СОВРЕМЕННЫЕ ТРЕНДЫ КОМПЬЮТЕРНОЙ ЛИНГВИСТИКИ В ОБЛАСТИ ЗАЩИТЫ ИНФОРМАЦИИ

Мельников Сергей Юрьевич,
д.ф.-м.н.,
ООО «Лингвистические и информационные технологии»
Пересыпкин Владимир Анатольевич,
д.т.н., чл.-корр. АК РФ
ФГУП «НТЦ «Орион», г. Москва

25 мая 2022 г., конференция КЗИ - 2022

АКТУАЛЬНОСТЬ

Стратегические направления научных исследований в области обеспечения информационной безопасности перечислены в «Доктрине информационной безопасности Российской Федерации», утвержденной Указом Президента РФ № 646 от 5 декабря 2016 г.

и в «Основных направлениях научных исследований в области обеспечения информационной безопасности Российской Федерации», утвержденными Секретарем Совета Безопасности Российской Федерации Н.П.Патрушевым 31 августа 2017 г.

Во многом эти направления связаны с развитием автоматических методов обработки информации, для чего необходимо совершенствовать, в том числе, вычислительно-лингвистические методы автоматической обработки текстов.

ПРИМЕРЫ ЗАДАЧ ИБ, ДЛЯ РЕШЕНИЯ КОТОРЫХ НЕОБХОДИМА АВТОМАТИЧЕСКАЯ ОБРАБОТКА ТЕКСТОВ

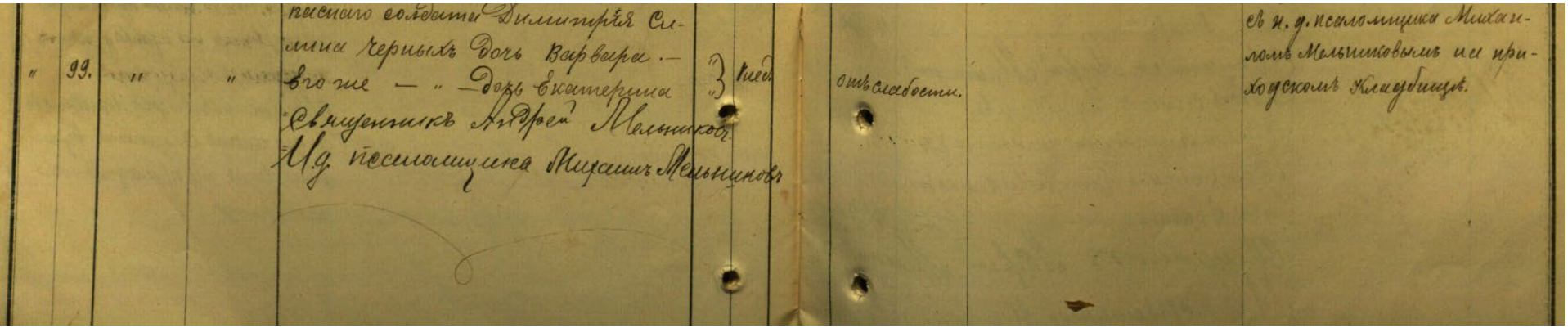
- выявление скомпрометированных аккаунтов в социальных сетях [3],
- построение фильтров для анализа фишинговых атак [4],
- непрерывная идентификация пользователя по потоку его сообщений [5],
- обнаружение фейковых новостей [6], фейковых обзоров товаров и услуг [7],
- оценка достоверности систем учета мнений и анализа отзывов пользователей [8],
- обнаружение атак на веб-ресурсы, использующих автоматически сгенерированные тексты с незначительными изменениями контента [9],
- задачи выявления материалов деструктивной направленности [10],
- круг задач, связанных с обнаружением искусственно сгенерированных текстов [11],
- задачи, связанные с использованием избыточности языка в криптографических [12] и стеганографических [13] приложениях,
- задачи анализа искаженных текстов, в случае, когда искажения имеют случайное [14], умышленное [15] или неумышленное [16] происхождение,
- и ряд других задач.

О РАСПОЗНАВАНИИ РУКОПИСНОГО ТЕКСТА

База genotek.ru содержит около двух миллионов сканов документов Главархива Москвы. В основном, это метрические книги Москвы и Московской губернии примерно с середины 18 века, но встречаются также такие документы, как исповедные ведомости и ревизские сказки. Особенности дореволюционной орфографии в виде букв, которых нет в современном русском языке (ять, фита и ижица), а также написание многих слов с буквами еръ (ъ) и ерь (ь).

Фрагмент скана Метрической книги Богородице-Казанской единоверческой церкви села Средне-Егвинского Пермского уезда за 1910 год.

ГА Пермского края (ГАПК), Фонд №37, опись №6, ед. хр. 855а, стр. 127



Распознанный фрагмент рукописного текста:

1 лед. 99. Его ше дочь Екатерина Священникъ Андрей Мельников 4д песаломщика Михаилъ Мельниковъ отъ слабости. ходскомъ кладбищъ.

ОПРЕДЕЛЕНИЕ ЯЗЫКА МУЛЬТИЯЗЫЧНОГО ТЕКСТА

Фрагмент переписки индийских студентов в Фейсбуке

хинди

бенгали

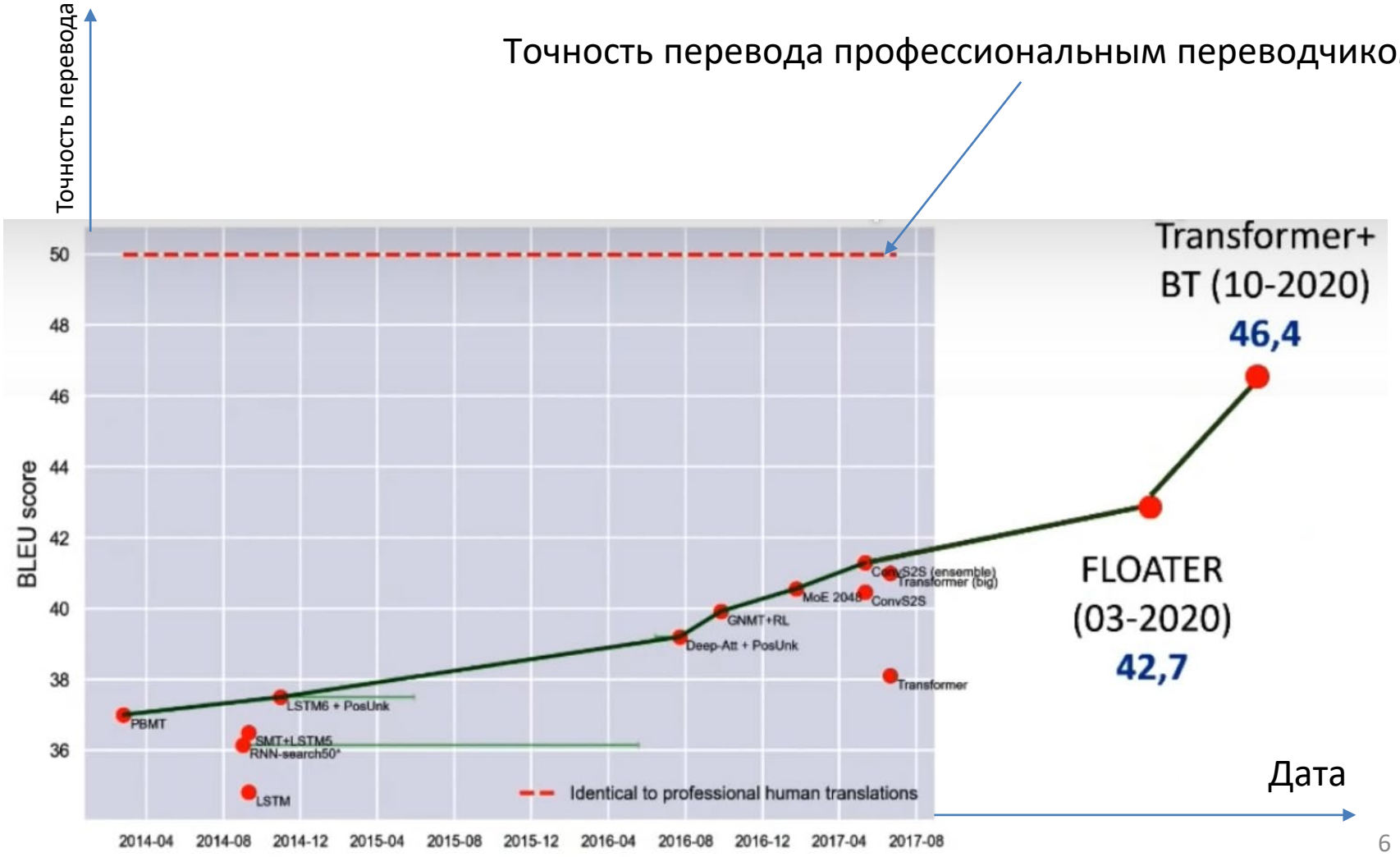
английский

Yaar tu to, GOD *hain*. **tui** JU **te ki korchis**? Hail u man!

Translation: Buddy you are GOD. What are you doing in JU? Hail u man!

Используется упрощенная романизация на фонетических принципах.

МАШИННЫЙ ПЕРЕВОД. РОСТ ТОЧНОСТИ СИСТЕМ



НЕКАЧЕСТВЕННЫЙ ОБУЧАЮЩИЙ КОРПУС ПРИВЕДЕТ К НЕКАЧЕСТВЕННОЙ МОДЕЛИ

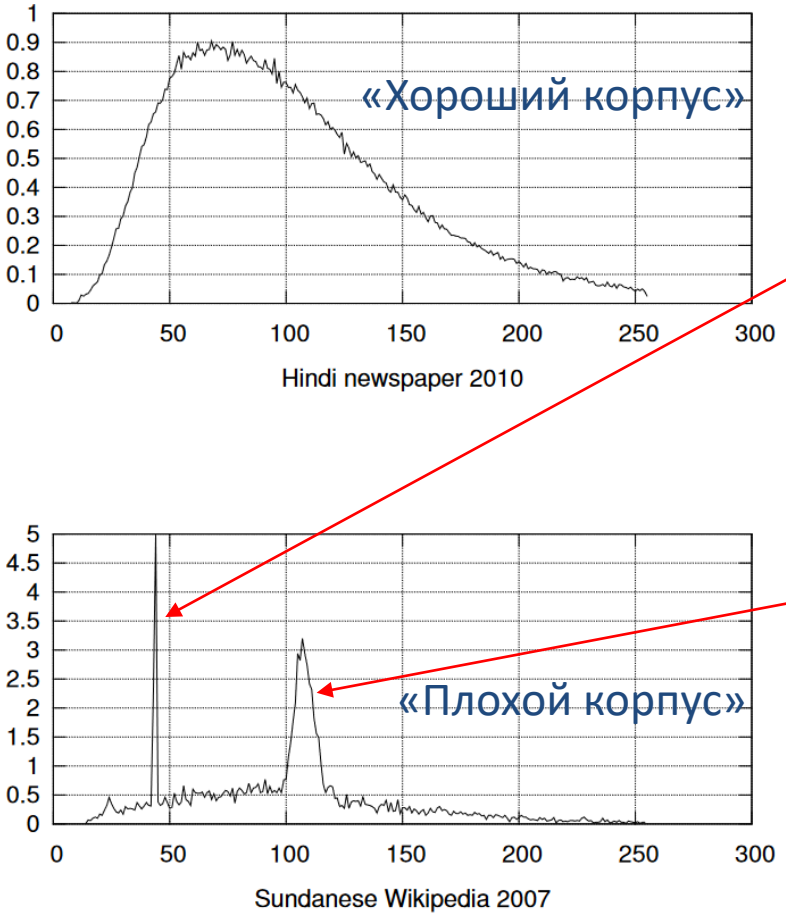


Figure 1: Sentence length distribution for two corpora (percentage for number of characters)

Фрагменты «плохого» корпуса, которые привели к пикам на графике.

- Taun ka-1118 Maś ehi dina Kaĺ ender Gŕ egorian.
- Taun ka-1119 Maś ehi dina Kaĺ ender Gŕ egorian.
- Taun ka-1120 Maś ehi dina Kaĺ ender Gŕ egorian.

11?? год нашей эры по григорианскому календарю.

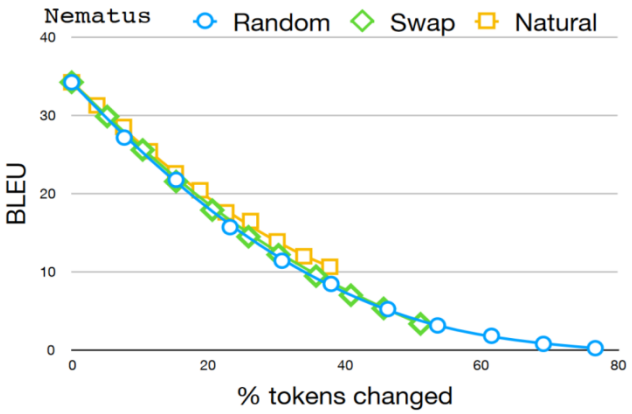
- Ancol nyáeta salasahiji d́esa di kacamatan Cińeam, Kabupat́en Tasikmalaya, Propinsi Jawa Barat, Indońesia.
- Babakan nyáeta salasahiji d́esa di kacamatan Wanayasa, Kabupat́en Purwakarta, Propinsi Jawa Barat, Indońesia.
- Bakung Lor nyáeta salasahiji d́esa di kacamatan Klangenan, Kabupat́en Cirebon, Propinsi Jawa Barat, Indońesia.

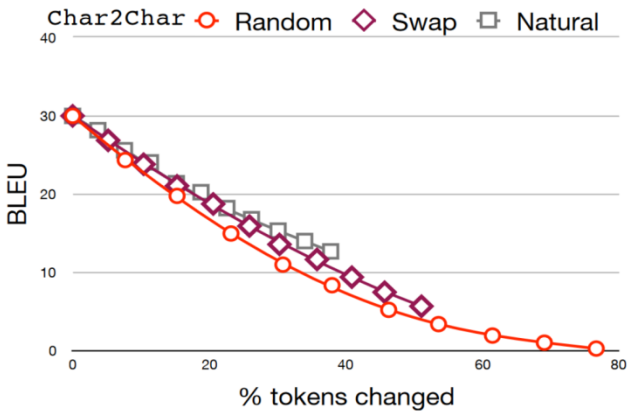
X — деревня в районе Y округа Z провинции Западная Ява, Индонезия.

Eckart T., Quasthoff U., Goldhahn D. The Influence of Corpus Quality on Statistical Measurements on Language Resources // *in: LREC'12*

ВЛИЯНИЕ ИСКАЖЕНИЙ В ТЕКСТЕ НА КАЧЕСТВО МАШИННОГО ПЕРЕВОДА

1. *Belinkov Y., Bisk Y.* Synthetic and natural noise both break neural machine translation, *arXiv:1711.02173, 2017*.
Рассмотрены четыре типа случайных искажений, связанных с перестановками букв внутри слова (1 – транспозиция соседних букв, 2 – случайное нарушение порядка следования букв в слове, за исключением первой и последней, 3 – случайная перестановка букв в пределах слова и 4 – случайная замена буквы на другую). Исследования проводились с текстами на французском, немецком и чешском языках, для оценки качества перевода использован показатель BLEU.





Результаты при использовании коррекции от Гугл.

Table 5: Google Translate's performance with natural errors and the gains from using spell checking.

French			German			Czech		
Vanilla	Nat	Spelling	Vanilla	Nat	Spelling	Vanilla	Nat	Spelling
43.3	16.7	21.4	38.7	18.6	25.0	26.5	12.3	11.2

2. *Khayrallah H., Koehn P.* On the Impact of Various Types of Noise on Neural Machine Translation, *In Proceedings of the 2nd Workshop on Neural Machine Translation and Generation,* 2018, pp. 74-83.
Показано, что нейросетевые системы автоматического перевода менее устойчивы к зашумлению обучающих данных, чем системы автоматического перевода, построенные на статистических принципах.

ЭФФЕКТИВНОСТЬ BERT В УСЛОВИЯХ ИСКАЖЕНИЙ ТЕКСТОВ

Предварительно обученные нейросетевые языковые модели, такие как BERT (Bidirectional Encoder Representations from Transformers, в настоящее время обеспечивают наивысшее качество решения многих задач в области вычислительной лингвистики, таких как аннотирование, распознавание поименованных сущностей, машинный перевод и др.

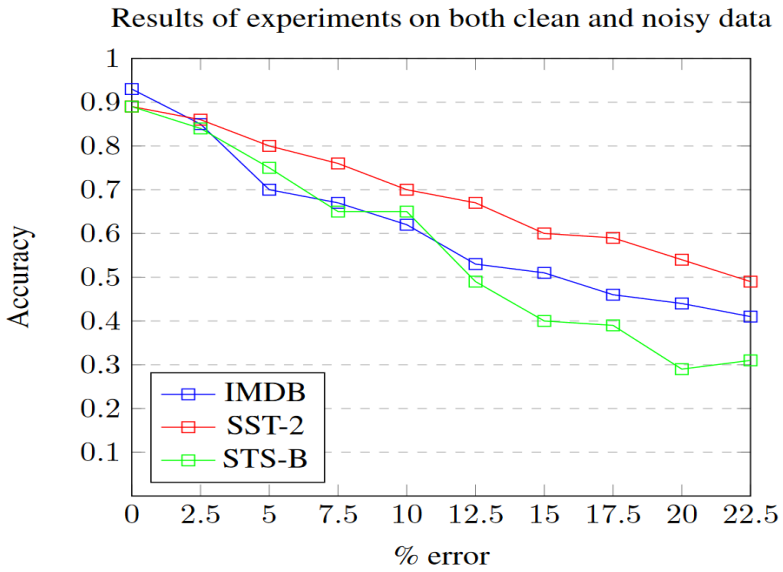


Figure 1: Accuracy vs Error

В качестве искажений выступают случайные опечатки, т.е. замена символа на другой символ, расположенный рядом на клавиатуре типа QWERTY – «fat finger problem». Рассматривался диапазон искажений от 0 до 22.5%. Показано, что с ростом уровня ошибок эффективность BERT резко падает. В частности, в задачах оценки тональности и определения близости предложений при уровне символьных ошибок в 15–17 % результат сопоставим со равновероятным выбором возможных ответов.

Kumar A., Makhija P., Gupta A. Noisy Text Data: Achilles' Heel of BERT, *Proceedings of the 2020 EMNLP Workshop W-NUT: The Sixth Workshop on Noisy User-generated Text,* pp. 16-21.

КОРРЕКЦИЯ ИСКАЖЕННЫХ ТЕКСТОВ

Уровень искажений – 7%

SUCH_A_CHANGE_IN_APPROACH_IN_THE_MIDDLE_OF_THE_PROCESS_IS_LIKELY_TO_HAVE_L
DEMOTIVATING_EFFECTION_THE_SEDARABIAN_LEADERSHIP_IS_DECLARED_DEMANDTECH_IS
CHANGEL_IN_BELARUS_HAS_NEVER_BEEN_EU__DETOX__THE_READINESS_OF_THE_SEPARABI
TO_ENGAGE_IN_CHANGESITORY_LIKELY_DEPENDS_ON_KEEPING_IT_CONVINCED_THAT_FULF
BOTH_DEMANDS_EXACTLY_DOES_NOTTENDANGER_ITS_REMAINING_IN_POWERI_TO_DEMAND_T
BELARUSIAN_CHANGES_IN_LEGISLATION_OR_ADMINISTRATIVE_MEASURES_OF_LIBREVERSI
SET_OF_CONDITIONS_THAT_ARE_IMPOSSIBLE_TO_FULFILLY_BY_THEIR_NATURE_TEACH_ME
ALMOST_ALWAYS_REVERSIBLE_ROE_IT_THE_ABDUCTION_OF_CERTAIN_ARTICLES_OF_THE_P

Уровень искажений – 18%

G_OUTH_CHANGE_IN_APPROACH_IN_THE_MIDDLE_OF_THE_PROCESS_TO_LIKELY_TO_HAVE_L
DEMOTIVATING_EFFECTION_THELPED_POSTAN_LEADERESIDE_A_DECLARGUIDE_ANOTECH_IS
CHANGELATE_SEPARDS_BAS_NEVER_BEEN_TO_POLICY__THE_FELLINESS_OF_THE_SEPARDST
TO_ENGAGE_IN_GRAMMEDITORY_LIKELY_DEPENDS_ON_KEEPING_IT_CONVING_BAIJAY_FULF
ESTER_EXANDS_EXACTLYINGESTHATTERDANGERATISTREMAINING_IN_POWERE_TO_DEMAND_T
BELLBUSIAN_CHANGES_IN_BEGISLATION_OR_ADMINISTRATIVEANHAGGREELSE_LIBREVERSI
EFT_OF_CONDITIONS_THAT_ARE_IMPOSSIBLE_TO_HOSPIDLY_BY_THEIR_NATUREDISCOW_ME
ALMOST_ALWAYS_REVERSIBLE_USE_IT_THE_ABOLITION_OF_CERTAIN_ARTICLES_OF_THE_R

Языковые вероятностные модели

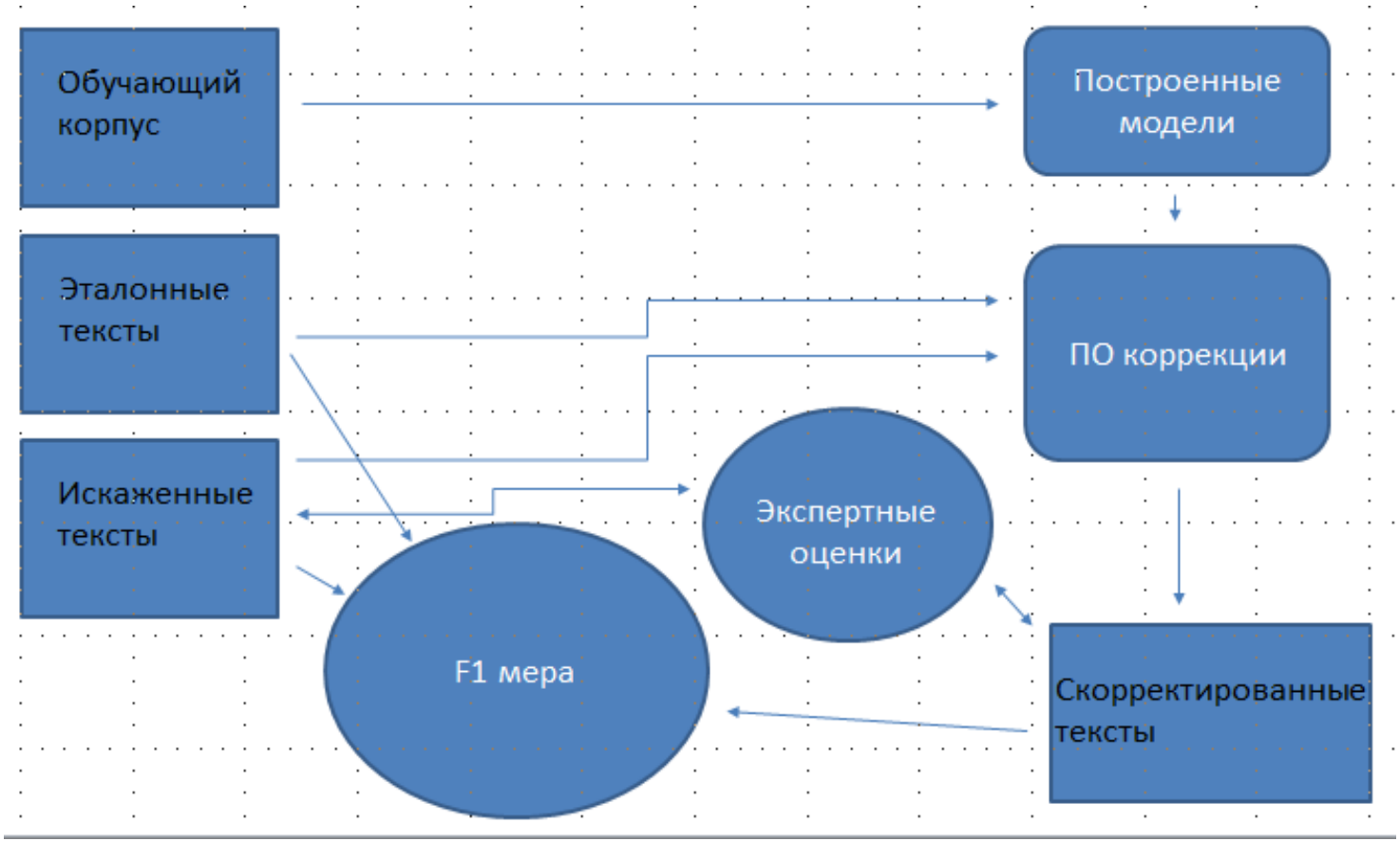
Алгоритмы поиска лучших вариантов восстановления

Реализовано для 10 языков, включая *английский, арабский, испанский, немецкий, и французский.*

Проект ФПИ «Семантика» 2018-2020 (арабский язык).

РАЗРАБОТКА СИСТЕМЫ КОРРЕКЦИИ

Рассматривается задача разработки системы коррекции искаженных текстов, точность работы которой сопоставима с точностью ручной коррекции, осуществляемой экспертом-лингвистом.



МЕТОДЫ И ТРУДОЕМКОСТЬ КОРРЕКЦИИ В ЗАВИСИМОСТИ ОТ УРОВНЯ ИСКАЖЕНИЙ

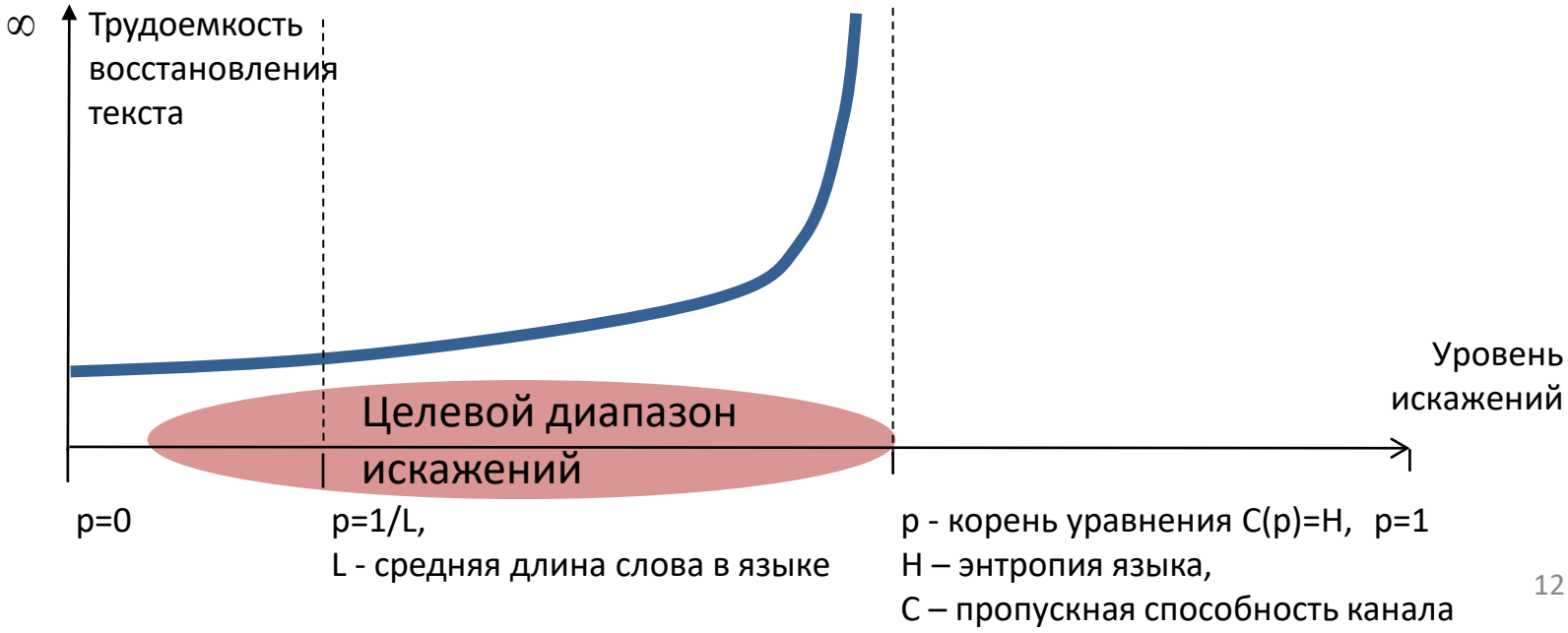
Методы

Коррекция отдельных ошибок по словарю

Восстановление текста методами поиска с использованием моделей языка

Восстановление невозможно

Трудоемкость



КОРРЕКЦИЯ ИСКАЖЕННЫХ ТЕКСТОВ

АНАЛИЗ ЭКСПЕРТНЫХ ОЦЕНОК ИСКАЖЕННЫХ И СКОРРЕКТИРОВАННЫХ ТЕКСТОВ (КОРПУС 4930 ТЕКСТОВ, 2 МЛН СЛОВОФОРМ) (ПРОЕКТ ФПИ «СЕМАНТИКА» 2018-2020)

1. Согласованность экспертных оценок

Проверка согласованности экспертных оценок читаемости проводилась с помощью вычисления коэффициента конкордации Кендалла и сравнения полученного значения со значением статистики предельного распределения при нулевой гипотезе с уровнем значимости 0.95. Подсчет проведен с помощью программы на Питоне с использованием библиотеки scipy. Вычисления показывают, что полученные оценки читаемости являются согласованными.

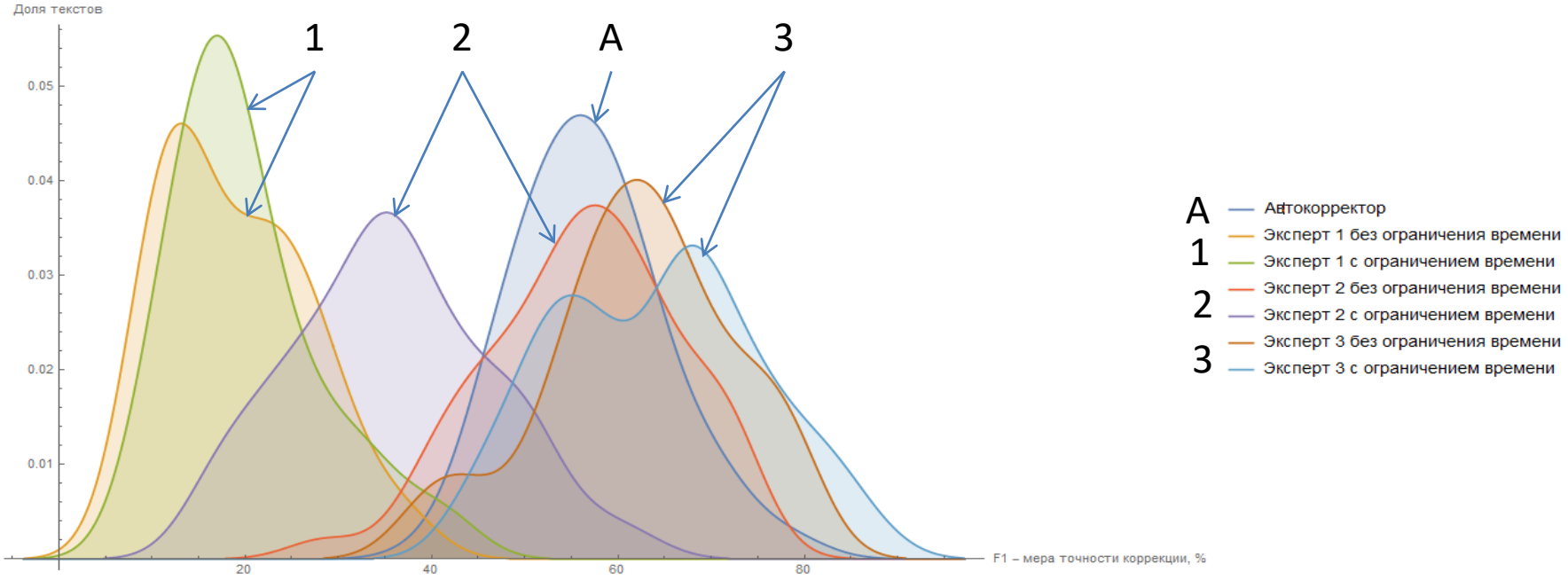
2. Статистические данные по средним оценкам лингвистов

	Лингвист 1	Лингвист 2	Лингвист 3
Искаженные тексты	3.36	3.45	3.25
Скорректированные тексты	3.85	4.16	4.13

3. Статистические данные по экспертным оценкам скорректированных текстов в зависимости от экспертных оценок искаженных текстов

Число текстов с указанной парой оценок				
Средняя оценка читаемости искаженных	Читаемость скорректированных			
	5	**4**	**3**	**2**
[2,3]	1925	2604	1356	301
(3,4]	4261	1787	502	56
(4,5]	1746	236	16	0

СРАВНЕНИЕ КАЧЕСТВА АВТОМАТИЧЕСКОЙ И РУЧНОЙ КОРРЕКЦИИ, ПРОВЕДЕННОЙ ЭКСПЕРТАМИ РАЗНОЙ СТЕПЕНИ КВАЛИФИКАЦИИ



Точность ручной коррекции существенно зависит от квалификации эксперта-лингвиста.
В случае **очень хорошего знания языка** эксперт корректирует искаженный текст точнее, чем автоматический корректор. При этом введенное ограничение на время проведения коррекции незначительно влияет на качество работы. В случае коррекции искаженного текста **квалифицированным экспертом** результаты его работы несколько ниже точности автоматического корректора. Однако если эксперту предоставить дополнительное время для работы, он заметно улучшает свои результаты. Для **эксперта-лингвиста с квалификацией ниже средней** точность коррекции искаженных текстов заметно хуже, чем точность работы разработанного ПО коррекции. Введенное ограничение на время проведения коррекции незначительно влияет на точность работы эксперта.

Германович А.В., Мельников С.Ю., Пересыпкин В.А., Сидоров Е.С., Цопкало Н.Н. Информационные измерения языка. Программная система оценки читаемости искаженных текстов // *Известия ЮФУ. Технические науки, № 8. — 2019. — С. 6–18.*

ОСОБЕННОСТИ ЗАДАЧИ ИНТЕРПРЕТАЦИИ СИСТЕМ ИИ ДЛЯ NLP

Необходимость интерпретации.

Аргументы «за»: доверие, безопасность, надежность.

Аргументы «против»: производительность.

Особенности в задачах NLP.

Нужно сравнить нейросетевые системы с традиционными т.н. «функциональными системами». В «функциональных системах» используются понятные и научно обоснованные в лингвистике признаки: морфологические свойства, лексические классы, синтаксические категории и т.п.

В нейросетевых системах сложно понять, что происходит в сквозной модели нейросети, на вход которой поступают например, слова (через встраивание), а на выходе – результат решения практической задачи из NLP, например, классификация предложений.

Как лингвистические признаки можно увидеть в нейронных сетях?

Обычно используется следующий подход. Модель нейросети обучается на какой-либо задаче, например, МТ, после чего ее веса фиксируются. Далее эта обученная модель запускается на корпусе, размеченным теми или иными лингвистическими признаками, и анализируется статистика работы сети (например, активации скрытых состояний) в зависимости от лингвистического признака.

Yonatan Belinkov and James Glass. Analysis Methods in Neural Language Processing: A Survey. *Transactions of the Association for Computational Linguistics, 7:49–72, 2019.*

ПРИМЕР ПОДХОДА К ОТСЛЕЖИВАНИЮ ЛИНГВИСТИЧЕСКИХ ПРИЗНАКОВ. СОХРАНЕНИЕ ЗАЛОГА В СИСТЕМЕ ПЕРЕВОДА

Обучаем англо-французскую систему NMT на 110 млн токенов двуязычных данных.

Возьмем 10 000 отдельных английских предложений и пометим их залог (активный или пассивный).

Преобразуем эти предложения в 1000-мерные вектора, используя обученный кодировщик NMT.

На корпусе из 9000 предложений обучим модель логистической регрессии для прогнозирования залога с использованием состояний кодирующих ячеек.

На оставшихся 1000 предложениях достигается точность 92,8%.

Это означает, что при сокращении исходного предложения до вектора фиксированной длины в системе NMT залог сохраняется.

Model	Accuracy
Majority Class	82.8
English to French (E2F)	92.8
English to English (E2E)	82.7

Table 1: Voice (active/passive) prediction accuracy using the encoding vector of an NMT system. The majority class baseline always chooses active.

Если провести тот же эксперимент с англо-английской (автокодирующей) системой, то обнаруживается, что вектор кодирования не содержит дополнительной информации о залоге предложения. В этом случае залог предсказывается только с точностью 82,7%, что соответствует соотношению частот использования активного и пассивного залогов в английском языке.

Xing Shi, Inkit Padhi, and Kevin Knight. Does String-Based Neural MT Learn Source Syntax? // *In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pages 1526–1534, Austin, Texas.*

СОСТЯЗАТЕЛЬНЫЕ ПРИМЕРЫ В ЗАДАЧАХ NLP

Задача. Имея модель нейросети f и входной пример x, построить состязательный пример x', на котором достигается

$$\min_{x'} ||x - x'||$$

при этом $$f(x) = l, f(x') = l', l \neq l'$$

Проблемы, характерные для задач NLP:

1 – неясно, что такое расстояние между текстами,

2 – неясно, как его минимизировать.

В работе (*) предложен генетический алгоритм для создания состязательных примеров. В алгоритме поддерживается популяция модификаций исходного предложения, а в каждом поколении оценивается пригодность модификаций.

Рассмотрены:

- *задача определения тональности на базе IMDB, состязательный пример обманывает систему в 97% случаев, но в 92% случаев проходит оценку экспертов.*
- *задача определения логической связи между текстами Stanford Natural Language Inference (SNLI), состязательный пример обманывает систему в 70% случаев.*

** Moustafa Alzantot, Yash Sharma, Ahmed Elgohary, Bo-Jhang Ho, Mani Srivastava, and Kai-Wei Chang.* Generating Natural Language Adversarial Examples // *In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*

ЗАКЛЮЧЕНИЕ

1. В условиях колоссального роста объемов анализируемого информационного контента альтернативы нейросетевым методам машинного обучения в задачах NLP нет.
2. Применение нейросетевых методов машинного обучения для построения моделей языка значимо снижает энтропию построенных моделей в сравнении с вероятностно-статистическими методами.
3. Применение нейросетевых методов машинного обучения в задачах NLP требует значительных вычислительных ресурсов, что ограничивает глубину зависимости при построении моделей, и объемов обучающих корпусов текста (критично для малоресурсных языков).
4. Достоверность построенных нейросетевых моделей языка зависит от чистоты, актуальности и тематической направленности обучающих корпусов текстов.
5. Эффективность нейросетевых методов машинного обучения в задачах NLP существенно снижается при наличии искажений в текстах.
6. Прямая интерпретация работы нейросетевых методов в задачах NLP весьма затруднена, однако в некоторых случаях возможна косвенная.
7. Разработка состязательных примеров для систем с нейросетевыми моделями текстов возможна в ряде систем. Для построения таких примеров перспективными представляются методы последовательной оптимизации.

БИБЛИОГРАФИЯ

1. http://www.scrf.gov.ru/security/information/document5
2. http://www.scrf.gov.ru/security/information/document155
3. Barbon S., Igawa R., Zarpelão B. Authorship verification applied to detection of compromised accounts on online social networks: A continuous approach // Multimedia Tools and Applications, 2017, 76(3), pp.3213–3233.
4. DumanS., Kalkan-Cakmakci K., Egele M., Robertson W., Kirda E. EmailProfiler: Spearphishing filtering with header and stylometric features of emails // In Proceedings of the 2016 IEEE 40th Annual Computer Software and Applications Conference, 2016, Vol. 1, pp. 408–416.
5. BrocardoM., Traore I., Woungang I., Obaidat M. Authorship verification using deep belief network systems // International Journal of Communication Systems, 2017, 30(12), e3259.
6. Третьяков А.О., Филатова О.Г., Жук Д.В., Горлушкина Н.Н., Пучковская А.А. Метод определения русскоязычных фейковых новостей с использованием элементов искусственного интеллекта // International Journal of Open Information Technologies, vol. 6, №12, 2018, pp. 99-105.
7. Layton R., Watters P., Ureche O. Identifying faked hotel reviews using authorship analysis // In Proceedings - 4th Cybercrime and Trustworthy Computing Workshop, CTC ’13, 2013, pp. 1–6.
8. Panicheva P., Cardiff J., Rosso P. Personal sense and idiolect: Combining authorship attribution and opinion analysis // In Proceedings of the International Conference on Language Resources and Evaluation, LREC 2010.
9. Shahid U., Farooqi S., Ahmad R., Shafiq Z., Srinivasan P., Zaffar F. Accurate detection of automatically spun content via stylometric analysis // In Proceedings of the 2017 IEEE International Conference on Data Mining (ICDM), 2017, pp. 425–434.
10. Iskhakova A., Iskhakov A., Meshcheryakov R. Research of the estimated emotional components for the content analysis // Journal of Physics: Conference Series, 2019, 1203. 012065.
11. Исхакова А.О. Метод определения искусственных текстов на основе расчета меры принадлежности к инвариантам // Труды СПИИРАН. 2016. № 6 (49). С. 104-121.
12. Teahan W., Cleary J. The Entropy Of English Using PPM-based Models, Proceedings of Data Compression Conference-DCC'96, IEEE Computer Society Press, 1996, pp. 53-62.
13. Alghamdi N., Berriche L. Capacity Investigation of Markov Chain-Based Statistical Text Steganography: Arabic Language Case // In Proceedings of the 2019 Asia Pacific Information Technology Conference (APIT 2019). ACM, New York, USA, pp. 37-43.
14. Германович А.В., Мельников С. Ю., Пересыпкин В. А., Сидоров Е. С., Цопкало Н. Н. Информационные измерения языка. Программная система оценки читаемости искаженных текстов // Известия ЮФУ. Технические науки, №8, 2019, С.6-18.
15. Северин Н.В. Методы нечеткого поиска в системах контроля нецелевого контента // Вісник Східноукраїнського національного університету ім.В.Даля, № 8 (179), Ч.2., 2012, С.199 – 205.
16. Бирин Д.А., Мельников С.Ю., Пересыпкин В.А., Писарев И.А., Цопкало Н.Н. Об эффективности средств коррекции искаженных текстов в зависимости от характера искажений // Известия ЮФУ. Технические науки, №8, 2018, С.104-114.
17. Кулай А.Ю., Леднов Д.А., Мельников С.Ю. О статистических методах идентификации языка, искаженных текстовых и речевых сообщений // Известия ЮФУ. Технические науки, №8, 2008, С. 177–183.
18. Iskhakova A., Kruglova S., Melnikov S., Sidorov E. The Approach to Minimize the Impostor Method Errors in the Author Identification Open Problem // Proceedings of the R. Piotrowski's Readings in Language Engineering and Applied Linguistics. S.Petersburg, Russia, November 27, 2019. CEUR Workshop Proceedings, V.2552, pp. 60-72.
19. Орлов Ю.Н., Осминин К.П. Определение жанра и автора литературного произведения статистическими методами // Прикладная информатика 2(26), 2010, С. 95-108.
20. Исхакова А.О. Выбор параметров для идентификации искусственно созданных текстов // Доклады Томского государственного университета систем управления и радиоэлектроники, 2013, № 2 (28), С. 126-128.
21. Вахлаков Д.В., Мельников С.Ю., Пересыпкин В.А. Многоэтапный метод автоматической коррекции искаженных текстов // Известия ЮФУ. Технические науки, №7, 2020, С.35-45.
22. Arjun M.Y., Chirag H. G. Language Identification from a Tri-lingual Printed Document: A Simple Approach // Int. Journal of Engineering Research and Applications. Vol. 4, Issue 9, September 2014, pp.132-136.
23. Elamine M., Mechti S., Belguith L. An Unsupervised Method for Detecting Style Breaches in a Document // 2019 IEEE/ACS 16th International Conference on Computer Systems and Applications (AICCSA), 2019, pp.1-6.
24. Сафин К. Ф., Кузнецов М. П., Кузнецова М. В. Определение заимствований в тексте без указания источника // Информ. и её примен., 11:3 (2017), С.73–79.
25. Agirre E. Semantic textual similarity / E. Agirre, D. Cer, M. Diab, A. Gonzalez-Agirre, G. Weiwei // 2nd Joint Conference on Lexical and Computational Semantics, 2013, pp. 32-43.

СПАСИБО ЗА ВНИМАНИЕ

Спасибо за внимание!

info@linfotech.ru

ООО «Лингвистические и информационные технологии»
Москва, ул. Образцова, 38.