МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ
РОССИЙСКОЙ ФЕДЕРАЦИИ
НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ТОМСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
Филологический факультет

АКТУАЛЬНЫЕ ПРОБЛЕМЫ ЛИНГВИСТИКИ И ЛИТЕРАТУРОВЕДЕНИЯ

Сборник материалов VII (XXI)
Международной научно-практической конференции
молодых учёных
(16–18 апреля 2020 г.)

Выпуск 21

Томск
Издательский Дом Томского государственного университета
2020

Литература

1. Информационное агентство «ТАСС» // Рубрика: спорт. [Б. с.], 2020. URL: <https://tass.ru/> (дата обращения: 05.04.2020).
2. Torres-Moreno J.M. Automatic Text Summarization. London : ISTE ; Hoboken, NJ : Wiley, 2014. 376 p.

DOI: 10.17223/978-5-94621-901-3-2020-31

ДИСКУРСИВНЫЕ МАРКЕРЫ КАК СРЕДСТВО АВТОМАТИЧЕСКОЙ КЛАССИФИКАЦИИ НОВОСТНЫХ ТЕКСТОВ МЕТОДАМИ МАШИННОГО ОБУЧЕНИЯ

Ершова Е.А.

Томский государственный университет, студент

DISCOURSE MARKERS AS FEATURES FOR AUTOMATIC CLASSIFICATION OF NEWS ARTICLES USING MACHINE LEARNING METHODS

Ershova E.A.

Tomsk State University, student

Информационная перегрузка в обществе обуславливает необходимость включения в процесс обработки текстовых данных автоматизированных механизмов, к числу которых относится автоматическая классификация текстов по заданным параметрам. В статье представлены результаты использования дискурсивных маркеров в решении задачи автоматической классификации новостных текстов.

Ключевые слова: дискурсивные маркеры, машинное обучение, классификация текстов, дискурс новостных текстов.

Information overload in modern society requires including computer-assisted mechanisms in the textual data processing, in particular, classification of texts. This research presents the results of using discourse markers as features for the automatic classification of news articles.

Key words: discourse markers, machine learning, text classification, news discourse.

Научный руководитель: З.И. Резанова, д-р филол. наук, профессор ТГУ.

Качество автоматической классификации текстов зависит, в частности, от языковых единиц, на основании которых классификация осуществляется. В статье в качестве таких единиц анализируются дискурсивные маркеры – формальные средства, указывающие на связи частей дискурса как на уровне макроструктуры, так и на уровне микроструктуры [1. С. 30]. Мы полагаем, что вариативность

использования таких единиц обуславливается своеобразием макроструктуры (обусловленная целью организация крупных блоков) и микроструктуры (организация минимальных элементов, определяющая связность) дискурса, которые в свою очередь зависят также от его модуса (устного или письменного) и жанрового воплощения.

Материал анализа, представленного в статье, – тексты одного из новостного жанра медийного дискурса, реализованного в письменном модусе. Ранее было установлено, что в письменном дискурсе частотность использования дискурсивных маркеров ниже, кроме того, в отличие от устного, в нём чаще встречаются маркеры со значением точности (напр., «именно») [1. С. 19]. Полагаем, что сочетание основных жанровых признака новости – содержательность, краткость, объективность – также предопределяет специфику используемых дискурсивных маркеров.

Объектом проведенного исследования является автоматическая классификация текстов с помощью машинного обучения. **Предметом** – использование дискурсивных маркеров в решении задачи автоматической классификации новостных текстов с помощью машинного обучения. **Цель** данного исследования – изучить возможность использования дискурсивных маркеров в решении задачи автоматической классификации новостных текстов.

Новизна данного исследования заключается, во-первых, в использовании заданных списков слов – дискурсивных маркеров как средства классификации текстов, тогда как в современных исследованиях к новостным текстам применяется либо классификация на основании самых значимых слов в текстах, либо кластеризация (в этом случае классы текстов не задаются исследователем заранее); во-вторых, в использовании портала «ТАСС» как источника **материала** исследования.

Было создано три тематических корпуса новостных текстов (тематики «Общественная жизнь», «Политика», «Происшествия»), которые составили обучающую выборку классификатора. Объём каждого корпуса – 100 текстов, объём каждого текста – около 1500 символов. Далее были созданы словари дискурсивных маркеров, на основании которых обучался классификатор. Списки «Общественная жизнь», «Политика», «Происшествия» формировались с использованием метода интроспекции, источником послужили тексты из созданных нами ранее тематических корпусов. «Общий список» состоял из дискурсивных маркеров, описанных в работах [2, 3], примеры распределения дискурсивных маркеров по тематикам представлены в табл. 1.

Т а б л и ц а 1

Примеры дискурсивных маркеров тематических корпусов новостных текстов

Политика	Социальная жизнь	Происшествия	Общий список
ведь	весьма	лишь	вон
в частности	якобы	безусловно	все-таки

Далее с помощью языка программирования R была проведена предобработка текстов, которая включала в себя лемматизацию, а также удаление ненужных символов. На следующем этапе были вычислены абсолютные, а также относительные величины, которые показали частотность употребления каждого дискурсивного маркера для каждого из текстов тематических корпусов. На основании данных величин был построен классификатор, а именно дерево решений. 80% данных использовались для тренировки модели, 20% – для её валидации. Классификация текстов была проведена сначала по всем трём тематикам, а далее – по двум тематикам попарно. В результате были получены следующие показатели параметра оценки качества классификатора «Ассигасу» (данный параметр показывает процент правильных решений модели), представленные в табл. 2.

Т а б л и ц а 2

Показатели параметра оценки качества классификатора по тематикам

	Ассигасу	
	Тренировка	Валидация
«Политика», «Происшествия», «Социальная жизнь»	0,55	0,775
«Политика», «Происшествия»	0,88125	0,925
«Происшествия», «Социальная жизнь»	0,84375	0,675
«Политика», «Социальная жизнь»	0,9	0,775

Полученные результаты были интерпретированы нами следующим образом. Качество работы модели повышается при классификации по двум тематикам. Самое высокое качество работы классификатора наблюдается при разбиении текстов на классы «Политика» и «Происшествия». Высокий показатель правильных решений на валидационной выборке относительно тренировочной может свидетельствовать о том, что данная модель характеризуется низкой вероятностью ошибки обобщения при её использовании на новых данных.

В связи с этим нами были сделаны следующие **выводы**: тексты тематик «Политика» и «Происшествия» характеризуются самыми большими

различиями в употреблении дискурсивных маркеров; следовательно, дискурсивные маркеры могут использоваться в качестве признаков для автоматической классификации новостных текстов по тематикам «Политика» и «Происшествия» с помощью модели дерева решений.

Перспективы исследования включают в себя использование других типов оценки классификатора (F-мера); использование другой модели классификации; проведение классификации на основании разных типов дискурсивных маркеров.

Литература

1. Кибрик А.А. Анализ дискурса в когнитивной перспективе : дис. ... д-ра филол. наук. М., 2003. 219 с.
2. Дискурсивные слова русского языка: опыт контекстно-семантического описания : сб. ст. / сост. К.Л. Киселева, Д. Пайар. М., 1998. 447 с.
3. Баранов А.Н., Плунгян В.А., Рахилина Е.В. Путеводитель по дискурсивным словам русского языка. М. : Помовский и партнёры, 1993. 207 с.

DOI: 10.17223/978-5-94621-901-3-2020-33

ПОЛЬЗОВАТЕЛЬСКИЕ СЛОВАРИ И КОРПОРАТИВНЫЕ ГЛОССАРИИ КАК МЕТОД ОПТИМИЗАЦИИ ПЕРЕВОДА ГУМАНИТАРНОГО НАУЧНОГО ТЕКСТА (НА МАТЕРИАЛЕ МЕТАДАННЫХ ЖУРНАЛА «РУСИН»)

Калинина Д.В.

Томский государственный университет, магистр

CUSTOM DICTIONARIES AND CORPORATE GLOSSARIES AS A METHOD TO OPTIMIZE THE PROCESS OF TRANSLATION OF HUMANITARIAN SCIENTIFIC TEXT (ON THE MATERIAL OF *RUSIN* METADATA)

Kalinina D.V.

Tomsk State University, master student

Статья посвящена проблеме использования средств автоматического и автоматизированного перевода при работе с научно-гуманитарными текстами. Материал исследования - метаданные научно-гуманитарных статей журнала «Русин». В ходе работы подтвердилась гипотеза о том, что использование современных средств автоматического перевода и корпоративных словарей может ускорить работу переводчика.