УДК 004.657 DOI 10.25730/VSU.0536.19.019

К вопросу устранения ЕЯ-неопределенностей в современных средствах обработки информации

В. С. Русов[1], Д. А. Захарова[2], В. Л. Клюкин[3]

[1]магистрант кафедры электронных вычислительных машин, Вятский государственный университет. Россия, г. Киров. E-mail: vyacheslavrusov96@gmail.com

[2] магистрант кафедры электронных вычислительных машин, Вятский государственный университет. Россия, г. Киров. E-mail: zakharovadaria96@gmail.com

[3]старший преподаватель кафедры электронных вычислительных машин, Вятский государственный университет. Россия, г. Киров. E-mail: klyukin@vyatsu.ru

Аннотация. В данной статье рассмотрена основная проблема устранения естественно-языковых неопределенностей текстов – устранение опечаток с учетом контекста в современных средствах обработки информации. Была разработана эффективная модель для решения данной проблемы. В ходе разработки модели сформулирована основная терминология проблемы, описаны примеры алгоритмов для реализации модели, даны советы по эффективному использованию модели на практике. При разработке модели были учтены краевые случаи, возникающие при обработке текстов с опечатками, описаны возможные шаблоны поведения при возникновении подобных ситуаций. Применение разработанной модели позволяет разрешать конфликтные ситуации неопределенностей опечаток в текстах на естественном языке, оставляя за конкретной реализацией выбор конкретных алгоритмов и подходов к решению каждого из этапов. В результате был рассмотрен реальный пример снятия неопределенности, позволяющий наглядно представить этапы работы с моделью.

Ключевые слова: естественно-языковой интерфейс, устранение языковых неопределенностей, опечатки.

Введение. Исторически автоматизация решения задач с помощью ЭВМ проводилась в тех областях, где задачи не требовали от исполнителя серьезной умственной либо творческой активности. Однако с ежегодным ростом мощностей компьютерной техники появляется возможность заменять большую часть работы человека автоматизированными системами. Так все больше интеллектуальных компьютерных систем входят в обиход современного человека: умные фитнес-трекеры, «умные» дома, боты-консультанты по юридическим вопросам. Большинство современных компьютерных систем объединяет тот факт, что для них пользователь является основным источником поступления информации [5].

Как известно, человеку гораздо проще выражать свои мысли на естественном языке (ЕЯ). Использование естественно-языковых интерфейсов (Natural language interface – NLI) со стороны клиента позволяет проводить общение с системой на понятном языке [7, с. 1164]. Основными требованиями, которые выдвигаются к NLI, являются глубокое понимание семантики языка, быстродействие, корректная обработка ситуаций, когда система не может дать однозначный ответ без дополнительных сведений. Таким образом, важной частью решения большого класса задач современного мира является анализ естественно-языковой информации, ее приведение к структурированному виду для дальнейшей обработки системой [4, с. 813].

В модулях обработки ЕЯ информации можно выделить множество проблем. К ним относятся зависимость от конкретного ЕЯ, различные виды неопределенностей, большая вычислительная сложность, быстрое изменение языковой среды. Для решения данных проблем могут быть применены следующие подходы: введение ограничений ЕЯ; привлечение профессиональных лингвистов для построения языковой модели; разработка и применение эффективных современных методов автоматизированной обработки информации и построения языковой модели [2, с. 306]. Основной проблемой NLI является разрешение языковых неопределенностей, самыми популярными из которых являются опечатки.

Целью данной работы является устранение проблем NLI, связанных с языковыми неопределенностями, возникающими из-за опечаток.

Задачи исследования. Для достижения данной цели необходимо решить следующие задачи.

1. Проанализировать предметную область, введя термин ЕЯ-неопределенности, рассмотреть основные проблемы и стратегии устранения опечаток.
2. Произвести анализ существующих решений для устранения неопределенностей опечаток.
3. Разработать формальную модель разрешения неопределенностей, вызванных опечатками.
4. Привести пример, демонстрирующий достоинства и недостатки формальной модели.

Предлагаемый подход. Будем понимать под неопределенностью, возникающей при анализе текста на естественном языке, ситуацию, когда система, находясь в контексте только данного текста, не может явно сопоставить вложенный автором смысл с предполагаемым смыслом текста.

Опечатки – неверное письменное обозначение желаемой информации – возникают из-за случайных ошибок в наборе текста, при отсутствии некоторых знаний языковых правил. Пример: «сонце» (правильно «солнце»). Основной проблемой обработки опечаток является ситуация, когда слово с опечаткой существует в словаре. В таких случаях единственным критерием неправильного употребления слова является его семантическая неупотребимость в данном контексте [6].

Проанализируем существующие средства разрешения неопределенностей опечаток. Самым популярным из методов решения неопределенностей, вызванных опечатками, является поиск слов в словаре. Основным достоинством данного метода является простота реализации и низкая вычислительная сложность [1, с. 8]. Многие системы хранят статистику запросов и используют ее для разрешения неопределенностей. Данный подход позволяет эффективно разрешать неопределенности, однако не может определить победителя из равнопопулярных вариантов. Система правил часто применяется для решения проблемы опечаток, но описание формальных правил языковой системы требует огромного количества человеческих ресурсов. Примером данного подхода может служить замена связок «жы» на «жи».

Стоит заметить, что существуют опечатки, которые не могут быть исправлены без семантического анализа контекста, а для полного понимания семантики предложения требуется избавиться от опечаток, следовательно, возникает противоречие – устранение опечаток требует анализа, проведение которого требует устранения опечаток.

Определим следующую формальную графовую модель, которая может быть применена для решения подобных задач. Вершинами являются слова, ребрами – связи между словами. Модель оставляет за конкретной реализацией подход к определению связи слов [3, с. 1588]. Возможные разновидности связей представлены на рис. 1 и рис. 2.

Рис. 1. Линейная семантическая связь слов в предложении

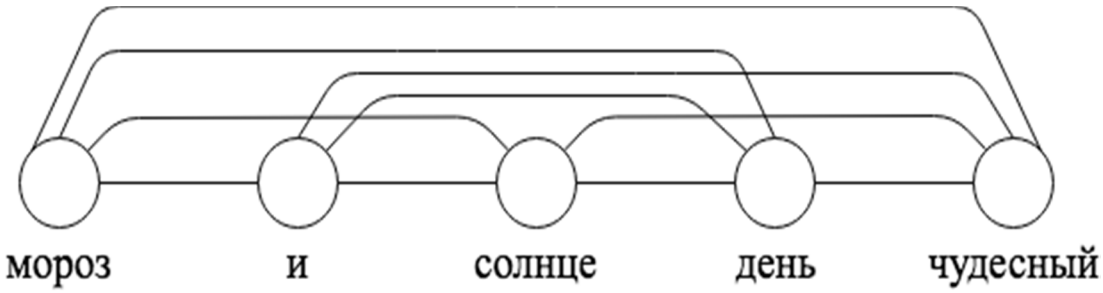


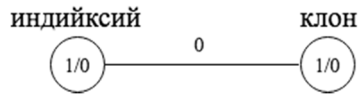
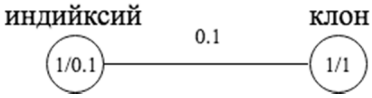
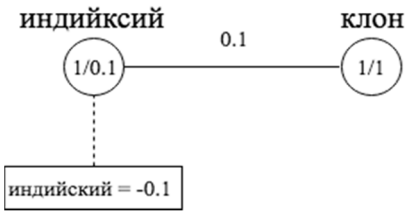
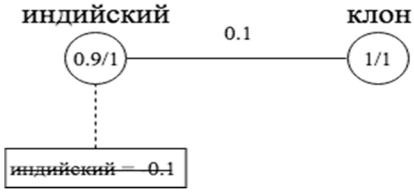
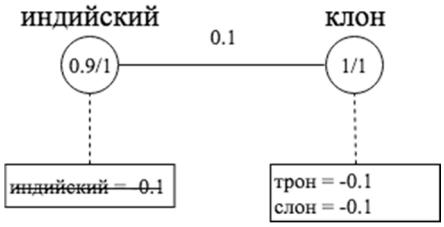
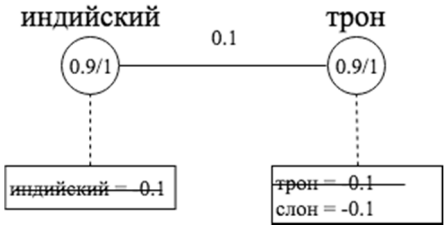
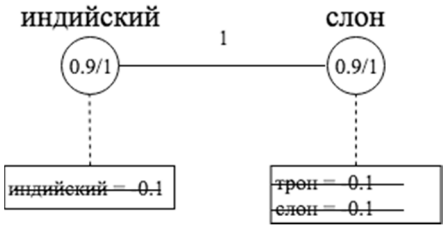
Рис. 2. Полносвязная семантическая связь слов в предложении

Будем понимать под *соответствием* с эталонным значением метрику, которая показывает, насколько заменяемое слово отличается от исходного. Каждое устранение опечаток относительно эталонного значения вычитает из этой величины значение, прямо пропорциональное количеству изменений и обратно пропорциональное вероятности появления этих изменений. Изначально соответствие с эталонным значением всех вершин равно единице.

Результаты исследования. Рассмотрим пример устранения неопределенностей словосочетания «индийксий клон». В данном примере возможны только два способа организации связей между словами. Случай, когда слова не имеют связи, неинтересен, поэтому линейно свяжем эти слова. Процесс устранения неопределенностей данного словосочетания приведен в табл 1.

Таблица 1

Устранения неопределенностей словосочетания «индийксий клон»

Состояние графа	Описание
	Инициализируем все соответствия с эталонным значением единицей, все релевантности нулями.
	Определяем реальную релевантность ребра. Имеется слишком мало информации по употребимости слов «индийксий» и «клон» – определяем новую релевантность ребра 0.1. Определяем релевантности вершин: для вершины «индийксий» = 0.1 – слова нет в словаре, для слова «клон» – 1, слово есть в словаре.
	Так как вершина «индийксий» имеет малую релевантность, подбираем похожих кандидатов. В данном случае он один, имеет цену погрешности 0.1.
	Новое значение соответствия с эталонным значением = 1 – 0.1 = 0.9. Новое значение релевантности ребра по-прежнему имеет малое значение (так как слова редко встречаются в контексте).
	Предполагаем, что в слове «клон» допущена опечатка, два равноправных кандидата. Выбираем первый «трон».
	Пересчитываем соответствие с эталонным значением вершины, релевантность ребра не изменяется.
	Пробуем использовать второго кандидата. Релевантность ребра возрастает до единицы. Среднее значение всех величин больше 0.5, следовательно, все неопределенности решены.

Выводы. В данной статье были рассмотрены основные проблемы современных средств обработки естественно-языковых запросов, выделен класс проблем, касающихся неопределенностей.

Проанализированы существующие решения и стратегии, которые позволяют бороться с ЕЯ-неопределенностями.

Разработана собственная модель, позволяющая эффективно разрешать неопределенности и устранять опечатки в тексте. Применимость, а также преимущества предложенного метода были продемонстрированы на конкретном примере.

Список литературы

1. *Адитья Б.* Грокаем алгоритмы : иллюстрированное пособие для программистов и любопытствующих. СПб. : Питер, 2019. 288 с.
2. *Мельцов В. Ю., Нечаев А. А.* Семантико-синтаксический анализ предложений на русском языке с помощью нейронных сетей // Фундаментальные исследования. 2017. № 11-2. С. 306–310.
3. *Мельцов В. Ю., Страбыкин Д. А.* Вывод следствий с построением схемы логического вывода // Фундаментальные исследования. 2013. № 11-8. С. 1588–1593.
4. *Мельцов В. Ю., Чистяков Г. А.* Эффективный метод построения оптимизированного дерева грамматического разбора формул темпоральной логики линейного времени // Вестник тамбовского государственного технического университета. 2012. Т. 18. № 4. С. 813–820.
5. Обработка естественного языка. URL: https://ru.wikipedia.org/wiki/Обработка_естественного_языка (дата обращения: 02.02.2019).
6. Правила русского языка. URL: http://www.fio.ru/pravila/leksika (дата обращения: 02.02.2019).
7. *Шипицына А. А., Крутиков А. К., Мельцов В. Ю.* Представление знаний в интеллектуальной CAT-системе с ЕЯ-интерфейсом // Общество, Наука, Инновации (НПК-2015). Всероссийская ежегодная научно-практическая конференция. Киров : ВятГУ. 2015. С. 1164–1167.

On the issue of eliminating ЕЯ-uncertainties in modern means of information processing

V. S. Rusov[1], D. A. Zakharova[2], V. L. Klukin[3]

[1]master student of the Department of electronic computing machines, Vyatka State University.
Russia, Kirov. E-mail: vyacheslavrusov96@gmail.com
[2]master student of the Department of electronic computing machines, Vyatka State University.
Russia, Kirov. E-mail: zakharovadaria96@gmail.com
[3]senior lecturer of the Department of electronic computing machines, Vyatka State University.
Russia, Kirov. E-mail: klyukin@vyatsu.ru

Abstract. In this article the main problem of elimination of natural language uncertainties of texts is considered – elimination of typos taking into account a context in modern means of information processing. An effective model has been developed to address this problem. During the development of the model, the basic terminology of the problem is formulated, examples of algorithms for the implementation of the model are described, tips on the effective use of the model in practice are given. During the development of the model the marginal cases arising in the processing of texts with typos were taken into account, possible patterns of behavior in the event of such situations were described. The application of the developed model allows to resolve conflict situations of uncertainty of typos in natural language texts, leaving the specific implementation of the choice of specific algorithms and approaches to solving each of the stages. As a result, a real example of removing uncertainty was considered, which allows to visualize the stages of working with the model.

Keywords: natural language interface, the elimination of language ambiguities, typographical errors.

References

1. *Adit'ya B. Grokaem algoritmy : illyustrirovannoe posobie dlya programmistov i lyubopytstvuyushhih* [Groote algorithms: an illustrated guide for programmers and curious people]. SPb. Piter. 2019. 288 p.
2. *Mel'cov V. Yu., Nechaev A. A. Semantiko-sintaksicheskij analiz predlozhenij na russkom yazyke s pomoshh'yu nejronnyh setej* [Semantic-syntactic analysis of sentences in Russian using neural networks] // *Fundamental'nye issledovaniya* – Fundamental research. 2017. No. 11-2. Pp. 306–310.
3. *Mel'cov V. Yu., Strabykin D. A. Vyvod sledstvij s postroeniem skhemy logicheskogo vyvoda* [Conclusion of consequences with the construction of a logical conclusion scheme] // *Fundamental'nye issledovaniya* – Fundamental research. 2013. No. 11-8. Pp. 1588–1593.
4. *Mel'cov V. Yu., Chistyakov G. A. Effektivnyj metod postroeniya optimizirovannogo dereva grammaticheskogo razbora formul temporal'noj logiki linejnogo vremeni* [An effective method of constructing an optimized tree of grammatical analysis of formulas of temporal logic of linear time] // *Vestnik tambovskogo gosudarstvennogo tekhnicheskogo universiteta* – Herald of Tambov State Technical University. 2012. Vol. 18. No. 4. Pp. 813–820.
5. *Obrabotka estestvennogo yazyka* – Natural language processing. Available at: https://ru.wikipedia.org/wiki/Обработка_естественного_языка (date accessed: 02.02.2019).
6. *Pravila russkogo yazyka* – Russian language rules. Available at: http://www.fio.ru/pravila/leksika (date accessed: 02.02.2019).
7. *Shipicyna A. A., Krutikov A. K., Mel'cov V. Yu. Predstavlenie znanij v intellektual'noj CAT-sisteme s ЕЯ-interfejsom* [Knowledge representation in intelligent CAT system with a ЕЯ-interface] // *Obshhestvo, Nauka, Innovacii (NPK-2015). Vserossijskaya ezhegodnaya nauchno-prakticheskaya konferenciya* – Society, Science and Innovation (Scientific and practical conference-2015). All-Russian annual scientific and practical conference. Kirov. VyatSU. 2015. Pp. 1164–1167.