УДК 004.6, 004.9, 004.827, 004.824

МЕТОДЫ ИНФОРМАЦИОННОЙ СЕМАНТИКИ

Цветков В.Я.

Московский государственный технический университет радиотехники, электроники и автоматики МГТУ МИРЭА, Москва, e:mail cvj2@list.ru

Статья описывает методы информационной семантики. Проводится сравнение естественного языка и языка информатики. Язык информатики использует информационные единицы и семантические информационные единицы. дается различие между информационной единицей и семантической информационной единицей. Показаны три смысловых критерия делимости семантических информационных единиц. Описаны основные функции языка информатики. Показано, что информационная семантика одной из главных целей ставит уменьшение зависимости интерпретации информационных конструкций от человека. Показана универсальность слова как лингвистической информационной единицы. Показана дифференциация информационных единиц по группам информационных технологий. Показано, что проблема анализа в информационной семантике разделяется на техническую и семантическую. Показано, что одной из главный задач информационной семантики является обработка больших объемов информации, а также слабоструктурированной информации, которые для обычного человеческого интеллекта создают неразрешимую проблему.

Ключевые слова: семантика, информационная семантика, информационные единицы, лексические информационные единицы, семантические информационные единицы, когнитивный анализ, семантический анализ, компьютерный анализ семантических единиц

SOLUTION TASKS OF THE SECOND KIND WITH THE USE OF THE INFORMATION APPROACH

Tsvetkov V.Y.

Moscow State Technical University of Radio Engineering, Electronics and Automation MSTU MIREA, Moscow, e:mail cvj2@list.ru

This article describes the methods of information semantics. Article compares natural language and artificial language of computer science. Language informatics uses information units and semantic information units. Article makes a distinction between information units and semantic information units. This article describes three semantic criteria of divisibility of semantic information units. This article describes the basic functions of the language of computer science. Article shows that one of the main objectives is to decrease the semantics of information, depending on the interpretation of the information structures of man. The universality of the word as a linguistic unit of information. Shows the differentiation of information units in groups of information technology. It is shown that the problem of semantic analysis information is divided into technical and semantic. Article shows that one of the main problems is the semantics of information processing large volumes of information and semistructured information. These problems for ordinary human intelligence creates an unsolvable problem

Keywords: semantics, information semantics, information units, lexical information units, semantic information units, cognitive analysis, semantic analysis, computer analysis of semantic units

Существует точка зрения, согласно которой «Информационная семантика – это направление в моделировании смысла фраз на естественном языке, основанное на анализе количества переданной информации» [1]. Недостатком этого подхода является то, что под количеством переданной информации понимают информационный объем передаваемых сообщений. Это обусловлено применением теории информации, созданной на основе работ К.Э. Шеннона [2] Эта теория информации, названая Н. Винером статистической [3], не рассматривает проблемы смысла и семантики. Об этом заявил сам К.Э. Шеннон «Проблемы передачи информации не релевантны семантическим проблемам» [3]. Поэтому использовать инструментарий для изучения семантики, который исключает эту возможность, выглядит странным.

Точка зрения, отражаемая в данной статье, звучит иначе. «Информационная семантика – это направление в информационном моделировании, основанное на применение семантических информационных моделей и семантических информационных единиц и анализа их смысловых и интерпретационных характеристик».

Основная часть. Проведем сравнение естественного языка и языка информатики. При этом примем во внимание то, что основой языка информатики являются информационные конструкции, семантические информационные единицы и структурные информационные единицы.

С когнитивной точки зрения естественный язык (ЕЯ) [4] представляет знаковую систему, отражающую жизненный опыт человека и его взаимодействия с окружением в форме, приспособленной для передачи другим людям и для организации собственного оптимального поведения. С формальной точки зрения естественный язык представляет знаковую систему, содержащую алфавит, совокупность лингвистических единиц-слов, совокупность словарей интерпретирующих эти слова, совокупность правил употребления и интерпретации этих слов.

С формальной точки зрения информационный язык [5] представляет знаковую систему, содержащую информационные единицы (алфавит), совокупность сложных семантических единиц – слов, совокупность тезаурусов интерпретирующих эти семантические единицы, совокупность правил построения и интерпретации семантических информационных единиц. Различие в том, что все слова в естественном языке переносят смысл и имеют информационный объем.

В языке информатики есть информационные единицы – носители информации. Их также называют структурные информационные единицы. И есть информационные единицы, содержащие смысл, которые называют семантические информационные единицы. Как элементы сложной системы – языка, эти информационные элементы характеризуются неделимостью, связанной с критерием делимости. Структурная неделимость приводит к элементу – символ, который специального смысла не имеет. В некоторых случаях структурная неделимость в ЯИ приводит к слову. Смысловая неделимость определяет семантические информационные единицы [6] (СИЕ).

Смысловая сигнификативная неделимость определяет семантическую информационную единицу слово Смысловая предикативная неделимость определяет семантическую информационную единицу предложение. Смысловая ассоциативная неделимость определяет семантическую информационную единицу фразу.

Между перечисленными информационными единицами существуют отношения иерархии. Слово есть совокупность символов. Интерпретация слова осуществляется с помощью словарей и тезаурусов.

Предложение – совокупность слов, выражающих законченную мысль. Интерпретация предложения осуществляется на основе соотнесения его смысла с действительностью.

Фраза совокупность предложений, выражающих законченную мысль, некоторые их которых не могут быть интерпретированы без других предложений в этой фразе. Все выше перечисленной относится в равной степени к естественному языку и к искусственному языку.

Различие в неделимой смысловой единице слово. В языке информатики слово как информационная единица может быть не лексическим объектом и даже структурным объектом. Например, машинное слово – единица обработки информации на компьютере имеющее определенную разрядность: 32 байта, 64 байта, 256 байтов и т.д. Это слово может переносить разную смысловую нагрузку, в некоторых случаях только совокупность машинных слов содержит смысловое значение.

Слова в ЕЯ, в первую очередь, ориентированы на семантическую обработку их человеком. Информационные единицы, в первую очередь, ориентированы на компьютерную обработку и во вторую на семантическую обработку компьютером или человеком.

Для человека ЕЯ выполняет две главные функции: служит средством коммуникации и средством моделирования явлений окружающего мира. Язык информатики (ЯИ) имеет следующие функции: служит средством формализации описаний окружающего мира на основе информационных моделей [7], средством формального построения информационных моделей, средством моделирования явлений окружающего мира, средством коммуникации, средством запоминания информационных моделей и опыта, средством анализа, средством репрезентации информационных моделей. Можно сказать, что язык информатики является более грубым как средство описания. Однако в условиях больших информационных объемов и информационных барьеров, он позволяет решать задачи, которые человек не в состоянии решить с помощью ЕЯ. То есть доминирующей функцией в ЯИ является анализ, в первую очередь, больших информационных массивов.

Моделирования явлений окружающего мира в ЕЯ осуществляется путем запоминания всего множества ситуаций, в которых оказывался человек, и организацией механизмов оперативного извлечения этой информации. ЕЯ позволяет хранить информацию в формализованном виде с помощью лексических единиц слов, что уменьшает искажения интерпретации смысла.

Язык информатики позволяет хранить информацию в формализованном виде с помощью информационных конструкций. Информационная конструкция специфическая формализованная (кодированнная) форма описания и хранения информации, которая обобщает [8]: информационную модель, информационный процесс, семантическую информационную единицу и структурную информационную единицу. Такая форма описания и хранения ориентирована в первую очередь на компьютерную обработку и также уменьшает искажения интерпретации смысла. При этом она позволяет использовать дополнительные (по отношению к человеческому интеллекту) скоростные технологии компьютерной обработки и анализа. Однако слова как лингвистические информационные единицы в ЕЯ являются универсальным средством, а информационные единицы в ЯИ являются специализированными. Слово в ЕЯ

неразрывно связано со смыслом. Слово как информационная единица в ЯИ может быть полисемическим или носителем любого смысла. Слова в ЕЯ слабо структурированы, информационные единицы в ЯИ хорошо структурированы. Однако информационные единицы не являются универсальными, а существуют группами под разные информационные технологии. То есть для репрезентации используют одни информационные единицы, для машинной обработки другие, для хранения в базах данных третьи, для описания четвертые и для коммуникации пятые – группы информационных единиц.

Когнитивный аспект ЕЯ состоит в том, что полноценное понимание ЕЯ достигается вместе с созреванием человека, когда его суммарный лингвистический опыт (СЛО) позволяет интерпретировать около 200 миллионов слов [1]. Это требует десятки лет. Язык информатики, хранимый на носителях информации, является межличностным. Он передается от человека к человеку и требует освоения 1–2 года.

Функционально осмысленными считают интерпретируемые фразы, которые связаны с поведением и целями носителя языка, с моделированием внешнего мира и коммуникацией. Первая функция зависит от субъекта, она является связана с его интеллектом, психическим состоянием, ситуацией в которой он находится и целями его действий. Вторая и третья функции представляют предмет изучения теоретической лингвистики.

Информационная семантика одной их целей ставит задачу снижения зависимости интерпретации от состояния субъекта. Объектом исследований в информационной семантике являются семантические информационные единицы, позволяющие передавать сведения, накапливать опыт и моделировать окружающий мир. Познавательная функция языка информатики также является предметом исследований информационной семантики.

Проблема анализа в информационной семантике разделяется на техническую и семантическую. К числу основных свойств информационных моделей, допускающих возможность обработки и анализа их человеком, относят [9]: обозримость, воспринимаемость, целевую определенность, ситуационную определенность, функциональность, полноту, информационное соответствие, актуальность, точность, регламентированность, ассоциативность, согласованность, надежность.

Остановимся на наиболее важных с точки зрения возможности семантического анализа. *Обозримость* – свойство моделей или информационных коллекций, состоящее в том, что человек (*в рамках своего человеческого интеллекта*) в состоянии обозреть совокупность параметров и связей, входящих в модель и *понять* данную модель как целое. Это свойство у виртуальных моделей значительно выше, чем у реальных объектов. Оно обусловлено возможностью масштабирования визуального пространства. Например, человек, находясь в городе, видит только окружающие его дома. Но, используя электронную карту, навигатор, космический снимок – он увеличивает обозримость и видит то, что в реальности увидеть не может. Соответственно принимаемое им решение более обосновано.

Воспринимаемость – свойство моделей или информационных коллекций, состоящее в том, что человек (*в рамках своего человеческого интеллекта*) в состоянии *воспринять и понять* данную модель как отражение объективной реальности или ее практическое назначение. Воспринимаемость связана с наличием базовых знаний. Чем больше базовых знаний, тем выше воспринимаемость.

Если модель необозрима или не воспринимаема**,** она, как правило, отвергается и не применяется человеком. Если модель воспринимаема одним человеком и не воспринимаема другим человеком, между ними появляется состояние информационной асимметрии.

Ассоциативность – свойство информационных моделей вызывать ассоциации в *когнитивной области и с* одной стороны создавать свободу выбора, с другой стороны развивать творческие начала в субъекта, работающего с такой моделью

Эти свойства связаны с когнитивной областью человека. Следует подчеркнуть, что обозримость и воспринимаемость виртуальных моделей выше, чем реальных моделей окружающего мира. Это создает определенный комфорт при работе с ними. Регламентированность виртуальных моделей делает более предсказуемой виртуальную информационную ситуацию по сравнению с реальной ситуацией. Это также создает ощущение комфорта.

Техническая проблема информационной семантики связана с техническими задачами информационнго поиска в больших массивах данных. Специфической задачей информационной семантики является работа с большими данными [10] и преодоление информационных барьеров [11]. Новое направление в области обработки данных большие данные (Big Data), связывают с «проблемой трех V»: большим информационным объемом (Volume), слабой структурированностью (Variety), требованием высокой скорости обработки (Velocity).

Текстовый контент, содержащийся в информационных потоках сети Интернет, соответствует двум первым характеристикам. Обработка текстового контента ведется статистическими методами в информационно – поисковых системах и при создании семантических сетей. Типичные методы обработки такой информации: кластерный анализ, семантический анализ, контент-анализ. Для многих методов анализа больших объемов информации, основными инструментами становятся высокопроизводительные вычислительные кластеры, которые, работая в многопоточном режиме, могут дать многократное ускорение за счет количества установленных в кластере процессоров и разделения задачи на части [12].

В последние годы широко применяют автоматические системы обработки текстов, основанные на методах и алгоритмах компьютерной лингвистики, которые выполняют лингвистический анализ текстов на естественном языке [13]. Классический лингвистический подход к анализу текста предполагает существование независимых уровней анализа: морфологического, синтаксического и семантического. Данный подход задает последовательность анализа: морфологический, синтаксический, семантический. Методы анализа текстов основываются на правилах, разработанных экспертами-лингвистами. Для создания автоматических систем на основе этих правил требуется разработка модели естественного языка, что в каждом отдельном случае требует больших трудозатрат высококвалифицированных лингвистов и системных операторов.

Альтернативным методом построения модели ЕЯ является метод на основе размеченных лингвистических «корпусов текстов» [13]. При использовании этого метода производится обогащение массивов текстов на естественном языке соответствующей лингвистической информацией, например, морфологической и синтаксической, разметкой именованных сущностей. Разработка таких лингвистических ресурсов менее трудоемка, чем разработка модели языка. При использовании «корпусного метода» автоматические лингвистические анализаторы конструируются с использованием методов машинного обучения. Корпус текстов – совокупность текстов, собранных в соответствии с определёнными принципами, размеченных по определённому стандарту [14].

В результате применения машинного обучения происходит обобщение частных примеров, представленных в лингвистическом корпусе текста, при этом конструируются общие, качественные и во многих случаях эффективные процедуры обработки и анализа текстов. В целом это направление информационной семантики больше связано с семантической теорией информации [15].

Заключение

Современное развитие информационной семантики происходит по разным направлениям. Одним из доминирующих является попытка статистического анализа информации, в частности с использованием энтропийных методов оценки информации по К.Э. Шеннону. Другое направление связано с постановкой задач в рамках семантической теории информации. оно использует понятия когнитивного моделирования [16] , семантического окружения [17] и понятия информационных единиц [7].

Список литературы

1. Информационная семантика – Викизнание http://www.wikiznanie.ru/ru-wz/index.php.

2. Shannon C.E. A Mathematical Theory of Communication, Bell System Technical Journal, vol. 27, 379–423 & 623–656, July & October, 1948.

3. Winner N. Cybernetics or Control and Communication in the Animal and the Mashine. The Technology Press and John Wiley & Soris Inc. New York – Herman et Cie, Paris, 1948. – 194 p.

4. Заболеева-Зотова А.В. Естественный язык в автоматизированных системах. Семантический анализ текстов. – Волгоград: РПК «Политехник», 2002.

5. Цветков В.Я. Язык информатики // Успехи современного естествознания. – 2014. – № 7. – С. 129–133.

6. Цветков В.Я. Информационные единицы сообщений // Фундаментальные исследования. – 2007. – № 12. – С. 123–124.

7. Цветков В.Я. Информационные единицы как средство построения картины мира // Международный журнал прикладных и фундаментальных исследований. – 2014. (Ч. 4) – № 8 – С. 36–40.

8. Tsvetkov V.Ya. Information Constructions // European Journal of Technology and Design, 2014. – Vol.(5). – № 3. – P. 147–152

9. Цветков В.Я. Когнитивные аспекты построения виртуальных образовательных моделей // Интеграция образования. − 2014. − № 3 (76). − С. 71−76.

10. Майер-Шенбергер В., Кукьер К. Большие данные: Революция, которая изменит то, как мы живем, работаем и мыслим. – Манн, Иванов и Фербер, 2014. – 240 с.

11. Цветков В.Я Маркелов В.М., Романов И.А. Преодоление информационных барьеров // Дистанционное и виртуальное обучение. – 2012. – № 11. – С. 4–7.

12. Сигов А.С., Кошкин Д.Е., Дробнов С.Е. Кластеризация текста на основе анализа слов с применением распределенных вычислений // Информатизация образования и науки». – 2011. – № 2(10). – С. 74–80.

13. Казенников А.О. Разработка моделей и алгоритмов для комплекса автоматической обработки и анализа потоков новостных сообщений на основе методов компьютерной лингвистики / Диссертации на соискание степени кандидата технических наук. Специальность 05.13.15. Вычислительные машины, комплексы и компьютерные сети. – М.: МИРЭА, 2014 – 138 с.

14. Jurafsky D., Martin M. Statistical Speech and Language Processing. Prentice Hall, 1999.

15. Шрейдер Ю.А. О семантических аспектах теории информации. Информация и кибернетика. – М.: Советское радио, 1967. – С. 15–47.

16. Tsvetkov V.Ya. Cognitive information models. Life Science Journal 2014. – № 11(4). – P. 468–471.

17. Tsvetkov V.Ya. Semantic environment of information units // European Researcher, 2014. – Vol.(76). – № 6-1. – P. 1059–1065.