НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ

Автоматическое определение типов вопросительных предложений русского языка

Николаев Кирилл,
ФГН, 15ФПЛ

Научный руководитель:
к.ф.н., доцент
Малафеев А.Ю.

Вопросительное предложение

- Вопросительное предложение – одна из важнейших языковых единиц;
- Вопрос имеет познавательное значение, но не заключает обычное атрибутивное суждение;
- Ключевой инструмент познания;

- Вопросы и вопросительные предложения представляют, таким образом, высокий кросс-дисциплинарный интерес.

Практическая значимость исследования

- Увеличивается количество вопросно-ответных систем на базе английского языка;
- **Вопросно-ответная система** – программный комплекс с естественно-языковым интерфейсом;
- В отличие от поисковых систем, ответы вопросно-ответных систем точны и конкретны, позволяют избежать информационной перегрузки.

Актуальность и научная новизна

- Q-A системы на базе английского языка: TEQUESTA (Monz, 2003), START (Katz et. al., 2006), OpenEphyra (van Zaanen, 2008), и т.д.
- Раскрытие темы в отечественной литературе невелико: Соснин (2007), Сулейманов (2001); Тихомиров (2006); Мозговой (2006); Соловьёв, Пескова (2010) вносят основной вклад в теоретический аспект;
- Данная работа в первую очередь направлена на решение практической задачи типологизации.

Цель исследования

- Создание инструмента автоматической типологизации вопросительных предложений русского языка с использованием языка программирования Python

Задачи

- Сформировать комплексную типологию вопросительных предложений русского языка с использованием уже существующих;
- Разработать baseline-инструмент автоматической типологизации вопросительных предложений русского языка посредством языка программирования Python с использованием метода регулярных выражений;
- Реализовать инструмент автоматической типологизации вопросительных предложений с использованием методов машинного обучения.

Объект и предмет исследования

- **Объект** исследования: вопросительные предложения русского языка;

- **Предмет** исследования: автоматическая типологизация вопросительных предложений русского языка

Методы исследования

- Описательный метод;
- Методы корпусной лингвистики;
- Экспериментальный метод;
- Метод измерения;
- Дистрибутивный анализ.

Вопросительное предложение

- «Вопросительными… называются предложения, в которых специальными языковыми средствами выражается стремление говорящего узнать что-либо или удостовериться в чем-либо». [Шведова, Н. Ю. (1980). Русская грамматика]

Функционально-семантические типы

- Объединяются на основе первичных и вторичных функций вопросительных предложений.
- В первичных вопрос направлен на поиск информации; во вторичных – на её передачу;
- Данные функции устанавливаются на основе:

1. Типа и объема той информации, которая ожидается в ответе;
2. Осведомленности говорящего о предмете вопроса;
3. Ожидаемого ответа.

Таксономия Грэсера [2]

Вопрос	Абстрактная спецификация	Пример
1. Подтверждение	Истинный ли факт? Произошло ли событие?	Гагарин – космонавт? Шел ли вчера дождь?
2. Сравнение	Чем X похож на Y? Как X отличается от Y?	Как Флорида похожа на Китай?
3. Выбор из множества вариантов	X или Y?	Он заказал курицу или свинину?
4. Завершение концепта	Кто? Что? Где? Когда?	Кто написал эту песню? Что украл ребенок?
5. Определение	Что означает X?	Что такое фрейм?
6. Пример	Что служит примером X? Каково конкретное обстоятельство категории?	Что служит примером группы? Какой эксперимент поддерживает это утверждение?

Наша типология

Тип	Численная метка	Пример
1. Общий	1	Что происходит в…?
2. Подтверждение	2	Правда ли, что…?
3. Определение	3	Что означает/такое…?
4. Пример	4	Приведи пример…?
5. Сравнение	5	Чем похожи / отличаются…?

Вопросно-ответные системы

- OpenEphyra [van Zaanen, 2008], TEQUESTA [Monz, 2003], START [Katz et al., 2006], IBM Watson [Ferucci et al., 2010], EAGLi [Gobeill et al., 2012]

- Работу любой из них можно условно разбить на **три этапа:**

1) Анализ запроса пользователя;
2) Поиск релевантной информации;
3) Формирование ответа.

Baseline-метод: шаблоны

Тип	Тэг	Номерной тэг	Регулярное выражение
Определение	Definition	3	/.*([чЧ]то означает|такое).*[\?\.]/
Пример	Example	4	/.*(([пП]риведи|[кК]акой пример|образец)|(([чЧ]то|[кК]то) ((может ((служить)|(выступ(ать|ить))))|((по?)служит|выступ(ает|ит))) (как)?((пример(ом?))|(образ(ец|цом))))).*[\?\.]/
Сравнение	Comparison	5	/.*(([чЧ]ем\S.* похо(ж|жи|жа|же)|отлича(ются|ется|))|[кК]ак\S.* похо(ж|жи|жа|же)| отлича(ются|ется|)).*[\?\.]/
Конкретизация свойств	Quality	8	/.*(([кК]ак(ой|ая|ое|ом|ие))|([кК]ак(им|ими|ие) ((свойств(ом|ами|а))|(качеств(ом|ами|а))) (наделен|обладает|характеризуется|отличается|имеет))).*[\?\.]/

Baseline-метод: регулярные выражения

- Пошаговое сравнение с шаблонными строками: от простейших (1, 2, 3, 6) к наиболее комплексным (10, 11, 12);
- При совпадении вопрос получает новый тип, при несовпадении тип остается неизменным; тип вопроса A определяется после проверки всех шаблонов;
- Проверку начинаем с шаблона Complex – типа вопроса, не обладающего какими-то дистинктивными характеристиками (/.*[\?\.]/)
- Точность алгоритма: 52,7%, 79/150 вопросов определены верно.

Baseline-метод: основные недостатки

- Всегда найдутся вопросы, не соответствующие шаблону;
- Априорная детальность категорий требует поиска компромисса между тем, насколько вопросы им соответствуют, и простоте построения тэггеров и классификаторов для них.

Таким образом, вероятностная модель, напрямую вычисляющая степень соответствия между потенциальным ответом и контекстом вопроса гораздо более эффективна.

Классические алгоритмы

- Набор в 2008 вопросов взят из Русского Интернет-корпуса (Sharoff, 2006);
- Ручная разметка;
- Два разных набора классов: с упрощёнными “Завершение концепта” и “Количество” (14 классов) и изначальной подробной классификацией (23 класса)
- Пять репрезентаций формата «мешок слов»: символьные биграммы, триграммы, словесные униграммы, биграммы, триграммы: всего 10 (подробная + упрощённая)
- Основной инструмент на данном этапе: RapidMiner

Соотношение данных по классам

- Три классических алгоритма машинного обучения: наивный байесовский классификатор, метод опорных векторов, логистическая регрессия

Set	1	2	3	4	5	6	7	
Training	153	424	64	10	15	76	579	
Test	15	34	6	3	2	10	25	
	8	9	10	11	12	13	14	Total
Training	186	49	59	9	116	201	67	2008
Test	17	10	1	2	5	16	4	150

Классические алгоритмы

Алгоритм и модель	Точных соответствий (полная кл.)	Точность (полная кл.)	Соответствий (упрощённая кл.)	Точность (упрощённая кл.)
Байес (НБ), сл-три	61	40.7%	62	41.3%
Метод опорных векторов (ОВ), сл-би	73	48.7%	82	54.7%
Логистическая (ЛР), сл-би	90	**59.3%**	92	61.3%
ЛР, сл-три	88	56%	97	**64.7%**
ОВ, сл-три, нормализованная пропорция	68	45.3%	81	54%
ОВ (линейное ядро), сл-три, нормализованная пропорция	98	**65.3%**	103	**68.7%**

Результаты классических алгоритмов

- Результат соответствует лучшему для классификации вопросов на материалах английского языка:
  - Loni B. A survey of state-of-the-art methods on question classification. – 2011., 95%, Linear SVM

- Максимальная микро-точность для русского языка – 68.7%;

- Результаты представлены на AIST-2017
  - Nikolaev, K. and Malafeev, A., 2017, July. Russian-language question classification: a new typology and first results. In *International Conference on Analysis of Images, Social Networks and Texts* (pp. 72-81). Springer, Cham.

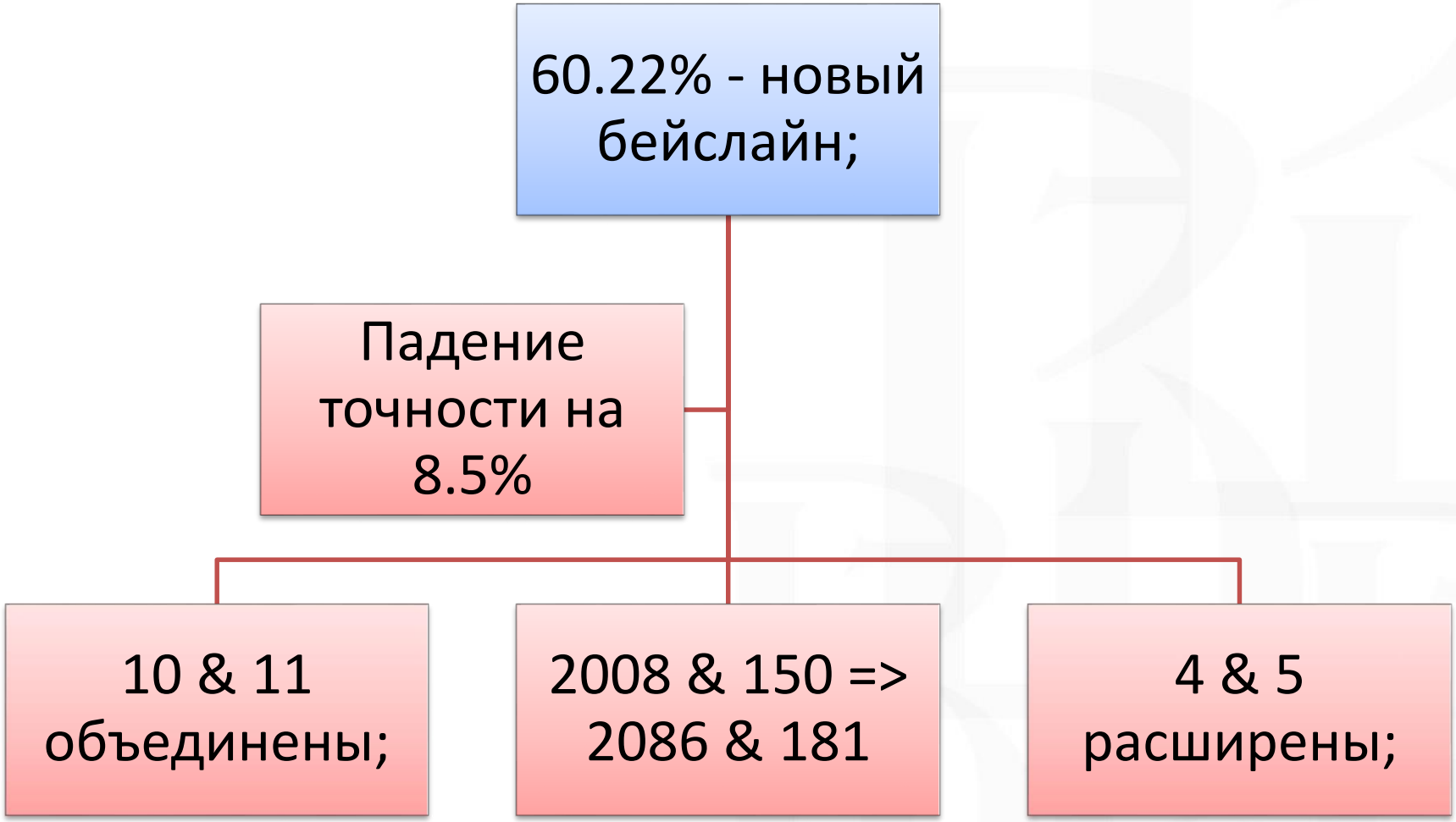
Следующий этап

- Lai et.al. RCNN for Text Classification. – AAAI, 2015 – применение рекуррентно-свёрточной нейронной сети в классификации текстов;

- Слишком маленький объём данных:
  - Всего 9% на данной архитектуре;

- Предложения можно рассматривать как структурные данные:
  - Было решено использовать свёрточную нейронную сеть.

Модификации датасета



Представление данных

- Векторные представления слов – дистрибутивная семантика;

- Word2Vec:
  - Предобученная модель НКРЯ, 250 миллионов слов, размерность 300;
  - Дополнительно – 40 бинарных признаков;

- Первые 8 слов в предложении;
  - Среднее количество – 7;
  - Итоговая размерность – 340x8

Архитектура

2-D Conv layer: 26 нейронов; размер ядра: 20x3

Leaky ReLU: alpha = 0.1

MaxPooling2D

Dropout(0.2)

Flatten()

Dense(13, activation='softmax') – 72.38% микро-точность, 0.67 F-мера

Выводы

- Рассмотрены вопросительные предложения, их синтаксический и функциональный аспекты;
- Изучены существующие решения в области типологизации вопросительных предложений на базе вопросно-ответных систем английского языка;
- Сформирована типология вопросительных предложений на основе существующих;
- Применён метод регулярных выражений, результат взят как бейслайн для следующего этапа (52.7%);

Выводы

- Обучены три классификатора на пяти разных наборах признаков: наиболее точный – 68.7% для упрощённой типологии, в отличие от базового результата в 52.7% для регулярных выражений;
- По результатам второго этапа, датасет изменён; новый бейслайн – 60.22%;
- Применена свёрточная нейронная сеть; достигнутая точность – 72.38%.

Выводы

- Достигнут максимальный результат для данного набора данных;
- Наиболее проблемные классы – 1 (общий) и 10-11 (инструмент / действие);
- Результат возможно улучшить:
  - Значительное увеличение объёма данных;
  - Дальнейшая балансировка классов;
  - Использование более сложных архитектур (RCNN) и алгоритмов (3D-представление данных).

Список использованных ресурсов

1. Bunescu R., Huang Y. (2010), Towards a general model of answer typing: Question focus identification, Proceedings of The 11th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing 2010), RCS Volume, pp. 231-242.
2. Burger J., Cardie C., Chaudhri V., Gaizauskas R., Harabagiu S., Israel D., ... & Moldovan, D. (2001), Issues, tasks and program structures to roadmap research in question & answering (Q&A), Document Understanding Conferences Roadmapping Documents, pp. 1-35.
3. Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., & Kuksa, P. (2011). Natural language processing (almost) from scratch. Journal of Machine Learning Research, 12(Aug), pp. 2493-2537.
4. Damljanovic D., Agatonovic M., Cunningham H. (2010), Identification of the Question Focus: Combining Syntactic Analysis and Ontology-based Lookup through the User Interaction, LREC.
5. Elman, Jeffrey L. "Finding structure in time." Cognitive science14.2 (1990): pp. 179-211.
6. Galea A. (2003), Open-domain surface-based question answering system, Proceedings of CSAW 3.
7. Gobeill J., Pasche E., Teodoro D., Veuthey A. L., Ruch, P. (2012), Answering gene ontology terms to proteomics questions by supervised macro reading in Medline, EMBnet. Journal, No. 18(B), pp. 29-31.
8. Goldberg, Y., & Levy, O. (2014). word2vec Explained: deriving Mikolov et al.'s negative-sampling word-embedding method. arXiv preprint arXiv:1402.3722.

Список использованных ресурсов

9. Katz B., Borchardt G. C., Felshin S. (2006), Natural Language Annotations for Question Answering, FLAIRS Conference, pp. 303-306.
10. Klinkenberg R. (ed.) (2013), RapidMiner: Data mining use cases and business analytics applications, Chapman and Hall/CRC.
11. Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems (pp. 1097-1105).
12. Lai, S., Xu, L., Liu, K., & Zhao, J. (2015, January). Recurrent Convolutional Neural Networks for Text Classification. In AAAI(Vol. 333, pp. 2267-2273).
13. Lauer T. W., Peacock E., Graesser A. C. (2013), Questions and information systems, Psychology Press.
14. Li X., Roth D. (2002), Learning question classifiers, Proceedings of the 19th international conference on Computational linguistics, Association for Computational Linguistics, Vol. 1, pp. 1-7.
15. Loni B. (2011). A survey of state-of-the-art methods on question classification. Literature Survey, Published on TU Delft Repository.
16. Monz C. (2003), Document retrieval in the context of question answering, European Conference on Information Retrieval, Springer Berlin Heidelberg, pp. 571-579.
17. Monz C. (2003), From document retrieval to question answering, Institute for Logic, Language and Computation.

Список использованных ресурсов

18. Nikolaev, K., & Malafeev, A. (2017, July). Russian-Language Question Classification: A New Typology and First Results. In International Conference on Analysis of Images, Social Networks and Texts (pp. 72-81). Springer, Cham.
19. Nikolaev, K., & Malafeev, A. (2018, July). Russian Q&A Method Study: From Naive Bayes to Convolutional Neural Networks. In *International Conference on Analysis of Images, Social Networks and Texts* (pp. 121-126). Springer, Cham.
20. Pereira F., Mitchell T., Botvinick M. (2009), Machine learning classifiers and fMRI: a tutorial overview, Neuroimage, Vol. 45, No. 1, pp. S199-S209.
21. Pinchak C., Lin D. A (2006), Probabilistic Answer Type Model, EACL.
22. Sharoff S. (2006), Creating general-purpose corpora using automated search engine queries, WaCky, pp. 63-98.
23. Silva J., Coheur L., Mendes A. C., Wichert A. (2011). From symbolic to sub-symbolic information in question classification. Artificial Intelligence Review, No. 35(2), pp. 137-154.
24. van Zaanen M. (2008), Multi-lingual Question Answering using OpenEphyra, CLEF (Working Notes).
25. Xu, Z., Yang, Y., & Hauptmann, A. G. (2015). A discriminative CNN video representation for event detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 1798-1807).

Список использованных ресурсов

26. Zhang D., Lee W. S. (2003), Question classification using support vector machines, Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval, pp. 26-32.
27. Мозговой М. В. Простая вопросно-ответная система на основе семантического анализатора русского языка //Вестник Санкт-Петербургского университета. Серия 10. Прикладная математика. Информатика. Процессы управления. – 2006. – №. 1.
28. Соловьёв А. А., Пескова О. В. Построение вопросно-ответной системы для русского языка: модуль анализа вопросов //Новые информационные технологии в автоматизированных системах. – 2010. – №. 13.
29. Соснин П. И. Вопросно-ответное моделирование в разработке автоматизованных систем. – 2007.
30. Сулейманов Д. Ш. Исследование базовых принципов построения семантического интерпретатора вопросно-ответных текстов на естественном языке в АОС //Образовательные технологии и общество. – 2001. – Т. 4. – №. 3.
31. Тихомиров И. А. Вопросно-ответный поиск в интеллектуальной поисковой системе Exactus //Труды четвертого российского семинара по оценке методов информационного поиска РОМИП. – 2006. – С. 80-85.
32. Невольникова С. В. Функционально-семантические разновидности русских вопросительных предложений и их роль в текстообразовании //Ростов-н/Д. – 2004.
33. Шведова Н. Ю. Русская грамматика. В двух томах. – 1980.

НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ

Автоматическое определение типов вопросительных предложений русского языка

Николаев Кирилл,
ФГН, 15ФПЛ

Научный руководитель:
к.ф.н., доцент
Малафеев А.Ю.

Текущая работа

Интересы (по https://vc.ru/marketing/25614-audience-of-telegram):