10. *Wu, Z.* Verb semantics and lexical selection / Z. Wu, M. Palmer // In Proceedings of the 32nd Annual meeting of the Associations for Computational Linguistics, Las Cruces, New Mexico. – 1994. – P. 133–138.
11. *Davide, B.* IRIT: Textual Similarity Combining Conceptual Similarity with an N-Gram Comparison Method / B. Davide, T. Ronan, A. Nathalie, M. Josiane // First Joint Conference on Lexical and Computational Semantics (*SEM), Montreal, Canada, June 7–8. – Association for Computational Linguistics, Stroudsburg, PA, USA. – 2012. – P. 552–556.
12. *Bar, D.* UKP: Computing Semantic Textual Similarity by Combining Multiple Content Similarity Measures / D. Bar, C. Biemann, I. Gurevych, T. Zesch // First Joint Conference on Lexical and Computational Semantics (*SEM), Montreal, Canada, June 7–8. – Association for Computational Linguistics, Stroudsburg, PA, USA. – 2012. – P. 435–440.
13. *Li, Y.* Sentence similarity based on semantic nets and corpus statistics / Y. Li, D. McLean, Z. Bandar, J. O'Shea, K. Crockett // IEEE Transactions on Knowledge and Data Engineering. – Vol. 18, No. 8. – IEEE Educational Activities Department Piscataway, NJ, USA. – 2006. – P. 1138–1149.
14. *Islam, A.* Semantic text similarity using corpus-based word similarity and string similarity / A. Islam, D. Inkpen // ACM Transactions on Knowledge Discovery from Data. – Vol. 2, No. 2. – ACM, New York, NY, USA. – 2008. – P. 1–25.
15. *Совпель, И. В.* Система автоматического извлечения знаний из текста и ее приложения / И. В. Совпель // Искусственный интеллект. ІПШІ МОН і НАН Украіни «Наука і освіта». – 2004. – № 3. – С. 668–677.
16. *Постаногов, Д. Ю.* Автоматическая обработка естественного языка в задаче инженерии знаний и доступа к ним: дис. ... к-та техн. наук: 05.13.17 / Д. Ю. Постаногов. – Минск, 2012. – 136 л.

This article explores some aspects of automatic recognition of semantically relevant texts. We propose a method to detect the semantic relevance of textual documents using their sematic indices, which are built as a result of deep analysis of semantics taking into account synonymous and hierarchical relationships. The paper also describes linguistic resources needed for the problem solution.

Поступила в редакцию 28.04.18

М. В. Чернышевич

ПРИНЦИПИАЛЬНАЯ СХЕМА РЕШЕНИЯ ЗАДАЧИ АСАТ И ЕГО ЛИНГВИСТИЧЕСКОЕ ОБЕСПЕЧЕНИЕ

В данной работе предложена принципиальная схема системы автоматического сентимент-анализа текстов на английском языке, определены необходимые ее компоненты и реализовано решение трех основных подзадач системы АСАТ. Подробно описано лингвистическое обеспечение, разработанное в рамках данной работы, включающее аннотированные корпуса, разнообразные лексические ресурсы и лингвистические паттерны.

Учитывая результаты проведенного анализа существующих подходов к решению задачи автоматического сентимент-анализа текста (АСАТ) и конкретных систем рассматриваемого типа, а также требования, предъявляемые к ее современному решению [1; 2], вполне обоснованно можно считать как наиболее приемлемую следующую его принципиальную схему (рисунок).

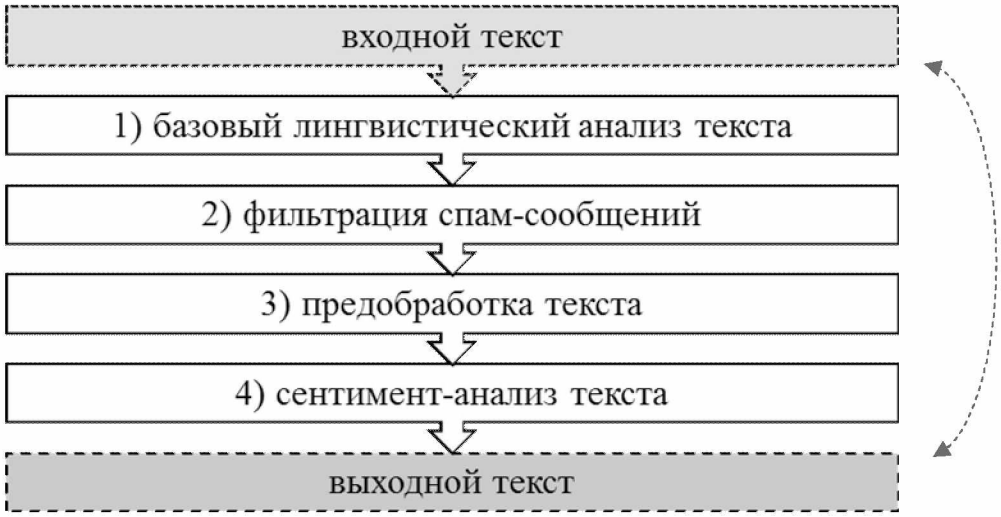

Рисунок. Принципиальная схема решения задачи АСАТ

Базовый лингвистический анализ (1) текста является необходимой, уже достаточно хорошо проработанной составной частью большинства задач его автоматической обработки и обычно реализуется так называемым базовым лингвистическим процессором (БЛП) [3]. В нашем случае в качестве такового используется известный многоязычный БЛП IHS Goldfire, который выполняет следующие типы анализа текста: лексический (графемный), лексико-грамматический, синтаксический и семантический анализ. Результаты всех этих этапов обработки входного текста образуют в совокупности его лингвистический индекс, необходимый для эффективного решения подзадач 2–4 (рисунок) целевой задачи АСАТ, которая заключается в распознавании в тексте предложений, содержащих мнения, а также всех их компонентов [2].

Описанная выше функциональность БЛП обеспечивается соответствующей лингвистической базой знаний (ЛБЗ), которая является неотъемлемым элементом современных систем автоматической обработки ЕЯ и содержит как декларативные (словари, классификаторы, статистические данные), так и процедурные (это прежде всего лингвистические правила, или иначе, паттерны) знания, описывающие состав и механизмы функционирования ЕЯ.

Решение отдельных подзадач 2–4 в рамках комплексной задачи АСАТ требует разработки дополнительных компонентов ЛБЗ, которые включают текстовые ресурсы (коллекции текстов и аннотированные корпуса), лексические ресурсы (различного рода словари и списки) и лингвистические паттерны.

Общим для всех подзадач ресурсом является коллекция текстов – неаннотированный текстовый материал из различных источников, который стал основой для данного исследования (далее исследуемый текстовый материал). Отбор материала осуществлялся исходя из поставленной задачи, предполагаемого приложения получаемых результатов и известных принципов построения репрезентативных корпусов текстов [4]. Во-первых, речь идет о текстах на английском языке. Во-вторых, система АСАТ должна войти в состав уже существующей индустриальной многофункциональной системы автоматической обработки как технической, так и статей и сообщений, публикуемых в СМИ, а также в таких актуальных на сегодняшний день источниках текстовых документов, как социальные сети. Такое разнообразие потенциальных источников и самого содержания текстовых данных обусловило создание универсальной коллекции текстов, которая включила следующий текстовый материал:

1) корпус сообщений пользователей (как показал анализ, именно эти текстовые документы содержат наибольшее количество мнений);

2) корпус новостных статей (текстовые документы этого типа содержат относительно небольшое количество мнений);

3) корпус технических текстов (такие тексты практически не содержат мнений).

Корпус сообщений пользователей содержит сообщения пользователей из трех источников: социальная сеть Twitter.com (далее корпус Твиттер), социальная сеть Facebook.com (далее корпус Фейсбук), а также интернет-платформ Fixya.com, Amazon.com и Tripadvisor.com (далее корпус форумов).

Корпусы новостных статей, собранных на новостном портале reaters.com, и технических текстов, содержащих патенты патентного фонда США и научного журнала *IEEE*, являются дополнительным материалом, необходимым для обеспечения высокой точности распознавания мнений. Ведь, как показал проведенный анализ, многие реализованные системы АСАТ не ориентированы на обработку технических текстов, содержащих в основном фактическую информацию, и ошибочно извлекают при их обработке большое количество несуществующих мнений [1].

Все тексты были проанализированы, а имеющиеся в них мнения аннотированы в полуавтоматическом режиме. Количественное распределение представлено в таблице.

Таблица

Количественное распределение предложений и мнений по корпусам

Корпус	Предложений	Мнений
Корпус сообщений пользователей	21740	8612
Корпус Твиттер	7580	2350
Корпус Фейсбук	6960	2287
Корпус форумов	8200	3975
Корпус новостных статей	3640	25
Корпус технических текстов	3238	0
Итого	29618	8637

Предварительный анализ данного текстового материала, особенно корпуса сообщений пользователей, позволяет заключить, что в нем есть ряд существенных особенностей, и их необходимо учитывать при разработке системы АСАТ. В первую очередь, это – спам-сообщения и сленговые слова и выражения. Данные особенности обуславливают наличие в разрабатываемой системе модулей фильтрации спам-сообщений (подзадача 2) и нормализации лексических единиц (подзадача 3), которые предшествуют модулю собственно сентимент-анализа текста (подзадача 4).

Задача идентификации и фильтрации спама в нашем случае решается как задача классификации с применением методов машинного обучения с учителем, а это требует наличия обучающей выборки (аннотированного корпуса текстов), представляющей собой совокупность описаний прецедентов (ситуаций, объектов и т.п.) с использованием зафиксированных показателей (признаков), измеряемых у всех прецедентов. По этим частным данным алгоритм автоматически, как правило, с приемлемой степенью точности, выявляет общие закономерности и взаимосвязи, присущие не только этой конкретной выборке, но вообще всем прецедентам, в том числе тем, которые еще не наблюдались. Аннотированный корпус при задаче идентификации спама состоит из сообщений пользователей, аннотированных метками класса «спам» и «не спам». Учитывая трудоемкость этого процесса, его желательно автоматизировать. С этой целью из исследуемого текстового материала были выбраны сообщения пользователей, принадлежащих классу «спам», по следующим критериям: повторяемость в выборке более 20 раз (полные или частичные дубликаты); отношение «мусорных» (не закрепленных в словаре ЕЯ и списков сленговых выражений) слов сообщения к общему количеству его слов – более чем 0,3; наличие устойчивых конструкций, указывающих на рекламный характер сообщений, например, *available online* ‘доступен онлайн’, *visit our website* ‘зайдите на наш сайт’ (всего разработано 26 паттернов, ориентированных на данный критерий). Отметим, что сообщение помещалось автоматически в формируемую выборку, если оно удовлетворяло хотя бы одному из представленных критериев.

Для формирования выборки сообщений, относящихся к классу «не спам», из множества сообщений пользователей, не попавших в первую выборку, было взято произвольным образом 8000 уникальных сообщений, т.е. повторяемость равна 1. Кроме того, из оставшегося множества вручную было размечено еще 10000 сообщений пользователей, одна часть которых попала в первую выборку, а другая – во вторую. Таким образом, было выбрано всего 25583 сообщений (10214 для класса «спам» и 15369 для класса «не спам»).

В качестве классифицирующего алгоритма в данной работе выбран метод опорных векторов (SVM [5]), так как он показал наилучшее качество и наибольшую скорость обработки выборки. Для обучения классификатора было исследовано большое количество признаков, предложенных в различ-

ных работах, например, [6; 7; 8]. В конечном счете, сформированное пространство признаков включило в себя такие лексические, стилистические, синтаксические и статистические признаки, как бинарные признаки наличия или отсутствия униграмм и биграмм в сообщении; количество хэш-тэгов в сообщении; число ссылок и смайлов в сообщении; количество слов; число отношений «Субъект-Акция-Объект»; количество личных местоимений в сообщении и другие.

Тестирование разработанного модуля фильтрации спам-сообщений показало следующие оценки эффективности его работы на контрольной выборке: точность – 98,76 %, полнота – 97,68 %.

Анализ результатов обработки данным модулем пользовательских сообщений показал, что спамом являются около 35 % сообщений из социальной сети Twitter (714 из 2059), около 20 % сообщений из сети Facebook (382 из 1986) и около 2 % – из форумов Fixya и Amazon (45 из 2533).

Модуль предварительной обработки производит обработку сообщений с целью их адаптации для решения задачи сентимент-анализа и сводится к лексической нормализации слов, не соответствующих нормам ЕЯ, а также удалению слов, выражений и предложений, которые не являются необходимыми для решения целевой задачи. Подробно его описание дано в работе [9], включая аннотированный корпус текстов из 2954 сообщений для обучения классификатора и другие лингвистические ресурсы.

Решение задачи собственно сентимент-анализа текста также основывается на методах машинного обучения с учителем. С учетом достоинств и недостатков, присущих как этим методам, так и методам, основанным на лингвистических паттернах, в данном случае упор сделан на комбинировании указанных методов, а именно дополнение метода машинного обучения с учителем процедурами оперирования, при необходимости, синтаксическими конструкциями и результатами более глубокого лингвистического анализа текста. Все это все позволяет максимально обобщить признаковое описание прецедентов и тем самым обеспечить высокое качество решения задачи ACAT. Требуемый классификатор здесь был построен на основании иерархической нейронной сети [10] с механизмом внимания на словах [11]. При этом определение того, является ли именная группа (объект-кандидат) объектом мнения, и если да, то установление тональности мнения происходят одновременно.

В качестве аннотированного корпуса взят корпус, данные о нем приведены в таблице, и во всех предложениях которого мнения, объекты и их тональность, при наличии в предложении, аннотированы с помощью xml-разметки, например:

Personally <opinion> I like <object value="positive">the margherita pizza </object> </opinion>, but they are all good.

Минимальный семантически значимый тонально окрашенный фрагмент текста заключен в теги <opinion> (начальный тег) и </opinion> (конечный тег), объект мнения – в теги <object> и </object>, значение тональности указано как параметр value тега <object>.

Признаковое описание прецедентов здесь полагается на возможности обработки текста с помощью БЛП, например, с целью определения лексико-грамматической категории слова и его семантической роли в САО-отношении [4; 12], и на развитую ЛБЗ, включающую большое количество лексических ресурсов (словарей) и лингвистических паттернов.

Поскольку тональность в тексте выражается, в большей степени, лексическими средствами, то необходимыми лингвистическими ресурсами при решении рассматриваемой задачи являются словари прилагательных, наречий, глаголов и существительных с положительной и отрицательной тональностью. Оценочные слова в тексте могут сопровождаться словами-модификаторами, к которым относятся так называемые эмотивные интенсификаторы, т.е. слова и выражения, которые усиливают семантику (в нашем случае тональность) других слов, например, *so much* ‘так сильно’, *very* ‘очень’, нейтрализаторы, ослабляющие их семантику, например, *barely* ‘едва’, и шифтеры, слова, которые изменяют тональность слов и конструкций на противоположную, например, *not* ‘не’, *I don't think* ‘я не думаю’.

Большое количество лексических единиц ЕЯ не могут быть однозначно отнесены к положительно или отрицательно окрашенным словам, однако для каждого слова можно вычислить тональную ориентацию, показывающую в каком контексте, позитивном или негативном, слово встречается чаще [13]. Тональная ориентация может быть определена на основании корпуса текстов. При решении задачи была произведена выборка сообщений пользователей корпуса Твиттер, в которых автор использовал позитивные и негативные эмотиконы (смайлы), и сообщений (отзывов), из корпуса форумов, в которых автор указал общий рейтинг товара, при этом использовались только отзывы с рейтингом 1 и 2 (отрицательные отзывы) и 5 (положительные отзывы), общим объемом около 300 млн слов. Следует отметить, что ориентация целого сообщения далеко не всегда соответствует тональной ориентации смайла или общего рейтинга, однако в целом такие данные можно использовать вполне эффективно для статистической оценки.

Для количественной оценки тональной ориентации слов применялась формула поточечной взаимной информации PMI [14]. Итоговый объем данного лингвистического ресурса составил 89380 слов со значением от $+\infty$ (позитивная тональность) до $-\infty$ (негативная тональность). Чем больше абсолютное значение оценки, тем сильнее степень тональности. Например, *magnificent* (+6,672), *easygoing* (+4,128), *excited* (+4,098), *overpowered* (–1.064), *malfunction* (–1,992), *cameraless* (–3,343), *devasted* (–5.125).

В состав лингвистических ресурсов в рассматриваемом случае включены также паттерны, описывающие сложные тонально окрашенные конструкции, например, *can't live without* ‘не могу жить без’, *can't stop thinking about* ‘не могу перестать думать о’, для формального описания которых применяется известный язык WRE [15]. С его использованием классы паттернов, к которым принадлежат конкретные паттерны из приведенных выше примеров, будут представлены в следующем виде:

1) “can” XNOT [RB] “live” “without”
2) “can” XNOT [RB] “stop” (“thinking” | “dreaming”) “about”

Здесь XNOT – лексико-грамматический тег (ЛГК), означающий отрицание (*not* или *n't*), RB – «любое наречие», квадратные скобки содержат опциональную часть, а круглые скобки обозначают объединение.

Еще одним лингвистическим ресурсом, используемым для решения задачи ACAT на данном этапе, является представленное на языке WRE описание именных групп, которые не могут выступать в роли объекта мнения, например, *no problems, many times.*

В целом, наполнение лингвистической базы знаний системы ACAT базировалось на применении электронной лексической базы данных английского языка WordNet [16], содержащей более 250 тыс. лексических единиц, на общедоступных онлайн-ресурсах [17; 18; 19; 20; 21; 22], коллекции текстов для сбора статистических данных, а также на исследуемом материале.

Указанные ресурсы, а также их производные в совокупности с БЗ БЛП составляют лингвистическое обеспечение задачи ACAT, благодаря которому нам удалось достичь высоких показателей полноты (81,80 %) и точности (86,09 %) работы системы ACAT, ориентированной на индустриальные системы обработки текстовой информации.

ЛИТЕРАТУРА

1. *Чернышевич, М. В.* Обзор существующих систем автоматического сентимент-анализа текста / М. В. Чернышевич // Вестн. МГЛУ. Сер. 1, Филология. – 2017. – № 6 (91). – P. 111–117.
2. *Чернышевич, М. В.* Актуальные аспекты решения задачи автоматического сентимент-анализа текста / М. В. Чернышевич // Вестн. МГЛУ. Сер. 1, Филология. – 2018. – № 1 (92). – P. 100–106.
3. *Чеусов, А. В.* Разработка алгоритмов и технологии построения многоязычного базового лингвистического процессора: дис. ... к-та техн. наук: 05.13.17 / А. В. Чеусов. – Минск, 2013. – 116 л.
4. *Совпель, И. В.* Инженерно-лингвистические принципы, методы и алгоритмы автоматической переработки текста / И. В. Совпель. – Минск : Выш. шк., 1991. – 120 л.
5. *Vapnik, V.* Statistical Learning Theory / V. Vapnik. – Chichester, GB: Wiley, 1998.
6. Making the most of tweet-inherent features for social spam detection on Twitter. / B. Wang [et al.] // he 5th Workshop on Making Sense of Microposts. – 2015. – Vol. 1395. – P. 10–16.
7. The DARPA Twitter bot challenge / V. S. Subrahmanian [et al.] // Computer. – 2016. – Vol. 49, № 6. – P. 38–46.
8. Twitter: who gets caught? observed trends in social micro-blogging spam / A. Almaatouq [et al.] // the 2014 ACM conference on Web science. – 2014. – P. 33–41.

9. *Чернышевич, М. В.* Автоматическая нормализация англоязычных сообщений пользователей социальных сетей для задачи их сентимент-анализа. / М. В. Чернышевич // Вестн. МГЛУ. Сер. 1, Филология. – 2017. – № 5 (90). – P. 66–73.
10. Hierarchical Attention Networks for Document Classification / Z. Yang [et al.] // Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. – San Diego, CA, USA, 2016. – P. 1480–1489.
11. *Bahdanau, D.* Neural machine translation by jointly learning to align and translate / D. Bahdanau, K. Cho, Y. Bengio // CoRR. – 2016.
12. *Todhunter, J.* System and method for automatic semantic labeling of natural language texts. US Patent Appl. 20100235165 / J. Todhunter, I. Sovpel, D. Pastanohau.
13. *Hatzivassiloglou, V.* Predicting the semantic orientation of adjectives. / V. Hatzivassiloglou, K. R. McKeown // ACL-1997 : Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics. – Madrid, Spain, 1997. – P. 174–181.
14. *Church, K.* Word association norms, mutual information and lexicography / K. Church, P. Hanks // ACL : Proceedings of the 27th Annual Conference of the ACL. – New Brunswick, NJ, USA, 1989. – P. 76–83.
15. *Cheusov, A.* The word-based regular expressions in computational linguistics / A. Cheusov // Pattrn Recognition and Information Processing (PRIP-2003): Proc. of 7th Intern. Conf. : Proceedings of the. – Minsk, 2003. – P. 208–212.
16. WordNet [Electronic resource]. – Mode of access : http://wordnet.princeton.edu/. – Date of access : 08.01.2017.
17. *Stone, D.* Harvard General Inquirer lexicon / D. Stone [Electronic resource]. – Mode of access : http://www.wjh.harvard.edu/~inquirer/. – Date of access : 06.05.2017.
18. *Strapparava, C.* WordNet-Affect: An affective extension of WordNet / C. Strapparava, A. Valitutti // LREC-2004 : Proceedings of the Fourth International Conference on Language Resources and Evaluation. – Lisbon, Portugal, 2004. – P. 1083–1086.
19. SenticNet: A Publicly Available Semantic Resource for Opinion Mining / E. Cambria [et al.] // AAAI Fall Symposium: Commonsense Knowledge. – Arlington, Virginia, USA, 2010. – P. 14–18.
20. *Mohammad, S. M.* NRC-Canada: Building the State-of-the-Art in Sentiment Analysis of Tweets / S. M. Mohammad, S. Kiritchenko, X. Zhu // Proceedings of the seventh international workshop on Semantic Evaluation Exercises (SemEval-2013). – Atlanta, USA, 2013.
21. *Esuli, A.* SentiWordNet: A publicly available lexical resource for opinion mining / A. Esuli, F. Sebastiani // LREC-2006 : Proceedings of the Fifth International Conference on Language Resources and Evaluation. – Genoa, Italy, 2006. – P. 417–422.

22. *Mohammad, S.* Emotions Evoked by Common Words and Phrases: Using Mechanical Turk to Create an Emotion Lexicon / S. Mohammad, P. Turney // NAACL HLT- 2010 : Proceedings of the Workshop on Computational Approaches to Analysis and Generation of Emotion in Text. – LA, California, USA, 2010. – P. 26–34.

This article describes the architecture of sentiment analysis system, its main components and essential linguistic resources. The proposed system is based on various machine learning methods and a rich set of features that came from a deep linguistic analysis of the text. The evaluation demonstrated great effectiveness of the proposed solution.

Поступила в редакцию 24.05.18