М. И. СВЯТОЩИК

СЕМАНТИЧЕСКАЯ РАЗМЕТКА КАК ОДНО ИЗ ПРИОРИТЕТНЫХ НАПРАВЛЕНИЙ СОВРЕМЕННОЙ АВТОМАТИЧЕСКОЙ ОБРАБОТКИ ЯЗЫКА

Семантический анализ языка – одна из наиболее важных и перспективных задач современной вычислительной лингвистики. В качестве решения данной задачи выступает процесс создания системы, способной понимать и порождать тексты на естественном языке – это большой шаг на пути к созданию искусственного интеллекта.

Во-первых, одним из основных и наиболее естественных для человека способов передачи информации является языковая коммуникация, а создание прозрачного языкового интерфейса между людьми и компьютерами позволило бы значительно повысить эффективность их взаимодействия.

Во-вторых, тексты традиционно используются для накопления и передачи знаний, и автоматический анализ текстов позволил бы достичь эффективности в оперировании этими знаниями. В контексте развития глобальной сети Интернет и быстрого роста объёмов информации, представленной в машиночитаемом виде, эта задача приобретает ещё большую актуальность.

На сегодняшний день не существует систем, способных к полноценному анализу, интерпретации и порождению текстов на естественном языке, однако учёные уже достигли значительного прогресса в этой области.

В современной инженерии знаний можно выделить две области, наиболее активно вовлечённые в исследование языка: автоматическая обработка языка и компьютерная лингвистика [1, с. 45].

Семантическая разметка в современном понимании возникла в начале XXI. Теоретической основой для направления служат **теория семантических падежей** Ч. Филлмора [2], **универсальный сеантический код** В. В. Мартынова [3] и его финализация в **теории автоматического порождения архитектуры знаний** А. Н. Гордея [4], которые описывали ролевые инвентари и задавали общую семантическую модель, на основе которой производится анализ ситуаций.

Первые системы автоматической (семантической) разметки были созданы для английского языка, который на тот момент обладал наиболее обширными ресурсами и развитой инфраструктурой. Со временем ресурсы стали создаваться и для других языков, однако английский язык до сих пор сохраняет первенство в плане качества разрабатываемых систем и их применения в реальных приложениях. Исторически многие методы автоматической обработки языка были созданы на базе английского и затем перенесены на другие языки [3, с. 87].

Теория семантических падежей, изложенная в трудах Ч. Филмора, предъявляет к набору семантических падежей следующие требования:

- **полнота** и **уникальность** – *каждый* аргумент глагола имеет *ровно один* падеж;

• **единственность заполнения** – каждый падеж может быть заполнен только один раз;

• **независимость** и **атомарность** – определение семантического падежа не должно зависеть от конкретного выбранного предиката и от других падежей, который имеет категориальную природу и не может быть разделен на компоненты.

На основании этих критериев Ч. Филлмор предложил следующий инвентарь семантических падежей:

• **агенс** – одушевленный инициатор события, способный по своей воле его прекратить;

• **пациенс** – партиципант, наиболее вовлеченный в событие и претерпевающий наиболее значительные изменения;

• **бенефактив** – участник, чьи интересы наиболее затронуты в ходе ситуации.

• **экспериенцер** – получатель информации при глаголах восприятия;

• **стимул** – источник информации при глаголах восприятия;

• **инструмент** – неодушевленный объект, с помощью которого осуществляется действие, но который при этом не претерпевает изменений;

• **адресат** – получатель сообщения при глаголах речи;

• **источник** – место, из которого осуществляется движение;

• **цель** – место, в которое осуществляется движение [2, с. 25].

Семантический падеж характеризуется, с одной стороны, синтаксическим оформлением, а с другой – лексическими ограничениями на его заполнение. Одно из наиболее важных свойств семантического падежа в прикладном отношении – устойчивость к трансформациям, например:

[Константин] Агенс *сломал [стул]* Пациенс;
[Стул] Пациенс *был сломан [Константином]* Агенс.

В ходе дальнейших исследований семантических падежей выяснилось, что предложенный Ч. Филлмором набор обладает ограниченными описательными возможностями и ни одно из указанных выше свойств не является абсолютным. Основные проблемы, с которыми столкнулась теория – это проблема **фрагментации падежей** и проблема **неструктурированности падежного инвентаря**. Проблема фрагментации падежей связана с тем, что для повышения описательной силы теории при соблюдении теоретических требований к ним приходится вводить новые семантические падежи, что, в свою очередь, приводит к снижению описательной силы теории. Рассмотрим следующие предложения:

[Иван] Агенс *готовит [мясо]* Пациенс *на [огне]* Инструмент;
[Иван] Агенс *готовит [мясо]* Пациенс *в [котле]* Инструмент;
* *[Иван]* Агенс *готовит [мясо]* Пациенс *в [котле]* Инструмент *на [сковородке]!* [Инструмент]!
[Иван] Агенс *готовит [мясо]* Пациенс *в [котле]* Инструмент *на [огне]* ??

Для того, чтобы принцип единственности заполнения падежа соблюдался, в данном случае требуется вводить новый падеж, который при этом поступает непосредственно в ролевой инвентарь и получает универсальное определение, в результате чего ролевой инвентарь растёт [5, с. 18].

Подобных проблем не возникает в теории автоматического порождения архитектуры знаний (ТАПАЗ) А. Н. Гордея, в которой семантические роли не вводятся декларативно (эмпирически), в выводятся алгоритмически при помощи специальной алгебры [6, p. 12–24], что позволят установить вектрый переход между ролями. «Упорядоченный ТАПАЗ-алгеброй векторный лист ролей индивидов (ролевой лист ТАПАЗ) представляет собой следующий набор: *субъекта (инициатора → вдохновителя → распространителя → вершителя) → инструмента (активатора → суппрессора → усилителя → преобразователя) → медиатора (ориентира → локуса → транспортёра → адаптера → материала → макета → фиксатора → ресурса → стимула → регулятора → хронотопа → источника → индикатора) → объекта (покрытия → корпуса → прослойки → сердцевины) → продукта (заготовки → полуфабриката → прототипа → изделия)*, где: *субъект* – инициатор акции, *объект* – реципиент акции, *продукт* – результат воздействия субъекта на объект (индивид, адаптированный к заданной роли в новой акции), *инструмент* – исполнитель акции (ближайший к субъекту индивид), *медиатор* – посредник акции (ближайший к объекту индивид). Разновидности субъекта: *инициатор* – инициирует акцию, *распространитель* – распространяет акцию, *вдохновитель* – вовлекает в акцию, *вершитель* – завершает акцию производством из объекта продукта. Разновидности инструмента: *активатор* – непосредственно воздействует на медиатор, *суппрессор* – подавляет сопротивление медиатора, *усилитель* – наращивает воздействие на медиатор, *преобразователь* – преобразует медиатор. Разновидности медиатора: *ориентир* – ориентирует воздействие на объект, *локус* – локализует объект в пространстве, *транспортёр* – перемещает объект, *адаптер* – приспосабливает инструмент к воздействию на объект, *материал* – используется в качестве объекта-сырья для производства продукта, *макет* – является исходным образцом для производства из объекта продукта, *фиксатор* – превращает переменный локус объекта в постоянный, *ресурс* – питает инструмент, *стимул* – проявляет параметр объекта, *регулятор* – служит инструкцией в производстве из объекта продукта, *хронотоп* – локализует объект во времени, *источник* – обеспечивает инструкциями инструмент, *индикатор* – отображает параметр воздействия на объект или параметр продукта как результата воздействия на объект. Разновидности объекта: *покрытие* – внешняя изоляция оболочки индивида, *корпус* – оболочка индивида, *прослойка* – внутренняя изоляция оболочки индивида, *сердцевина* – ядро индивида. Разновидности продукта: *заготовка* – превращённый в сырьё объект, *полуфабрикат* – наполовину изготовленный из сырья продукт, *прототип* – опытный образец продукта, *изделие* – готовый продукт» [7, с. 10, 15–16]. Мы также оказали посильную помощь в интерпретации некоторых формул ролевого листа ТАПАЗ [1].

Автоматическая обработка языка как раздел искусственного интеллекта возникла примерно в то же время. Непосредственной задачей автоматической обработки языка было обеспечение понимания и интерпретации текстов на естественном языке. Первые системы автоматической обработки языка были основаны на правилах и часто представляли собой формализацию

[1] Нашу работу по данной теме см. [8].

той или иной лингвистической или логической теории. Со временем стало понятно, что подобные системы обладают рядом недостатков, в числе которых высокая стоимость разработки, трудность адаптации к новым языкам и новым типам текстов, недостаточная гибкость. На сегодняшний день большинство систем автоматической обработки языка основываются на статистике и машинном обучении.

Основные задачи автоматической обработки языка – создание и оценка модулей анализа языка (лемматизация, морфологический, синтаксический и дискурсивный анализ). Многие задачи языкового анализа на сегодняшний день переформулированы как общие задачи машинного обучения и выполняются практически без использования знаний об исходных лингвистических моделях, на основе которых они были сформулированы.

Автоматическая разметка семантических ролей – одно из приоритетных направлений в современной автоматической обработке языка. Это тип высокоуровневого анализа текста, при котором для исходного текста на естественном языке порождается поверхностная интерпретация на основе теории семантических ролей [9, с. 65].

Предположим, что дано предложение на естественном языке, и в этом предложении выбран некоторый **предикат** (например, глагол). Задача семантической разметки состоит в том, чтобы найти участников ситуации описанной данным предикатом, и приписать им **семантические роли**.

Например, предложение *«Стас купил грушу за 10 рублей»* в ТАПАЗ–2 будет проанализировано следующим образом:

[Стас] Субъект *[отдал]* Акция / Макропроцесс (73) ‘Подведение’ *[продавцу]* Ориентир *[10 рублей]* Объект → *[Продавец]* Субъект *[принял]* Акция / Макропроцесс (60) ‘Присоединение’ *[10 рублей]* Объект → *[Продавец]* Субъект *[отдал]* Акция / Макропроцесс 73 ‘Подведение’ *[Стасу]* Ориентир *[грушу]* Объект → *[Стас]* Субъект *[принял]* Акция / Макропроцесс (60) ‘Присоединение’ *[грушу]* Объект.

Синтаксический анализ – строгая процедура, которая опирается на контекстносвободные и контекстнозависимые грамматики того или иного естественного языка и предполагает конечный однозначный результат. Семантическая разметка – гораздо более сложная процедура, в которой главную роль играет реконструкция ролевой структуры события, глубина которой зависит от конкретной прикладной задачи.

ЛИТЕРАТУРА

1. *Батура, Т. В.* Семантический анализ и способы представления смысла текста в компьютерной лингвистике / Т. В. Батура // Программные продукты и системы. – 2016. – № 4. – С. 45–57.
2. *Fillmore, Ch. J.* Frame semantics and the nature of language / Ch. J. Fillmore // Annals of the New York Academy of Sciences : Conference on the Origin and Development of Language and Speech. – Berkley, California, 1976. – 43 p.
3. *Мартынов, В. В.* Основы семантического кодирования. Опыт представления и преобразования знаний / В. В. Мартынов. – Минск : ЕГУ, 2001. – 140 с.
4. *Гордей, А. Н.* Принципы исчисления семантики предметных областей / А. Н. Гордей. – Минск : Белгосуниверситет, 1998. – 156 с.
5. *Маркова, М. В.* Автоматическая семантическая разметка предложений английского языка / М. В. Маркова // Альманах современной науки и образования. – 2013. – №6 (73). – 65 с.

6. *Hardzei, A.* Theory for Automatic Generation of Knowledge Architecture: TAPAZ–2 / A. Hardzei; transl. from Rus. I. M. Boyko. – Rev. English edn. – Minsk : Republican Institute of Higher Education, 2017. – 50 p. URL: http://tapaz.by.
7. *Гордей, А. Н.* Семантическая разметка события и её отображение средствами китайского и русского языков / А. Н. Гордей // Иностранные языки в высшей школе / ФГБОУВО «Рязанский государственный университет имени С. А. Есенина». – Рязань : Редакц.-изд. центр РГУ им. С. А. Есенина, 2021. – Вып. 2 (57). – С. 5–26.
8. *Святощик, М. И.* Перевод многообъектной семантики в однообъектную как способ минимизации семантических вычислений / М. И. Святощик // Учёные записки Витебского государственного университета им. П. М. Машерова : сб. науч. тр. – Витебск, 2020. – Т. 31. – С. 180–184.
9. *Рассел, С.* Искусственный интеллект : современный подход / С. Рассел, П. Норвиг.– М. : ООО «И.Д. Вильямс», 2006. –1408 с.