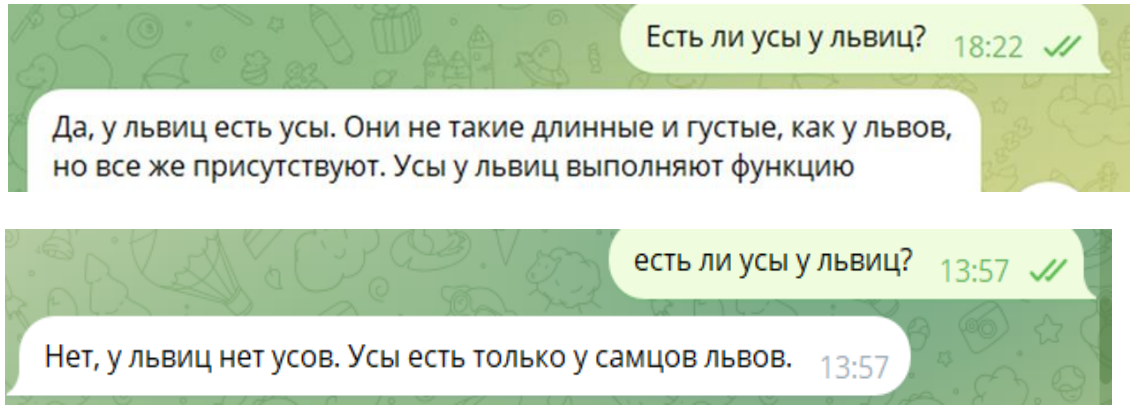
Type	Model Name	#Parameters	Release
Encoder-Only	BERT	110M, 340M	2018
	RoBERTa	355M	2019
	ALBERT	12M, 18M, 60M, 235M	2019
	DeBERTa	-	2020
	XLNet	110M, 340M	2019
Decoder-only	GPT-1	120M	2018
	GPT-2	1.5B	2019
Encoder-Decoder	T5 (Base)	223M	2019
	MT5 (Base)	300M	2020
	BART (Base)	139M	2019
GPT Family	GPT-3	125M, 350M, 760M, 1.3B, 2.7B, 6.7B, 13B, 175B	2020
	CODEX	12B	2021
	WebGPT	760M, 13B, 175B	2021
	GPT-4	1.76T	2023
LLaMA Family	LLaMA 1	7B, 13B, 33B, 65B	2023
	LLaMA 2	7B, 13B, 34B, 70B	2023
	Alpaca	7B	2023
	Vicuna-13B	13B	2023
	Koala	13B	2023
	Mistral-7B	7.3B	2023
	Code Llama	34	2023
	LongLLaMA	3B, 7B	2023
	LLaMA-Pro-8B	8.3B	2024
	TinyLlama-1.1B	1.1B	2024
PaLM Family	PaLM	8B, 62B, 540B	2022
	U-PaLM	8B, 62B, 540B	2022
	PaLM-2	340B	2023
	Med-PaLM	540B	2022
	Med-PaLM 2	-	2023
Other Popular LLMs	FLAN	137B	2021
	Gopher	280B	2021
	ERNIE 4.0	10B	2023
	Retro	7.5B	2021
	LaMDA	137B	2022
	ChinChilla	70B	2022
	Galactia-120B	120B	2022
	CodeGen	16.1B	2022
	BLOOM	176B	2022
	Zephyr	7.24B	2023
	Grok-0	33B	2023
	ORCA-2	13B	2023
	StartCoder	15.5B	2023
	MPT	7B	2023



Информация о несуществующих понятиях:



Одна версия модели в разное время:



Minaee S. et al. Large language models: A survey //arXiv preprint arXiv:2402.06196. – 2024.

Преимущества

- Универсальность
- Высокое качество
- Решение нестандартных задач

Zhou L. et al. Larger and more instructable language models become less reliable //Nature. – 2024. – С. 1-8.

Проблема доверия к LLM

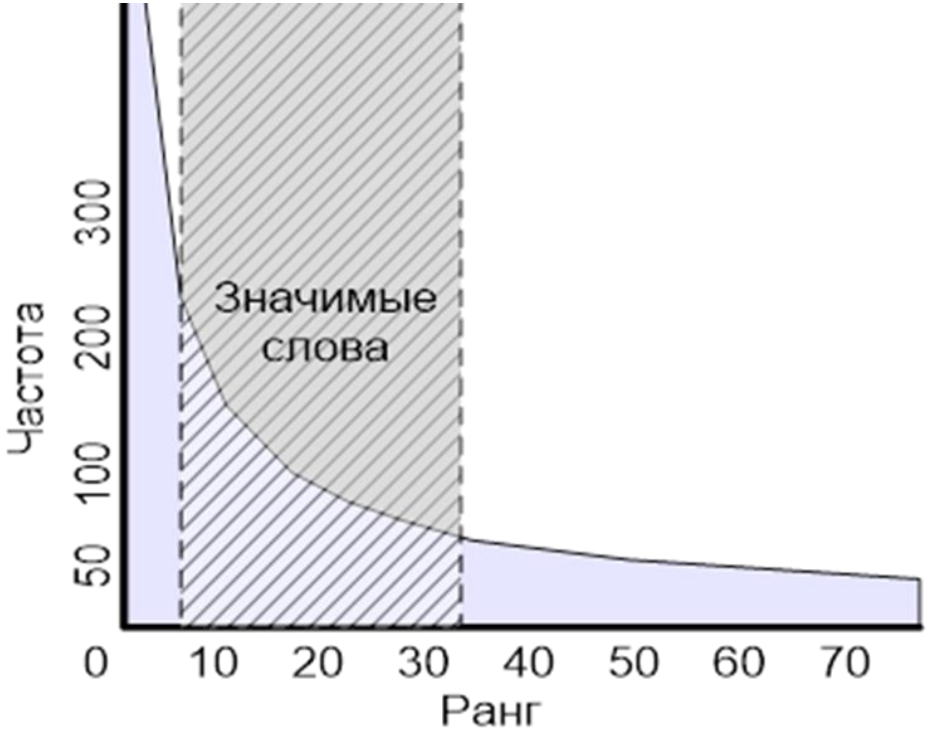
- неинтерпретируемость
- требовательность к вычислительным ресурсам
- недостоверность
- «галлюцинации», чувствительность к обману (атакам)
- идеологические смещения
- принадлежность большим корпорациям

Лексические подходы

I love this movie! It's sweet, but with satirical humor. The dialogue is great and the adventure scenes are fun... It manages to be whimsical and romantic while laughing at the conventions of the fairy tale genre. I would recommend it to just about anyone. I've seen it several times, and I'm always happy to see it again whenever I have a friend who hasn't seen it yet!

it	6
I	5
the	4
to	3
and	3
seen	2
yet	1
would	1
whimsical	1
times	1
sweet	1
satirical	1
adventure	1
genre	1
fairy	1
humor	1
have	1
great	1

https://ogre51.medium.com/nlp-explain-bag-of-words-3b9fc4f211e8



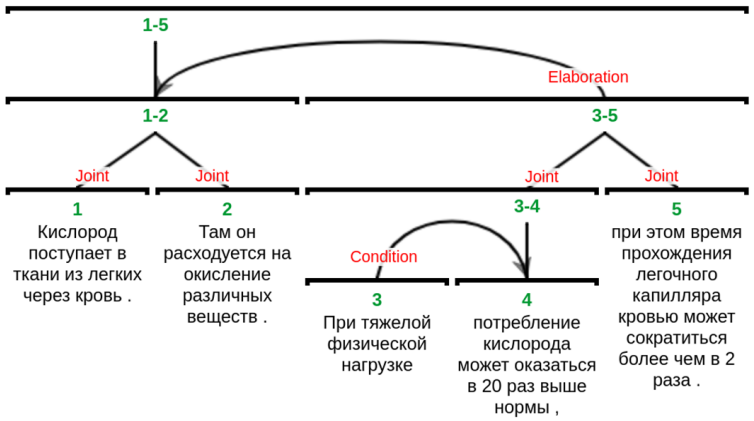
$$w_{ik} = tf_{ik} \cdot idf_k \qquad idf_{ik} = \log(\frac{N}{N_k})$$

- Легко реализуются
- Представляют текст в виде множества (взвешенных) слов
- Словари, правила, машинное обучение
- Не учитывают связанность слов в предложении и структуру текста
- Могут не обеспечивать надлежащего качества решения прикладных задач

Лингвистические подходы

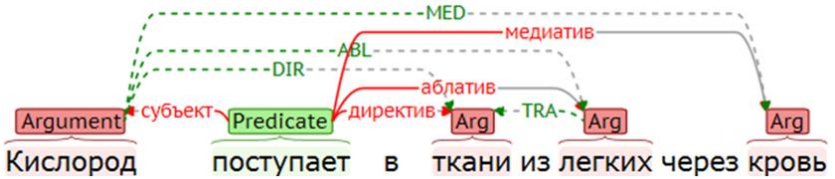
Методы разноуровневого анализа текста

Методы дискурсивного анализа



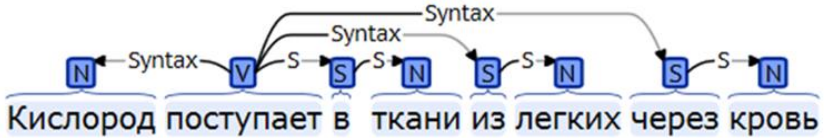
Позволяет учитывать связность текста, моделировать рассуждения и аргументацию

Методы семантического анализа



Позволяет учитывать значения слов и семантические отношения между словами

Методы синтаксического анализа



Позволяет учитывать словосочетания

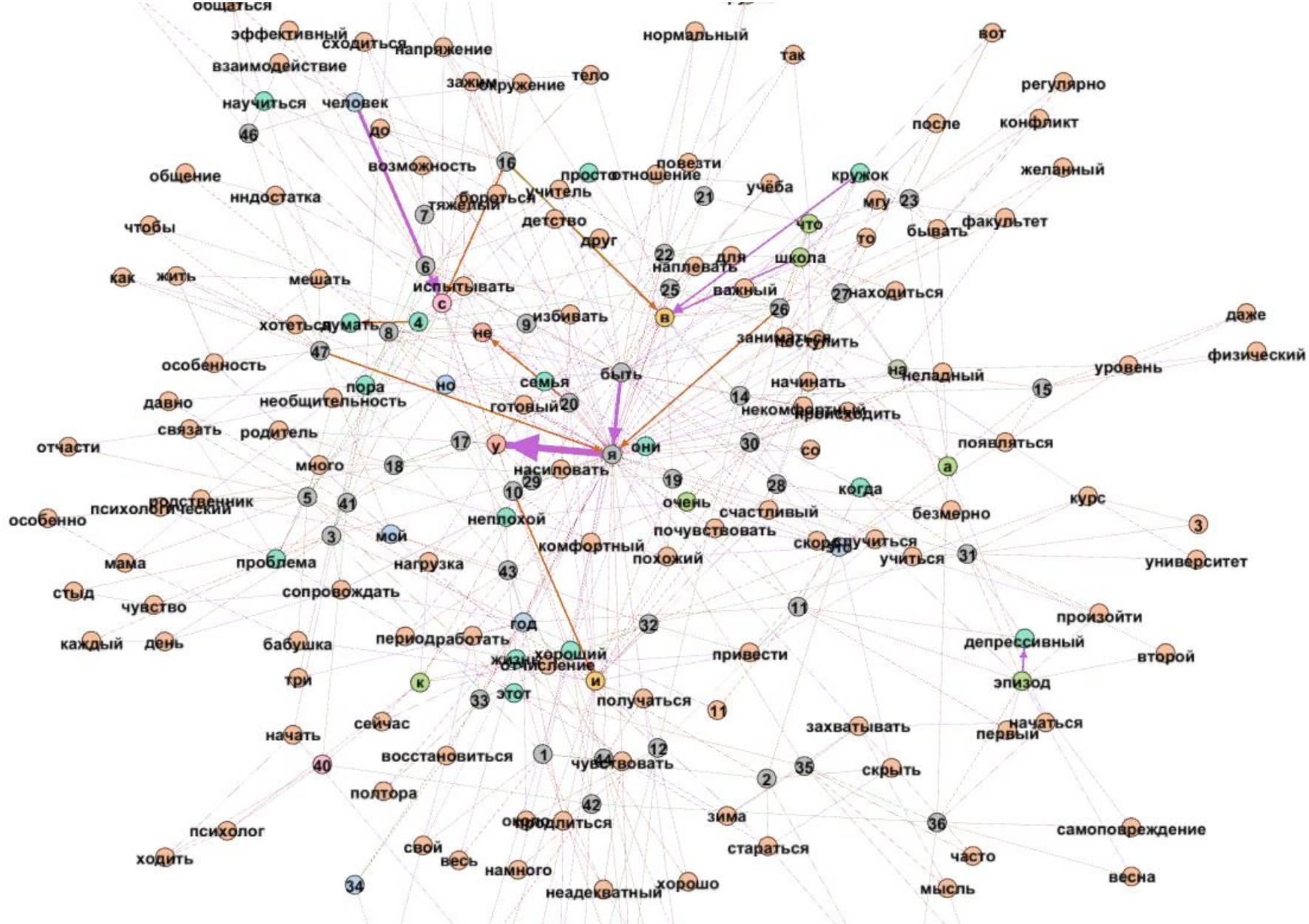
Методы морфологического анализа

Морфологическая структура

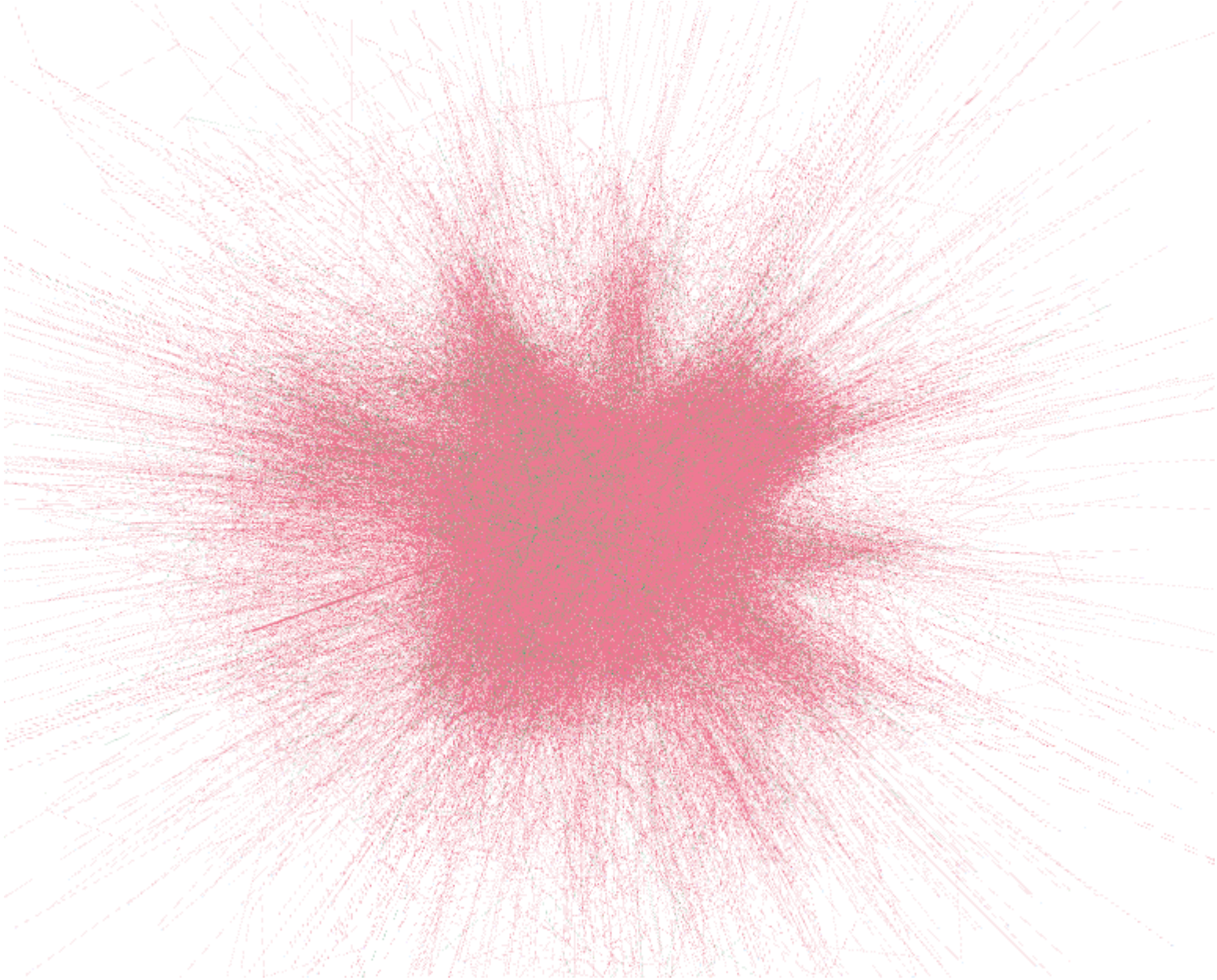
ткани - часть_речи: сущ., род: жен., падеж: им.вин.предл., кск: предметное

Отражает грамматические свойства слов

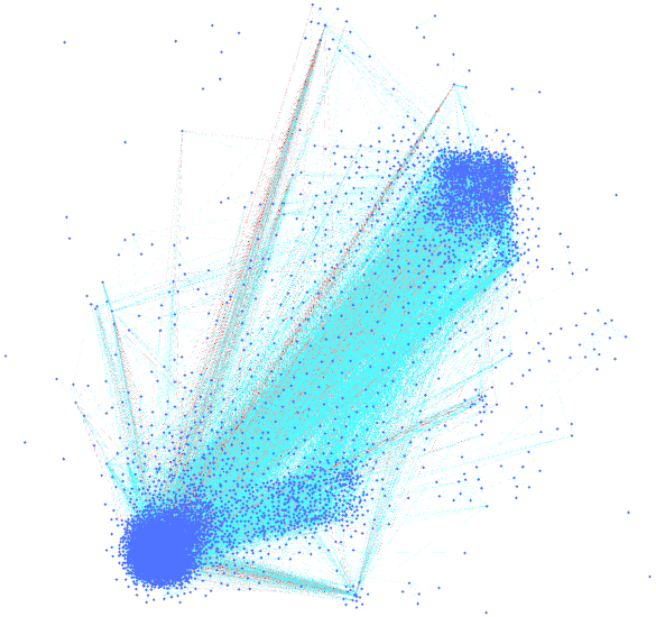
- Вклад в моделирование семантики внесли: G. Lakoff, R.Montague, R.Schank, N.Chomsky, C.Fillmore, R.Mooney, D.Jurafsky, M.Lapata, И.А. Мельчук, Ю.Д. Апресян, И.М. Кобозева, Г.Г.Белоногов, И.П. Кузнецов, В.А. Тузов, Г.С.Осипов, О.Н. Ляшевская и другие
- Вклад в моделирование дискурса и дискурсивный анализ внесли: W.Mann, S.Thompson, D.Marcu, L.Polanyi, N.Asher, A.Lascarides, А.А. Кибрик, С.Ю. Толдова и другие



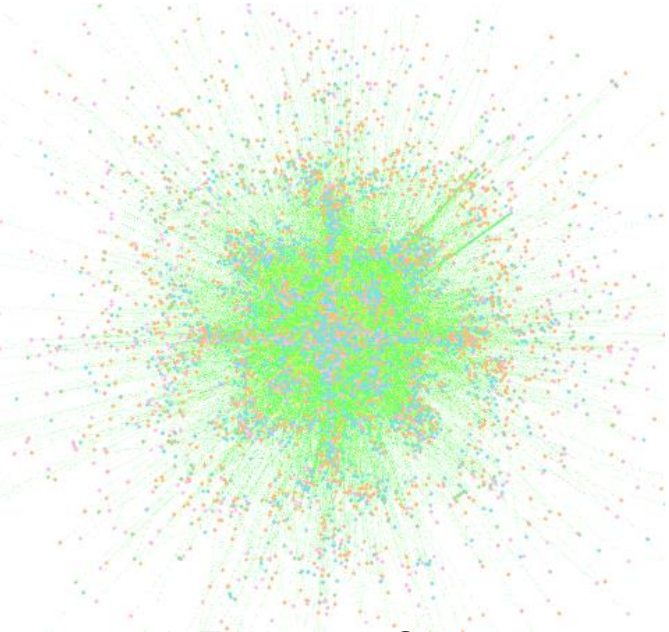
Сочинение на тему «Я, другие, мир». Пациент психиатрической клиники



И.С. Тургенев. Отцы и дети



Н.С. Лесков. Островитяне



А.П. Чехов. Степь

Визуализация с помощью Gephi

Реляционно-ситуационный анализ текста

Существующие подходы: СМЫСЛ-ТЕКСТ, падежные грамматики, семантика Монтегю и др.

Реляционно-ситуационная модель текста (Осипов Г.С.)

Семантические роли

- **Субъект** – производитель действия (*исследование показало перспективность…*)
- **Каузатив** – причина (*гипертония приводит к поражению артерий)*

Семантические отношения (связи)

- **DIR** – директивное отношение, в котором один компонент обозначает путь, направление второго компонента (*Владимир Путин отправился* в *США*)
- **CAUS** – каузальное отношение, один компонент которого обозначает причину проявления другого компонента спустя какое-то время (*Казнокрадство приводит к обнищанию населения*)

$$C_{synt} =< C_{gr}, R_{syntW} >,$$

$$R_{syntW} = \{\langle (w_1, w_2), t_{synt} \rangle | t_{synt} \in Type_{synt}, w_i \in C_{gr} \cup root, i = 1,2\}.$$

$$T_{Synt} = \left(S_{synt\ 1}, S_{synt\ 2}, \ldots S_{synt\ l}\right), S_{synt\ h} < S_{synt\ h+1}, h = \overline{1, l-1}.$$

Синтаксема – элементарная единица смысла в высказывании. Задаётся морфологической формой и значением.

Пример: <из-за, род.п., пространственное, аблатив>: *из-за синих гор*

ШАГ 1. Выделить в тексте предложения и слова.

ШАГ 2. Морфологический анализ.

ШАГ 3. Синтаксический анализ.

ШАГ 4. Семантический анализ.

4.1 Определить предикатное слово.

4.2 Установить значения синтаксем и отношений на них с использованием словаря предикатных слов.

Леммы предикатных слов	Отправить, отправлять, направить, направлять, послать, посылать, сослать, выслать, слать (класс: акциональные).	
Семантические роли (синтаксические значения)		
Субъект (инициатор действия)	**КСК**	**Морфологические признаки**
	Личное	**Предлог** / **Падеж**
		- / Именительный
Объект (компонент, подвергающийся действию)	**КСК**	**Морфологические признаки**
	Любой	**Предлог** / **Падеж**
		- / Винительный
Директив (направление движения, ориентированного действия или положения предмета)	**КСК**	**Морфологические признаки**
	Локатив Предметное	**Предлог** / **Падеж**
		В / Винительный
		За / Винительный
		На / Винительный

Семантические связи

Семантическая связь	**Роль 1**	**Роль 2**
OBJ	Субъект	Объект
DIR	Объект	Директив

Осипов Г.С., Смирнов И.В., Тихомиров И.А. Реляционно-ситуационный метод поиска и анализа текстов и его приложения // Искусственный интеллект и принятие решений. – 2008. – №2. – С. 3-10.

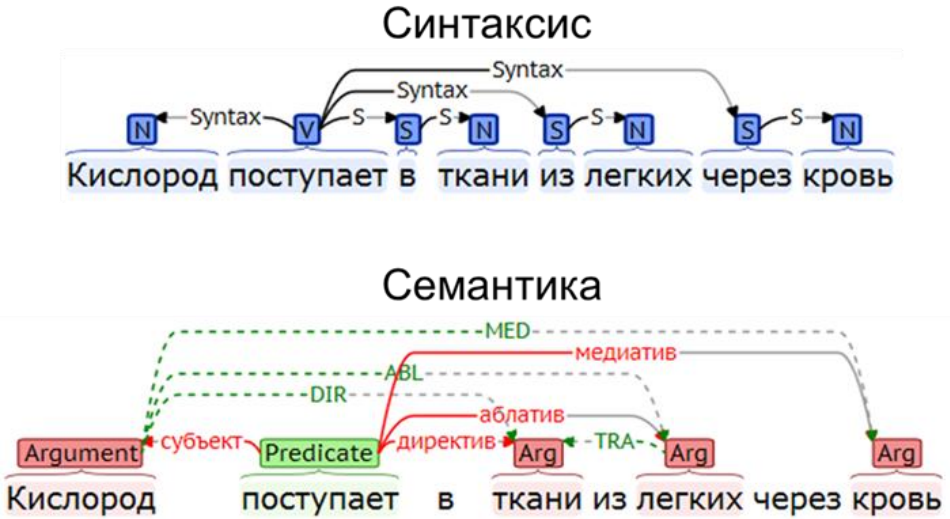
Семантико-синтаксический анализ текста

Принцип: устанавливаем семантические роли совместно с синтаксическим анализом в одной процедуре на одних и тех же структурах данных

Используем для этого метод синтаксического анализа, основанный на переносах (transition-based, MaltParser), способный назначать типы синтаксических связей, соответствующие семантическим ролям. Метод обучается по размеченным корпусам текстов. Его сложность линейна относительно количества слов, а разбор выполняется за один проход по словам предложения.

Создан семантически размеченный подкорпус «СинТагРус»
- более 1 700 предложений
- около 3 000 предикатных конструкций
- около 4 000 аргументов, с установленными ролями

Преимущества
- работает быстрее
- повышает полноту установления ролей
- устанавливает семантические роли у аргументов при неизвестных предикатах

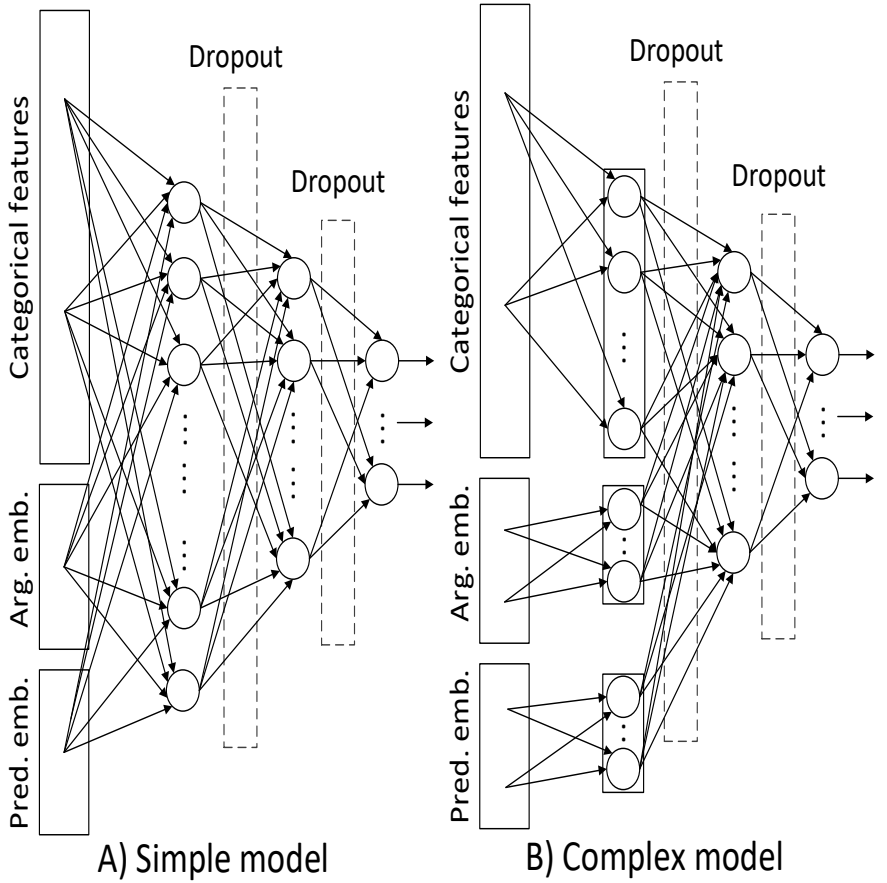


Метод	P, %	R, %	F1, %
Раздельный на основе словаря	89,6	61,0	72,6
Совместный семантико-синтаксический	**89,6**	**62,7**	**73,8**

Shelmanov A. O., Smirnov I. V. Methods for semantic role labeling of Russian texts // Computational Linguistics and Intellectual Technologies. Proceedings of International Conference Dialog. – 2014. – V. 13. – №20. – pp. 607-620.

Нейросетевые подходы к установлению ролей

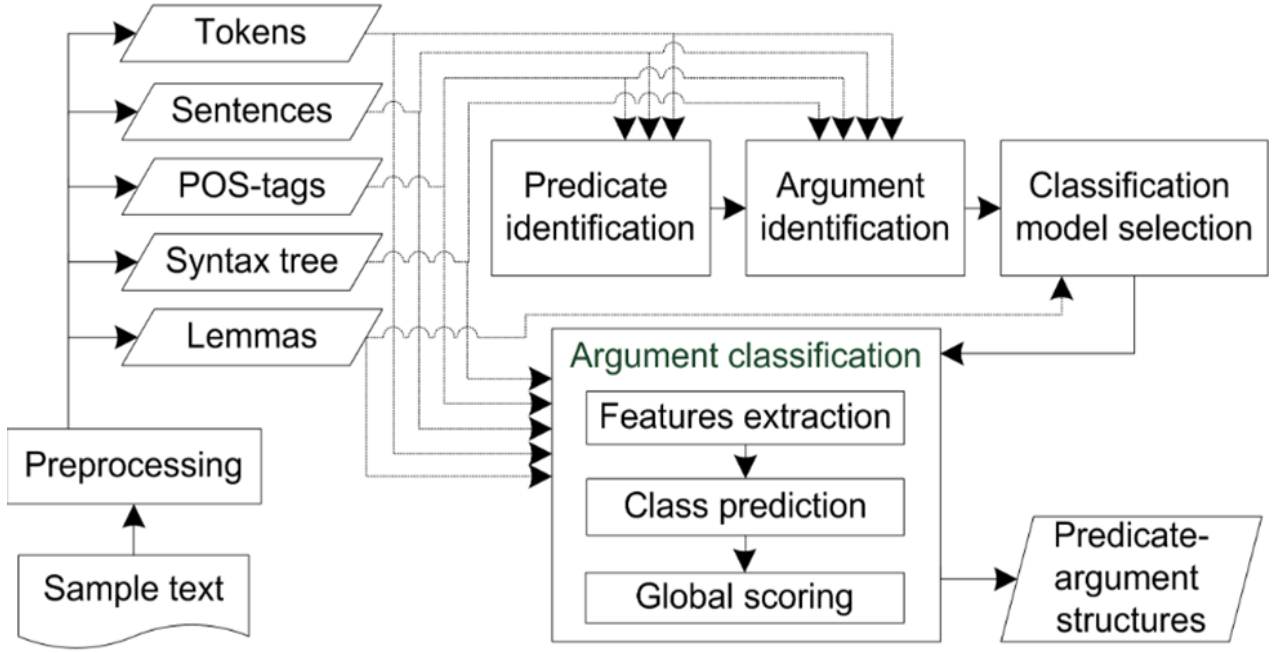
- Neural networks allow to use atomic features
- Categorical features:
  - Various types of morphological features of both an argument and a predicate: part of speech, grammar case, animacy, verb form, time, passiveness, and others (“morph”)
  - Relative position of an argument in a sentence with respect to a predicate (“rel_pos”)
  - Predicate lemma (“pred_lemma”)
  - Preposition of an argument extracted from a syntax tree (“arg_prep”)
  - Name of a syntax link from an argument to its parent in a syntax tree (“synt_link”)
- Embeddings:
  - Embedding of an argument lemma (“arg_emeddings”)
  - Embedding of a predicate lemma (“pred_emeddings”)



Model	Macro F_1-score,%	Micro F_1-score,%
LinearSVC	74.3 ± 0.2	77.6 ± 0.1
LogReg	75.1 ± 0.1	78.2 ± 0.3
LightGBM	71.3 ± 0.4	76.0 ± 0.1
Random Forest	69.7 ± 0.4	71.9 ± 0.1
Top neural network	**79.2 ± 0.3**	**82.3 ± 0.2**

Shelmanov A., Devyatkin D. Semantic role labeling with neural networks for texts in Russian // Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference "Dialogue" (2017). — Vol. 1. — 2017. — P. 245–256.

Предобученные языковые модели для установления ролей

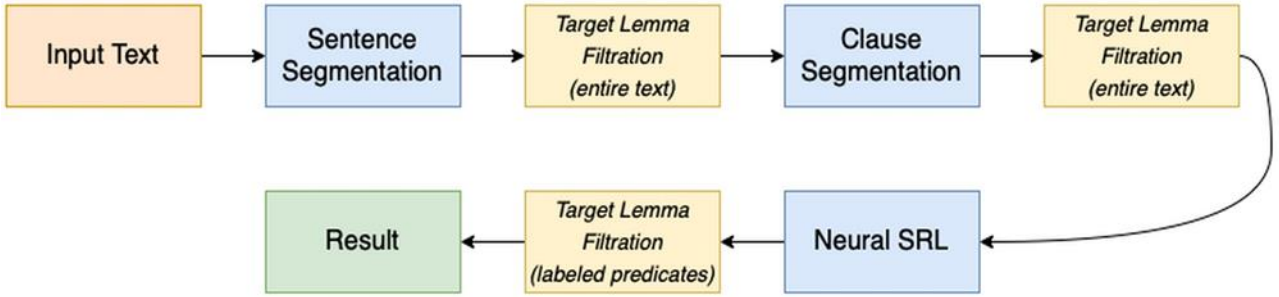


Model	Micro F1	Macro F1
Categ. features only	76.96 ± 0.67	73.63 ± 0.61
Word2Vec UPOS	79.87 ± 0.34	76.70 ± 0.77
FastText	80.60 ± 0.51	77.39 ± 0.36
ELMo	**83.42** ± 0.60	79.91 ± 0.40
BERT-Multiling	79.04 ± 0.63	75.68 ± 0.72
RuBERT	83.12 ± 0.60	**80.12** ± 0.62

Table 1. Performance of models on the corpus with "known" predicates

Model	Micro F1	Macro F1
ELMo (for known pred.)	45.51 ± 0.50	29.31 ± 0.82
Word2Vec UPOS	53.97 ± 0.21	37.29 ± 0.74
FastText	49.37 ± 0.43	37.26 ± 0.29
ELMo	**55.50** ± 0.51	**37.64** ± 0.41
BERT-Multiling	31.81 ± 0.51	21.04 ± 0.13
RuBERT	43.68 ± 0.50	30.84 ± 0.55

Table 2. Performance of models on the corpus with "unknown" predicates



Role	Rule-based Pipeline	Deep Learning Pipeline
Cause	50.5	70.7
Experiencer	26.2	63.6
Predicate	73.7	97.9
Overall	50.1	77.4

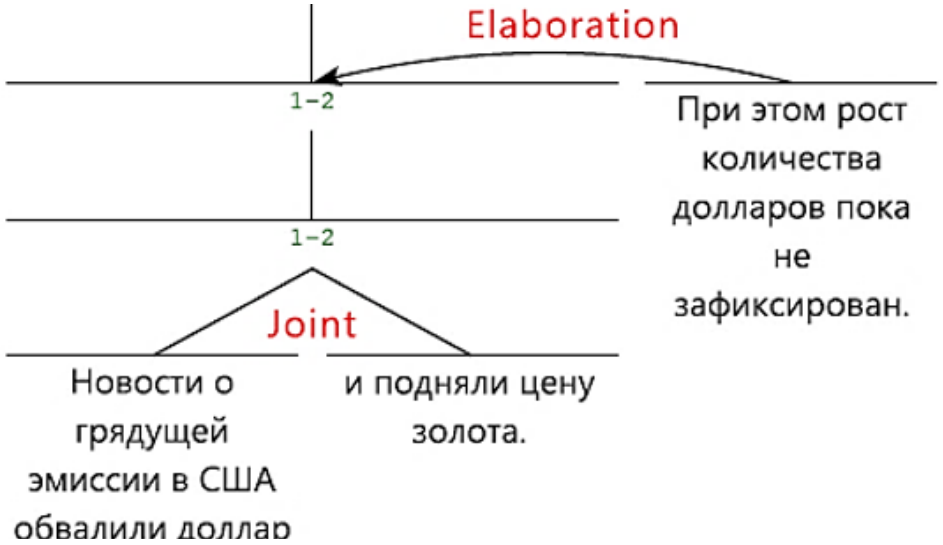
Larionov D., Shelmanov A., Chistova E., Smirnov I. Semantic role labeling with pretrained language models for known and unknown predicates // Proceedings of International Conference on Recent Advances of Natural Language Processing. – 2019. – pp. 619-628.

Дискурсивный анализ текста на русском языке

Существующие модели дискурса: LDM, SDRT, PDT, RST

Для представления связности текста на русском языке предложено использовать теорию риторических структур (RST, Mann W. C. & Thompson S. A.)

Пример отношения: Elaboration (Детализация). Сателлит содержит ту же информацию, что и ядро, но информация в сателлите более развернутая и подробная



$$P_{disc} = < U, R_{discU} >,$$

$$R_{discU} = \{\langle (u_1, u_2), t_{disc} \rangle \mid t_{disc} \in Type_{disc}, u_i \in U, i = 1,2\}.$$

Метод дискурсивного анализа

ШАГ 1. Дискурсивная сегментация

- маркировка последовательностей слов с помощью модели BiLSTM-CRF

ШАГ 2. Построение дискурсивного дерева

Для каждой пары дискурсивных единиц:

2.1 Определить, есть ли между ними отношение

- структурный классификатор на основе метода симметричного сравнения текстов BiMPM

2.2 Если есть отношение, установить его тип

- классификатор отношений на основе метода симметричного сравнения текстов BiMPM

Input: List of discourse units $[e_1, e_2, ..., e_n]$
Output: Discourse trees
$Trees \leftarrow [e_1, e_2, ..., e_n]$
$Scores \leftarrow \emptyset$
for $i \leftarrow 1$ *to* $n-1$ **do**
  $Scores[i] = getScore(e_i, e_{i+1})$
end
while $|Trees| > 1$ *and* $any(Scores) > confidenceThreshold$ **do**
  $j = argmax(Scores)$
  $NewDU = mergeNodes(j, j+1)$
  $NewDU.relation = getRelation(NewDU)$
  Replace $Trees[j]$ and $Trees[j+1]$ with $NewDU$
  if $j \neq 0$ **then**
    $Scores[j-1] = getScore(L[j-1], NewDU)$
  end
  if $j \neq length(Scores)$ **then**
    $Scores[j] = getScore([NewDU, L[j+1]])$
  end
end
return $Trees$

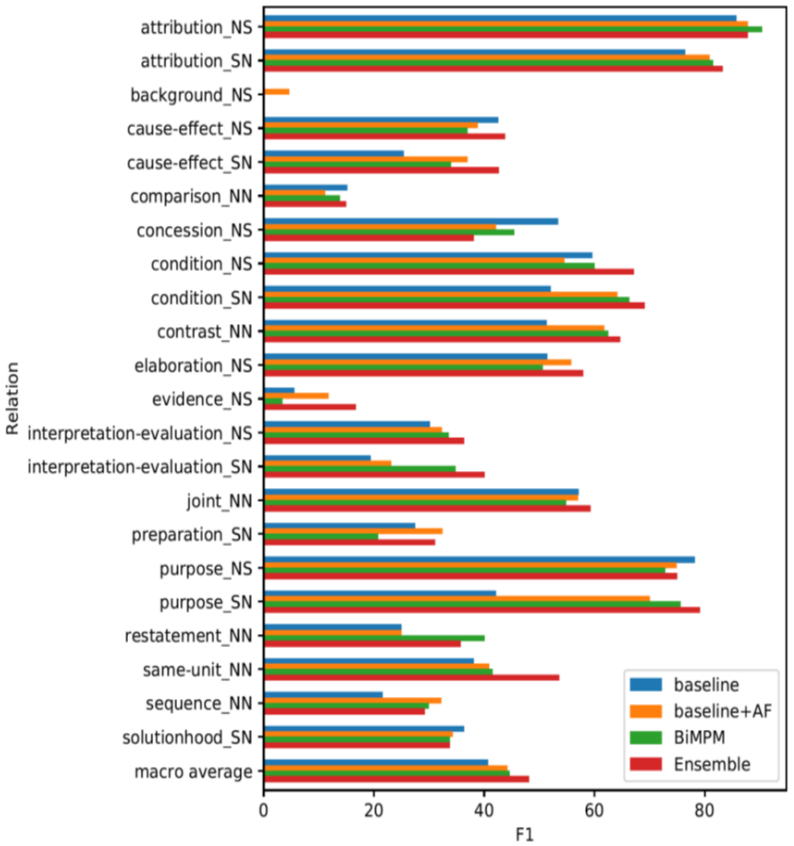
Chistova E. V., Shelmanov A. O., Kobozeva M. V., Pisarevskaya D. B., Smirnov I. V., Toldova S. Yu. Classification models for RST discourse parsing of texts in Russian // Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialogue 2019”. – 2019. – pp. 163-176.

Оценка качества дискурсивного анализа

Создан открытый корпус текстов с дискурсивной разметкой Ru-RSTreebank

- 330 текстов, 435 000 словоупотреблений
- 27683 элементарных дискурсивных единиц
- 24960 отношений

Этап	Метод	Признаки	Test		
			P	R	F1
Сегментация	Baseline	BERT-M	**89,66**	85,56	87,56
	BiLSTM+ CRF	BERT-M	87,80	**88,99**	88,39
		ELMo	89,09	87,86	**88,42**
Структурная классификация	Baseline	Baseline +AF	**58,42**	76,38	66,21
	BiMPM	ELMo	54,54	82,82	65,77
	Baseline+ BiMPM	Baseline+AF, ELMo	57,66	**83,06**	**68,07**
Классификация риторических отношений	Baseline	Baseline	42,48	41,32	40,63
		Baseline+AF	46,54	44,18	44,19
	BiMPM	ELMo	47,35	45,40	44,64
	Baseline+ BiMPM	Baseline+AF, ELMo	**49,89**	**47,73**	**47,50**



Нейросетевой классификатор показывает наилучший результат для распознавания большинства риторических отношений по сравнению с классификатором на основе лингвистических признаков

Комбинация базового метода и BiMPM-классификатора позволяет улучшить макроусредненный показатель F-меры на 2,86%.

Chistova E., Shelmanov A., Pisarevskaya D., Kobozeva M., Isakov V., Panchenko A., Toldova S., Smirnov I. RST Discourse Parser for Russian: an Experimental Study of Deep Learning Models // International Conference on Analysis of Images, Social Networks and Texts. Lecture Notes in Computer Science. – 2021. – V. 12602. – pp. 105-119.

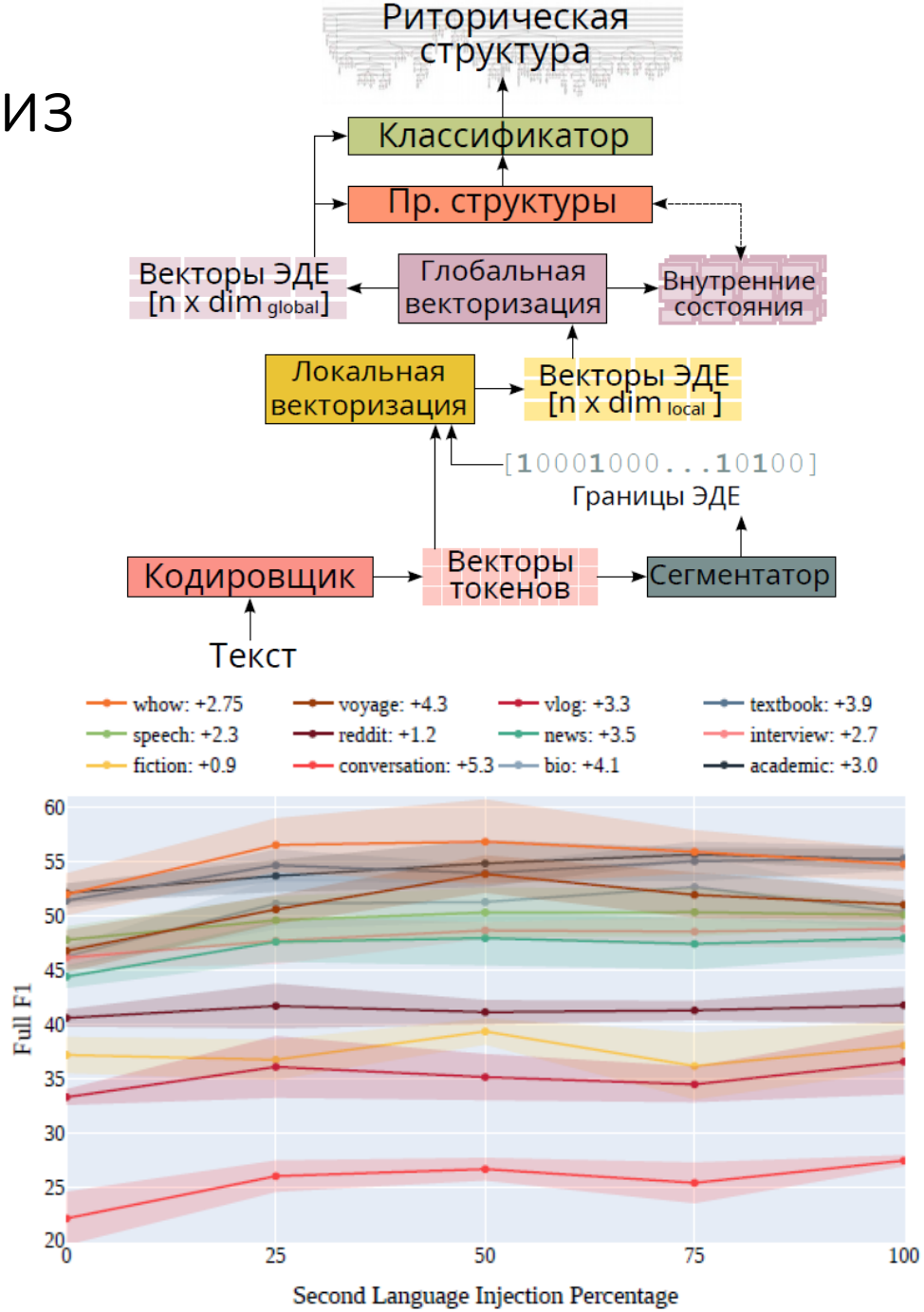
Полнотекстовый дискурсивный анализ

На основе открытого корпуса GUM RST v9.1 разработана полнотекстовая разметка для русского языка

Выводы:

- Высокое качество непосредственного переноса
- Двуязычная модель анализирует русский язык качественнее, чем модель, обученная только на RRG
- Качества, близкого к оптимальному, можно достичь даже с 25% параллельной разметки для каждого жанра.

Смешение разных корпусов текстов позволило достичь F1=65.4% для русского языка



Chistova Elena. End-to-End Argument Mining over Varying Rhetorical Structures // Findings of the Association for Computational Linguistics: ACL 2023. — Toronto, Canada: Association for Computational Linguistics, 2023. — Pp. 3376–3391.

Учет семантической структуры в информационном поиске и сопоставлении текстов

«Ситуативные» значения слов

- купил **подарок**[*объект*]
- обрадовался **подарку**[*каузатив_эмоц*]
- потратиться на **подарок**[*дестинатив*]
- потратил **подарок**[*объект*]
- разбил **подарок**[*деструктив*]
- разбил **подарком**[*инструмент*]
- упал на **подарок**[*директив*]
- мотивировал **подарком**[*медиатив*]
- расплатился за **подарок**[*предмет_обмена*]
- …

Запрос

Гипертония [*каузатив*] приводит к нарушению [*результатив*] кровоснабжения

Документ

Нарушение [*каузатив*] кровоснабжения при гипертонии приводит к инсульту [*результатив*]

Надо учитывать не только лексемы, но и их значения в высказываниях

Corcoglioniti F. et al. Knowledge extraction for information retrieval // The Semantic Web. Latest Advances and New Domains: 13th International Conference, ESWC 2016, – 2016

Tymoshenko K., Moschitti A. Assessing the impact of syntactic and semantic structures for answer passages reranking // Proceedings of the 24th ACM International on Conference on Information and Knowledge Management. – 2015. – pp. 1451-1460

Mohebbi M., Razavi S. N., Balafar M. A. Computing semantic similarity of texts based on deep graph learning with ability to use semantic role label information // Scientific reports. – 2022. – V.12. – №1. – pp. 1-11

Sinoara R. A., Rossi R. G., Rezende S. O. Semantic role-based representations in text classification // 23rd International Conference on Pattern Recognition (ICPR). – IEEE, 2016. – C. 2313-2318

Семантический поиск на основе реляционно-ситуационной структуры текста

Принцип:

- При вычислении степени релевантности документа запросу учитываем не только слова, но и разноуровневые структуры

Алгоритм вычисления релевантности по семантическим ролям

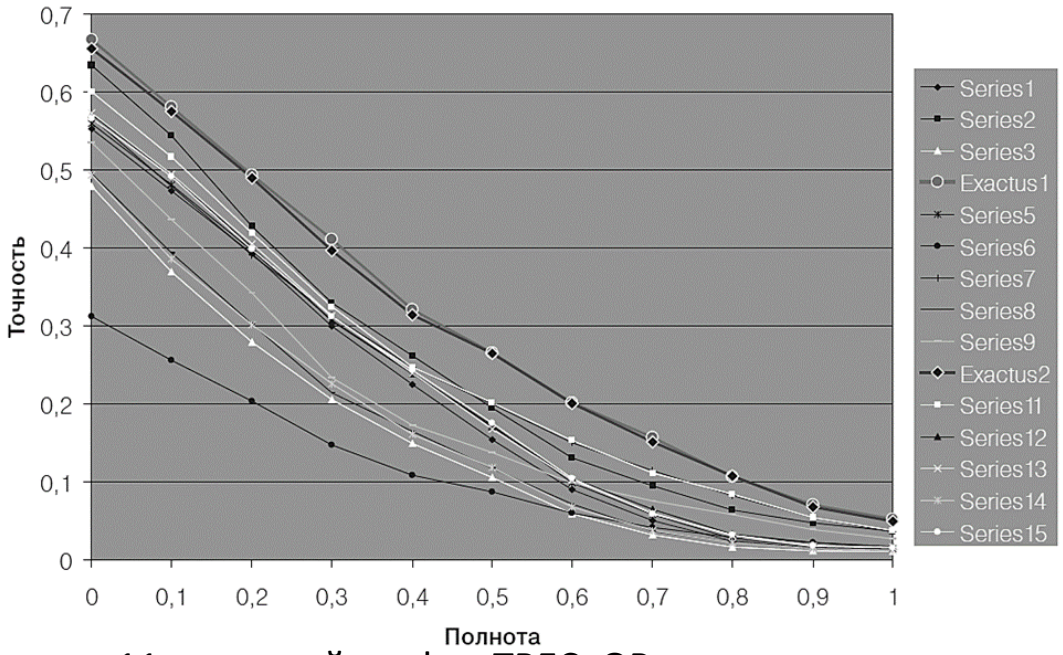
$$REL_{SEMROLE}(q,d) = \frac{|S^q \cap_{val} S^d|}{|S^q|},$$

$$S^q \cap_{val} S^d = \{s^q \in S^q | (w_N^q = w_N^d) \,\&\, (r_s^q = r_s^d), w_N^q \in s^q, w_N^d \in s^d\},$$

1. Цикл по синтаксемам запроса.
1.1 Цикл по синтаксемам документа.
1) Если совпали семантические роли синтаксем запроса и документа, то:
а) Сравниваем главные слова синтаксем.
б) Если совпадают главные слова синтаксем, увеличиваем релевантность документа на 1,0. Выходим из цикла по синтаксемам документа.
2. Нормируем релевантности по ролям путем деления на число синтаксем в запросе.

Типы релевантности:

- По словам – совпадение слов и словоформ
- По словосочетаниям – совпадение групп синтаксически связанных слов
- По семантическим ролям – совпадение семантических ролей
- По семантическим связям – совпадение семантических связей
- Общая семантическая релевантность – свертка остальных



11-точечный график TREC. OR-оценка

Смирнов И.В., Соченков И.В., Муравьев В.В., Тихомиров И. А. Результаты и перспективы поискового алгоритма Exactus // Труды российского семинара по оценке методов информационного поиска РОМИП'2007-2008. – 2008. – С. 66-76.

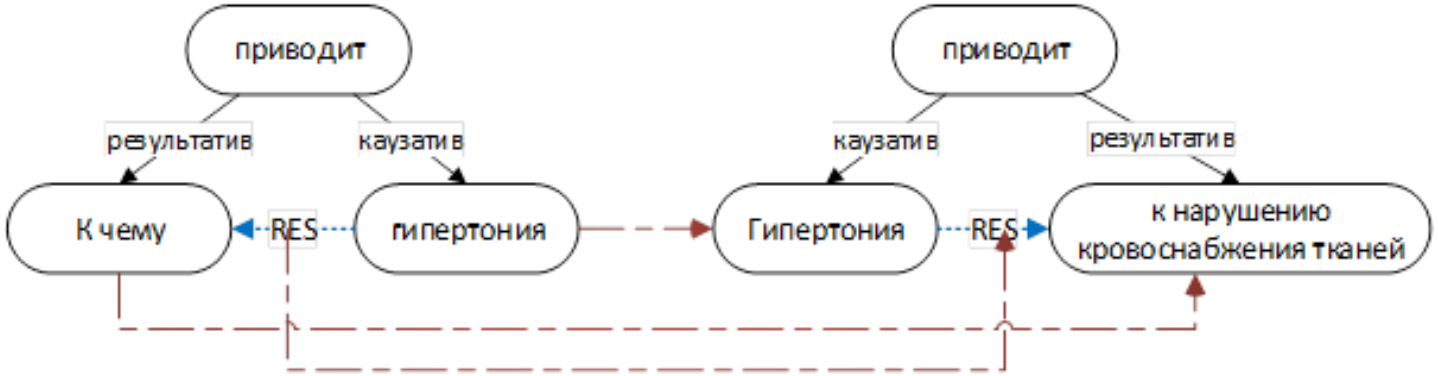
Вопросно-ответный поиск на основе семантических структур

Принцип: слово в тексте с той же семантической ролью, что и вопросительное слово запроса, является ответом, при совпадении остальной структуры

Разработаны:

- Алгоритм ранжирования результатов вопросно-ответного поиска
- Алгоритм определения лексико-семантической оценки релевантности предложения текста запросу
  - лексическая оценка r_l^s – близость запроса и предложения текста по лексике
  - оценка семантических ролей r_{sr}^s – близость запроса и предложения текста по семантическим ролям
  - оценка семантических отношений r_{sn}^s – близость запроса и предложения текста по семантическим отношениям

$$r_{ls}^s = \gamma_l \times r_l^s + \gamma_{sr} \times r_{sr}^s + \gamma_{sn} \times r_{sn}^s$$



Вопрос: к чему приводит гипертония?
Ответ: к нарушению кровообращения тканей

Представим семантический аргумент в виде четверки
a=<role,pred,syn,neg>

Оценка сходства семантических аргументов по ролям:

$$r_{sr}^{pa}(a^q, a^s) := r_{sr}^{pa}(a^q, a^s) + w_p$$

$$r_{sr}^{pa}(a^q, a^s) := r_{sr}^{pa}(a^q, a^s) + w_a \times sim(syn^s, syn^q)$$

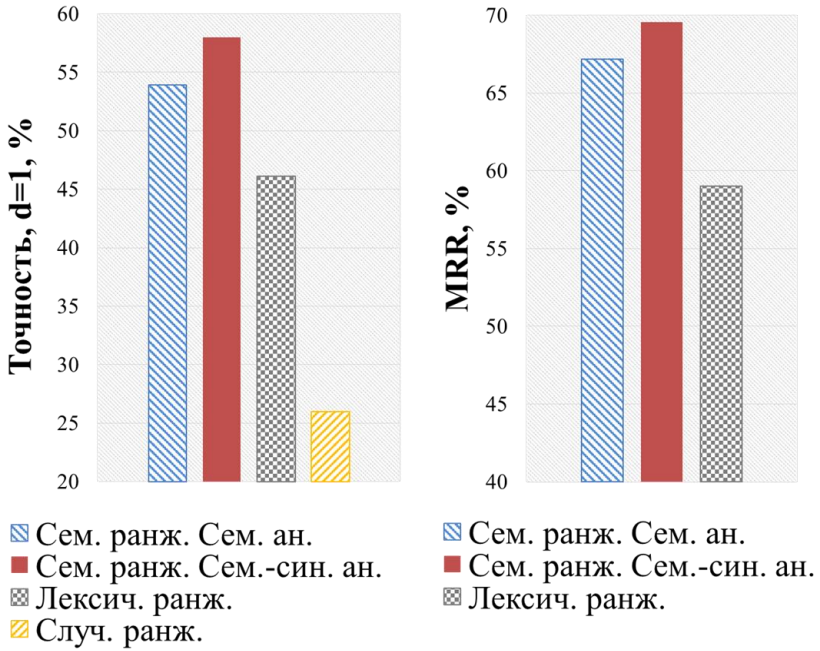
$$r_{sr}^{p}(p_i, p_j) = \sum_{k=1}^{K_i} \max_m r_{sr}^{pa}(a_k^q, a_m^s)$$

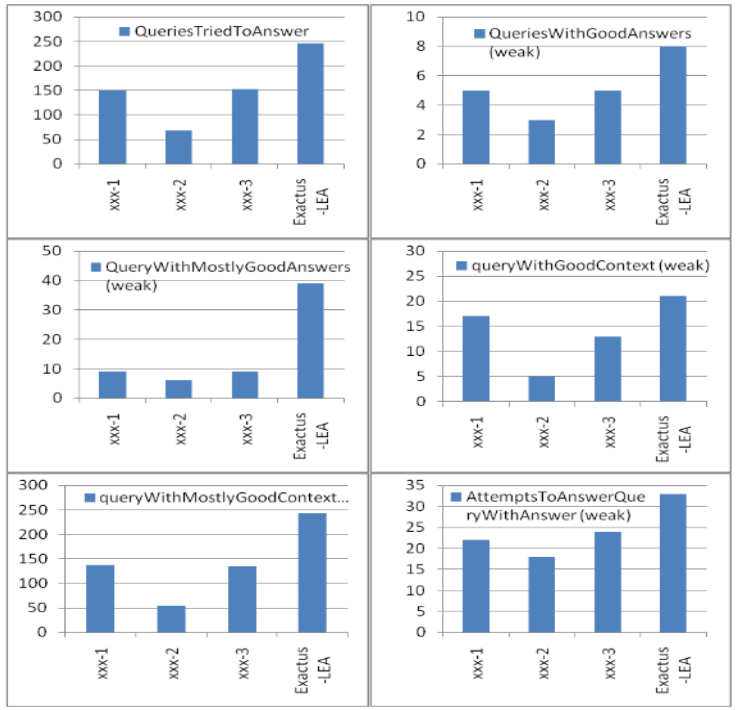
$$r_{sr} = \delta_{sr} \frac{\sum_{i=1}^{P_q} \max_j r_{sr}^{p}(p_i, p_j)}{Z_{sr}}.$$

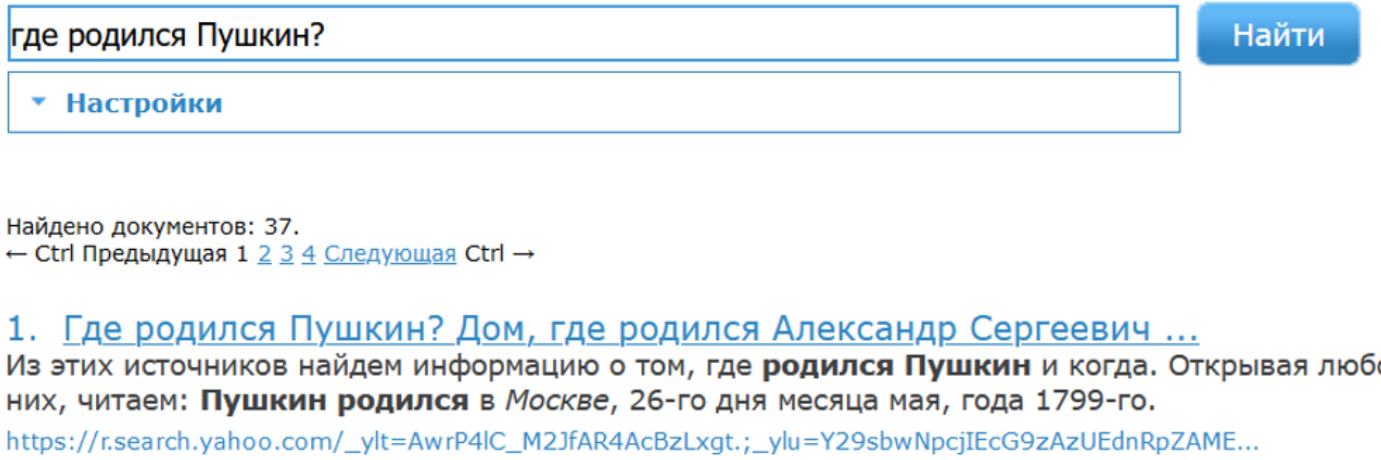
Экспериментальная проверка вопросно-ответного поиска на основе семантических структур

Проверка на:

- Специально созданном наборе пар вопрос-ответ
- Семинар РОМИП, дорожка вопросно-ответного поиска





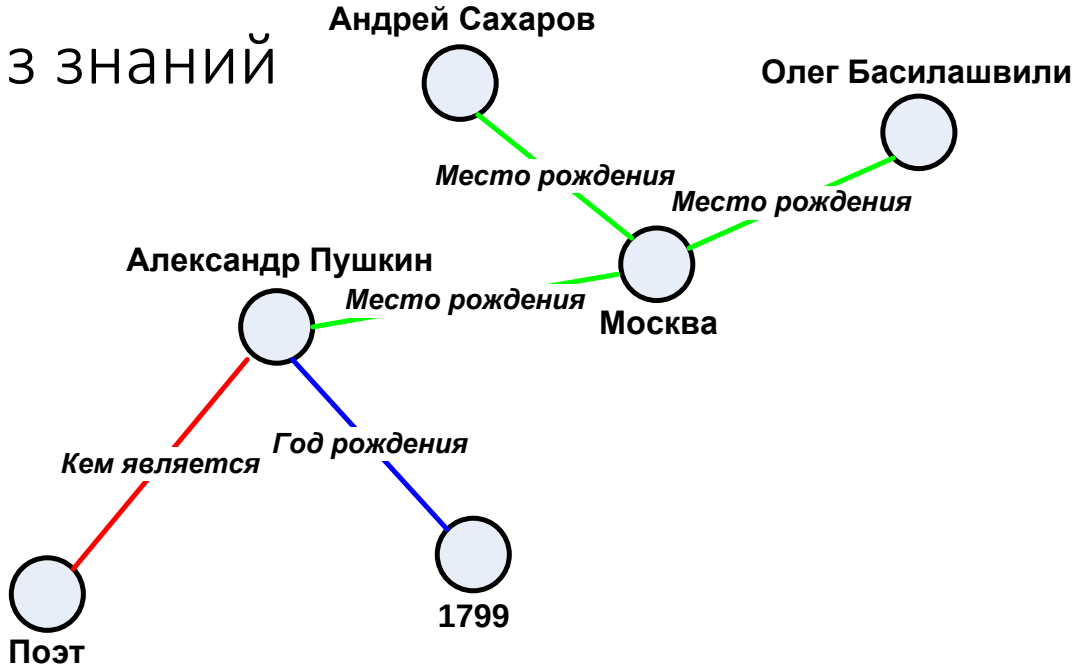


Шелманов А.О., Каменская М.А., Ананьева М.И., Смирнов И.В. Семантико-синтаксический анализ текстов в задачах вопросно-ответного поиска и извлечения определений // Искусственный интеллект и принятие решений. – 2016. – №4. – С. 47-61.

Вопросно-ответный поиск на основе баз знаний

Предложен метод автоматического построения баз знаний из текстов на основе открытого извлечения информации из текстов (Oren Etzioni, Anthony Fader, Janara Christensen)

- ШАГ 1. Извлекаем все именные группы из синтаксических деревьев
- ШАГ 2. Для каждой именной группы вычисляем её вес C-value
- ШАГ 3. Ранжируя именные группы по убыванию C-Value формируем список сущностей
- ШАГ 4. Формируем триплеты <сущность 1, предикат, сущность 2>
- ШАГ 5. Группируем триплеты с помощью глубокой кластеризации методом IDEC
  - Инициализация на аргументах
  - Инициализация на предикатах
  - Полная инициализация
- ШАГ 6. По номеру кластера назначить каждому триплету смысловую метку – семантическое отношение



Оценка качества группировки триплетов

Подход	F_1, %
Тривиальный	10,7
Случайный	25,2
Глубокая кластеризация на векторных представлениях глаголов (инициализация на предикатах)	46,3
Глубокая кластеризация на всех признаках (полная инициализация)	53,0

Предложен способ вопросно-ответного поиска на основе разработанного метода открытого извлечения информации

Чистова Е. В., Ларионов Д. С., Шелманов А. О., Латыпова Е. А., Смирнов И. В. Открытое извлечение информации из текстов. Часть III. Система вопросно-ответного поиска // Искусственный интеллект и принятие решений. – 2021. – №4. – С. 35-49.

Извлечение информации из текстов на основе семантических структур

Используются лексико-синтаксические шаблоны (Большакова) и машинное обучение (Zadgaonkar A. V., Agrawal)

Предложены лексико-семантические шаблоны

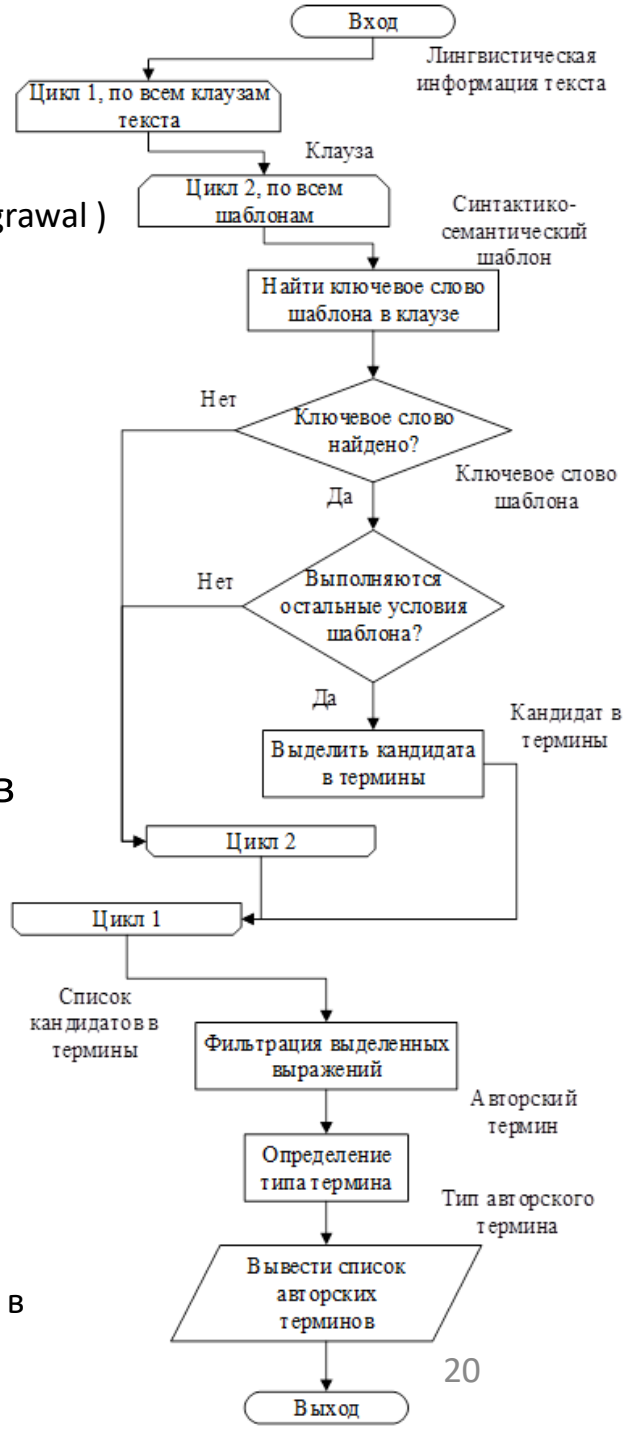
Примеры:

- **ЧР(Сущ.) && Сем.роль(эстиматив) + ПС(«называться»)**

- **ЧР(Сущ.) && Сем.роль(делибератив) + ПС(«определять») + Л(«как»)**

Позволяют создавать более простые и более общие правила извлечения информации

Разработан алгоритм извлечения информации из текстов на основе шаблонов



Оценка качества извлечения дефиниций из научных текстов

Подход	Precision,%	Recall,%	F_1,%
Семантический анализ раздельный	80,6	67,4	73,4
Семантико-синтаксический совместный	**80,7**	**67,6**	**73,6**
Без семантики	76,7	52,5	62,3

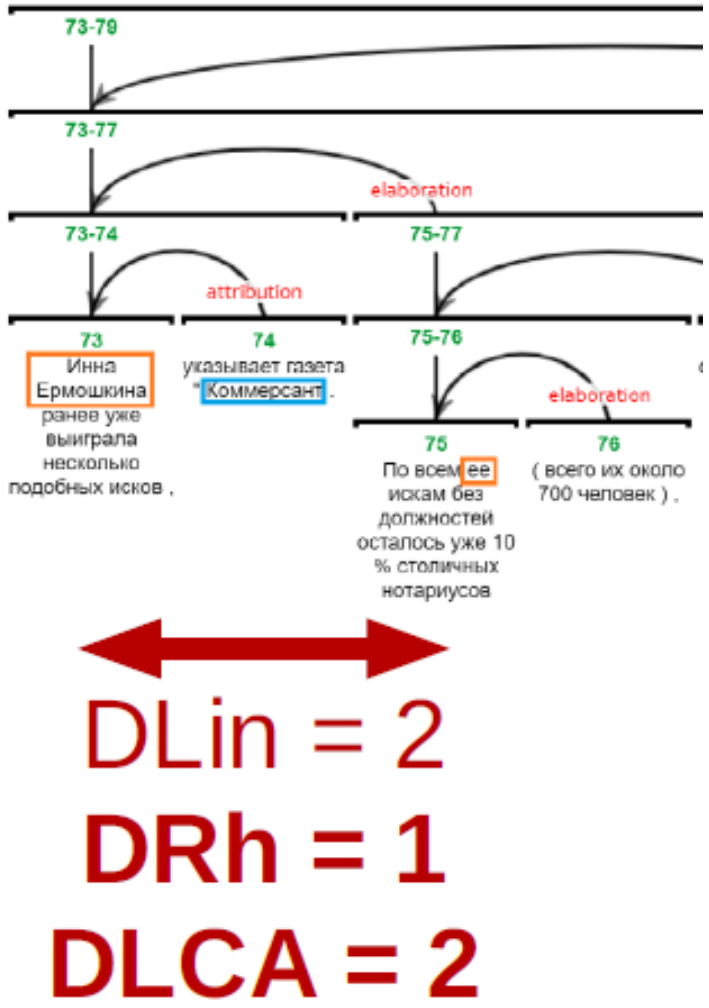
Шелманов А.О., Каменская М.А., Ананьева М.И., Смирнов И.В. Семантико-синтаксический анализ текстов в задачах вопросно-ответного поиска и извлечения определений // Искусственный интеллект и принятие решений. – 2016. – №4. – С. 47-61.

Дискурсивная структура для разрешения кореференции

Космонавт [**антецедент**] вернулся на борт станции. Он [**анафор**] сообщил, что чувствует себя нормально.

«Риторическое расстояние» - количество узлов в риторической структуре между анафором и потенциальным антецедентом

- Линейное расстояние D_Rh – расстояние между антецедентом и референтом в элементарных дискурсивных единицах;
- Риторическое расстояние D_Lin – расстояние между антецедентом и референтом в графе риторической структуры;
- Расстояние до наименьшего общего предка в риторическом дереве D_LCA.



Проверка на наборе данных RuCoCo-2023

	Precision	Recall	F1	Top-1 F1 (leaderboard)
Baseline	79.1 ± 0.8	66.9 ± 0.6	**72.5 ± 0.3**	72.8
$+D_{Rh}$	79.3 ± 1.6	66.6 ± 1.9	72.4 ± 0.5	**73.3**

We observed a marginal improvement using the rhetorical distance feature.
The model that uses this feature got the best result on the Shared Task development and test sets.

Chistova E., Smirnov I. Light Coreference Resolution for Russian with Hierarchical Discourse Features //Proceedings of the International Conference "Dialogue". – 2023. – pp. 34-41.

Дискурсивная структура для классификации текстов

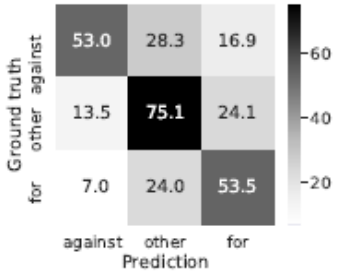
- Если соблюдать карантин месяц, то вирус будет остановлен *#за карантин, с аргументацией*
- Любители масок, не ужели вы думаете, что эта косметическая тряпочка поможет от вируса?! *#против масок, без аргументации*

Двухэтапная классификация:

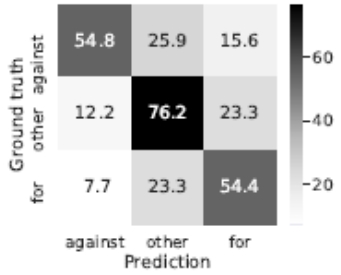
- На первом этапе используется модель классификации как последовательности токенов (например, на основе языковой модели). Обучается на всем корпусе, на этапе предсказания применяется ко всем ДЕ в риторическом дереве.
- На втором этапе классификатор на основе рекуррентной нейронной сети использует предсказания первой модели в каждом узле риторического дерева для формирования общего класса текста.

Метод	Тип текстов	Маски	Вакцины	Карантин	Mean
		Позиция автора			
BERT	Неэлементарные	59,8 ± 2,7	62,4 ± 3,4	54,5 ± 3,4	58,9 ± 2,3
	Все	60,6 ± 2,6	64,4 ± 2,2	56,4 ± 2,8	60,5 ± 1,9
+ RST-LSTM	Неэлементарные	**61,3 ± 2,7**	**63,4 ± 4,2**	**55,6 ± 2,7**	**60,1 ± 2,3**
	Все	**61,7 ± 2,6**	**65,1 ± 3,0**	**57,5 ± 2,4**	**61,4 ± 1,8**
		Аргументация			
BERT	Неэлементарные	66,4 ± 2,9	**61,7 ± 4,3**	56,4 ± 2,8	61,5 ± 2,2
	Все	66,0 ± 2,4	**62,6 ± 2,7**	57,0 ± 2,3	61,9 ± 1,6
+ RST-LSTM	Неэлементарные	**68,1 ± 2,1**	60,4 ± 3,3	**57,6 ± 2,0**	**62,0 ± 1,3**
	Все	**67,5 ± 1,9**	61,5 ± 2,3	**58,3 ± 2,1**	**62,4 ± 0,9**

Таблица 24 — Оценки качества на кросс-валидации. F1, в %.



(а) BERT



(б) BERT + RST-LSTM

Chistova Elena, Smirnov Ivan. Discourse-aware text classification for argument mining // Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference “Dialogue”. — No. 2022. — 2022. — Pp. 93–105.

Дискурсивная структура для анализа аргументации

Предложены два варианта анализатора:
- BAP: биафинный анализатор структуры аргументации в коротком тексте-рассуждении как дерева зависимостей.
- DBAP: введены коэффициенты наличия риторических отношений между дискурсивными единицами в риторическом дереве.

Поскольку одна структура аргументации может соответствовать разным риторическим структурам, предложено выявлять общие закономерности из множества вариантов риторической структуры.

1 In fact, it would be justified if all German universities charged tuition fees.
2 As long as it is guaranteed that the funds really benefit the universities directly, we can continue to regard it as social justice.
3 In any case, the question of further training must be decided in advance.
4 You can always take a student loan or get a scholarship.
5 However, it is unfair to oblige people who do not belong to scientific circles to pay for someone else's education by collecting additional taxes.

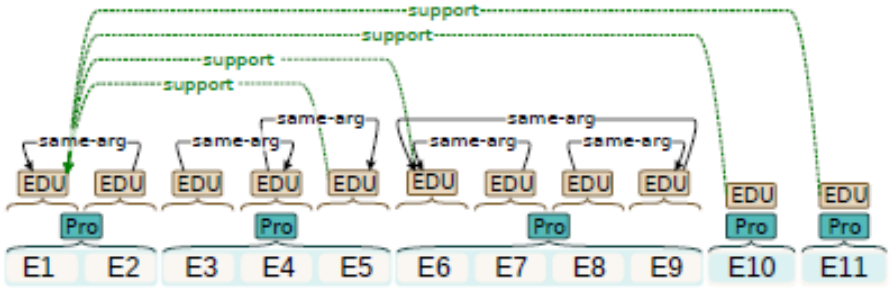


Figure 3: Argument tree representation in the end-to-end parser, `micro_k002:En`.

Для экспериментов использовался англо-русский паралелльный корпус Microtexts

Lang	Method	Augmented	cc	ro	fu	at	UAS	LAS
En	BAP	No	88.3 ± 4.9	**71.1 ± 5.7**	77.1 ± 4.6	53.8 ± 6.8	59.1 ± 6.8	52.9 ± 6.3
		Yes	88.9 ± 4.7	69.2 ± 3.9	**78.3 ± 4.9**	56.2 ± 5.9	61.2 ± 5.8	55.1 ± 5.9
	DBAP	No	**90.3 ± 3.3**	68.8 ± 6.9	77.3 ± 3.2	59.7 ± 7.4*	64.5 ± 6.6*	56.2 ± 5.3*
		Yes	89.5 ± 4.3	68.8 ± 7.6	76.5 ± 3.1	**60.1 ± 4.3****	**64.6 ± 4.1***	**56.6 ± 3.2***
Ru	BAP	No	**90.5 ± 5.7**	69.3 ± 7.8	78.9 ± 4.2	56.1 ± 6.3	61.7 ± 6.6	55.2 ± 6.7
		Yes	90.3 ± 2.8	66.9 ± 6.9	77.5 ± 4.3	56.1 ± 5.1	61.6 ± 4.7	53.9 ± 5.7
	DBAP	No	90.3 ± 5.7	68.9 ± 2.5	**79.8 ± 3.6**	59.8 ± 5.3	**64.6 ± 5.8**	**58.0 ± 3.6**
		Yes	88.3 ± 6.4*	**69.9 ± 5.4**	77.2 ± 6.1	**60.6 ± 4.9***	**64.6 ± 5.8**	57.0 ± 5.8

Chistova E. End-to-End Argument Mining over Varying Rhetorical Structures // Findings of the Association for Computational Linguistics: ACL 2023. – 2023. – Pp. 3376-3391.

Приложение в психолингвистических исследованиях

Задачи:

- выявление взаимосвязи между психологическими особенностями, психическими отклонениями человека и его письменной речью
- предсказание по тексту психологических особенностей автора или психологического неблагополучия у него

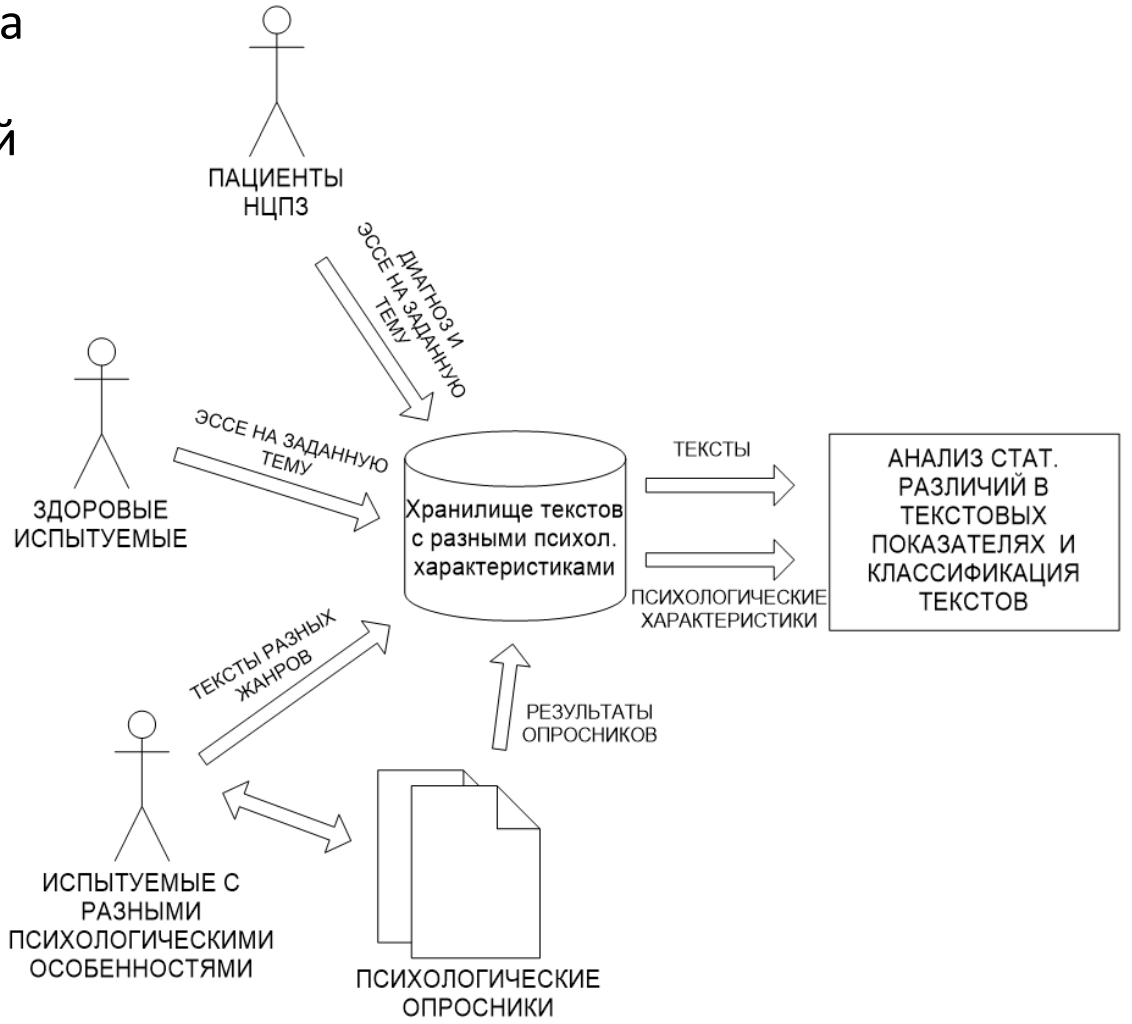
Используются преимущественно лексические подходы

Предложены разноуровневые психолингвистические показатели текста, отражающие взаимосвязь текстов и психологических особенностей авторов на всех уровнях:

- Морфо-стилистические
- Лексические
- Синтаксические
- Семантические
- Дискурсивные

Валидизация на текстах сочинений пациентов психиатрической клиники и здоровых

Схема валидизации показателей



Морфо-стилистические текстовые показатели

- Коэффициент логической связности = общее количество служебных слов (союзов и предлогов) / общее количество предложений

$$P_{Coh}(T) = \frac{P_{wcount}(T,"часть\_речи:союз") + P_{wcount}(T,"часть\_речи:предлог")}{P_{scount}(T) \times 3}.$$

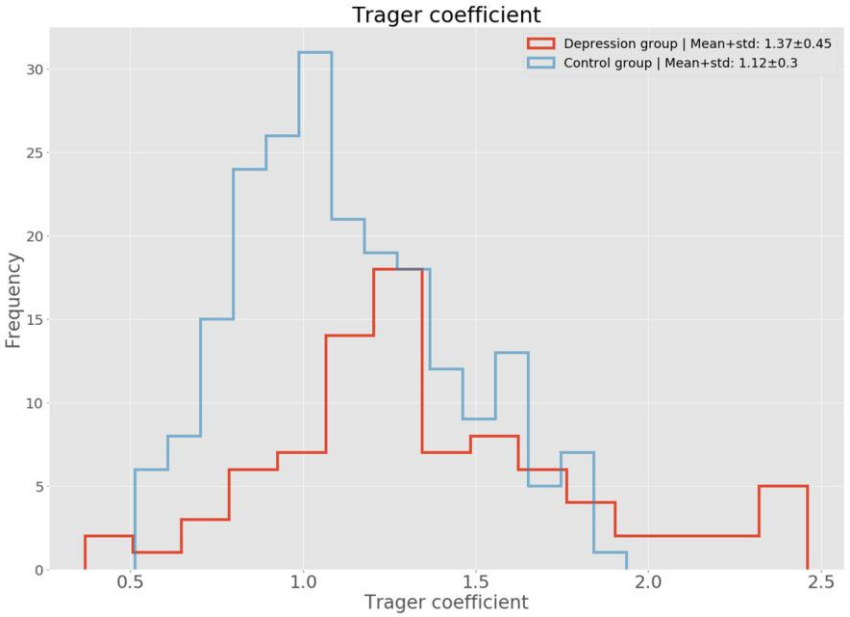
- Коэффициент Трейгера = отношение количества глаголов к количеству прилагательных в единице текста

$$P_{Traiger}(T) = \frac{P_{wcount}(T,"часть\_речи:глагол") - P_{wcount}(T,"форма:причастие") - P_{wcount}(T,"форма:деепричастие")}{P_{adjw}(T) + P_{wcount}(T,"часть\_речи:мест.-прилаг.") + P_{wcount}(T,"часть\_речи:числит.-прилаг.")}$$

- Коэффициент опредмеченности действия = отношение количества глаголов к количеству существительных в единице текста

$$P_{Obj}(T) = \frac{P_{wcount}(T,"часть\_речи:глагол")}{P_{wcount}(T,"часть\_речи:сущ.") - P_{wcount}(T,"часть\_речи:мест.-сущ.")}.$$

- И другие
- **Всего 56 показателей**



Психолингвистические показатели	Здоровые	Больные	Значимость различий
Морфо-стилистические показатели			
Кол. инфинитивов относительно кол. глаголов	0,3118±0,0874	0,2363±0,1351	**
Местоимения 1 лица мн. числа	0,0109±0,0112	0,0051±0,0102	**
Местоимения 1 лица единств. числа	0,0277±0,0222	0,0598±0,0288	**
Местоимения 3 лица мн. числа	0,0053±0,0058	0,0043±0,0066	*
Коэффициент Трейгера	0,8056±0,2200	0,6909±0,2697	*
Длина слов в символах	4,4506±0,3359	4,2794±0,3456	*

* – p <0,05; ** – p < 0,001

Лексические (словарные) текстовые показатели

- Лексика стенических негативных эмоций
- Положительная эмоциональная оценка
- Безысключительная и усилительная лексика
- Лексика разрушения и насилия
- Инвективы и отрицательная экспрессия
- Молодежный жаргон
- Мягкие инвективы
- Обсценная лексика
- Политическая лексика и канцеляризмы
- Положительная рациональная оценка и когнитивные действия
- Лексика протестного поведения
- Отрицательная рациональная оценка
- Лексика социальной разобщенности
- Лексика страдания

$$P_{wcount}\ (T, D) = \sum_{S_{gr} \in T_{gr}} \sum_{C_{gr} \in S_{gr}} |C_{gr}^{D}|$$

где $C_{gr}^{D} = \{w \in C_{gr} | \ w_N \cap D \neq \emptyset\}$,

w_N – лемма слова в клаузе C_{gr}

$$P_{wfreq}\ (T, D) = \frac{P_{wcount}\ (T, D)}{P_{wcount}\ (T)}$$

$$P_{sfreq}\ (T, D) = \frac{P_{wcount}\ (T, D)}{P_{scount}\ (T)}$$

Всего 21 словарь ~50.000 слов

Лексика страдания: *бедствие, болезнь, невезение….*
Лексика стенических негативных эмоций: *беспокоить, ненавидеть, обидеть, сердиться….*
Инвективные обозначения отрицательных черт характера, моральных качеств, поведения и образа жизни: *хитрость злобный злодейство язвительность грубость …*

Психолингвистические показатели	Здоровые	Больные	Значимость различий
Лексические показатели			
Словарь: Дети и образование	0,0026±0,0043	0,0005±0,0016	**
Словарь: Здравоохранение и медицина	0,0002±0,0009	0,0018±0,0049	**

** – p < 0,001

Синтаксические и семантические текстовые показатели

- Средняя глубина синтаксического дерева
- Максимальная глубина синтаксического дерева
- Минимальная глубина синтаксического дерева
- Отношение простых и сложных предложений
- Среднее количество клауз (простых предложений и оборотов) в предложении
- Частотность синтаксических связей
- Частотность семантических ролей и отношений
- И другие
- **48 синтаксических и 98 семантических показателей**

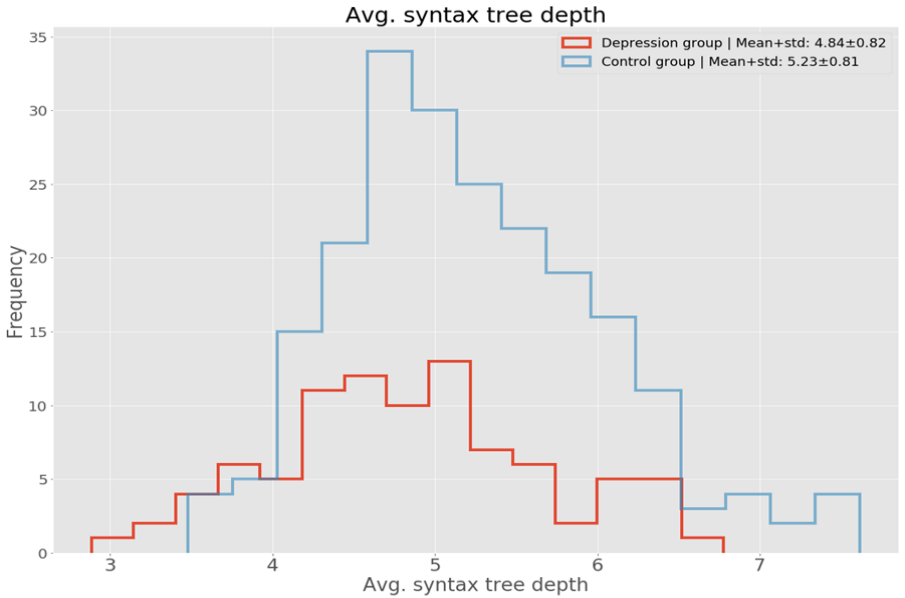
$$sc = (<(w^i_1, w^i_2), t^i_{synt}>) \mid w^i_2 = w^{i+1}_1, i = \overline{1, n-1}$$

$$P_{syntlenmax}\ (T) = \max_{S_{synt} \in T_{synt}} \max_{C_{synt} \in S_{synt}} \max_{sc \in CS_{synt}} len(sc)$$

$$P_{syntlenaver}\ (T) = \frac{\sum_{S_{synt} \in T_{synt}} \max_{C_{synt} \in S_{synt}} \max_{sc \in CS_{synt}} len(sc)}{P_{scount}\ (T)}$$

$$P_{SemRole}\ (T, \quad srole) = \frac{P_{semrolecount}\ (T, srole)}{P_{semrolecount}\ (T)}$$

Средняя глубина синтаксического дерева



Психолингвистические показатели	Здоровые	Больные	Значимость различий
Семантические роли			
Агенс	0,0012±0,0023	0,0004±0,0012	*
Каузатив	0,0048±0,0041	0,0040±0,0055	*
Генератив	0,0007±0,0015	0,0002±0,0009	*
Инструментатив	0,0079±0,0064	0,0058±0,0079	*
Предикат	0,0081±0,0053	0,0054±0,0051	*
Субъекты содействия	0,0010±0,0019	0,0002±0,0012	*
Субъект	0,0410±0,0124	0,0494±0,0191	*

* – p <0,05

Дискурсивные текстовые показатели

Характеризуют связность текста и позволяют определить, например, каким образом автор аргументирует свою позицию

- Количество дискурсивных единиц, дискурсивных деревьев
- Средняя длина дискурсивной единицы в словах
- Средняя длина элементарной дискурсивной единицы в словах
- Средняя глубина дискурсивного дерева
- Доля мультиядерных (NN) риторических отношений
- Частотность риторических отношений
- **34 показателя**

$$P_{ducount}(T) = \sum_{P_{disc} \in T_{disc}} |U| \qquad P_{treescount}(T) = |T_{disc}|$$

$$ds = \left(\langle \left(u_1^i, u_2^i\right), t_{disc} \rangle\right) |\ u_2^i = u_1^{i+1}, i = \overline{1, n-1}$$

$$P_{disc\,lenmax}\ (T) = \max_{P_{disc} \in T_{disc}} \max_{dc \in DC_{disc}} len(dc)$$

$$P_{disc\,len\,aver}\ (T) = \frac{\sum_{P_{disc} \in T_{disc}} \max_{dc \in DC_{disc}} len(dc)}{P_{trees\ count}\ (T)}$$

$$P_{discrelcount}(T, drel) = \sum_{P_{disc} \in T_{disc}} |R_{discU}^{drel}|$$

Риторическое отношение	**Частотность в корпусе больных**	**Частотность в корпусе здоровых**	**p-value**
contrast	0,34210	0,20788	0,00293
joint	0,21052	0,17164	0,01507
elaboration	0,01315	0,16211	0,01816
evaluation	0,0	0,06993	0,02083
concession	0,0	0,03305	0,03479

Определение психологического неблагополучия по текстам эссе

Решена задача классификации эссе на два класса – «депрессивные» и «не депрессивные» с выявлением значимых для классификации психолингвистических признаков

Исследуемые признаки:

- DM – предложенные психолингвистические показатели.
- Unigrams – униграммы слов. Отдельные слова, взвешенные по методу tf-idf.
- Bigrams – биграммы слов. Последовательности двух слов, взвешенные по методу tf-idf.

Random Forest				
Набор признаков	**Recall**	**Precision**	**F1**	**Accuracy**
DM	65,53 ± 8,31	**77,52** ± 4,91	70,65 ± 5,39	**84,16** ± 2,15
Unigrams	69,83 ± 9,36	70,87 ± 7,31	69,69 ± 4,51	82,27 ± 2,4
Unigrams + DM	72,01 ± 10,03	70,7 ± 10,18	70,48 ± 5,36	82,28 ± 3,4
Bigrams	**76,26** ± 7,37	70,42 ± 5,69	72,72 ± 2,44	83,21 ± 1,77
Bigrams + DM	74,18 ± 3,11	72,12 ± 4,16	**73,01** ± 2,11	83,85 ± 1,46
SVM				
Набор признаков	**Recall**	**Precision**	**F1**	**Accuracy**
DM	78,66 ± 14,27	52,78 ± 5,63	62,96 ± 8,12	73,11 ± 5,44
Unigrams	49,59 ± 16,11	69,04 ± 5,8	56,6 ± 12,67	78,81 ± 4,29
Unigrams + DM	72,28 ± 15,01	61,0 ± 11,09	66,05 ± 12,51	78,21 ± 8,17
Bigrams	64,67 ± 11,74	72,42 ± 8,89	68,01 ± 9,38	82,29 ± 5,04

Морфо-стилистические и синтаксические показатели обеспечивают лучшую точность классификации по сравнению с лексикой

Другие определяемые состояния по сочинению:

- Агрессивность

Stankevich M., Smirnov I., Kuznetsova Y., Kiselnikova N., Enikolopov S. Predicting Depression from Essays in Russian // Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialogue 2019”. – 2019. – pp. 637–647.

Определение личностных особенностей по текстам социальных сетей

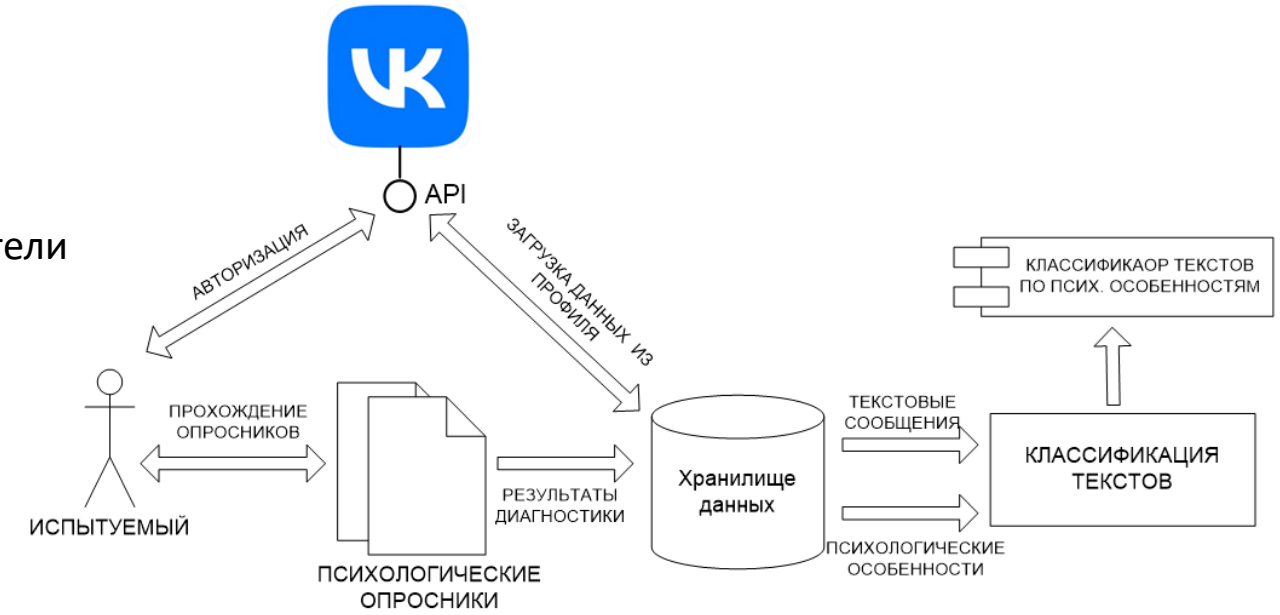
Решена задача классификации текстов социальных сетей на два класса – «депрессивные» и «здоровые»

Исследуемые признаки:

- MS – психолингвистические показатели
- UG – униграммы слов
- BG – биграммы слов
- D – словари
- MS-r – информативные психолингвистические показатели

Random Forest			
Признаки	**Precision**	**Recall**	**F1**
MS	59,80±6,21	59,80±6,21	54,47±3,66
UG	51,68±9,89	57,17±3,70	53,84±6,35
BG	49,64±6,67	58,47±6,06	53,12±3,16
D	46,21±5,52	56,30±7,20	50,66±5,80
MS-r	**62,60**±7,77	53,26±7,88	56,59±2,20
SVM			
Признаки	**Precision**	**Recall**	**F1**
MS	55,43±1,99	72,82±1,88	62,92±1,51
UG	45,63±7,94	83,69±13,53	57,57±3,41
BG	44,38±6,07	**85,86**±11,24	57,60±2,76
D	55,68±9,49	55,43±8,34	55,53±8,85
MS-r	58,40±2,99	77,17±1,88	**66,40**±1,33

Схема исследования



Другие определяемые состояния по текстам соцсетей*:*

- Личностные черты

Stankevich M., Latyshev A., Kuminskaya E., Smirnov I., Grigoriev O. Depression Detection from Social Media Texts // Data Analytics and Management in Data Intensive Domains: XXI International Conference DAMDID/RCDL'2019: Conference Proceedings. – 2019. – pp. 352-362.

Выявление типа реакции на фрустрацию по Розенцвейгу

Типы реакций:

- Экстрапунитивные реакции (extrapunitive, E) - реакция направлена на внешнее окружение, человек осуждает внешнюю причину фрустрации.

  Какой ужас! Вы должны заплатить!

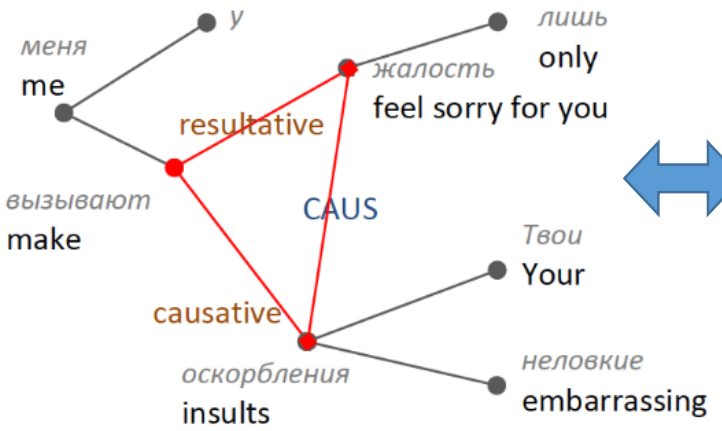
- Интропунитивные реакции (intropunitive, I) - реакция направлена на себя, человек испытывает чувство вины и ответственности за исправление сложившейся ситуации.

  Не надо было мне приходить. Как-нибудь выкручусь

- Импунитивные реакции (impunitive, M) - сложившаяся ситуация фрустрации рассматривается человеком как что-то неизбежное, незначительное, он никого не обвиняет.

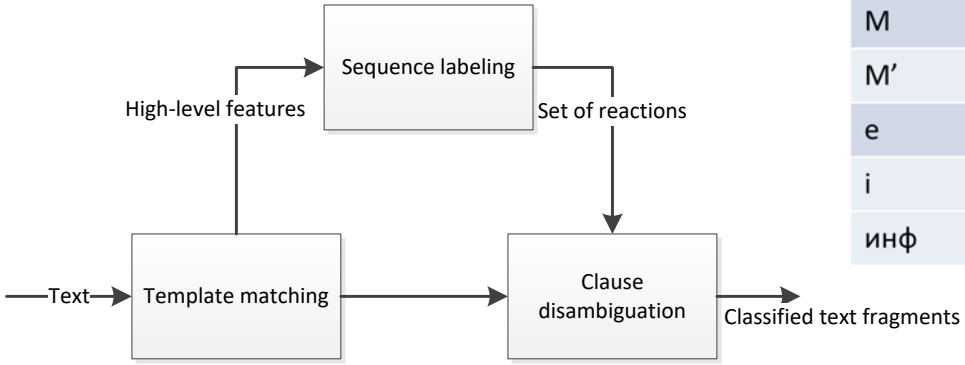
  Я не хотел вас обидеть. Всё будет хорошо

Реакция на фрустрацию - это ответ на ситуацию препятствия, когда планируемое или привычное поведение не может быть реализовано



{Case(Acc)+POS(NOUN)}<=CAUS=>{Case(Nom)+POS(NOUN)}<-{IS_PRED(True)+L_cause}

	P	R	F1
E	0.68	0.68	0.68
E’	0.74	0.88	0.80
I’	0.94	1.00	0.97
M	0.79	0.70	0.74
M’	0.80	0.88	0.84
e	0.78	0.85	0.81
i	0.87	0.53	0.66
инф	0.69	0.63	0.66

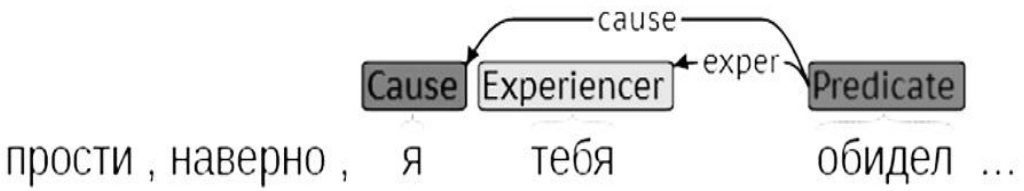


Выявление субъекта и причины упоминаемых в тексте эмоций (по авторской оценке)

- В тексте выявляется ***кто*** [субъект переживания] и от ***чего*** [причина переживания] испытывает эмоции по оценке автора, например,
  - *Мы* [субъект] *обрадовались* (эмоция: радость) *подарку* [причина]
  - *Ребёнок* [субъект] *боится* (эмоция: страх) *собак* [причина]

- Причину эмоции выражает семантическая роль ***каузатив***, субъекта эмоции выражает семантическая роль ***экспериенцер***

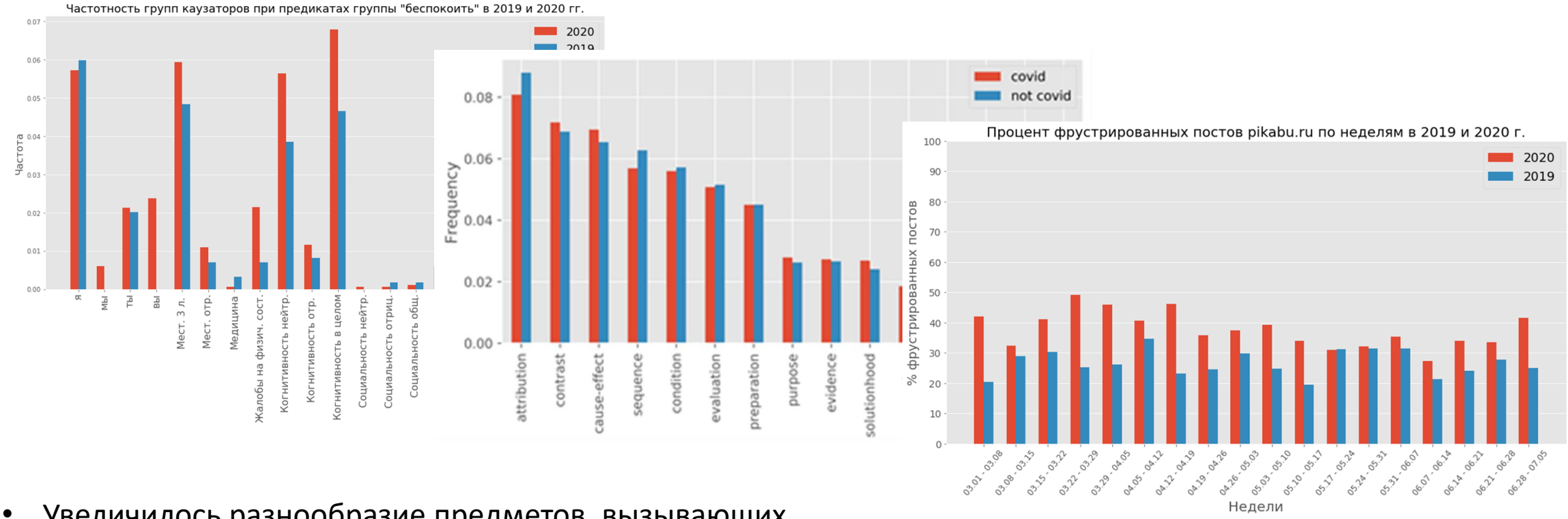
Задачи:
Поиск эмотивного предиката
Выделение предикатно-аргументной структуры
Установление семантических ролей



Выделяемая роль	Precision	Recall	F_1
Cause	76,77	83,50	80,00
Experiencer	90,48	96,94	93,60
Predicate	91,37	98,45	94,78
Overall	86,12	93,54	89,68

Campagnano C., Conia S., Navigli R. SRL4E – Semantic Role Labeling for Emotions: A Unified Evaluation Framework //Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). – 2022. – pp. 4586-4601

Анализ реакции общества на COVID-19 (самоизоляцию)

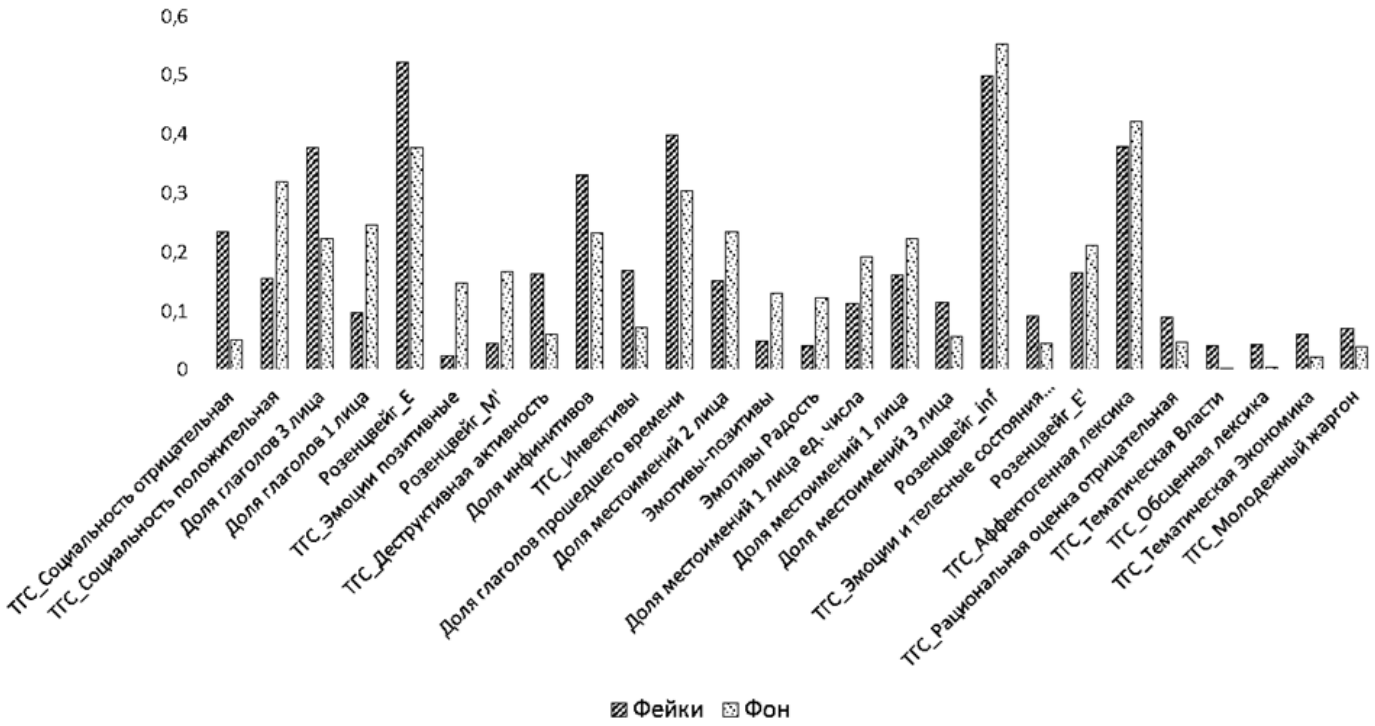


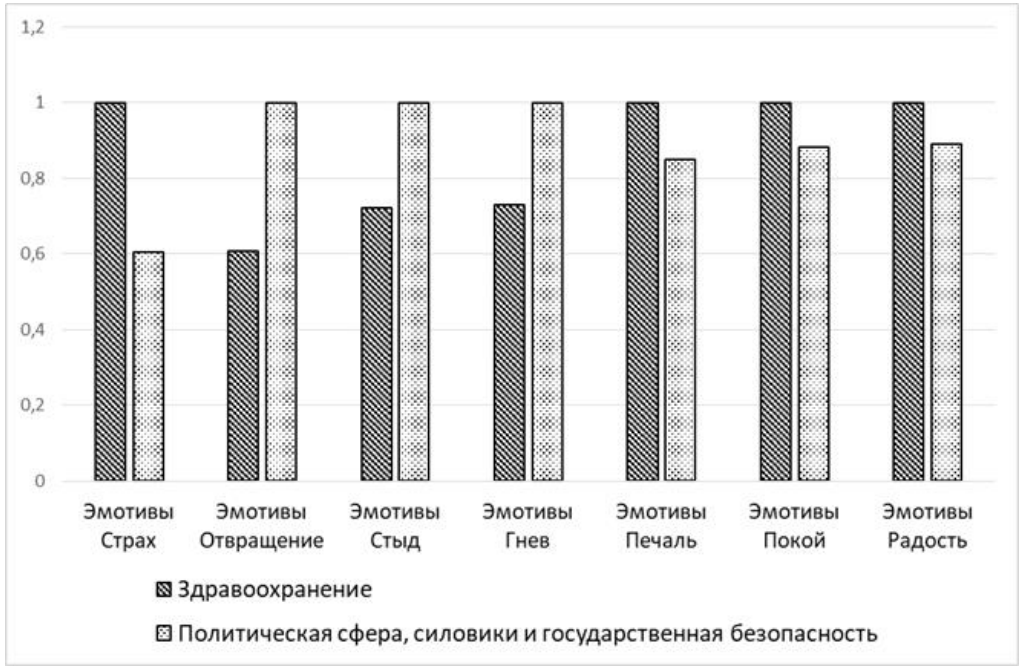
- Увеличилось разнообразие предметов, вызывающих беспокойство
- Прирост частотности аргументов из группы «жалобы на физическое состояние»
- Прирост упоминаний в качестве предмета беспокойства аргументов из групп «государственные финансы» и «внешняя политика»

- Участились обращения к аудитории, вопросы, в том числе и риторические
- Фрустрированность выросла на 30%

Анализ реакции на фейковые сообщения в социальных сетях

Исходные данные – комментарии к фейковым сообщениям в сети Вконтакте (288 тыс. комментариев)





Авторы комментариев к фейкам используют лексику негативных эмоций, негативной эмоциональной и рациональной оценки, лексику деструктивных действий и отрицательных явлений социальной жизни, оскорбления

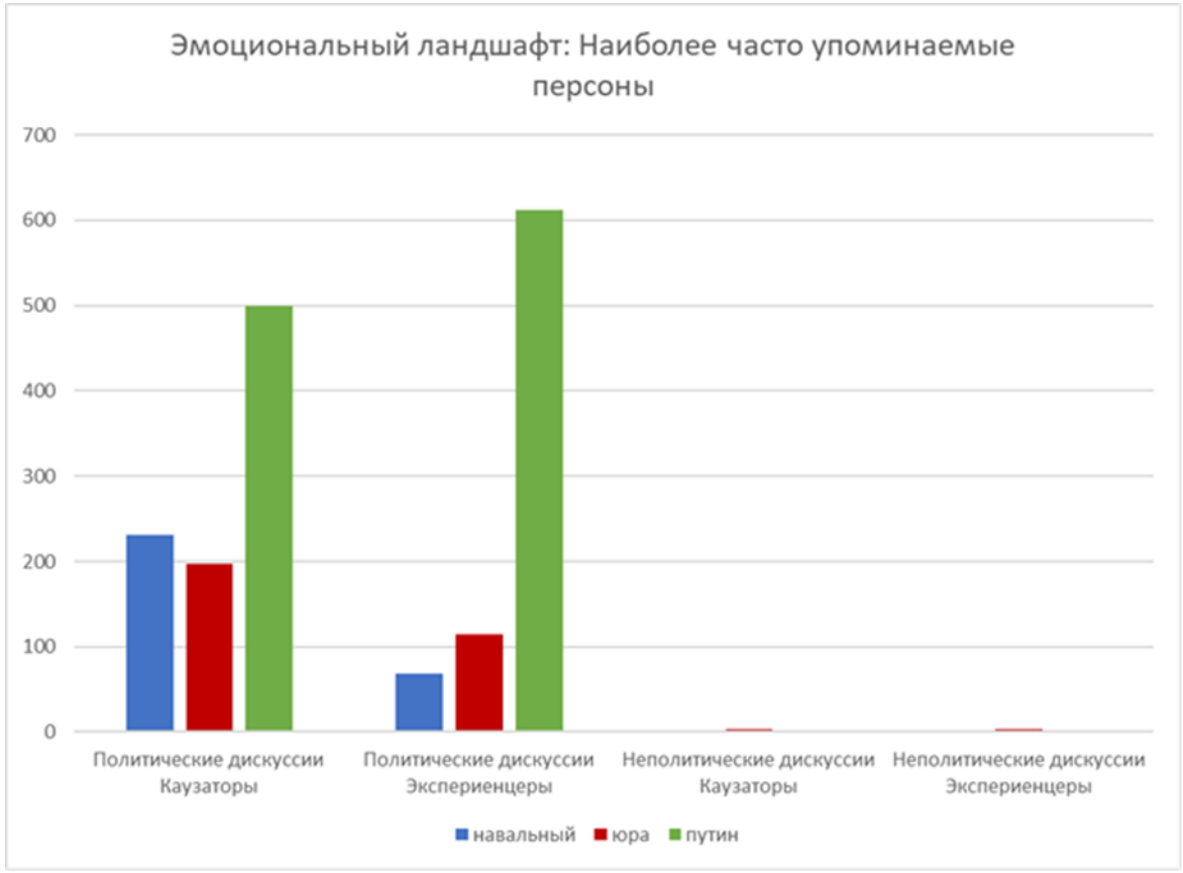
Комментариям к фейкам по теме «Здравоохранение» характерны больше эмоции «страх», «печаль», «покой» и «радость», политическая тема характеризуется эмоциями «отвращение», «стыд» и «гнев»

В комментариях к фейковым сообщениям на любые темы значимо чаще встречается агрессия (экстрапунитивные реакции)

Анализ сетевого общественно-политического дискурса

Комментарии к видео YouTube каналов и плейлистов политического и неполитического формата

Политические: *2629 видео и ~5.000.000 комментариев.* **Неполитические:** *2178 видео и ~1.200.000 комментариев*



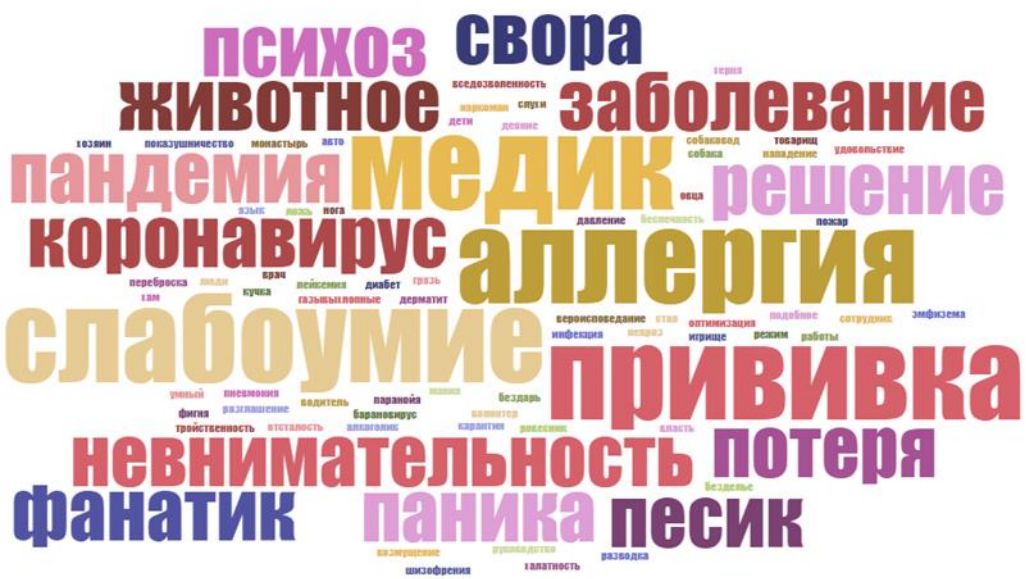
В политических дискуссиях «Я» является экспериенцером чаще

Указанные персоны (Навальный, Юра, Путин) в неполитических дискуссиях не являются каузаторами

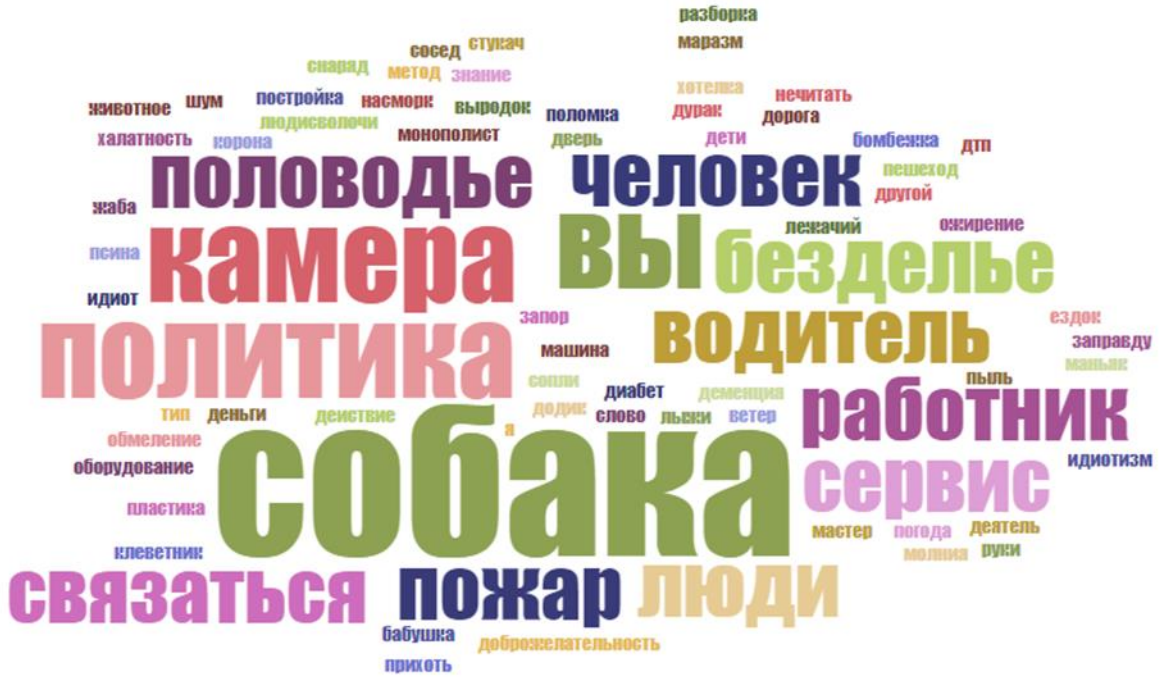
Эмотивный анализ текстовых сообщений сетевых сообществ г. Саров относительно медицины

Страдать: реально переживаемые угрозы

Дискуссии на медицинские темы



Дискуссии на немедицинские темы



- В «медицинских» дискуссиях причины реальных страданий в большей степени по сравнению со страхами отражают негативный опыт взаимоотношений с представителями системы здравоохранения и общее недовольство социальным окружением и его когнитивными и поведенческими особенностями.
- В обоих типах дискуссий собаки упоминаются как причина реальных страданий.
- Причины реальных страданий в немедицинских дискуссиях связаны чаще всего с управлением транспортными средствами и фиксацией правонарушений, сферой производственных отношений, политикой, а также с природными катастрофами, что, по-видимому, отражает местную специфику.

Системы текстовой аналитики

TextAppliance – система интеллектуального поиска и анализа больших массивов текстов

- **Модуль анализа текстов.** Модуль выполняет разноуровневый анализ текстов документов и поискового запроса с помощью методов реляционно-ситуационного и семантико-синтаксического анализа текстов.
- **Модуль семантического поиска.** Модуль выполняет семантический анализ поискового запроса с помощью модуля анализа текстов, выборку семантических структур документов из индексного хранилища и ранжирование результатов с помощью методов семантического и вопросно-ответного поиска.

Внедрения

Открытые поисковые системы **Exactus Expert, Exactus Patent, Exactus Like**

НИЦ «Институт имени Н.Е. Жуковского» - система поиска, анализа и учета знаний в авиастроении

Научно-технический институт межотраслевой информации - информационно-аналитическая система поддержки экспертной деятельности

ООО «ЗНАНИУМ» - электронно-библиотечная система «Знаниум»

Дирекция научно-технических программ МИНОБРНАУКИ, Российский центр научной информации, Сколтех, НИУ «Высшая школа экономики»

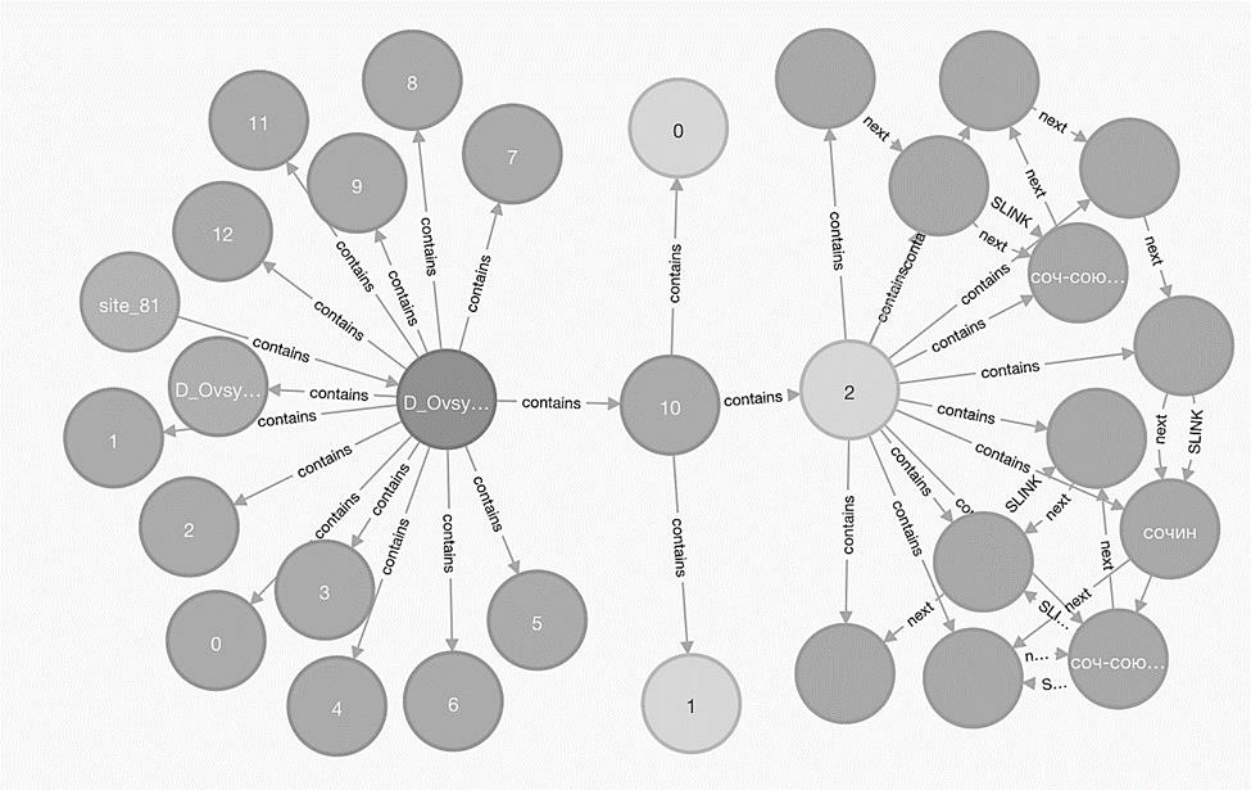
Системы текстовой аналитики

Машина РСА – инструмент лингво-статистических корпусных исследований

Функции машины РСА

- Поиск для заданного слова или списка слов из словаря всех предикатных слов, для которых слово является аргументом и заполняет ту или иную семантическую роль в тексте
- Поиск всех семантических ролей, установленных для заданного слова
- Поиск слов из заданных словарей
- Поиск и отображение в тексте предикатных слов с заданным значением и их аргументов.
- Вычисление частотных и статистических характеристик для результатов поиска, корпусов или отдельных текстов
- Сравнение корпусов по различным характеристикам

Компонентная архитектура



Разноуровневые текстовые структуры хранятся в графовой базе данных

Использование:

Институт психологии РАН – психолингвистические исследования, ФИЦ ИУ РАН – корпусные исследования

Кузнецова, Ю. М., Смирнов, И. В., Станкевич, М. А., Чудова, Н. В. Создание инструмента автоматического анализа текста в интересах социо-гуманитарных исследований. Часть 2. Машина РСА и опыт ее использования // Искусственный интеллект и принятие решений. – 2019. – №. 3. – С. 40-51.

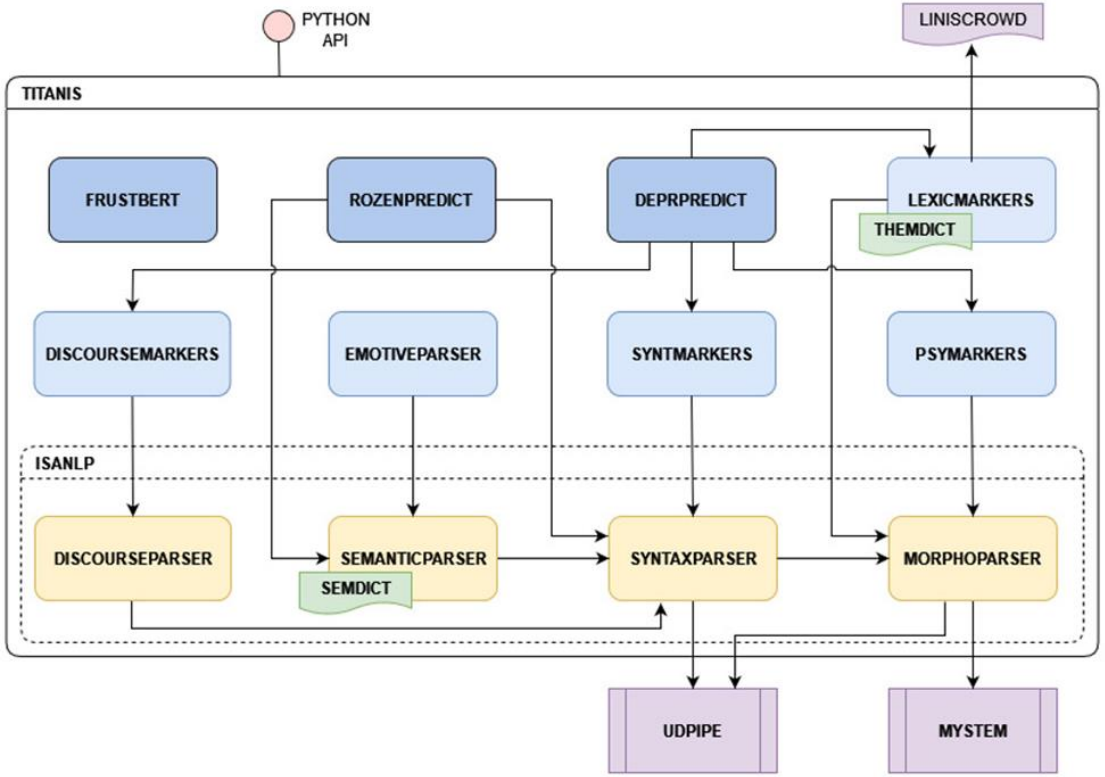
ISANLP https://github.com/IINemo/isanlp

TITANIS – инструмент психоэмоционального анализа текстов

Функции TITANIS

- Вычисление психолингвистических маркеров
- Выявление эмоциональной направленности текста и типов эмоционального состояния
- Выявление субъекта и причины эмоций
- Предсказание депрессии у автора текста по его небольшому сочинению
- Предсказание депрессивности у автора текста по текстовым сообщениям социальных сетей
- Предсказание наличия состояния фрустрации у автора текста
- Выявление типа реакции на фрустрацию по Розенцвейгу
- Сравнительный анализ текстовых показателей для нескольких наборов текстов

Компонентная архитектура TITANIS



Использование:

Институт психологии РАН – Психолингвистические исследования

ФИЦ ИУ РАН – Системы здоровьесбережения (ИнСиз, ИИ-Гиппократ)

ООО «НИИ МВУС» – Системы контроля состояния операторов АЭС

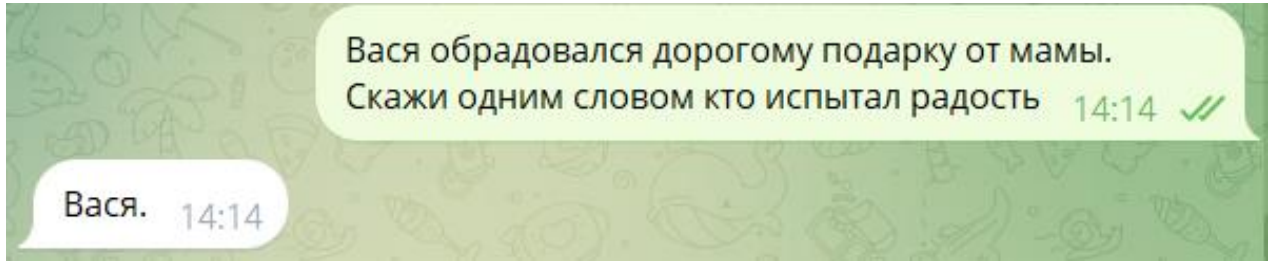
- Основан на открытых библиотеках обработки текстов со свободными лицензиями
- Параллельная обработка текстов
- Открытая версия с ограниченными возможностями: https://github.com/tchewik/titanis-open

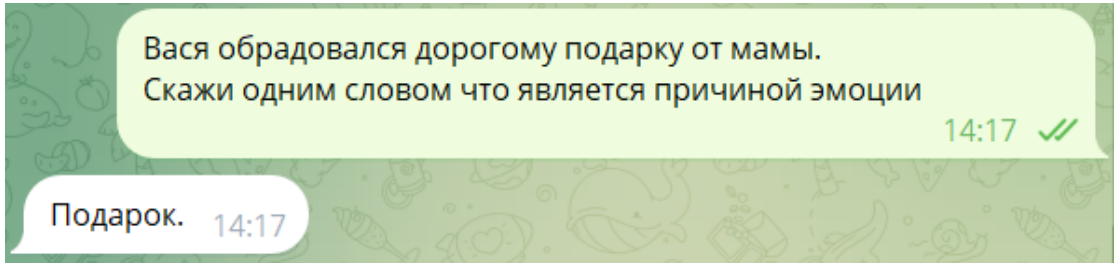
Smirnov I., Stankevich M., Kuznetsova Y., Suvorova M., Larionov D., Nikitina E., Savelov M., Grigoriev O. TITANIS: A Tool for Intelligent Text Analysis in Social Media // In Artificial Intelligence. RCAI 2021. Lecture Notes in Computer Science. – 2021. – V. 12948. – pp 232-247.

Перспективы использования разноуровневых структур текста в эпоху LLM

- Доверенные системы поиска и анализа текстов
  - Требовательные к надежности ответа и интерпретируемости процесса получения результата
- Научные исследования
  - Там, где требуется объяснение и возможность получения нового знания
  - Исследования в социо-гуманитарной сфере
- Гибридизация с LLM
  - Yang L. et al. Give us the facts: Enhancing large language models with knowledge graphs for fact-aware language modeling //IEEE Transactions on Knowledge and Data Engineering. – 2024.
- Попытки проинтерпретировать LLM
  - Zhao H. et al. Explainability for large language models: A survey //ACM Transactions on Intelligent Systems and Technology. – 2024. – Т. 15. – №. 2. – С. 1-38.

Использование chatGPT для установления семантических ролей





- Системная инструкция: You are a helpful assistant that extracts semantic roles from sentences. You do not extract implied arguments. If an argument is a multi-word phrase, you extract only the main word.
- 15 примеров типа:"text": "Мария любит цветы.", "roles": ["Мария#Experiencer", "любит#predicate", "цветы#Object"]
- Выход для текста "Многим сотрудникам нравится новый офис компании": Многим#Experiencer, офис#Object

Точность установления ролей > 90%

LLM для выявления ментальных расстройств (депрессивности и тревожности)

System: You play the role of a psychologist's assistant who helps diagnose the presence or absence of a depressive disorder. You will be given a text written by a person. Determine the author's depression level from the text, where 0 is no depression, 1 is depression, and then write why you chose this answer.

User: Text: {input_text}

Answer (0 or 1):

0-shot, 5-shot

Corpus	Mode	Model	Precision healthy	Recall healthy	F1-healthy	Precision pathology	Recall pathology	F1-pathology	F1-macro
DE	SFT	Linguistic features	94.0±0.8	95.2±1.0	94.6±0.8	79.3±4.0	75.0±3.5	77.0±3.3	85.8±2.1
	SFT	TF-IDF	91.6±2.1	94.4±2.2	93.0±1.2	74.8±7.0	64.4±10.3	68.5±6.6	80.7±3.8
	5-shot	Vikhr 7B IT 5.4	87.06±0.0	82.22±0.0	84.57±0.0	40.74±0.0	50.0±0.0	44.9±0.0	64.73±0.0
	5-shot MMLU	Gemma2 9B IT	92.96±0.0	73.33±0.0	81.99±0.0	41.46±0.0	77.27±0.0	53.97±0.0	67.98±0.0
	LoRA	VikhrGemma 2B IT	94.74±0.88	96.48±1.62	95.59±0.66	84.96±5.72	78.03±4.08	**81.13±2.42**	**88.36±1.52**
	SFT	RuBERT	94.45±1.16	91.11±1.7	92.74±1.04	68.41±4.04	78.03±4.85	72.8±3.43	82.77±2.21
	0-shot	Gemma2 9B IT	93.33±0.0	62.22±0.0	74.67±0.0	34.62±0.0	81.82±0.0	48.65±0.0	61.66±0.0
DSM	SFT	Linguistic features	62.8±1.8	59.3±7.4	60.7±4.0	43.7±2.7	47.2±7.7	45.1±3.8	52.9±2.0
	SFT	TF-IDF	62.2±2.6	51.2±13.9	55.3±8.5	42.9±3.6	53.7±11.4	46.9±4.7	51.1±3.5
	5-shot	SaigaLlama3 8B	100.0±0.0	3.7±0.0	7.14±0.0	40.91±0.0	100.0±0.0	**58.06±0.0**	32.6±0.0
	5-shot	Vikhr 7B IT 0.4	70.97±0.0	81.48±0.0	75.86±0.0	64.29±0.0	50.0±0.0	56.25±0.0	**66.06±0.0**
	5-shot MMLU	SaigaLlama3 8B	100.0±0.0	3.7±0.0	7.14±0.0	40.91±0.0	100.0±0.0	**58.06±0.0**	32.6±0.0
	5-shot MMLU	Vikhr 7B IT 5.4	68.75±0.0	81.48±0.0	74.58±0.0	61.54±0.0	44.44±0.0	51.61±0.0	63.09±0.0
	LoRA	Qwen2 7B IT	68.87±6.39	72.22±8.21	70.4±6.75	55.33±11.08	50.93±10.84	52.81±10.26	61.61±8.32
	SFT	RuBioRoBERTa	62.1±2.98	62.96±6.05	62.46±4.3	43.66±4.43	42.59±4.14	43.01±3.7	52.74±3.67

Ответ: Депрессия.
Причины: **Тяжелое детство и семейные проблемы**: Алина описывает сложные отношения в семье, в том числе развод родителей, конфликты с бабушкой, которая была авторитарной и контролирующей. Это может привести к развитию депрессии и тревожных расстройств. **Трудности в общении и социальной адаптации**: Алина испытывала трудности в общении с детьми в школе, была изолирована и чувствовала себя одинокой. Это может указывать на социофобию или тревожные расстройства. **Самооценка и чувство вины**: Алина говорит о том, что она не понимает, почему люди вмешиваются в ее жизнь, и чувствует себя виноватой за то, что не может контролировать эти ситуации. Это может указывать на низкую самооценку и чувство вины.
Все эти факторы указывают на то, что у Алины есть депрессивные и тревожные симптомы, которые требуют внимания и профессиональной помощи.

Спасибо за внимание!