СОСТОЯНИЕ И ПЕРСПЕКТИВЫ РАЗВИТИЯ КОМПЬЮТЕРНОЙ ЛИНГВИСТИКИ

Хайрова Н.Ф.
Национальный технический университет "Харьковский политехнический институт",
г. Харьков, ул. Пушкинская, 79/2, тел. 707–64–60,
e-mail:nina_khajrova@yahoo.com

В пятидесятые годы прошлого столетия на стыке компьютерных наук, искусственного интеллекта и лингвистики выделилось новое научное направление Natural language processing (NLP), изучающее взаимодействие между компьютером и естественным языком. Но активное развитие компьютерной лингвистики началось после 1954 года, когда ученые университета Georgetown и компании IBM впервые показали систему машинного перевода, которая успешно осуществила автоматический перевод нескольких предложений с русского языка на английский.

Такой успех Джорджтаунского эксперимента (Georgetown experiment) позволил ученым утверждать, что в ближайшие несколько лет произойдет полный переход от интеллектуального к машинному переводу. И хотя до настоящего времени это не произошло, и вряд ли произойдет в ближайшем будущем, сегодняшние успехи прикладной и компьютерной лингвистики в области разработки систем автоматической обработки текстов на естественном языке очевидны. Среди задач, связанных с использованием NLP следует выделить:

- классификацию и кластеризацию текстов;
- информационный поиск текстовых документов;
- тематическое индексирование и рубрицирование;
- извлечение фактов и понятий;
- автоматическое реферирование и квазиреферирование и др.

Существующий уровень реализаций приложений NLP можно разделить на три больших категории: направления, по которым есть коммерческие продукты и системы; направления, в реализации которых задействованы работающие алгоритмы; и практически не решенные на сегодня задачи лингвистического процессора.

Так, например, существует достаточное количество коммерческих приложений по определению спама в получаемом пользователем потоке e-mail сообщений. И, хотя эти приложения продолжают принимать достаточное количество ошибочных решений, они включены в современные почтовые приложения. Аналогичным образом, достаточно коммерциализированы системы синтеза речи (speech synthesis), имеющие успешно разработанные приложения, подобно спецификации SALT (Speech Application Language Tags) [1].

В то же время, сегодня существует большая группа задач обработки текстов, в которых достигнут достаточный прогресс, но полностью они пока не решены. К таким задачам относятся системы извлечения информации

(Information extraction), включающие задачи Opinion Mining. Такие системы используются, в том числе при маркетинговом исследовании, для извлечения позитивной или негативной информации о том или ином продукте или сервисе в Веб. Методы Opinion Mining, позволяющие классифицировать тексты по тональности, появились сравнительно недавно, и во многом используют подходы Text Mining и Information Retrieval [2].

К этой же группе задач относится уже упомянутое приложение лингвистического процессора — машинный перевод, подразумевающий полностью автоматический переводчик. В автоматически получаемых переводах, несмотря на достигнутый в последнее время прогресс в развитии технологии парсинга и статистических параллельных переводных баз данных, до сих пор встречается достаточное количество ошибок.

Последнее, третье направление задач лингвистического процессора является на сегодня практически не решенным. Так, вопросно-ответные системы, которые используются для автоматического ответа на вопросы любого вида, имеют реальные алгоритмы только для формулировки самых простых фактографических вопросов. Простые, но общие вопросы, по-прежнему остаются тяжелой проблемой.

Аналогичным образом вопросно-ответные системы, используемые в интеллектуальном интерфейсе современных автоматизированных информационных систем, существенно отстают от уровня практической востребованности [3] и относятся к третьей группе практически не решенных задач лингвистического процессора.

Еще одним, практически не решенным направлением лингвистического процессора, является перефразирование (paraphrases). Система, использующая перефразирование, должна обладать мощными средствами семантического смыслового анализа, практически не существующими на сегодняшний день. Еще одна задача, которая к сложностям перефразирования добавляет проблемы смыслового обобщения — задача автоматического реферирования (summarization). В настоящее время по-прежнему на рынке присутствуют только системы квазиреферирования, использующие по большей части статистико-позиционные подходы, а алгоритмов «истинного» смыслового обобщения практически нет [4].

Наибольшую сложность и наименьшее количество решений представляют диалоговые системы, которые могут отвечать на вопросы и интерпретировать имеющуюся ситуацию. Это системы, для работы которых существуют только первичные варианты алгоритмов [5].

Проведенный анализ показывает, что, несмотря на то, что работы в направлении автоматизации обработки естественного языка продолжаются более 50 лет и интенсифицировались в последние годы, когда накоплены огромные полнотекстовые информационные массивы, и лингвистические технологии из средств, занимающихся разработкой формальных моделей языка, переходят в фактор производства, в настоящий момент можно выделить четко очерченный класс приложений лингвистического процессора, которые не решены или решены лишь частично (для узких предметных областей).

Невозможность быстрого решения проблемы автоматической обработки текстов на естественном языке обусловлена двумя основными (фундаментальными) причинами:

- задача разработки лингвистического процессора относится к сложно формализуемым задачам, связанным с неопределенностью;
- практически все существующие проблемы обработки текстов на сегодня связаны с проблемой смыслового анализа и необходимостью формализации понимания смысла текстовой информации.

Для решения задач NLP кроме лингвистических знаний необходимо обладать знаниями о мире и уметь формализовать объединение этих знаний. С этим связана необходимость подхода к задачам NLP с точки зрения системного анализа, рассматривая язык как сложно формализуемую иерархическую систему, в которой для формализации семантики нужно моделировать функции человеческого интеллекта по пониманию и идентификации знаний в естественно-языковых текстах.

Список литературы

1. *Daniel Jurafsky, James H. Martin*. Speech and language processing. An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. Second Edition. – Prentice Hall, 2008. – 988 p.

2. *Nozomi Kobayashi, Kentaro Inui and Matsumoto Yuji*. Opinion Mining from Web Documents: Extraction and Structurization. Journal of Japanese society for artificial intelligence, special issue on data mining and statistical science, 2007. – Vol. 22. – № 2. – P. 227–238.

3. *Юсупов Р. М.* Научно-методологические основы информатизации / Р. М. Юсупов, В. П. Заболотский. – СПб.: Наука, 2000. – 455 с

4. Ландэ Д. Глубинный анализ текстов технология эффективного анализа текстовых данных [Электронный ресурс] //CHIP Ukraine. – №10. – 2003. – Режим доступа: http://www.visti.net/~dwl/art/dz/.

5. *Кокорева Л.В.* Диалоговые системы / Л. В. Кокорева, О. Л. Перевозчикова, Е. Л. Ющенко. – Ин-т кибернетики АНУ. – К.: Наук. думка, 1993.– 448с