ИНФОРМАТИКА
ИНФОРМАЦИОННЫЕ СИСТЕМЫ

УДК 025.4.03

Методы автоматической классификации текстов по функциональным стилям

Л. М. Ермакова, М. А. Абашев, Р. В. Никитин, Р. И. Ушаков
Пермский государственный национальный исследовательский университет
Россия, 614990, Пермь, ул. Букирева, 15
liana87@mail.ru; +73422396298

Основной задачей информационного поиска является извлечение из массива данных неструктурированной документальной информации, релевантной информационной потребности пользователя. Однако зачастую пользователи хотят найти документы определенного функционального стиля, например научные или публицистические тексты. При этом возникает необходимость автоматической классификации документов в зависимости от функционального стиля. Статья нацелена на определение формализуемых признаков функциональных стилей, а также на обзор существующих методов автоматической классификации.

Ключевые слова: *классификация; категоризация; рубрикация; функциональный стиль; функциональная разновидность языка; функциональный тип речи; машинное обучение; информационный поиск.*

Введение

Информационный поиск нацелен на извлечение из массива данных неструктурированной документальной информации, релевантной информационной потребности пользователя. Зачастую пользователи хотят получить не просто любые документы, релевантные их запросу, но отфильтровать их по функциональному стилю. Примером такой информационной потребности может послужить поиск научных статей или новостных сюжетов. Кроме того, классификация стилей способствует повышению качества информационного поиска [1]. Специфика функциональных стилей может быть использована в задаче автоматического реферирования (например, при определении наиболее/наименее важных абзацев, предложений и т.д.) [2]. Знание функционального стиля и жанра способствует значительному повышению качества систем обработки естественного языка, в т.ч. разметка частей речи, синтаксический анализ, снятие омонимии [3].

Данная статья нацелена на определение формализуемых признаков функциональных стилей, а также на обзор существующих методов автоматической классификации.

1. Понятие функционального стиля. Виды функциональных стилей

Словарь лингвистических терминов Т.В. Жеребило дает следующее определение функционального стиля:

"*Функциональный стиль – исторически сложившаяся, общественно осознанная речевая разновидность, обладающая речевой системностью, специфическим характером, сложившимся в результате реализации особых принципов отбора и сочетания языковых средств, разновидность, соответствующая*

сфере общения и деятельности, соотносимой с определенной формой сознания: наукой, искусством, правом и т.п" [4].

Функциональный стиль также иногда называется функциональной разновидностью языка, функциональным типом речи или языковым жанром. Не следует путать функциональный стиль с литературным жанром, представляющим собой "исторически складывающийся и развивающийся тип литературного произведения (художественного, публицистического, научного и др.), напр., роман, монография, репортаж и т.д." [5].

В отечественной лингвистике принято выделять научный, публицистический, официально-деловой и религиозный стили. Кроме того, противоречивый статус имеют художественный и разговорный стили.

Научный стиль обслуживает сферу научного общения [5]. Научному стилю свойственна строго оформленная композиция текста. Научный стиль характеризуется номинализацией, абстрактной лексикой, терминологичностью, глаголами широкой семантики и десемантизированными глаголами, выступающими в роли связок. Преимущественно используются глаголы несовершенного вида в настоящем времени и существительные в единственном числе. Повышена частота среднего рода. Преобладают деагентивные синтаксические структуры, а именно обобщенно-личные и безличные предложения, а также пассивные конструкции с процессуальным значением. При этом союзные предложения (в особенности сложноподчиненные) преобладают над бессоюзными. Прииненный родительный падеж доминирует над приглагольным. Широко используются вводные слова и словосочетания и обороты-связки, а также обособленные согласованные определения, в том числе причастные обороты. Обычно тема предшествует реме. Диалогичность выражается в использовании вопросительных предложений, цитат и императива. Научному стилю не свойственно наличие разговорной лексики и эмоционально-экспрессивных средств. Ограничено использование синонимичных конструкций.

Официально-деловой стиль обслуживает, прежде всего, правовую сферу [5]. Особенностью данного стиля является использование инфинитивов и форм настоящего времени глаголов в значении предписания, кратких прилагательных модального характера, специальной терминологии и клише, лексических повторов, собирательных, отглагольных существительных. Кроме того, ограничено использование синонимических конструкций. На грамматическом уровне отмечается высокая частота сложных предлогов, цепей родительных падежей, придаточных условия, предложений с однородными членами, страдательных конструкций. Обычно отсутствуют глаголы в форме 1-го и 2-го лица, а также эмоционально-экспрессивные речевые средства. Преобладает прямой порядок слов.

Публицистический стиль обслуживает область общественных отношений: политических, экономических, культурных, спортивных и др. [5]. Преобладает общелитературная лексика, а также публицистическая лексика, образованная путем переносного использования специальной лексики с развитием в ней социально-оценочной окраски.

Художественный стиль является "инструментом художественного творчества и сочетающий в себе языковые средства всех других стилей речи" [6]. Художественный стиль часто называют стилем художественной литературы.

Разговорный стиль обслуживает сферу бытового общения [6]. Разговорному стилю свойственны разговорные слова и фразеологизмы, лексика с эмоционально-экспрессивной окраской, уменьшительно-ласкательные суффиксы, суффиксы субъективной оценки, а также субстантивация. Ограничено использование абстрактной, книжной и иноязычной лексики.

Приведенная классификация функциональных стилей не принимается единодушно всеми исследователями. Так, в советские времена не принято было выделять религиозный стиль. В настоящее время идут споры о правомерности выделения художественного и разговорного стилей, обусловленные, прежде всего, их неоднородностью.

Мы постарались исключить слабоформализуемые признаки, такие как логичность и последовательность изложения, доказательность, образность. Современные средства обработки текста на естественном языке позволяют анализировать текст на морфологическом и синтаксическом уровнях [7–9]. Существуют проекты по формализации семантики [10]. Разработаны методы и подходы определения эмоциональной и оценочной лексики [11–13]. Таким образом, лингвистическая теория предоставляет возможность производить классификацию текстов не только в рамках модели мешка слов, но и учитывая особенности функциональных стилей на различных языковых уровнях.

2. Постановка задачи классификации

Задача фильтрации является частным случаем задачи классификации (или рубрикации), а именно представляет собой бинарную классификацию. Бинарная классификация предполагает наличие двух непересекающихся категорий.

Приведем формальную постановку задачи классификации. Пусть дано конечное множество категорий $C = \{c_1, c_2, ..., c_{|C|}\}$ и конечное множество документов $D = \{d_1, d_2, ..., d_{|D|}\}$. Целевая функция $\phi : D \times C \to \{0,1\}$, которая для каждой пары $< документ, категория >$ определяет, соответствуют ли они друг другу, неизвестна. Необходимо найти классификатор ϕ', т.е. функцию, максимально близкую к функции ϕ [14]. Результаты решения классификационной задачи могут быть: точными, когда документ однозначно относится к той или иной категории $\phi' : D \times C \to \{0,1\}$, или ранжированными, если документ относится к категории с некоторой вероятностью $\phi' : D \times C \to [0,1]$. Понятия классификация и кластеризации не тождественны, так как классы заранее задаются пользователем или экспертом, а кластеры формируются автоматически при анализе коллекции.

Решение задачи классификации предполагает выбор дифференцирующих признаков, который может осуществляться на основе эвристики, лингвистических критериев, знаний предметной области, статистических характеристик текстов. Поскольку наиболее распространенными моделями документов являются варианты моделей множества слов (bag-of-words), а именно – бинарная модель и модель с весами терминов (первая учитывает только наличие или отсутствие слова в документе, тогда как во взвешенной модели каждому термину ставится в соответствие его вес) во многих случаях в качестве признаков выступают слова. Однако не все они релевантны для решения данной задачи. Поэтому многие классификаторы игнорируют слова из заранее заданных списков или же термины, встречающиеся как слишком часто, так и слишком редко. При этом пороговое значение выбирается на основе эвристик и может зависеть от корпуса и решаемой задачи. Обычно алгоритмы выбора признаков работают по следующей схеме: для каждого термина вычисляется мера различия между классами, после чего термины сортируются в порядке убывания этой величины и выбираются лучшие признаки.

3. Существующие методы классификации текстов по функциональным стилям

Несмотря на существование работ, опирающихся на прямое сопоставление текста с уже классифицированными документами или псевдо-документом, представляющим собой жанр [15], классическим подходом классификации текстов стало использование методов машинного обучения. Машинное обучение предполагает наличие обучающей и контрольной выборки, т.е. дана начальная коллекция документов $\Omega = \{d_1, d_2, ..., d_{|\Omega|}\} \subset D$, где значения целевой функции ϕ известны для $\forall < d_i, c_j > \in \Omega \times C$, коллекция разбивается на два непересекающихся множества. Классификатор ϕ' обучается индуктивно на основе выявленных характеристик документов [14]. При этом каждому документу соответствует вектор признаков.

Традиционно в качестве признаков используются частоты слов. Так, [16] используют модель bag-of-words для представления документов, а в качестве классифицирующего алгоритма – машину опорных векторов. Помимо частот слов, может быть учтена частота знаков препинания [17]. В [15] авторы предлагают использовать в качестве признаков наиболее частотные n-граммы. Качество классификации может быть повышено при использовании адаптированной метрики TF-IDF [18, 19].

Лексические признаки коррелируют с тематикой документа, что является существенным ограничением в их применимости [20]. Ограниченная репрезентативность тематики в обучающей выборке может значительно снизить качество классификации. В связи с этим, помимо лексических признаков целесообразным является использование грамматических характеристик, как морфологических, так и синтаксических. Например, метод, предложенный в [21], базируется на дискриминантном анализе по частоте местоимений третьего лица, глубине деревьев синтаксического разбора, длине предложений и слов. В [3] роль признаков играют части речи, знаки препинания, слова, используемые для обозна-

чения дат, а также средняя длина предложения и дисперсия, а в качестве классификаторов выступают нейронные сети и логистическая регрессия. Гистограммы частей речи являются признаками в [22]. Частотность частей речи в русскоязычных текстах анализируется в [1]. В методе, представленном в [23], помимо частот слов учитываются следующие признаки: частота использования прилагательных, вариативность глагольных форм, глубина синтаксических деревьев, использование знака табуляции, новой строки и пробелов, а также не буквенно-цифровых символов. В роли классификаторов выступают деревья решений, наивный байесовский классификатор и машина опорных векторов.

Широко распространенными остаются квантитативные признаки текста. В [24] применяются деревья решений, а в качестве классифицирующих признаков используются количественные характеристики текста (например, длина текста в символах, слогах и т.д.).

Следует отметить, что Интернет породил новую форму текста – гипертекст. Гипертекст обладает характеристиками, не свойственными другим текстам, что также может быть использовано для жанровой классификации. Так, в [25] в качестве признаков предлагается использовать метаданные HTML-страниц, теги, а также лингвистические характеристики. Авторы [26] используют комбинацию квантитативных и грамматических признаков с анализом HTML-разметки и ссылок, в т.ч. индекс удобочитаемости Флеша, использование вопросительных предложений, пассивных конструкций, служебных частей речи, количество параграфов, слова в ссылках и заголовке HTML-страницы. В работе [27] также используется комбинация квантитативных признаков (количество слов, доля числительных, доля слов в верхнем регистре, средняя длина слова, предложения, частота использования Интернет-аббревиатур и т.д.), эмоционально-оценочной лексики, частеречных характеристик (в т.ч. доля служебных слов), графематических признаков (например, частотность специальных символов), а также анализ HTML-тегов и HTML-структуры.

Обособлено стоят методы, не опирающиеся на лингвистические признаки. Примером может послужить работа [28], где в качестве классифицирующих признаков предлагается использовать последовательность из 4 символов. Классификатором служит машина опорных векторов с двумя типами расстояний между классами: расстояние на основе анализа пути в иерархии классов и на основе анализа содержимого. Вспомогательной характеристикой является частота жанра.

Перечисленные выше методы применимы в рамках одного языка. В свою очередь, [29] предлагают использовать сопоставительные корпуса для автоматической жанровой классификации. Авторы выделяют 3 типа характеристик:

- характеристики, зависящие от национального языка;
- характеристики, не зависящие от национального языка;
- сопоставимые характеристики, общие для двух или более языков.

Заключение

В данной статье была произведена попытка обобщить формализуемые признаки функциональных стилей. Основной акцент был сделан на морфемные, морфологические, синтаксические и лексические особенности текстов, принадлежащих разным функциональным стилям.

В настоящее время для классификации текстов преимущественно используются методы машинного обучения, где в качестве признаков традиционно применяются лексемы (или стемы). Однако все большее распространения получают методы, опирающиеся на квантитативные характеристики текста, а также учитывающие его морфологические и синтаксические особенности.

Современные средства обработки текста на естественном языке, разработанные алгоритмы классификации, а также теоретическое языкознание дают возможность усовершенствовать классическую модель мешка слов, традиционно используемую при классификации текстов.

Список литературы

1. *Браславский П.* Морфологический строй функциональных стилей (на материале документов Internet) // Известия Уральского государственного университета. 2001. № 21. Р. 9–17.
2. *Емашова О.А., Мальковский М.Г.* Функциональные стили русского языка и их влияние на задачу автоматического реферирования текста // Компьютерная лин-

гвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции "Диалог". 2007. Р.167–171.
3. *Kessler B., Nunberg G., Schutze H.* Automatic Detection of Text Genre // CoRR. 1997.
4. *Жеребило Т.В.* Словарь лингвистических терминов: Изд. 5-е, испр-е и дополн. Назрань: Изд-во "Пилигрим". 2010.
5. *Кожина М.Н. и др.* Стилистический энциклопедический словарь русского языка: Изд-во "Флинта". 2003.
6. *Белокурова С.П.* Словарь литературоведческих терминов. Паритет, 2006.
7. Manning C.D. et al. The Stanford CoreNLP Natural Language Processing Toolkit // Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations. 2014. P. 55–60.
8. *Nivre J., Boguslavsky I.M., Iomdin L.L.* Parsing the SynTagRus Treebank of Russian // Proceedings of the 22Nd International Conference on Computational Linguistics. Vol. 1. Stroudsburg, PA, USA: Association for Computational Linguistics, 2008. P.641–648.
9. *Segalovich I.* A Fast Morphological Algorithm with Unknown Word Guessing Induced by a Dictionary for a Web Search Engine. // MLMTA. Citeseer, 2003. P.273–280.
10. *Miller G.A., Beckwith R., Fellbaum C.* Introduction to WordNet: An On-line Lexical Database. 1993.
11. *Ermakov S., Ermakova L.* Sentiment Classification Based on Phonetic Characteristics // Advances in Information Retrieval / ed. Serdyukov P. et al. Springer Berlin Heidelberg, 2013. Vol. 7814. P. 706–709.
12. *Kim S.-M., Hovy E.* Identifying and Analyzing Judgment Opinions // Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL. 2006. P. 200–207.
13. *Pang B., Lee L.* Opinion Mining and Sentiment Analysis // Foundations and Trends in Information Retrieval. 2008. Vol. 2, № 1–2. P. 1–135.
14. *Лифшиц Ю.* Классификация текстов [Electronic resource]. 2005. URL: http://yury.name/internet/ (accessed: 10.10.2011).
15. *Mason J.E., Shepherd M., Duffy J.* An n-gram based approach to automatically identifying web page genre // System Sciences, 2009. HICSS'09. 42nd Hawaii International Conference on. IEEE, 2009. P. 1–10.
16. *Freund L., Clarke C.L., Toms E.G.* Towards genre classification for IR in the workplace // Proceedings of the 1st international conference on Information interaction in context. ACM, 2006. P. 30–36.
17. *Stamatatos E., Fakotakis N., Kokkinakis G.* Automatic text categorization in terms of genre and author // Computational linguistics. 2000. Vol. 26, № 4. P. 471–495.
18. *Lee Y.-B., Myaeng S.H.* Text genre classification with genre-revealing and subject-revealing features // Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 2002. P. 145–150.
19. *Snyman D.P., Van Huyssteen G.B., Daelemans W.* Automatic Genre Classification for Resource Scarce Languages // Proceedings of the 22nd Annual Symposium of the Pattern Recognition Association of South Africa. 2011. P. 132–137.
20. *Petrenz P., Webber B.* Stable classification of text genres // Comput. Linguist. Vol. 37, №2. P. 385–393.
21. *Karlgren J., Cutting D.* Recognizing Text Genres with Simple Metrics Using Discriminant Analysis // Proceedings of the 15th Conference on Computational Linguistics - Volume 2. Stroudsburg, PA, USA: Association for Computational Linguistics, 1994. P.071–1075.
22. *Feldman S. et al.* Part-of-speech histograms for genre classification of text // Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on. IEEE, 2009. P. 4781–4784.
23. *Dewdney N., VanEss-Dykema C., MacMillan R.* The form is the substance: Classification of genres in text // Proceedings of the workshop on Human Language Technology and Knowledge Management-Volume 2001. Association for Computational Linguistics, 2001. P. 7.
24. *Шевелев О.Г., Петраков А.В.* Классификация текстов с помощью деревьев решений и нейронных сетей прямого распространения // Вестник Томского государственного университета. 2006. Т. 290.
25. *Rehm G.* Towards Automatic Web Genre Identification // Proceedings of the 35th Annual Hawaii International Conference on System Sciences (HICSS'02)-Volume 4 -

Volume 4. Washington, DC, USA: IEEE Computer Society, 2002. P. 101.
26. *Boese E.S., Howe A.E.* Effects of web document evolution on genre classification // Proceedings of the 14th ACM international conference on Information and knowledge management. ACM, 2005. P. 632–639.
27. *Abramson M., Aha D.W.* What's in a URL? Genre Classification from URLs // Conference on Artificial Intelligence. 2012. P. 262–263.
28. *Wu Z., Markert K., Sharoff S.* Fine-grained genre classification using structural learning algorithms // Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, 2010. P. 749–759.
29. *Petrenz P., Webber B.* Robust cross-lingual genre classification through comparable corpora // The 5th Workshop on Building and Using Comparable Corpora. 2012. P. 1.

Automatic methods of text genre identification

L. M. Ermakova, M. A. Abashev, R. V. Nikitin, R. I. Ushakov
Perm State University, Russia, 614990, Perm, Bukireva st., 15
liana87@mail.ru; +73422396298

In the time of exponential growth of digital documents and particularly texts, the problem of automated text classification has become of the key interest of computer and information science as well as library science. Text classification is used in application to a big variety of related problems such as spam filtering, sentiment analysis, language identification, genre classification etc. In this paper we concentrate on the scientific texts detection which can be considered as a part of genre classification problem.

Key words: *classification; genre; machine learning; information retrieval.*