ОСНОВНЫЕ ЗАДАЧИ АВТОМАТИЧЕСКОЙ ОБРАБОТКИ ТЕКСТОВ И ПОДХОДЫ К ИХ РЕШЕНИЮ

Н. К. Рубашко

Белорусский государственный университет
Минск, Беларусь
E-mail: roubashko@bsu.by

Данная статья посвящена анализу основных подходов к решению задач автоматической обработки текстов, возникающих при создании высокотехнологичных интеллектуальных систем, обеспечивающих замену человеческого труда в интеллектуальной сфере, опирающейся на использование естественного языка.

Ключевые слова: естественный язык, автоматическая обработка текстов, компьютерная лингвистика.

ВВЕДЕНИЕ

Когда речь идет о создании перспективных информационных технологий, то проблемы автоматической обработки текстовой информации выступают на передний план. Это определяется тем, что естественный язык (ЕЯ) является не только инструментом мышления, но и универсальным средством общения – средством восприятия, накопления, хранения, обработки и передачи информации [1]. Более того, ЕЯ становится также универсальным средством описания действительности и коммуникации с вычислительной системой. В наше время, когда пользователем может оказаться практически каждый, проблема взаимодействия человека с ЭВМ на естественном языке стала важной практической задачей.

Сегодня автоматическая обработка ЕЯ (в том числе и текста ЕЯ) – это бурно развивающаяся область научных исследований и коммерческих разработок, ставящих целью разработку промышленных систем обработки ЕЯ, которые должны быстро и эффективно обрабатывать в режиме реального времени огромные потоки информации, циркулирующие в информационных сетях [2, 3].

АНАЛИЗ ПОДХОДОВ К РЕШЕНИЮ ЗАДАЧ АВТОМАТИЧЕСКОЙ ОБРАБОТКИ ТЕКСТОВ

Автоматическая обработка текстов (АОТ) предполагает решение многих задач, которые условно можно разбить на два уровня. Задачи высокого уровня представлены задачами распознавания речи, реферирования текстов, генерации документов, машинного перевода, извлечения информации, обучения языку, т. е. приложениями. К задачам низшего уровня относят грамматический разбор, снятие смысловой многозначности, корректировку орфографии и синтаксический разбор, т. е. задачи собственно лингвистической обработки ЕЯ [4, 5]. К настоящему времени этот круг задач значительно расширился и в целом охватывает всю индустрию развития и поддержки компьютерной формы существования ЕЯ.

Главная проблема при решении указанных задач состоит в необходимости обрабатывать неструктурированные тексты. Единый типовой алгоритм их автоматической обработки создать не удается, поскольку конкретный вид алгоритма, в первую очередь, определяется строем языка [3].

Существует два подхода к решению задач АОТ [6]. Первый разрабатывается в рамках искусственного интеллекта и носит название *ИИ-подхода.* Он основан на предположении, что вычислительная система для успешной обработки ЕЯ должна быть в состоянии привлекать обширные ресурсы знаний о мире и делать логические выводы на основании этих знаний. Второй подход, *инженерно-лингвистический (ИЛ-подход),* сформировался в компьютерной лингвистике и основан на концепции воспроизводящей инженерно-лингвистической модели. Эта модель предполагает прохождение всего пути лингвистического исследования: от общего, приближенного, основанного на изучении больших объемов реальных текстов ЕЯ описания языкового явления через его структурное, формальное описание к алгоритмическому воспроизведению описываемого явления в рамках той или иной системы автоматической обработки ЕЯ [7]. В литературе ИИ-подход называется также традиционным, рационалистическим, алгебраическим, а ИЛ-подход – эмпирическим, вероятностно-статистическим, информационно-статистическим [3, 4, 8].

Разработка методов решения задач АОТ фактически до конца 1980-х гг. находилась под прямым влиянием точки зрения исследователей в области ИИ, что привело к распространению и укреплению мнения о преемственности только ИИ-методологии обработки ЕЯ. Ведь автоматическая обработка (переработка) и ее теоретический фундамент – структурно-математическое языкознание – родились и формировались в тот период, когда ведущие лингвисты-теоретики были увлечены идеей дискретно-атомарного строения языка и его знаков [3].

В 1950-е гг. предпринимались попытки применить эмпирические (статистические) методы к анализу ЕЯ [9]. Целью подобных исследований было получение алгоритмической методологии для выведения структуры языка на основе анализа лексической и синтаксической информации по реальным текстам. Главное внимание уделялось использованию распределительной информации как средству для изучения языка (например, окружения, в котором может появляться слово).

Интерес к распределенным лингвистическим исследованиям начал слабеть после важных работ N. Chomsky 1957 г. [10] и 1959 г. [11], посвященных разработке основ лингвистической теории трансформационных порождающих грамматик (ТГ-грамматик). Развитие N. Chomsky генеративной лингвистики и его критика существующих эмпирических подходов к обработке ЕЯ сместили фокус к традиционным (рационалистическим) методам, с их акцентом на символических грамматиках и врожденных лингвистических знаниях, т. е. *универсальной грамматике.* Математический, а точнее, механистический взгляд на язык основывался на предположении, что ЕЯ представляет собой исчисление, описываемое с помощью аппарата теории множеств, алгебры отношений и математической логики [3]. Это дало основание надеяться, что язык может быть описан набором конечных правил порождения цепочек символов – предложений – и соответственно правилами определения верности порождаемых конструкций с точки зрения их принадлежности к системе языка. При этом в стороне остался вопрос приемлемости таких правильных конструкций с точки зрения смысла. Создаваемые на основе предложенных N. Chomsky ТГ-грамматик алгоритмы оказались эффективными для анализа лишь ограниченных «подмножеств» ЕЯ, обнаружив свою недостаточность для языка в целом. Выявились принципиальные ограничения на описание структуры и механизмов языка и непригодность ТГ-грамматик для описания его семантики [12].

Попытки построить сильные системы ИИ с лингвистическим обеспечением на основе универсальных закрытых семантических языков и ТГ-грамматик не дали положительных результатов [13]. На практике оказалось, что все далее продвигающаяся и требующая все больших человеческих усилий формализация отдельных деталей языка и речи не столько улучшала, сколько «зашумляла» конечный результат. Это привело к кризису размерности и постепенному свертыванию исследований [3].

Тем не менее в течение 1970-х гг. развивались системы ИИ, в большинстве явно или неявно основанные на порождающих грамматиках, которые уже заранее ориентировались на обработку определенного класса всех «хорошо сформированных» предложений и только их. Ведь при достаточно ограниченной предметной области семантические ограничения всегда могут приводить к однозначному результату.

В тот период также получили развитие более сложные механизмы анализа, такие как обобщенная структурно-фразовая грамматика и расширенные сети переходов. Однако разработка подобных систем оставалась трудоемкой, требующей значительных усилий со стороны инженерии знаний, зависящих от предметной области. Тестирование осуществлялось на искусственно смоделированных лингвистических

ситуациях. Но в реальном языковом материале непрерывно появляются различного рода отклонения от существующих стандартных правил [6], поэтому эти системы не могли адекватно функционировать вне ограниченных задач, для которых они были разработаны.

В начале 1980-х гг. наблюдается постепенный рост числа работ по автоматическому извлечению лингвистических знаний непосредственно из текста, учитывающих статистические характеристики лексических и грамматических единиц языка и его разновидностей. Эти работы базировались в значительной степени на двух широко доступных корпусах текстов: Brown Corpus и Lancaster-Oslo-Bergen (LOB) Corpus [4]. Первые успехи в применении корпусов текстов были получены при выполнении морфологического анализа – назначения подходящего лексико-грамматического класса каждому слову в предложении с достаточно высокой точностью (более 95 %). К концу 1980-х гг. успех статистических методов распространился и на другие области обработки ЕЯ: синтаксический и семантический анализ, синтез речи, информационный поиск. Применение этих методов позволило успешно решить проблемы снятия смысловой многозначности, разрешения проблемы анафор (например, интерпретация местоимений), сегментации дискурса и других.

Следует отметить, что именно доступность большого количества данных явилась основополагающей причиной для возрождения интереса к эмпирическим исследованиям ЕЯ. И если в начале 1980-х гг. Brown Corpus со своим 1 млн слов считался очень большим, то уже сегодня в наличии имеются выборки текстов размером до сотен миллионов и даже миллиардов слов. Эти тексты сегодня широко доступны благодаря достижениям по сбору этих данных таких организаций, как Association for Computational Linguistics' Data Collection Initiative (ACL/DCI), European Corpus Initiative (ECI), ICAME, British National Corpus (BNC), Linguistic Data Consortium (LDC), Consortium for Lexical Research (CLR), Electronic Dictionary Research (EDR) и достижениям Text Encoding Initiative (TEI) по стандартизации представления текстовой информации [14].

Эмпирические исследования обусловили потенциальную возможность решать следующие важные и связанные между собой проблемы [4]:

– *сбора знаний* – идентификации и кодирования всех необходимых знаний автоматически;

– *покрываемости* – объяснения всех явлений в данной области или приложении;

– *робастности* – приспособления реальных данных, которые содержат шум и различные неучтенные аспекты, к выбранной модели;

– *расширяемости* – легко выполняемого расширения или переноса системы на новое множество данных или на новую задачу или предметную область.

Эмпирические методы позволяют применять реальные тексты для целевого обучения и, следовательно, автоматически настраивать выполнение на конкретную задачу. Автоматизированные обучающие средства обеспечивают извлечение релевантных знаний непосредственно из данных. Если обучающие данные избыточны и представляют все релевантные явления, эмпирические методы, пытающиеся оптимизировать работу над полным обучающим множеством, помогают обеспечить адекватную покрываемость. Эти методы вырабатывают вероятностную оценку для каждого случая, классифицируя все возможные альтернативы. Этот более гибкий подход улучшает робастность с помощью аккумулирования шума и обеспечивает выбор привилегированного результата даже в том случае, когда основная модель неадек-

ватна. Эмпирические методы делают ставку на автоматическое повторное обучение на дополнительных данных или данных из различного распределения или новой области, что способствует обеспечению расширяемости [4].

Еще одна характеристика, отличающая эмпирические методы, касается типа данных, требуемых для обучения. Многие системы используют *контролируемые* методы, основанные на *аннотированных* текстах, в которых экспертом каждому слову приписан соответствующий код части речи или семантического смысла, а каждому предложению дан соответствующий синтаксический разбор или семантическое представление. В других системах применяются *неконтролируемые* методы на основе неаннотированных текстов. Обучение с использованием неконтролируемых методов обычно более сложное, поскольку необходимо наличие косвенной обратной связи, подтверждающей различные предположения, например, что все предложения в тексте – грамматически правильные.

В процессе разработки систем АОТ приходится работать с трудно наблюдаемыми, а иногда вообще ненаблюдаемыми лингвистическими связями и объектами типа семантических множителей, вероятностно-информационных характеристик слова и особенно системных связей плана содержания. Поэтому единственно возможным приемом здесь является метод моделирования, позволяющий создавать *воспроизводящие инженерно-лингвистические модели*, которые должны не только объяснять и предсказывать лингвистические явления, но и воспроизводить языковые и речевые объекты. Воспроизводящее инженерно-лингвистическое моделирование является наиболее развитой и завершенной формой «микромоделирования» языка и механизмов текстообразования [7]. Каждая модель должна быть подвергнута проверке с точки зрения ее строгости и объяснительной силы. Поэтому теоретическим фундаментом ИЛ-подхода к решению задач АОТ должна стать теория моделей, воспроизводящих именно *реальные* объекты языка и речи.

ЗАКЛЮЧЕНИЕ

В заключение следует отметить достижения в развитии вычислительной техники и информационных технологий, которые делают реальным проведение исследований с помощью ИЛ-подхода [15]:

– доступность относительно недорогих автоматизированных рабочих мест, оборудованных компьютерами с достаточным быстродействием и объемом памяти для анализа большого количества данных;

– существование больших корпусов лингвистических и лексических данных (словари, тезаурусы, тексты) для обучения и тестирования систем;

– возрастание спроса на коммерческие системы, обрабатывающие огромные объемы электронных текстов, в связи с развитием информационных сетей и повышением их мобильности при работе в многоязычной среде;

– появление информационных технологий, способных при решении определенного класса задач АОТ достигать довольно высокого уровня точности при работе с реальными данными.

Ориентация на природу ЕЯ позволяет преодолевать сбойные и тупиковые ситуации, возникающие на том или ином уровне обработки текста ЕЯ.

ЛИТЕРАТУРА

1. *Белоногов, Г. Г.* Компьютерная лингвистика и перспективные информационные технологии / Г. Г. Белоногов [и др.] // Научно-техническая информация. Сер. 2. 2004. № 8. С. 30–43.

2. *Забежайло, М. И.* Интеллектуальный анализ данных – новое направление развития информационных технологий / М. И. Забежайло // Научно-техническая информация. Сер. 2. 1998. № 8. С. 6–17.

3. *Пиотровский, Р .Г.* Автоматическая переработка текста: теория и практика к концу XX в. / Р. Г. Пиотровский // Научно-техническая информация. Сер. 2. 1998. № 5. С. 26–36.

4. *Brill, E.* An Overview of Empirical Natural Language Processing / E. Brill, R. J. Mooney // AI magazine. 1997. Vol. 18, № 4. P. 13–24.

5. *Grishman, R.* Computational Linguistics: An Introduction / R. Grishman. Cambridge etc.: Cambridge University Press, 1986. 193 p.

6. *Совпель, И. В.* Инженерно-лингвистические принципы, методы и алгоритмы автоматической переработки текста / И. В Совпель. Минск: Вышэйшая школа, 1991. 118 с.

7. *Пиотровский, Р. Г.* Инженерная лингвистика и теория языка / Р. Г. Пиотровский. Л.: Наука, 1979. 112 с.

8. *Шереметьева, С. О.* Теоретические и методологические проблемы инженерной лингвистики / С. О. Шереметьева // Научно-техническая информация. Сер. 2. 1998. № 2. С. 1–9.

9. *Stolcke, A.* Linguistic Knowledge and Empirical Methods in Speech Recognition / A. Stolcke // AI magazine. 1997. Vol. 18, № 4. P. 25–32.

10. *Хомски, Н.* Синтаксические структуры / Н. Хомски // Новое в лингвистике. М.: Изд-во иностр. лит-ры, 1962. Вып. 2. С. 412–527.

11. *Chomsky, N.* Review of Skinner's Verbal Behavior / N. Chomsky // Language 35. 1959. P. 26–58.

12. *Котов, Р. Г.* Лингвистика и информационная технология / Р. Г. Котов // Лингвистические вопросы алгоритмической обработки сообщений. М.: Наука, 1983. С. 84–96.

13. *Пиотровский, Р. Г.* Инженерная лингвистика и проблемы «искусственного интеллекта» / Р. Г. Пиотровский //Лингвистические проблемы «искусственного интеллекта». М., 1980. С. 157–189.

14. *Church, Kenneth W.* Introduction to the Special Issue on Computational Linguistics Using Large Corpora / Kenneth W. Church, Robert L. Mercert // Computational Linguistics. Special Issue on Using Large Corpora: I. 1993. Vol. 19, № 1. P. 1–24.

15. Software Infrastructure for Natural Language Processing / Cunningham H., Humphreys K., Gaizauskas R., Wilks Y. // Proc. of the 5th Conference on Applied Natural Language Processing. 1997. 8 p. – Mode of access: http://xxx.lanl.gov/ps/cmp-lg/9702005.