https://doi.org/10.30853/filnauki.2018-9-1.43

Яцко Вячеслав Александрович

КЛАССИФИКАЦИЯ ЛИНГВИСТИЧЕСКИХ ТЕХНОЛОГИЙ

Выделяются различные виды лингвистических технологий в зависимости от типа входных и выходных данных, уровня системы языка, средств и форм коммуникации, контингента пользователей. Описываются особенности функционирования лингвистических программ, систем и приложений, на входе у которых - текст на естественном языке. Особое внимание уделяется приложениям для статистического анализа текстовых документов. Показано, что выделенные виды технологий могут использоваться в качестве конфигурационных характеристик лингвистического программного обеспечения.

Адрес статьи: www.gramota.net/materials/2/2018/9-1/43.html

Источник

Филологические науки. Вопросы теории и практики

Тамбов: Грамота, 2018. № 9(87). Ч. 1. С. 193-196. ISSN 1997-2911.

Адрес журнала: www.gramota.net/editions/2.html

Содержание данного номера журнала: www.gramota.net/materials/2/2018/9-1/

УДК 81’33.2 Дата поступления рукописи: 06.06.2018
https://doi.org/10.30853/filnauki.2018-9-1.43

Выделяются различные виды лингвистических технологий в зависимости от типа входных и выходных данных, уровня системы языка, средств и форм коммуникации, контингента пользователей. Описываются особенности функционирования лингвистических программ, систем и приложений, на входе у которых – текст на естественном языке. Особое внимание уделяется приложениям для статистического анализа текстовых документов. Показано, что выделенные виды технологий могут использоваться в качестве конфигурационных характеристик лингвистического программного обеспечения.

Ключевые слова и фразы: лингвистические технологии; критерии классификации; лингвистические программы, системы, приложения; конкордансы; компьютерная лингвистика; информационная наука.

Яцко Вячеслав Александрович, д. филол. н., профессор
Хакасский государственный университет имени Н. Ф. Катанова, г. Абакан
iatsko@gmail.com

КЛАССИФИКАЦИЯ ЛИНГВИСТИЧЕСКИХ ТЕХНОЛОГИЙ

Исследование выполнено при поддержке гранта РФФИ 16-07-00014.

Последние десятилетия характеризуются выделением и интенсивным развитием в рамках научных дисциплин информационных направлений, таких как биоинформатика, медицинская информатика, химическая информатика, правовая информатика, историческая информатика [4]. Основная задача этих направлений – разработка информационных технологий с учётом специфики данных дисциплин с целью поддержки исследований, разработок и информационного обслуживания специалистов.

Не осталось в стороне от этого процесса и языкознание. Создание в 1964 г. в Брауновском университете США первого электронного корпуса текстов и выход в 1967 г. статьи *Компьютерный анализ современного американского варианта английского языка*, которую подготовили создатели корпуса Г. Кучера (H. Kucera) и У. Френсис (W. Francis) [6], ознаменовали появление корпусной лингвистики как нового направления, специально предназначенного для поддержки лингвистических исследований. С тех пор во многих странах были созданы текстовые корпуса разных видов (национальные, исторические, тематические), а использование информационных технологий стало неотъемлемой частью лингвистических исследований. Если раньше на поиск исследуемых лингвистических единиц в бумажных источниках уходили часы и дни, то сейчас на их поиск с помощью технологий, реализованных в корпусах, уходят секунды. Это позволяет увеличивать объём анализируемого материала и получать более достоверные выводы о свойствах лингвистических единиц и языка в целом. Статистическая информация о частотностях лингвистических единиц и конструкций, получаемая с помощью современных технологий, позволяет эффективно решать задачи классификации и типологии и имеет непосредственное значение для лингводидактики. Фактографические поисковые системы, используемые в корпусах, позволяют выполнять поиск по моделям, отражающим структуру единиц языка и конструкций, что стимулирует разработку метаязыков описания в различных направлениях языкознания и формализации проводимых исследований.

Современные лингвистические технологии, однако, не исчерпываются разработками в рамках корпусной лингвистики. В 60-х гг. XX века в связи с созданием автоматизированных информационно-поисковых систем и систем машинного перевода были разработаны алгоритмы морфологического анализа и взвешивания терминов. В 80-е гг. начались разработки в области автоматического распознавания устной речи. Развитие Интернета в 90-е гг. привело к глобализации технологий информационного поиска и обусловило необходимость создания систем автоматической классификации текстовых документов. В настоящее время для обозначения дисциплины, изучающей проблемы разработки лингвистических технологий, не связанных с поддержкой лингвистических исследований, используется термин «компьютерная лингвистика» [13]. Вместе с тем в рамках данной дисциплины до сих пор не разработано однозначных критериев классификации лингвистических технологий, что определяет актуальность данного исследования.

Цель настоящей статьи – выявить возможные критерии классификации и на их основе впервые дать всестороннее системное описание особенностей различных видов современных лингвистических технологий.

Одним из критериев системной классификации лингвистических технологий является уровень системы языка, к которому относятся обрабатываемые лингвистические единицы. На графемном уровне используются программы и технологии оптического распознавания символов, позволяющие с помощью сканеров и специализированного программного обеспечения выполнять перевод бумажных текстов в машиночитаемый формат. Эти программы используются при создании электронных библиотек, корпусов текстов. На фонетическом уровне используются программы распознавания звуков, обеспечивающие конвертацию речи в текст и лежащие в основе различных видов программного обеспечения для распознавания речи. На морфологическом уровне применяются программы распознавания основ слов (стемм и лемм), суффиксов и окончаний, которые позволяют отождествлять слова с одной основой, что необходимо для адекватного взвешивания терминов,

а также идентификации частей речи. На лексическом уровне используются программы лексической декомпозиции, аннотирования слов условными обозначениями (тегами), которые могут указывать на часть речи, семантический или прагматический признак. На синтаксическом уровне используются программы синтаксической декомпозиции для разбивки текста на предложения, словосочетания и клаузы, а также синтаксический парсинг, позволяющий получать граф иерархической структуры предложения. На дискурсивном уровне используются программы сегментации текста на единицы больше предложения, а также методы разрешения анафоры и кореференции (см. подробное описание лингвистических алгоритмов и программ, выделяемых по уровням системы языка в [3]).

Данные программы лежат в основе функционирования лингвистического программного обеспечения, приложений и систем, основной особенностью которых является обработка текстов на естественном языке. На входе таких систем – текст на естественном языке. На выходе может выдаваться информация различных типов, что также может служить критерием классификации лингвистических технологий и систем. У информационно-поисковых систем и систем автоматического реферирования на выходе – вторичный документ, отражающий содержание текста. В первом случае – ссылка на веб-ресурс с его описанием, во втором – реферат входного текста. У систем машинного перевода на выходе – эквивалентный текст на другом языке. У систем автоматической классификации на выходе – имя класса, к которому относится текст. В случае авторской атрибуции таким именем является имя автора текста. В системах распознавания спама и плагиата применяется бинарная классификация и, соответственно, на выходе указывается, является или не является входной текст спамом / плагиатом. В системах автоматической категоризации на выходе приводится название тематической категории или жанра входного текста.

Лингвистические системы следует отличать от указанных выше и классифицированных по уровням системы языка лингвистических программ, а также приложений. Лингвистические программы не используются самостоятельно, а служат подсистемами, модулями систем. Лингвистические приложения выдают пользователям статистическую информацию о распределении единиц текста (обычно слов, словосочетаний) по частотностям. Общепринятым термином для обозначения приложений для статистического анализа текстовых документов является «конкорданс». Существующие конкордансы различаются по функциональности. К стандартным требованиям относятся: функция просмотра слова в контексте (*key word in context*); функция *wordlist*, позволяющая получать списки слов с указанием их рангов и частотностей; функция получения коллокаций – статистически значимых слов, сочетающихся с ключевым словом, указанным пользователем. Для коллокаций выводится частотность их использования в определённой позиции слева или справа от ключевого слова. Также выводится информация о количестве уникальных слов во входном тесте (без учётов повторов) и количестве токенов (сумме частотностей слов).

Обычно создатели конкордансов наряду со стандартными функциями предлагают дополнительную функциональность, расширяющие возможности их использования. *KWIC Concordance* [7], разработанный японским специалистом Сатору Цукамото (Satoru Tsukamoto), также предоставляет возможность обрабатывать тексты, аннотированные тегами частей речи с использованием набора тегов, разработанного в Пенсильванском университете США. Соответственно, пользователи могут получать статистические данные и о распределении частей речи. Конкорданс *AntConc*, созданный Лоренсом Энтони (Laurence Anthony) [9], позволяет получать статистические данные о распределении словосочетаний (n-грам), а также списки ключевых слов. Платное приложение *WordSmith* [8] предусматривает возможность получения статистики по отдельным символам и сочетаниям символов, а также возможность находить дубликаты текстовых файлов, объединять текстовые файлы в один файл, разбивать большие текстовые файлы на несколько отдельных файлов.

Выходная информация может классифицироваться не только по типу текстовых данных, но и по степени интеллектуальности. Соответственно, можно выделить стандартные и интеллектуальные лингвистические технологии. Если первые отражают содержание входного текста, то в результате применения вторых генерируется новая информация, которая имплицитно содержится во входном тексте. Такой информацией могут быть, например, числовые коэффициенты, отражающие интенсивность положительной или отрицательной оценки некоторых товаров или деятельности некоторой личности [11]. С помощью интеллектуальных технологий могут выявляться новые, ранее неизвестные зависимости, например, при анализе больших объёмов текстов историй болезней можно выявить зависимость какого-то заболевания от индивидуальных характеристик личности. Интеллектуальный анализ может применяться не только в новых направлениях, таких как анализ мнений пользователей или покупателей, размещённых в Интернете, но и в традиционных направлениях обработки текстовых документов. В результате анализа текста в процессе реферирования могут генерироваться и включаться в реферат предложения, выражающие обобщения, которых нет в исходном документе. К интеллектуальным можно отнести и системы автоматической классификации документов, поскольку их задачей как раз и является раскрытие информации, имплицитно содержащейся в тексте: распознавание плагиата, спама, тематической категории. Еще более высокий уровень анализа представляют технологии искусственного интеллекта, благодаря которым возможно установление логических отношений между суждениями, выражаемых высказываниями, и моделирование логико-семантической структуры связного текста: аргументации, нарратива, описания как основных типов речи.

Наряду с типом выходных данных в качестве критерия классификации можно использовать и типы входных текстов. Основная сложность анализа текстовых документов состоит в том, что они содержат неструктурированные данные, что обуславливает необходимость задания определённой структуры в процессе

их обработки. Наиболее распространённым способом структурирования текстовых данных в настоящее время является графовая архитектура [10], в соответствии с которой в составе текстов выделяются объекты, ярлыки и отношения. В качестве объектов могут выступать списки слов, предложений, словосочетаний, абзацев и других единиц текста, которые связываются между собой отношениями типа *ContainedIn* (содержится в), *IdentifierOf* (идентификатор), *TagOf* (тег). Применение графовой архитектуры позволяет хранить в отдельных файлах списки единиц текста, а также метаинформацию (библиографические данные) о текстовом документе. Совокупность объектов и отношений составляет объектную модель текста, создание которой является результатом предварительной обработки документа.

Предварительная обработка может существенно усложняться, если на входе находится нередактированные документы, к которым относятся размещённые в Интернете тексты блогов, форумов, чатов. Их можно отнести к нередактированным, поскольку они не вычитываются и не правятся рецензентами и редакторами, что имеет место в случае публикации, например, научных изданий. Качество этих текстов зависит в основном от уровня образования и языковой компетенции их авторов. Данные виды текстов представляют собой ценный источник информации, на который ориентированы современные лингвистические технологии, такие как анализ мнений покупателей и пользователей, распознавание экстремистского контента. Трудность их автоматической обработки и анализа обусловлена двумя факторами. Во-первых, данные веб-ресурсы относятся к диалогической речи как форме коммуникации; для понимания текста, созданного одним автором, необходимо установить его связи с тестами других участников, что предусматривает выполнение сложных алгоритмов анализа связного текста и диалогической речи. Во-вторых, они могут содержать как орфографические ошибки, так и орфографические и лексико-грамматические варианты, характерные для данного жанра текста. Это требует создания специальных словарей и дополнительных правил предварительной обработки [5]. Особую сложность представляет обработка чатов, поскольку они создаются спонтанно, в то время как тексты, размещаемые пользователями блогов и форумов, могут предварительно готовиться авторами. Соответственно, нередактированные документы можно разделить на подготовленные и неподготовленные (спонтанные).

По средству коммуникации входные тексты можно разделить на два вида: письменные и устные; соответственно, выделяются технологии обработки текста и речи. Алгоритмы анализа текста разрабатываются с 50-х годов XX века, в то время как алгоритмы анализа устной речи стали разрабатываться и широко внедряться в 80-х годах XX века и сейчас делятся на два направления: распознавание речи и синтез речи. Распознавание речи широко применяется в системах голосового управления техническими объектами, системах распознавания индивидуальных характеристик личности, таких как возраст, пол и даже уровень алкогольного опьянения [12]. Голосовое управление телефонами позволяет совершать, принимать или отклонять звонки, отправлять короткие сообщения, управлять системами навигации, настройками телефона, напоминаниями и будильниками, фото- и видеокамерой, прослушиванием музыки, просмотром видео.

Синтез речи предусматривает конвертацию текста в речь, которая применяется в приложениях, предназначенных для озвучивания текста. Такие приложения особенно важны для людей с расстройствами речи и зрения. В современных вопросно-ответных системах [2] применяются как технологии распознавания, так и технологии синтеза речи. К таким системам относятся автоответчики, посредством которых пользователь может, например, заказать билет на самолёт, в театр, кинотеатр. При разработке приложений-автоответчиков выделяются ключевые слова и фразы и с помощью информантов устанавливается возможный диапазон их произношения. При подборе информантов учитываются различные возрастные, социальные, этнические характеристики. Ещё один распространённый тип гибридного программного обеспечения – голосовые ассистенты, устанавливаемые на смартфоны и компьютеры [1]. Голосовые ассистенты разработаны всеми ведущими ИТ-компаниями и поддерживают наиболее популярные платформы. Компанией *Google* для платформы Андроид разработан *Google Assistant*; компанией *Microsoft* для платформы *Windows* – Кортана; компанией *Apple* для платформ *iOS*, *macOS* – *Siri*. Компания *Amazon* создала голосовой ассистент *Alexa* для аппаратной платформы *Amazon Echo*, который поддерживает три указанные выше программные платформы. В *Google Assistant* вопрос или реплика пользователя конвертируются в текстовый запрос, который выполняется поисковой системой *Google*. Найденные поисковой системой результаты (письменные тексты) озвучиваются. Также *Google Assistant* выполняет описанные выше функции голосового управления, включая управление умными домашними устройствами с помощью голосового динамика *Google Home*.

Заметим, что разграничение двух видов технологий достаточно условно: один и тот же вид лингвистических технологий может применяться и в системах обработки текста, и в системах обработки речи. Информационно-поисковые системы могут проводить поиск как по письменному запросу, так и по устному. Системы реферирования могут проводить анализ как устных, так и письменных текстов. При обработке и текста, и речи используются единые алгоритмы лексической и синтаксической декомпозиций, а для идентификации единиц и устного, и письменного текста применяются методы вероятностно-статистического анализа, такие как байесовские модели и скрытые марковские модели.

Еще одним критерием классификации лингвистических технологий может выступать целевая категория пользователей. По этому критерию можно выделить глобальные, специальные, специализированные технологии и системы. Глобальные технологии разрабатываются с учётом потребностей любых пользователей, независимо от возраста, социального положения, уровня образования. Типичный пример – универсальные информационно-поисковые системы, такие как *Гугл* и *Яндекс*. Специальные технологии ориентированы на специалистов предметной области. К ним относятся технологии, разрабатываемые в рамках корпусной

лингвистики и предназначенные в первую очередь для поддержки лингвистических исследований. К специализированным относятся технологии, предназначенные для поддержки принятия решений.

Нами были рассмотрены различные виды лингвистических технологий, выделенных по шести основным критериям: уровню системы языка; типу входных данных; типу выходных данных; средству коммуникации; форме коммуникации; контингенту пользователей. Выделяемые по основным критериям виды лингвистических технологий, в свою очередь, могут быть классифицированы по дополнительным критериям. Так, по типу входных текстов можно выделить технологии обработки редактированных и нередактированных документов, а последние – на технологии обработки подготовленных и неподготовленных документов по критерию спонтанности речи. В одной и той же лингвистической системе могут реализовываться разные виды технологий. Например, в системах анализа мнений применяются технологии интеллектуального анализа, поскольку они выдают информацию, имплицитно присутствующую в документах. В этих системах также применяются технологии обработки письменных диалогических нередактированных подготовленных текстов, причём на разных уровнях системы языка: морфологическом, лексическом, синтаксическом, дискурсивном. Таким образом, выделенные нами в данной статье виды технологий позволяют определить конфигурацию лингвистического программного обеспечения, что имеет существенное значение для системного представления электронных ресурсов предметной области.

Список источников

1. **Косач Д. И., Жидкова Л. О., Белехов А. Н.** Виртуальные голосовые помощники с элементами искусственного интеллекта [Электронный ресурс] // Научный альманах. 2016. № 3-3 (17). URL: http://ucom.ru/doc/na.2016.03.03.083.pdf (дата обращения: 06.06.2018).
2. **Соловьёв А. А., Пескова О. В.** Построение вопросно-ответной системы для русского языка: модуль анализа вопросов [Электронный ресурс]. URL: https://cyberleninka.ru/article/v/postroenie-voprosno-otvetnoy-sistemy-dlya-russkogo-yazyka-modul-analiza-voprosov (дата обращения: 06.06.2018).
3. **Яцко В. А.** Алгоритмы и программы автоматической обработки текста [Электронный ресурс]. URL: https://cyberleninka.ru/article/v/algoritmy-i-programmy-avtomaticheskoy-obrabotki-teksta (дата обращения: 06.06.2018).
4. **Яцко В. А.** Принципы исследования исторического развития информатики // Научно-техническая информация. Серия 1. Организация и методика информационной работы. 2017. № 9. С. 1-9.
5. **Яцко В. А., Клещ А. В.** Особенности автоматической обработки чатов // Естественные и технические науки. 2011. № 5. С. 319-327.
6. **Francis W. N., Kucera H.** Computational analysis of present day American English. Providence, R. I.: Brown University Press, 1967. 424 p.
7. **http://dep.chs.nihon-u.ac.jp/english_lang/tukamoto/kwic_e.html** (дата обращения: 06.06.2018).
8. **http://lexically.net/LexicalAnalysisSoftware** (дата обращения: 06.06.2018).
9. **http://www.laurenceanthony.net/software/antconc** (дата обращения: 06.06.2018).
10. **Ide N., Keith Suderman K.** GrAF: a graph-based format for linguistic annotations [Электронный ресурс] // Proceedings of the Linguistic annotation workshop. Prague, 2007. P. 1-8. URL: https://www.cs.vassar.edu/~ide/papers/LAW.pdf (дата обращения: 06.06.2018).
11. **Kim Y., Jeong S. R., Ghani I.** Text opinion mining to analyze news for stock market prediction [Электронный ресурс] // International Journal of Advances in Soft Computing and Its Applications. 2014. Vol. 6. № 1. Special issue. P. 1-13. URL: http://home.ijasca.com/data/documents/Paper-ID-424-IJASCA_Formated.pdf (дата обращения: 06.06.2018).
12. **Schiel F., Heinrich C.** Laying the foundation for in-car alcohol detection by speech [Электронный ресурс] // Proceedings of Interspeech. Brighton, 2009. P. 983-986. URL: https://pdfs.semanticscholar.org/934c/b738afe2fa6ab2231c189c49cd84fa7331e7.pdf (дата обращения: 06.06.2018).
13. **The handbook of computational linguistics and natural language processing** [Электронный ресурс] / ed. by A. Clark, C. Fox, S. Lappin. Oxford: Wiley-Blackwell, 2010. 802 p. URL: http://course.duruofei.com/wp-content/uploads/2015/05/Clark_Computational-Linguistics-and-Natrual-Language-Processing.pdf (дата обращения: 06.06.2018).

CLASSIFICATION OF LINGUISTIC TECHNOLOGIES

Yatsko Vyacheslav Aleksandrovich, Doctor in Philology, Professor
Katanov Khakass State University, Abakan
iatsko@gmail.com

The paper identifies different types of linguistic technologies depending on the input and output data type, language system level, means and forms of communication, contingent of users. The author describes the peculiarities of the functioning of linguistic programs, systems and applications, the input of which is the text in the natural language. Special attention is paid to the applications for the statistical analysis of text documents. It is shown that the identified technologies can be used as the configurational characteristics of linguistic software.

Key words and phrases: linguistic technologies; classification criteria; linguistic programs, systems, applications; concordances; computational linguistics; informational science.