УДК 004.054

В.С. Юсков, И.В. Баранникова

СРАВНИТЕЛЬНЫЙ АНАЛИЗ ПЛАТФОРМ ОБРАБОТКИ ЕСТЕСТВЕННОГО ЯЗЫКА

Рассмотрены сферы применения инфокоммуникационных технологий. Даны определения «естественный язык» и «обработка естественного языка». Показаны критерии сравнения платформ обработки естественного языка. Изучены актуальные и популярные платформы обработки естественного языка, так или иначе представленные в сфере. Показаны критерии сравнительного анализа платформ. Приведено обоснование выбора критериев сравнения. В качестве примера показан рекомендуемый порядок выбора, учитывающий важность критериев. Дополнительно были описаны факторы, изначально не включенные в сравнение, вероятность влияния которых, тем не менее, не исключается.

Ключевые слова: инфокоммуникационные технологии, автоматизация, естественный язык, обработка естественного языка, инструментарий обработки естественного языка, платформа, оцениваемость, критерии сравнения.

Сегодня инфокоммуникационные технологии являются неотъемлемой частью многих областей человеческой деятельности и отраслей производства: промышленность (предприятия по добыче горно-химического сырья, электроэнергетика и т.д.), транспорт, реклама, информационно-вычислительное обслуживание, наука, финансы, а также геодезия, геология и разведка недр. Естественно, ИТ-специалистам приходится иметь дело с так называемыми «большими данными», то есть структурированными и неструктурированными данными огромных объемов и значительного многообразия, в том числе информацией на естественных языках (ЕЯ). Поэтому автоматизация обработки ЕЯ становится важным аспектом развития современного программного обеспечения.

Естественным языком является любой язык, используемый для общения людей и не созданный целенаправленно (в отличие от искусственных языков и других типов знаковых систем). Под термином «обработка естественного языка» понимается компьютерный анализ и синтез естественных языков. В общем смысле решение этих проблем означает создание более удобной формы взаимодействия компьютера и человека.

ISSN 0236–1493. Горный информационно-аналитический бюллетень. 2017. № 3. С. 272–278.
© 2017. В.С. Юсков, И.В. Баранникова.

Применение инструмента обработки ЕЯ в компьютерных системах достаточно широко. Это машинный перевод (процесс перевода текстов с одного естественного языка на другой с помощью специализированного программного обеспечения или веб-сервиса), распознавание речи (голосовое управление) или анализ текста (категоризация текстов, извлечение информации, информационный поиск или разработка инструментов представления информации для ЛПР/пользователя). Также примерами являются: генерация текста (на данном этапе развития находится на стадии исследований и экспериментов, таких как программы, предназначенные имитировать общение), синтез речи (искусственное воспроизведение человеческой речи на основе текста, образов или программных команд – на сегодняшний день активно внедряется в пользовательские операционные системы) и изменение текста, например упрощение или аннотирование.

Разработка любой крупной компьютерной системы или программного пакета – это достаточно трудоемкий и длительный процесс. Затраченное время на такие проекты варьируется от нескольких месяцев до года (и более), также влияет на процесс создания и величина финансовых вложений. Инструментарием, который значительно облегчает влияние этих факторов на разработку, является использование вспомогательных компонентов, которые обычно включают различные наборы библиотек, фреймворков и пакетов. Инструментарий обработки ЕЯ не исключение, он также представляет собой различные библиотеки и фреймворки для символического и статистического анализа естественного языка и обработки речи, а также готовые платформы, с помощью которых сразу можно работать с данными. В сравнительном анализе были рассмотрены как платформы, так и отдельные компоненты, в виде наборов библиотек.

Среди разработчиков, представленных в табл. 1 и 2, как и в любой другой коммерческой среде, существует определенная конкуренция. Ее наличие следует из двух логичных фактов: во-первых, разработчиков несколько; во-вторых, некоторые из них владеют платным вариантом собственного «решения». В табл. 1 приведены популярные готовые платформы, а в табл. 2 актуальные подключаемые компоненты обработки ЕЯ (наборы библиотек). В качестве критериев сравнительного анализа (оцениваемости), были выбраны следующие: вид лицензии (модель распространения), разработчик или автор. Также в сравнение включен вид предоставляемого решения, то есть, яв-

Таблица 1

Платформа	Разработчик/ автор	Лицензия	Вид предоставляемого решения	Язык программирования (для наборов библиотек)	Дополнительные сервисы
Alchemy API	Orchestr8	бесплатная демо-версия/ коммерческая	платформа	–	Обширная поддержка по продукту. Готовые решения для: «облачных» сервисов, обработки файлов документов и e-mail, а также веб-сайтов. Корпоративная версия коммерческой лицензии.
Carabao	Digital Sonata Pty Ltd	бесплатная демо-версия/ коммерческая	платформа	–	Автоматизированный многоязычный поиск, компоненты для мониторинга веб-ресурсов.
Ellogon	Georgios Petasis	свободное ПО (LGPL)	платформа	–	Открытое внедрение собственных словарей, инструменты для обработки и визуализации текстов HTML/XML и наличие базового графического интерфейса.

Таблица 2

Платформа	Разработчик/автор	Лицензия	Вид предоставляемого решения	Язык программирования (для наборов библиотек)	Дополнительные сервисы
Free Ling	Политехнический университет Каталонии	свободное ПО (GNU)	набор библиотек	C++	–
Ling Pipe	Alias-i	бесплатная демо-версия / коммерческая	набор библиотек	Java	–
Lingua Stream	Университет Кан (Франция)	академическая	набор библиотек	Java	поддержка и помощь разработчикам, использующим это решение, в виде веб-сообщества на соответствующем форуме
modular Audio Recognition Framework	The MARF Research and Development Group	свободное ПО (BSD)	набор библиотек	Java	расширенный инструментарий по обработке звуковой информации
UIMA	Apache	коммерческая	набор библиотек	Java, C++	качественная коммерческая техническая поддержка

ляется ли компонент готовой платформой, с помощью которой сразу можно заниматься обработкой данных или это подключаемая библиотека, содержащая только алгоритмы обработки ЕЯ. В критерии включены: поддерживаемый язык программирования (если это библиотека) и наличие дополнительных сервисов, которые разработчик предоставляет клиенту.

При определении важности критериев выбора платформы возникает много дополнительных аспектов, которые важно разъяснить.

Определение модели распространения собственного продукта

Определение соответствия архитектуры и совместимости с нужными ОС

Определение соответствия языка программирования (для набора библиотек) с выбранным для разработки собственного продукта

Анализ дополнительных сервисов, предоставляемых разработчиком и определение их важности при выборе

Порядок выбора платформы с учетом важности критериев

Во-первых – лицензия или модель распространения продукта. Существует три основных вида лицензирования: бесплатный, коммерческий и свободное ПО. Использование противоположной собственному продукту модели распространения может повлечь за собой проблемы у пользователя. В основном это касается двух видов лицензий: коммерческой и свободной. Причины просты: пересечение этих лицензий в одном программном продукте нарушает правила использования обеих моделей распространения, однако это уже юридическая сфера вопроса.

Второе – это поддержка языков программирования, что актуально для наборов библиотек. Естественно необходимо, чтобы язык соответствовал тому, который используется в собственной разработке. Большое же количество поддерживаемых языков значительно расширяет разработчику базу клиентов. Дополнительные сервисы, включаемые разработчиком в свою платформу или компонент, всегда положительно влияют на выбор клиентов, так как это так же сокращает и расходы, и время на разработку.

Еще одним немаловажным фактором является кроссплатформенность – адаптация алгоритмов для работы в различных операционных системах и архитектурах. Поскольку все представленные платформы поддерживают актуальные ОС, этот критерий не был внесен в таблицу сравнения.

Ниже представлен порядок выбора (рисунок), соответствующий важности критериев, по которому можно ориентироваться разработчикам при ведении проектов с использованием автоматизации обработки ЕЯ.

Проведенный анализ позволяет выстроить оптимальный порядок, при котором выбор подходящей платформы или компонента обработки ЕЯ не вызывает трудностей. Однако не стоит забывать о различных внешних факторах, влияющих на выбор: политика роста цен, столкновение законодательств, экономические и политические факторы и многие другие.

СПИСОК ЛИТЕРАТУРЫ

1. *Большакова Е. И.*, *Клышинский Э. С.*, *Ландэ Д. В.*, *Носков А. А.*, *Пескова О. В.*, *Ягунова Е. В.* Автоматическая обработка текстов на естественном языке и компьютерная лингвистика: учебное пособие. – М.: МИЭМ, 2011. – 272 с.
2. *Попов Э. В.* Общение с ЭВМ на естественном языке, изд. 2. – М.: Эдиториал УРСС, 2004. – 360 с.
3. *Кристофер Д. Маннинг, Прабхакар Рагхаван, Хайнрих Шютце.* Введение в информационный поиск. – М.: Вильямс, 2011. – 528 с.
4. *Daniël de Kok, Harm Brouwer* Natural Language Processing for the Working Programmer. – 2011. URL: https://www.researchgate.net/publication/259572969_Draft_Natural_Language_Processing_for_the_Working_Programmer (дата обращения 01.01.2016)
5. *Statistical* Natural Language Processing / In M. Lothaire, editor, Applied Combinatorics on Words. Cambridge University Press, 2005.
6. *Википедия* – Свободная энциклопедия. Обработка естественного языка. URL: https://ru.wikipedia.org/wiki/Обработка_естественного_языка (дата обращения 01.01.2016)
7. *Соммервилл И.* Инженерия программного обеспечения. – М.: Изд-во «Вильямс», 2002. – 624 с.
8. *Батоврин В. К.* Толковый словарь по системной и программной инженерии. – М.: ДМК Пресс, 2012. – С. 280.
9. *Джек Гринфилд, Кит Шорт, Стив Кук, Стюарт Кент, Джон Крупи.* Фабрики разработки программ: потоковая сборка типовых приложений, моделирование, структуры и инструменты. – М.: Изд-во «Диалектика», 2006. – С. 592.
10. *Поляков В. Н., Дубравина Т. В.* Лингвистические основы информатики. Электронный учебник. – М.: Изд-во ИНФО-Рутения, 2007. – 144 с.
11. *Бодров Д. А.*, *Кожитов С. Л.*, *Поляков В. Н.* Задачи интерактивной обработки поисковых запросов в теоретико-множественной постановке // Известия Саратовского университета Новая серия. Серия Математика. Механика. Информатика. – 2007. – т. 7. – Вып. 1. – С. 78–83.
12. ГОСТ Р ИСО/МЭК 12207-2010 Информационная технология. Системная и программная инженерия. Процессы жизненного цикла программных средств. ГИАБ

КОРОТКО ОБ АВТОРАХ

Юсков Виталий Сергеевич[1] – студент, e-mail: v.yuskov@gmail.com,
Баранникова Ирина Владимировна[1] – кандидат технических наук, доцент,
[1] НИТУ «МИСиС».

Gornyy informatsionno-analiticheskiy byulleten'. 2017. No. 3, pp. 272–278.

UDC 004.054

V.S. Yuskov, I.V. Barannikova

COMPARISON OF PLATFORMS OF NATURAL LANGUAGE PROCESSING

Scopes of infocommunication technologies are considered. Definitions «natural language» and «natural language processing» are given. Criteria of comparison of platforms of natural language processing are shown. The actual and popular platforms of natural language processing anyway presented in the sphere are studied. Criteria of the comparative analysis of platforms are shown. Justification of a choice of criteria of comparison is given. As an example the recommended choice order considering importance of criteria is shown. In addition the factors which are initially not included in comparison which probability of influence, nevertheless, isn't excluded were described.

Key words: infocommunication technologies, automation, natural language, natural language processing, tools of natural language processing, platform, estimation, criteria of comparison.

AUTHORS

Yuskov V.S.[1], Student, e-mail: v.yuskov@gmail.com,
Barannikova I.V.[1], Candidate of Technical Sciences, Assistant Professor,
[1] National University of Science and Technology «MISiS», 119049, Moscow, Russia.

REFERENCES

1. Bol'shakova E. I., Klyshinskiy E. S., Lande D. V., Noskov A. A., Peskova O. V., Yagunova E. V. *Avtomaticheskaya obrabotka tekstov na estestvennom yazyke i komp'yuternaya lingvistika*: uchebnoe posobie (Automatic processing of texts in a natural language and computational linguistics, Educational aid), Moscow, MIEM, 2011, 272 p.

2. Popov E. V. *Obshchenie s EVM na estestvennom yazyke*, izd. 2 (Communication with the COMPUTER in a natural language, 2nd edition), Moscow, Editorial URSS, 2004, 360 p.

3. Kristofer D. Manning, Prabkhakar Ragkhavan, Khaynrikh Shyuttse. *Vvedenie v informatsionnyy poisk* (Introduction to information search), Moscow, Vil'yams, 2011, 528 p.

4. Daniël de Kok, Harm Brouwer. *Natural Language Processing for the Working Programmer.* 2011, https://www.researchgate.net/publication/259572969_Draft_Natural_Language_Processing_for_the_Working_Programmer (accessed 01.01.2016).

5. *Statistical Natural Language Processing*. In M. Lothaire, editor, Applied Combinatorics on Words. Cambridge University Press, 2005.

6. Vikipediya Svobodnaya entsiklopediya. Obrabotka estestvennogo yazyka, https://ru.wikipedia.org/wiki/Обработка_естественного_языка (accessed 01.01.2016)

7. Sommervill I. *Inzheneriya programmnogo obespecheniya* (Software engineering), Moscow, Izd-vo Vil'yams, 2002, 624 p.

8. Batovrin V. K. *Tolkovyy slovar' po sistemnoy i programmnoy inzhenerii* (The explanatory dictionary on system and program engineering), Moscow, DMK Press, 2012, pp. 280.

9. Dzhek Grinfild, Kit Short, Stiv Kuk, Styuart Kent, Dzhon Krupi. *Fabriki razrabotki programm: potokovaya sborka tipovykh prilozheniy, modelirovanie, struktury i instrumenty* (Software Factories: Assembling Applications with Patterns, Models, Frameworks, and Tools), Moscow, Izd-vo «Dialektika», 2006, pp. 592.

10. Polyakov V. N., Dubravina T. V. *Lingvisticheskie osnovy informatiki*. Elektronnyy uchenik (Linguistic fundamentals of informatics. Electronic textbook), Moscow, Izd-vo INFO-Ruteniya, 2007, 144 p.

11. Bodrov D. A., Kozhitov S. L., Polyakov V. N. *Izvestiya Saratovskogo universiteta Novaya seriya. Seriya Matematika. Mekhanika. Informatika*. 2007, vol. 7, issue 1, pp. 78–83.

12. *Informatsionnaya tekhnologiya. Sistemnaya i programmnaya inzheneriya. Protsessy zhiznennogo tsikla programmnykh sredstv. GOST R ISO/MEK* (Information technology. System and program engineering. Processes of life cycle of software. State Standart R ISO/MEK 12207-2010).