УДК 004.9

Е. Е. Чехарин

Информационная модель семантического окружения

Статья исследует информационную модель семантического окружения. Рассмотрены различные подходы семантического и когнитивного анализа. Показано значение моделей информационных единиц при семантическом анализе. Отмечены особенности анализа терминов и предложений на естественном языке. Описаны модели интерпретируемости и не интерпретируемости терминов в теоретико-множественном представлении.

Ключевые слова: философия информации, информация. информационные модели, информационные единицы, семантика, когнитивная семантика, интерпретация

E. E. Chekharin

Information model of the semantic environment

Paper explores the semantic information model environment. Various approaches and cognitive semantic analysis. Shows the importance of models of information units in the semantic analysis. The features of the analysis of terms and sentences in natural language. The models interpretability and interpretability of terms in the set-theoretic representation

Keywords: philosophy of information, information. information models, information units, semantics, cognitive semantics, the interpretation

Введение

Исследованием семантических проблем в информационных объектах занимается семантическая теория информации. В семантической теории информации основной целью ставиться анализ содержательности информационных сообщений и информационных объектов. Этот подход развит в работах Н. Виннера [1], Р. Карнапа [2], Л. Флориди [3] и др. [4]. В этом подходе под информацией понимают содержание информационного объекта безотносительно к его объему. В последнее время в теории информации стали широко применять понятие информационных единиц [5]. В семантической теории информации рассматривают семантический разрыв и семантические информационные единицы [6].

Различные подходы к семантическому анализу

Научное направление, которое занимается анализом языковых единиц, называется когнитивная семантика [7, 8]. Когнитивная семантика занимается исследованием преимущественно естественных языков. Когнитивная семантика предлагает модели фрагментов языковой картины мира [9]: Семантические и когнитивные информационные [8] модели могут быть разными для разных языков, но сопоставимыми по смыслу. Поэтому целесообразно применять семантические единицы в когнитивной семантике.

Статистическая семантика – направление анализа, изучающее смысловое содержание слов и фраз естественного языка (ЕЯ), использующее статистические методы. В 1960 г. Е. Делавней [10] предложил термин «информационная семантика» и определил ее как статистическое изучение смысла слов по их частотности и порядку следования. Информационная семантика заключается в моделировании смысла фраз на естественном языке основанное на анализе количества переданной информации. В ней также слабо применяют информационные единицы.

Следует разделять познавательную (когнитивную), содержательную (семантическую), описательную (информационную) функции ЕЯ. Содержательная функция ЕЯ является главным предметом семантических исследований. Статистический подход предполагает относительный характер смысловых отношений в зависимости от расположения слов. Соответственно можно говорить о теоретической семантике языка, о семантике языка индивидуальных носителей языка и о семантике языка книг и документов.

Большинство современных исследователей семантики ЕЯ связывают смысловые отношения с разбиением речи на предложения. Считается, что предложения ЕЯ выделяют отдельные ситуации, сценарии, episodes, отличающиеся активным началом и замкнутым действием (подлежащим и сказуемым). Связи между двумя разными предложениями (дискурс) считаются слабыми. Каждое предложение устанавливает свою смысловую связь между его компонентами (словами, фразами). Такую семантику называют сентенциальной.

Более глубокую семантику порождает разбиение предложений на семантически обособленные фразы (Синтагма). Это приводит к фразеологической семантике. ЕЯ представляет картину окружающего мира [11] из информационного поля. С этой точки зрения информационное содержание смысла фраз на ЕЯ заключается именно в заведомо неслучайных комбинациях слов. Эти комбинации называют структурными элементами (СЭ). Неделимые СЭ представляют атомы смысла, из которых строятся фразы ЕЯ. Отдельные слова также могут быть структурными элементами, если они повторяются в тексте достаточно часто. При информационном подходе смысл текста на ЕЯ связывается с закономерностями чередования слов и фраз.

Множество структурных элементов, входящих в текст, образует семантическое представление текста. Оно очищено от шумовой компоненты, и сохраняет только неслучайные (статистически значимые, понятийные) элементы описания накопленного опыта – ассоциации. О семантике текстов, основанной на анализе неслучайных цепочек символов, можно говорить как об ассоциативной семантике .

Множество понятий, соединенных между собой, образует семантическую сеть [12]. Понятия могут быть представлены словами или фразами, а их связи могут обозначаться другими словами или фразами, иметь числовое выражение, или могут быть неспецифицированы. Множество понятий, связанных таким образом с данным понятием, называют семантическим окружением [13] или семантическим полем понятия. Число шагов продвижения по семантической сети, необходимое для установления связи с некоторым словом или фразой, можно назвать порядком семантического поля слова или фразы.

Обычно семантические сети конструируют из отдельных слов. Такая сеть содержит значительный элемент случайности и описывает семантику вероятностно-статистического характера. Конструирование семантической сети из структурных элементов позволяет освободиться от случайной составляющей. Сеть, составленная из СЭ, представляет неслучайную информационную модель лингвистического опыта человека. При выделении смысловых связей между словами можно ограничиться связями первого порядка, которые связывают данное слово ''x'' со словами или фразами в предложениях, содержащих ''x''. В случае необходимости учета более глубокой семантики, можно характеризовать эти связи по пересечению семантических полей первого или высших порядков. При этом, однако, будет возрастать размер семантических полей и роль "зашумляющей" общеязыковой семантики

С точки зрения практических приложений для смыслового анализа документов хорошо зарекомендовал себя анализ скрытой семантики, называемый термином «Латентный семантический анализ» (Latent semantic analysis) [14,15].

Этот анализ основан на линейном алгебраическом подходе, и использует приведение матриц к каноническому виду. Его трудоемкость растет кубично с длиной текстов. Рассматривается прямоугольная матрица данных, с числом столбцов n, равным числу разных слов, и со строками, которые представляют семантически обособленные фрагменты текста (называемые концепциями), представленные предложениями, фразами или синтагмами.

Число повторений слова в "концепциях" характеризует их статистическую значимость, и интерпретируется как мера смысла. На столбцах и строках могут быть введены априорные целевые функции (функции интереса) и изучены условия диффузии интереса при движении по матрице.

Далее применяется алгебраическая процедура, которая формирует сингулярное разложение прямоугольной матрицы (Singular value decomposition) . Это разложение разбивает оптимальным образом матрицу на сумму декартовых произведений векторов строк на векторы слов с весами, равными собственным значениям матрицы. Тем самым в неявной форме решается задача кластеризации в пространстве слов и "концепций", что позволяет дать формальное решение для целого ряд задач смыслового анализа.

В их число входит характеризация смысла отдельных слов и фраз, определение смыслового расстояния между ними, выделение слов и фраз, несущих наибольшую смысловую нагрузку, вычисление меры смыслового сходства документов, выбор наиболее значимых частей документа и формирование рефератов по заданному интересу. Основным недостатком этого метода является его формально-математический подход, отсутствие прозрачной интерпретации численных характеристик и основанных на них заключений. Несмотря на то, что выделение СЭ освобождает текст от случайных (шумовых) вкраплений, та информация, которую несут СЭ, может быть неинтересной, если этот СЭ не менее часто используется в более широких контекстах, представляет субъективно авторское изложение или типовую фразу (штамп),

В рамках информационной концепции смысл каждой фразы, каждого предложения и документа определяется лишь только на фоне предыдущего (или объемлющего) текста и измеряется количеством новой информации, которую этот фрагмент несет.

В последние годы значительно развились методы автоматического лингвистического анализа текстов на естественном языке [16]. Классический лингвистический подход к анализу текста предполагает существование относительно независимых уровней анализа; в том числе: морфо-

 ISSN 2307-2447

логического, синтаксического и семантического. Кроме того, он предполагает определенную последовательность анализа, в начале - морфологического, затем синтаксического, и наконец, семантического.

Лингвистические методы автоматического анализа текстов основываются на правилах, разработанных экспертами-лингвистами. Их модели разработки лингвистических ресурсов очень трудоемки, поскольку для создания автоматических систем необходима разработка модели представления значительной части естественного языка, что требует больших трудозатрат высококвалифицированных лингвистов и системных операторов.

Завершая этот анализ следует отметить, что слабым местом семантических теорий является недостаточное использование понятий информационное поле и информационные единицы, включая семантические. Использование этих понятий позволяет по новому представить семантический анализ.

Информационные единицы как инструмент семантического анализа

Информационные единицы – это единицы, которые переносят порции информации безотносительно к содержанию или характеризуют содержание порции информации безотносительно к информационному объему. Как базовые элементы теории, информационные единицы (ИЕ) обладают свойством неделимости по какому-либо признаку [17]. Информационные единицы служат основой построения сложных: языковых описаний, информационных конструкций и многоцелевого управления [18].

Выделяют составные и простые информационные единицы. Простые ИЕ не включают в свой состав другие единицы. Составные информационные единицы включают в свой состав другие информационные единицы. Например, информационная единица «предложение» включает информационные единицы «слова» [5]. Информационная единица «слово» включает информационные единицы «символы».

Для многих составных информационных единиц имеет место характеристика – структурная вложенность. Структурная вложенность информационных единиц – это не структура, а отношение иерархии компонент единицы и ее окружения.

Для многих составных информационных единиц имеет место характеристика – окружение информационной единицы. Окружение информационной единицы – это другие, связанные с ней информационные единицы и характеристики, необходимые для однозначной интерпретации информационной единицы и ее информационной определенности. Информационное окружение единицы проявляется при ее непосредственном использовании [13]. Например, информационным окружением информационной единицы «слово» в предложении или во фразе, будут все связанные с этим словом символы и другие слова, а также такие информационные характеристики как позиция слова и вид его написания.

Информационное окружение единицы является информационной моделью семантического окружения [13] или семантического поля понятия.

Поле понятий является частью глобального информационного поля [4]. Как показано в [9, 19] информационное поле обладает разрывностью в отличие от физических непрерывных

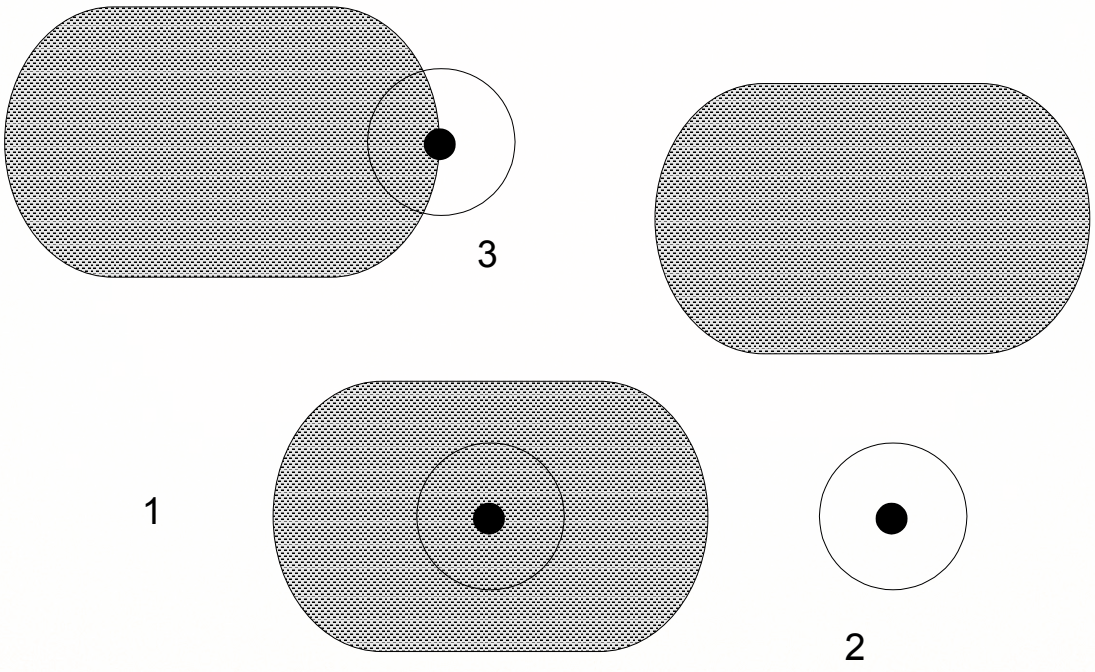

Рис.1. Информационное окружения информационной единицы и ее интерпретируемость.

полей. Если построить иерархию сущностей, связанных с информационными единицами, то получится такая последовательность: информационное поле; информационные совокупности; информационные объекты; информационные единицы. Между этими сущностями существуют различные информационные отношения

Семантические информационные единицы – это единицы, рассматриваемые в аспекте семантической содержательности [20] информационной совокупности или информационного объекта. Выделяют следующие семантические информационные единицы: слово, предложение, фраза. Для этих информационных единиц характерно расслоение или стратификация.

Окружение информационной единицы бывает локальным и глобальным. Глобальное окружение информационной единицы существует в глобальном информационном поле. Локальное окружение определяет семантическое окружение или семантическое поля понятия. Локальное окружение информационной единицы легко моделируется средствами когнитивной графики.

На рис.1 приведена графическая иллюстрация применения информационного окружения семантической информационной единицы. Разрывность информационного поля показана заштрихованными областями. Это области интерпретации информационной единицы. Точка обозначает информационную единицу, в качестве которой может быть семантическая информационная единица, логическое высказывание или термин. Кружок обозначает: информационное окружение; семантическое окружение или область интерпретации.

Показаны три информационные ситуации: 1 – полная интерпретируемость информационной единицы.; 2 не интерпретируемость ИЕ; 3 – частичная интерпретируемость ИЕ.

Область интерпретируемости (1) определиться как множество, для которого информационное окружение ИЕ является подмножеством. Частичная область интерпретируемости (3) определиться как область к которой семантическое окружение принадлежит частично. Область не интерпретируемости (2) определиться как множество, для которого информационное окружение ИЕ не является подмножеством и не принадлежит этому множеству.

Заключение

Примененная методика описания семантического информационного окружения дает возможность разрабатывать алгоритмы поиска области истинности при заданной информационной единице и ее окружении. Приведенная модель информационного окружения позволяет формировать представлять информационные конструкции любой сложности в виде совокупности связанных информационных единиц. Совокупности связанных информационных единиц дают возможность оценки морфологической и смысловой сложности языковых конструкций. В отличие от классического системного анализа данный подход допускает разные критерии делимости. Разные критерии делимости контента влекут появление разных информационных единиц. Семантический й анализ контента целесообразно выполнять с использованием локального семантического окружения информационных единиц.

ЛИТЕРАТУРА

1. 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. 99 p.
2. Carnap R. et al. An outline of a theory of semantic information. Research Laboratory of Electronics, Technical Report №247, MIT, 1952. 49 p.
3. Floridi L. Semantic Conceptions of Information. Available at: http://plato.stanford.edu/entries/information-semantic (accessed 25.08.2014).
4. Цветков В.Я. Семантика информации // Дистанционное и виртуальное обучение. 2012. № 10. С. 4-7.
5. Иванников А.Д., Кулагин В.П., Мордвинов В.А, Найханова Л.В., Овезов Б.Б., Тихонов А.Н. Цветков В.Я. Получение знаний для формирования информационных образовательных ресурсов. М.: ФГУ ГНИИ ИТТ «Информика», 2008. 440с
6. V. Y. Tsvetkov. Information Interaction as a Mechanism of Semantic Gap Elimination // European Researcher, 2013, Vol.(45), № 4-1, p.782- 786.
7. Gries S. T. Corpus-based methods and cognitive semantics: The many senses of to run // Trends in linguistics studies and monographs. 2006. V. 172. C. 57.
8. Tsvetkov V.Y. Cognitive information models // Life Sci J. 2014; 11(4). pp. 468-471
9. V. Y. Tsvetkov. Worldview Model as the Result of Education // World Applied Sciences Journal. 2014. 31(2). p.211-215.
10. Delavenay E. An Introduction to Machine Translation, New York, Thames and Hudson, 1960.
11. Цветков В.Я. Естественное и искусственное информационное поле // Международный журнал прикладных и фундаментальных исследований. 2014. №5. С.178-180.
12. Тихонов А.Н., Иванников А.Д., Цветков В. Я. Терминологические отношения // Фундаментальные исследования. 2009. № 5. С.146-148.
13. Цветков В.Я., Черкашин Е.Е. Окружение информационных единиц // Вестник МГТУ МИРЭА. 2014. №2. С.36-42.
14. S.Deerwester, S.Dumas, G.Furnas, T.Landauer, and R.Harshman, Indexing by Latent Semantic Analysis, J.Amer. Soc. For Information Science, 1990.
15. Thomas K., Landauer T., Harshman R. Latent semantic analysis, J. Amer. Soc. of Information Science, 1990, 41(6).
16. Nivre J. Algorithms for Deterministic Incremental Dependency Parsing. Computational Linguistics. 2008. 4 (34). pp. 513-553.
17. Поляков А.А., Цветков В.Я. Прикладная информатика: учебно-методическое пособие для студентов, обучающихся по специальности «прикладная информатика» (по областям) и другим междисциплинарным специальностям: В 2-х частях: / Поляков А.А., Цветков В.Я.; Под общ.ред. А.Н. Тихонова. М.: МАКС Пресс. 2008.
18. V. Ya. Tsvetkov. Multipurpose Management // European Journal of Economic Studies 2012, Vol.(2), № 2. pp.140-143.
19. Tsvetkov V.Y. Information field // Life Science Journal. 2014. №11(5). pp.551-554.
20. V. Ya. Tsvetkov. Semantic Information Units as L. Floridi's Ideas Development // European Researcher, 2012, Vol.(25), № 7, p.1036-1041.

REFERENCES

1. 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. 99 p.
2. Carnap R. et al. An outline of a theory of semantic information. Research Laboratory of Electronics, Technical Report no.247, MIT, 1952. 49 p.
3. Floridi L. Semantic Conceptions of Information. Available at: http://plato.stanford.edu/entries/information-semantic (accessed 25.08.2014).
4. Tsvetkov V.Ia. Semantics of the information. *Distantsionnoe i virtual'noe obuchenie - Distance and virtual learning*, 2012, no.10, pp.4-7 (in Russian).
5. Ivannikov A.D., Kulagin V.P., Mordvinov V.A, Naikhanova L.V., Ovezov B.B., Tikhonov A.N. Tsvetkov V.Ia. *Poluchenie znanii dlia formirovaniia informatsionnykh obrazovatel'nykh resursov* [Obtaining knowledge to develop information and educational resources]. Moscow, Informika, 2008. 440 p.
6. V.Y.Tsvetkov. Information Interaction as a Mechanism of Semantic Gap Elimination. *European Researcher*, 2013, Vol.(45), no.4-1, pp.782-786.
7. Gries S.T. Corpus-based methods and cognitive semantics: The many senses of to run. *Trends in linguistics studies and monographs*, 2006, V. 172, p.57.
8. Tsvetkov V.Y. Cognitive information models. *Life Sci J*, 2014, no.11(4), pp.468-471.
9. V.Y.Tsvetkov. Worldview Model as the Result of Education. *World Applied Sciences Journal*, 2014, no.31(2), pp.211-215.
10. Delavenay E. An Introduction to Machine Translation. New York, Thames and Hudson, 1960.
11. Tsvetkov V.Ia. Natural and artificial information field. *Mezhdunarodnyi zhurnal prikladnykh i fundamental'nykh issledovanii - International journal of applied and fundamental research*, 2014, no.5, pp.178-180.
12. Tikhonov A.N., Ivannikov A.D., Tsvetkov V.Ia. Terminological relations. *Fundamental'nye issledovaniia - Fundamental research*, 2009, no.5, pp.146-148 (in Russian).
13. Tsvetkov V.Ia., Cherkashin E.E. Environment information items. *Vestnik MGTU MIREA - Vestnik MSTU MIREA*, 2014, no.2, pp.36-42 (in Russian).
14. S.Deerwester, S.Dumas, G.Furnas, T.Landauer, and R.Harshman, Indexing by Latent Semantic Analysis. *J.Amer. Soc. For Information Science*, 1990.
15. Thomas K., Landauer T., Harshman R. Latent semantic analysis. *J. Amer. Soc. of Information Science*, 1990, 41(6).
16. Nivre J. Algorithms for Deterministic Incremental Dependency Parsing. *Computational Linguistics*, 2008, no.4(34). pp.513-553.
17. Poliakov A.A., Tsvetkov V.Ia. *Prikladnaia informatika Poliakov A.A., Tsvetkov V.Ia.: Uchebno-metodicheskoe posobie dlia studentov, obuchaiushchikhsia po spetsial'nosti «prikladnaia informatika» (po oblastiam) i drugim mezhdistsiplinarnym spetsial'nostiam* [Applied Informatics: Textbook for students majoring in "applied Informatics" (by regions) and other interdisciplinary majors]. Moscow, MAKS Press. 2008.
18. V.Ya.Tsvetkov. Multipurpose Management. *European Journal of Economic Studies*, 2012. Vol.(2), no.2, pp.140-143.
19. Tsvetkov V.Y. Information field. *Life Science Journal*, 2014, no.11(5), pp.551-554.
20. V.Ya.Tsvetkov. Semantic Information Units as L. Floridi's Ideas Development. *European Researcher*, 2012, Vol.(25), no.7, pp.1036-1041.

Информация об авторе

Чехарин Евгений Евгеньевич
(Москва, Россия)
Старший преподаватель. Московский государственный технический университет радиотехники, электроники и автоматики
E-mail: cvj7@mail.ru

Information about the author

Chekharin Evgenii Evgen'evich
(Russia, Moscow)
Senior lecturer
Moscow State Technical University of Radio Engineering, Electronics and Automation
E-mail: cvj7@mail.ru