УДК 004.912

АВТОМАТИЧЕСКОЕ ИЗВЛЕЧЕНИЕ СИТУАЦИЙ ЗРИТЕЛЬНОГО ВОСПРИЯТИЯ С ИСПОЛЬЗОВАНИЕМ ПОДХОДОВ НА ОСНОВЕ МОРФО-СИНТАКСИЧЕСКИХ ПРИЗНАКОВ

Р. Б. Рыбка*, А. В. Грязнов*, А. А. Кретов, А. Г. Сбоев***

**НИЦ Курчатовский институт*
***Воронежский государственный университет*

Поступила в редакцию 28.12.2016 г.

Аннотация. В статье описаны два подхода к автоматическому извлечению ситуаций зрительного восприятия из текста для расчёта индекса приятия мира. Описаны реализации этих подходов и результаты тестирования. Показано преимущество подхода с использованием обученной CRF модели и синтаксических признаков предложения.
Ключевые слова: машинное обучение, лингвистические правила, индекс приятия мира, анализ текста.
Annotation. The presented article describes two approaches to automatic retrieval of visual perception situations for calculation of the acceptance index of world. We describe the implementation of these approaches, and test results. The approach on the basis of conditional random fields model and syntactic features of sentence demonstrates the best performance.
Keywords: machine learning, rule based systems, natural language processing, conditional random fields (CRF).

ВВЕДЕНИЕ

В настоящее время получили широкое развитие методы определения тональности текстов различной длины, а также выделения объектов тональности. Методы анализа эмоционального состояния автора текста проработаны не в полной мере, так в ряде работ [1, 8], предложены признаки и автоматизированные средства для их выделения.

В работе 2012 года предложен подход к анализу текстов и расчёту индекса приятия мира (ИнПриМ). ИнПриМ позволяет измерять и сравнивать в сопоставимых величинах удовлетворённость авторов текстов окружающей действительностью. Отрицательные значения ИнПриМа свидетельствуют о преобладании своего рода недоверия к окружающему миру, неудовлетворённости действительностью [7, 9].

Так как работ по автоматизации расчёта индекса не проводилось, целью работы стало создание автоматизированных средств вычисления ИнПриМа.

ВХОДНЫЕ ДАННЫЕ

ИнПриМ рассчитывается на основе выделенных в тексте ситуаций зрительного восприятия (ЗВ). Формула расчёта следующая: ИнПриМ = $ФН_{вид}$ – $ФН_{смотр}$, где $ФН_{вид}$ – доля ситуаций видения в общем числе ситуаций зрения, а $ФН_{смотр}$ – доля ситуаций смотрения.

Так в тексте Аксёнова Василия Павловича «Остров Крым» есть 393 ситуаций зрения, из них 200 смотрения, 124 видения, 69 созерцания, ИнПриМ = 0,316 – 0,509 = –0,193. В тексте Бестужева Николая Александровича, «Русский в Париже» 1814 года: 147 ситуаций смотрения, 129 ситуаций видения, 65 ситуаций созерцания, ИнПриМ = 0,378 – 0,431 = –0,053. В тексте Глинки Фёдора Николаевича, «Письма русского офицера»: смотрения 49, видения – 108, созерцания – 49, ИнПриМ = 0,524 – 0,238 = 0,286.

Лингвистами были выделены 6262 предложения содержащего ситуации зрения (3608

смотрения/ 2154 видения/ 1255 созерцания). Предложения, содержащие ситуации зрения, взяты из художественной литературы. Для каждого предложения указана ситуация зрения и глагол, определяющий эту ситуацию. Примеры предложений:

• «И чем больше глядел он на егеря, тем крепче уверялся, что скорей надо избавлять его от возможной пагубы.» (ГЛЯДЕЛ – смотрение).

• «Я только два раза в жизни видел бедную девушку, которую вы преследуете, и решился защищать ее из одного сострадания.» (ВИДЕЛ – видение).

• «посмотрим памятник погибшим и порт.» (ПОСМОТРИМ – созерцание).

МАТЕРИАЛЫ И МЕТОДЫ

Основной задачей при разработке средств автоматизированного расчёта ИнПриМа является автоматизированное выделение ситуаций зрения и видения. Выделяют 2 подхода к решению подобного рода задач:

• с использованием шаблонов (правил);

• с использованием методов машинного обучения.

Первый вариант предполагает создание экспертами базы правил, позволяющих определить наличие нужных ситуаций в тексте. Такой подход предполагает наличие уже составленных лингвистами правил. Второй вариант требует достаточной коллекции размеченных примеров ситуаций созерцания и видения. В обоих случаях решается задача классификации слов по типам: 1 – если это слово глагол ситуации смотрения, 2 – видения, 3 – созерцания и 0 для всех остальных слов.

ПОДХОД НА ОСНОВЕ ПРАВИЛ

Был автоматизирован поиск правил для определения ситуаций смотрения (657 правил), видения (136 правил), созерцания (268 правил). Примеры правил:

• «*бегать* глазами поД. / сР.- наВ.», тип глагола: «СМОТРЕНИЕ»,

• «*рассмотреть*, кто», тип глагола: «ВИДЕНИЕ»;

• «*вести* наблюдение», тип глагола: «СОЗЕРЦАНИЕ».

В правилах всегда есть глагол, принадлежность которого к определенной группе зрительного восприятия определяется наличием связанных с ним слов. Также в правилах могут быть указаны: форма зависимого слова (предлог + падеж); подчинительные союзы и дополнительные слова в подчинённом предложении; конкретные зависимые слова (*взор*, *глаза*); слова, связанные предикативной связью с глаголом; тип объекта зрения слова (*зрелище*, *движение*).

Для нахождения любого из правил необходимо, чтобы в предложении нашлись все его составляющие. Обработка текста происходит следующим образом:

• Текст обрабатывается морфо-синтаксическим сервисом [6] для получения морфологических и синтаксических признаков слов.

• Из предложений выбираются глаголы, для каждого из них выбирается группа правил, содержащих выбранные глаголы.

• Проверяется наличие ситуации, описываемой каждым правилом.

• Если было найдено несколько правил, из них выбирается обладающее самой длинной цепью узлов. Иначе, если не было найдено подходящих правил, ситуация ЗВ не найдена.

• После обработки всех предложений рассчитывается ИнПриМ на основе количества найденных ситуаций зрения.

ПОДХОД НА ОСНОВЕ МАШИННОГО ОБУЧЕНИЯ

В данном подходе анализируется каждое слово предложения и определяется его класс (тип ситуации ЗВ) с использованием алгоритмов машинного обучения, решающих задачу классификации. При этом необходимо выбрать набор параметров описания каждого слова предложения. В рамках этого подхода анализируется возможность использования синтаксических конструкций предложения. Поэтому все размеченные предложения были разобраны с использованием морфо-синтаксического сервиса [6] для установления синтаксической структуры целого предложения.

На основе полученной структуры был создан набор признаков описания каждого слова, который включает: порядковый номер слова (id), номер главного слова (dom), тип входной синтаксической связи (link), длину слова (len), форму слова (forma), набор морфологических признаков (grm), его часть речи (pos), указание, является ли слово началом/концом предложения (isFirstWord, isLastWord), состоит ли слово целиком из букв или цифр (isalnnum), только из букв (isalpha), только из цифр (isdigit), идентификатор заглавной буквы в начале слова (istitle), все строчные буквы слова (islower) или заглавные (isupper), N букв в начале слова и конце слова (prefix_N, postfix_N, при N = 1,...,4). Для описания слова также использовался аналогичный набор признаков для:

- его соседей в окне +–3 слова от рассматриваемого (+1, +2, +3, –1, –2, –3);
- главного слова для него по синтаксической структуре, и главного для главного (+1_dom, +2_dom);
- самого левого и правого зависимых для слова по синтаксической структуре (Lchild, Rchild).

В итоге для описания каждого слова используется 575 признаков. В ходе работы проводится оценка релевантности признаков методами information gain, gini index [5].

В качестве классификаторов используются: метод условных случайных полей (CRF) [4], деревья gradient boosting (GBT) [3], метод случайного леса (RFC) [2].

КРИТЕРИИ ТОЧНОСТИ

Для оценки точности обоих методов использовались следующие критерии:

Precision или мера точности – характеризует, сколько полученных от классификатора положительных ответов являются правильными. Чем больше точность, тем меньше число ложных попаданий. Рассчитывается как $P = \frac{tp}{tp + fp}$, где tp (true positive) – количество объектов, которые объект правильно отнёс к данному классу, fp (false positive) – количество объектов, которые классификатор отнёс к данному классу, но они таковыми не являются.

Recall или мера полноты – характеризует способность классификатора «угадывать» как можно большее число положительных ответов из ожидаемых. Рассчитывается как $R = \frac{tp}{tp + fn}$, где fn (false negative) – количество объектов, которые относятся к данному классу, но классификатор отнёс их к другому.

F1_score это среднее гармоническое величин Precision и Recall, считается как $F1 = \frac{2(P * R)}{P + R} = \frac{2 * tp}{2tp + fp + fn}$.

Все оценки также посчитаны с тремя вариантами усреднения micro, macro, weighted.

При micro усреднении для каждого класса считаются tp_i, fp_i, tn_i, fn_i, затем считаются суммарные tp_s, fp_s, tn_s, fn_s, по которым и рассчитываются precision, recall и f1. Таким образом: $F1_{micro} = \frac{2\sum tp_i}{2\sum tp_i + \sum fp_i + \sum fn_i}$, где i – индекс класса.

При macro усреднении для каждого класса считаются precision, recall и f1. Затем для них вычисляются средние арифметические. Таким образом: $F1_{micro} = \frac{1}{n} * \sum \frac{2tp_i}{2tp_i + fp_i + fn_i}$, где i – индекс класса, n – количество классов.

При weighted усреднении для каждого класса считаются precision, recall и f1. Затем для них вычисляются средне арифметические с учётом представительности каждого класса.

ЭКСПЕРИМЕНТЫ

Результаты тестирования метода основанного на правилах (Rules) показаны в табл. 1. Задачей системы было найти глаголы ЗВ и провести их классификации во всех предложениях размеченного корпуса.

Для апробации подхода на основе машинного обучении выборка предложений была разделена на 2 части – обучающая (70 %) и тестовая (30 %).

Далее было проведено исследование по выбору признаков описания слов, т. к. сформированный набор признаков изначально был составлен избыточно. С использованием заранее выбранных алгоритмов оценки релевантности были рассчитаны и отранжирова-

Таблица 1

Оценки по классам для подхода, основанного на правилах

	precision	F1	f1-score	Кол-во слов
Без типа	0,99	1	0,99	39053
Смотрение	0,90	0,71	0,79	1203
Видение	0,81	0,77	0,79	719
Созерцание	0,66	0,56	0,61	412
Средние значения:				
Macro	0,84	0,76	0,80	41387
Micro	0,98	0,98	0,98	41387
Weighted	0,98	0,98	0,98	41387

Таблица 2

Точности классификаторов в зависимости от алгоритма выборов признаков и порога

Классификатор	Алгоритм оценки признаков	Порог	Кол-во признаков	F1 Macro	F1 Micro	Precision	Recall
CRF	information gain	0,012	17	0,89	0,99	0,99	0,99
CRF	information gain	0,007	20	0,89	0,99	0,99	0,99
GBT	information gain	0,012	17	0,87	0,99	0,99	0,99
RFC	information gain	0,036	6	0,87	0,98	0,98	0,99
GBT	gini index	0,107	159	0,86	0,99	0,99	0,99
CRF	gini index	0,107	163	0,86	0,98	0,98	0,98
RFC	gini index	0,107	159	0,62	0,97	0,97	0,97

Таблица 3

Оценки по классам для CRF модели

	precision	recall	F1	Кол-во слов
Без типа	0,99	1.00	1.00	39053
Смотрение	0,92	0,88	0,90	1203
Видение	0,91	0,92	0,91	719
Созерцание	0,84	0,66	0,74	412
Средние значения				
Macro	0,92	0,87	0,89	41387
Micro	0,98	0,99	0,99	41387
Weighted	0,99	0,99	0,99	41387

ны значения информативности для каждого признака из сформированного набора. Далее изменяется порог и выбирается группа признаков, релевантность которых больше установленного порога. Выбранные признаки используются для описания множества эталонных примеров для обучения и тестирования классификатора. В табл. 2 представлены точности классификаторов при использовании той или иной комбинации алгоритмов классификации и оценки релевантности признаков.

Лучшие результаты достигаются при использовании алгоритмов Information gain и классификатора CRF. Точности классификации по отдельным классам этой моделью представлены в табл. 3. Точности не улучшаются при добавлении признаков. Минималь-

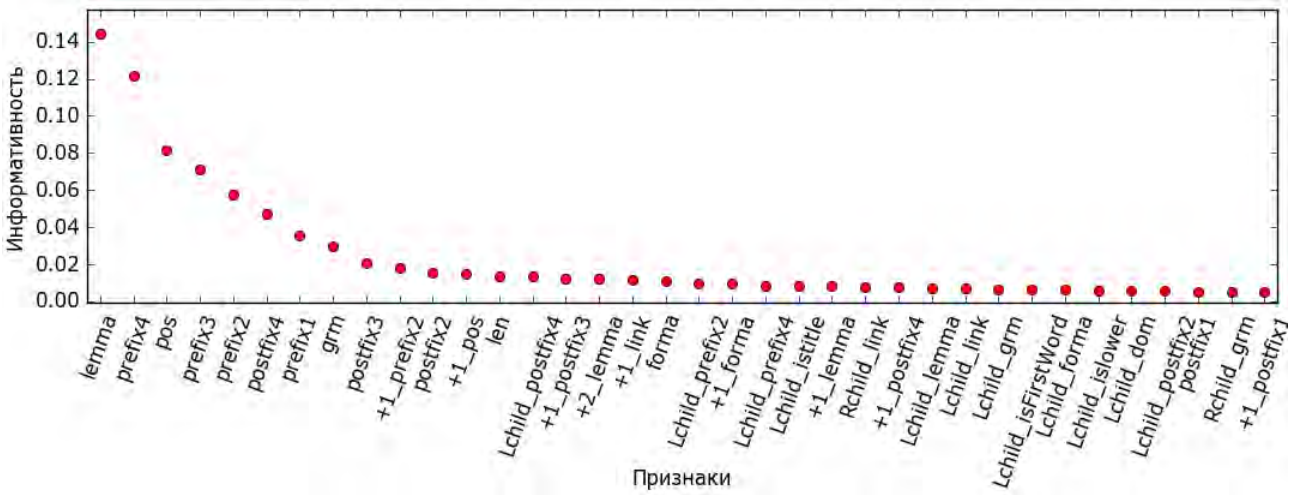

Рис. 1. Информативности признаков по information gain (больше – информативней)

ный набор параметров (см. рис. 1), который в комплексе с CRF классификатором продемонстрировал наилучшие точности содержит: лемму слова, последние буквы слова, его часть речи, набор морфологических признаков, часть речи следующего слова в предложении и тип его входной синтаксической связи, последние буквы левого зависимого для слова по синтаксической структуре.

Далее было проведено сравнение по точности определения ИнПриМа подходов на основе правил (Rules), на основе машинного обучения (CRF) с эталонным разбором, сделанным экспертами. Для этого использовался метод кросс-валидационной проверки, когда исходное множество размеченных предложений разбивалось 100 раз на 2 непересекающиеся подвыборки: обучающую (70 % от исходного количества предложений) и тестировочную (30 %). Сравнение проводилось по средним значениям и среднеквадратичным отклонениям от среднего при определении ИнПриМа для тестовых подвыборок. Для модели на основе машинного обучения 100 раз проводилось обучение и тестирование классификатора CRF. Результаты представлены в табл. 4.

Таблица 4

Расчет ИнПриМа

Метод	Среднее значение ИнПриМа (± ср.кв.откл.)
Manual	–20,7 (± 1,5)
Rules	–14,5 (± 1,8)
CRF	–19,5 (± 1,6)

ЗАКЛЮЧЕНИЕ

В данной работе впервые разработана система автоматического расчета индекса приятия мира, позволяющая определять дополнительную характеристику автора письменного текста. В результате проведенного исследования показано, что метод классификации на основе случайных полей с использованием набора морфо-синтаксических признаков слов и предложений позволяет достичь лучшей точности определения отдельных значений слов, лежащих в основе расчета искомого индекса, а именно – ситуаций зрительного восприятия.

Работа выполнена при поддержке гранта 16-37-50069 мол_нр

СПИСОК ЛИТЕРАТУРЫ

1. *Sboev A., Gudovskikh D., Rybka R., Moloshnikov I.* A Quantitative Method of Text Emotiveness Evaluation on Base of the Psycholinguistic Markers Founded on Morphological Features // Procedia Computer Science. – 2015. – V. 66.

2. *Breiman Leo* RANDOM FORESTS // Machine Learning. – Kluwer Academic Publishers, 2001. – V. 45.

3. *Friedman J. H.* Greedy Function Approximation: A Gradient Boosting // The Annals of Statistics. – 2001. – V. 29.

4. *Lafferty John, McCallum Andrew, Pereira Fernando C.* N Conditional Random Fields: Probabilistic Models for Segmenting and La-

beling Sequence Data // 01 Proceedings of the Eighteenth International Conference on Machine Learning. – Morgan Kaufmann Publishers Inc., June 2001.

5. *Li Jundong [et al.]* Features Selection: A Data Perspective // Cornell University Library. – Jan 29, 2016. – v4. – https://arxiv.org/abs/1601.07996.

6. *Rybka Roman [et al.]* Morpho-Syntactic Parsing Based on Neural Networks and Corpus Data // Artificial Intelligence and Natural Language and Information Extraction, Social Media and Web Search FRUCT Conference (AINL-ISMW FRUCT). – 2015.

7. *Кретов А. А.* Лингвистическая прогностика: побочные результаты за XX лет // Проблемы лингвистической прогностики: Сборник научных трудов / под ред. А. А. Кретова. – Вып. 5. – Воронеж, 2012. – С. 50–58.

8. *Литвинова Т. А. Литвинова О. А., Середин П. В.* Частоты встречаемости последовательностей частей речи в тексте и психофизиологические характеристики его автора: корпусное исследование // Вестник Иркут. гос. лингв. ун-та. – 2014. – № 2.

9. *Силина Ю. А.* Динамика и прогностика номинаций зрительного восприятия во французских нарративных текстах : монография / Ю. А. Силина, А. А. Кретов ; Воронежский государственный университет ; [под ред. А. А. Кретова]. – Воронеж : Издательско-полиграфический центр Воронежского государственного университета, 2011. – 187 с. – (Серия : Библиотека лингвистической прогностики. Том 6).

Рыбка Роман Борисович – канд. техн. наук, инженер-исследователь, НИЦ «Курчатовский институт».
Тел.: 8-926-344-6135
E-mail: rybkarb@gmail.com

Rybka Roman Borisovich – research engineer, National Research Center «Kurchatov Institute».
Tel.: 8-926-344-6135
E-mail: rybkarb@gmail.com

Грязнов Артём Викторович – лаборант-исследователь, НИЦ «Курчатовский институт».
E-mail: artem.official@mail.ru

Gryaznov Artem Victorovich – assistant researcher, National Research Center «Kurchatov Institute».
E-mail: artem.official@mail.ru

Кретов Алексей Александрович – д-р филол. наук, профессор, заведующий кафедрой теоретической и прикладной лингвистики факультета романо-германской филологии, Воронежский государственный университет.
E-mail: tipl@rgph.vsu.ru

Kretov Alexey Alexandrovich – head of the department, professor, Voronezh State University.
E-mail: tipl@rgph.vsu.ru

Сбоев Александр Георгиевич – канд. физ.-мат. наук, ведущий научный сотрудник, НИЦ «Курчатовский институт».
Тел.: 8-926-253-7217
E-mail: sag111@mail.ru

Sboev Alexandr Georgievich – leading researcher, National Research Center «Kurchatov Institute».
Tel.: 8-926-253-7217
E-mail: sag111@mail.ru