УДК 81′322.2

ИНФОМАЦИОННЫЕ РЕСУРСЫ КОМПЬЮТЕРНОЙ ЛИНГВИСТИКИ

Д. С. Семак

Учреждение образования «Гомельский государственный технический университет имени П. О. Сухого», Республика Беларусь

Представлен обзор современных методов и технологий, используемых для компьютерной обработки естественного языка. Описаны различные лингвистические ресурсы, включая лингвистические словари, базы словосочетаний, онтологии и тезаурусы, грамматики и корпусы текстов. Подчеркнута значимость лингвистических ресурсов для создания высококачественных систем автоматической обработки текстов и извлечения информации.

Ключевые слова: лингвистические ресурсы, компьютерная лингвистика, лингвистические словари, корпус текстов.

LINGUISTIC RESOURCES OF COMPUTER LINGUISTICS

D. S. Semak

Sukhoi State Technical University of Gomel, the Republic of Belarus

The article presents an overview of modern methods and technologies used for computer processing of natural language. Various linguistic resources are described, including linguistic dictionaries, phrase bases, ontologies and thesauri, grammars and text corpora. The importance of linguistic resources for the creation of high-quality automatic text processing and information extraction systems is emphasized.

Keywords: linguistic resources, computer linguistics, linguistic dictionaries, text corpus.

Лингвистические информационные ресурсы (ЛИР) являются центральным элементом компьютерной лингвистики. Они используются для разработки алгоритмов и прикладных программ для обработки языковой информации. Постоянный рост объема информации различных баз знаний требует современных средств их автоматической обработки на основе новых подходов и технологий. Актуальным вопросом становится разработка соответствующих современных технологий обработки языковой информации.

Разработка лингвистической информационной системы требует определенного представления языковой информации о необходимом обрабатываемом языке. Данная информация отражается в разнообразных лингвистических ресурсах, к которым обращаются на всех этапах разработки и использования любой информационной системы.

Цель статьи – обзор основных видов лингвистических информационных ресурсов в компьютерной лингвистике, перспективы их развития.

Информационный ресурс представляется интеллектуальным результатом коллективного творчества, который можно разделить на пассивную (словари, книги, журналы и др.) и активную (программы, базы данных, модели и др.) формы. Рассмотрим некоторые примеры пассивной и активной форм лингвистических информационных ресурсов.

Так, традиционной формой представления лексической информации являются разные виды словарей. Например, один из популярных в компьютерной лингвистике видов – ***морфологические словари***, которые предназначены для систематизации и

описания основных морфологических свойств слов языка. Они включают в себя информацию о грамматических категориях, склонении и спряжении слов, а также указания на их синтаксическое использование. Основное назначение морфологических словарей заключается в том, чтобы помочь пользователям разобраться в лексической и грамматической структуре слова. С его помощью можно узнать о возможности использования слова в различных контекстах.

Довольно часто используются ***словари синонимов*** в лингвистических информационных системах, которые предназначены для систематического описания синонимических групп и рядов. В данных словарях указываются смысловые и стилистические различия между синонимами, условия их взаимозаменяемости в различных контекстах.

Пользуется популярностью и ***словари паронимов***, в которых описывают слова, имеющие сходство в морфологическом составе и, следовательно, в звучании, но различающиеся по значению. Данный тип словарей является относительно новым видом, поскольку изучение паронимов началось значительно позже. С употреблением паронимов в речи связаны многочисленные ошибки, что подчеркивает практическую значимость данного типа словаря.

Тезаурусы и онтологии принято относить к более сложным видам лексических ресурсов. ***Тезаурус*** – это семантический словарь, в котором представлены смысловые связи слов по линии синонимических отношений, родо-видовых, ассоциативных и др. ***Онтология*** – это формальная спецификация сущностей в определенной области знаний. Она описывает определенный класс объектов, их свойства и взаимосвязи. Онтология является форматом, который позволяет структурировать знания об определенном объекте или области, описывая его сущность, свойства и отношения между ними. При создании онтологии на базе имеющейся в языке лексики они будут относиться к лингвистическим.

Примером подобной лингвистической онтологии можно назвать систему WordNet [1, с. 281] – это лексическая база данных, разработанная в 1985 г. с целью облегчения исследований области семантики, а именно: для создания большого количества связей между словами. Каждое слово в Wordnet имеет набор синонимов, показывающих его возможные значение для конкретной положенной в основу группы слов. Кроме того, каждое слово имеет свойство галереи, которая является списком ассоциированных с ним слов, связанных с его значением. Wordnet дает возможность исследовать и анализировать исходные значения слова, исследовать между ними различные связи, такие, как разницу между значениями двух слов, отношение между значением слова и его синонимами и даже выявление каких-либо неявных свойств, не упомянутых в лексических данных.

Схема английского WordNet была использована при создании аналогичных лексических ресурсов для других европейских языков, объединенных под общим названием EuroWordNet.

Еще один вид лексических ресурсов – ***базы словосочетаний***, в которые отбираются наиболее типичные словосочетания конкретного языка. Такая база словосочетаний русского языка (около миллиона единиц) составляет ядро системы «Кросс-Лексика».

В связи с нарастающей автоматизацией процессов сбора и хранения информации появились весьма объемные текстовые базы данных, которые поддаются систематизации. Применение автоматизированного подхода к созданию словаря позволяет говорить не только об ускорении процесса его создания, но и о значительном улучшении полноты, качества и корректности представленной в нем информации.

Словари могут быть различны своими единицами, структурой, охватом лексики (словари терминов конкретной проблемной области, словари общей лексики и т. д.). Необходимость решения широкого круга задач вызывает необходимость использовать различные виды словарей.

Лингвистические ресурсы включают еще и грамматики естественных языков, которые бывают совершенно разных видов, в зависимости от синтаксической модели, используемой процессором. ***Грамматика*** – это набор правил, представляющих общие синтаксические свойства слова или группы слов. Общее количество грамматических правил также зависит от синтаксической модели и варьируется от десятков до сотен. По существу, проблема заключается в соотношении модели языка грамматики и лексики: чем больше информации представлено в словаре, тем короче может быть грамматика, и, наоборот.

Обратим внимание на то, что процесс создания компьютерных словарей, тезаурусов и грамматик – довольно сложный и трудоемкий. Поэтому одной из прикладных задач компьютерной лингвистики является автоматизация построения лингвистических ресурсов [2, с. 520].

Формирование компьютерных словарей зачастую происходит путем конвертации текстовых словарей. Однако в случае необходимости создания словаря для стремительно развивающихся научных областей необходимо прибегнуть к использованию ресурсов корпусов текстов.

Корпус текстов – это совокупность текстов, собранная в единое целое по определенному принципу (по жанру, авторской принадлежности и т. д.). Отличительной особенностью является разметка текста, соответственно все тексты обладают определенными лингвистическими аннотациями (морфологической, синтаксической и др.). На сегодняшний день существует более ста корпусов для разных естественных языков и с различной разметкой.

С развитием новых технологий, таких, как машинное обучение и искусственный интеллект, создание и использование лингвистических ресурсов становится более доступным и эффективным. Появляются новые методы создания ресурсов, такие как веб-скрэппинг, и новые источники данных, такие, как социальные сети и чатботы. Возникают новые типы ресурсов, такие, как корпусы социальных сетей и голосовые корпусы.

Важный аспект развития лингвистических ресурсов – стандартизация их создания и использования. В настоящее время существует ряд стандартов, таких, как ISO 24617-1, который определяет общие принципы и методы создания лингвистических ресурсов. Однако необходима дальнейшая работа в этом направлении, чтобы обеспечить более эффективное использование ресурсов и более точные сравнения между различными системами и методами.

Таким образом, лингвистические ресурсы являются необходимым компонентом компьютерной лингвистики, обеспечивая основу для автоматического анализа и обработки языковых данных. Однако для их создания и поддержания требуется значительное количество времени, усилий и финансовых ресурсов. Важным направлением развития в данной области становится создание больших корпусов для машинного обучения и использование технологий искусственного интеллекта для улучшения качества языковых моделей и инструментов. В целом развитие лингвистических ресурсов по-прежнему остается одной из ключевых задач компьютерной лингвистики для достижения высокоэффективной обработки естественного языка.

Литература

1. Синтаксический анализ предложения в системе англо-русского вероятностного машинного перевода / М. В. Данейко [и др.] // Частн. вопр. автомат. анализа текстов / М. В. Данейко [и др.]. – Минск, 1972. – С. 279–289.
2. Kushal, D. Mining the peanut gallery: opinion extraction and semantic classification of product reviews / D. Kushal, L. Steve, M. David // Proceedings of WWW. – 2003. – P. 519–528.