Инженерный вестник Дона, №3 (2015)
ivdon.ru/ru/magazine/archive/n3y2015/3096

Анализ метода выявления синонимических рядов, соответствующих одинаковым понятиям

Ю.А. Киселёв

Уральский федеральный университет имени первого Президента России Б.Н. Ельцина, Екатеринбург

Аннотация: В статье анализируется метод, позволяющий выявлять синонимические ряды, описывающие одинаковые понятия. Получена оценка качества этого метода на основании опроса носителей русского языка; точность анализируемого метода составляет 73 %. Применение данного метода к данным открытого тезауруса русского языка YARN выявило необходимость в повышении качества синонимических рядов этого ресурса.
Ключевые слова: лексический ресурс, словарь, Викисловарь, краудсорсинг, тезаурус, синонимия, синонимический ряд, семантические отношения, мера сходства, русский язык

Введение

Два выражения являются *синонимичными*, если замена одного выражения на другое никогда не меняет истинность утверждения, в котором была произведена такая замена. Понятно, что таких выражений (слов) существует немного. Поэтому понятие синонимии обычно уточняется: два выражения являются синонимичными в лексическом контексте *C*, если замена одного на другое в *C* не меняет истинности выражения [1]. Это позволяет более конструктивно интерпретировать слова и устанавливать синонимические отношения.

Отношение синонимии является чрезвычайно важным в языке. Например, оно служит основой для построения тезаурусов (далее под *тезаурусом* будем понимать особый вид словаря, отражающий семантические отношения между словами) – основной структурной единицей тезаурусов является *синсет*[1], т.е. синонимический ряд.

Считается, что синонимический ряд задаёт смысл, т.е. определяет некоторую концепцию [2]. В отличие от многих других лексических ресурсов входом в тезаурус является не слово, а понятие (или синсет).

[1] Синсет (от англ. synset – set of synonyms) – множество синонимов.

Понятно, что в тезаурусе концепции должны быть представлены уникальным образом. Однако в случае, если ресурс достаточно большой, убедиться в этом может быть весьма сложно. Тем не менее, существует предположение, что пара синонимов задаёт некоторый смысл [2]. Его использование может помочь выявить сходные синсеты с целью дальнейшей очистки лексических ресурсов от синсетов с одинаковыми значениями.

В этой связи в настоящей статье предлагается метод, позволяющий на основе анализа слов синсетов, сделать вывод о сходстве их значений, и анализ его точности.

Обзор ресурсов, содержащих в своём составе синонимические ряды

В словарях синонимов, как следует из названия, лексика сгруппирована с использованием отношения синонимии: слова формируют синонимические ряды, и каждому ряду, как определяющему некоторый смысл, даётся определение. Такая форма ресурса позволяет эффективно вводить отношения между понятиями (то есть синсетами), так как каждое понятие встречается в ресурсе только один раз. В отличие, например, от толковых словарей, где слова, описывающие одинаковые концепции, встречаются в разных статьях.

Словари синонимов создаются профессиональными коллективами экспертов-лексикографов, поэтому в них отсутствует проблема того, что одно понятие может встретиться несколько раз. Однако для многих других ресурсов эта проблема является чрезвычайно актуальной. Рассмотрим такой лексикографический ресурс как *Викисловарь* [3]. Это многоязычный электронный словарь и тезаурус. Он содержит разную словарную информацию о заголовочных словах: морфологические сведения, семантические свойства, в том числе, определения, синонимы и другие.

Важно отметить принцип наполнения этого ресурса: Викисловарь является *краудсорсинговым* ресурсом, то есть его пользователи совместно вносят изменения, связанные как с содержимым, так и со структурой ресурса.

И хотя Викисловарь наполняется обычными пользователями, не-экспертами, он обладает достаточно высоким качеством, что было проанализировано в работе [4]. Русская версия Викисловаря по многим аспектам качества также не уступает традиционным лексическим ресурсам [5].

Тем не менее, анализ этого ресурса показал, что синсеты, входящие в его состав, могут описывать одинаковые понятия и при этом различаться. Например, в статье «Малодушие» есть синсет «*малодушие*, *трусость*, *нерешительность*, *безволие*». При этом в статье «Трусость» приведён другой синсет: «*трусость*, *боязливость*, *малодушие*». Эти синсеты описывают одинаковые понятия, но состоят из разного количества слов, часть из которых отличается. Очевидно, что на основании только этих двух имеющихся синсетов нужно было сформировать следующий более полный «*трусость*, *боязливость*, *малодушие*, *нерешительность*, *безволие*», который можно было включить в обе статьи.

Викисловарь не единственный ресурс, обладающий отмеченным недостатком. Рассмотрим другой краудсорсинговый ресурс: большой открытый электронный тезаурус (ЭТ) русского языка *YARN* [6]. Он создается в Уральском федеральном университете совместно с Высшей школой экономики с 2013г. Авторы ресурса разработали интерфейс, позволяющий его пользователям самим формировать синсеты [7]. Из-за особенностей интерфейса и того, что ресурс наполняется не-экспертами, в тезаурусе YARN тоже могут быть синсеты, описывающие одинаковые понятия.

В отличие от Викисловаря, наличие в YARN таких синсетов является существенной проблемой из-за того, что это осложняет дальнейшее введение семантических отношений в ресурс. (В Викисловаре отношения вводятся между отдельными словами (статьями), а не понятиями, поэтому наличие неполных синсетов, как и их дубликатов, хотя и является недостатком, но не столь существенным). При этом качество синонимических рядов само по

себе важно для решения многих задач, где могут применяться синонимы, например, в задаче тематической классификации документов (см., напр., [8, 9]).

В этой связи необходимо разработать метод, который позволил бы выявлять синсеты, описывающие одинаковые понятия, и оценить его качество.

Метод выявления синонимических рядов, соответствующих одинаковым понятиям, и его оценка

Авторы первого ЭТ *Princeton Wordnet* [10], анализируя подходы к определению слов, отмечают, что в большинстве случаев для идентификации слова достаточно одного его синонима [2], то есть пара синонимов задаёт смысл. Это наблюдение широко используется в толковых словарях, где в качестве определений часто приводится ровно один синоним.

Следуя данному утверждению, сформулируем *критерий эквивалентности синсетов*: если синсеты содержат хотя бы два одинаковых слова, то они разделяют общий смысл, т.е. являются *эквивалентными*. Этот критерий использовался в работе [5] для определения доли общих смыслов, представленных в различных лексических ресурсах, таких как ЭТ и словари синонимов. Однако верификация данного критерия не была проведена.

Отметим, что данный критерий имеет практическую значимость, так как его применение может выявлять сходные синсеты в ресурсах. Это может помочь в удалении идентичных концепций, представленных различных образом и повысить полноту отображения соответствующих концепций за счёт слияния действительно сходных синсетов в один.

В этой связи задача проверки верности данного критерия является актуальной и осуществляется в статье далее. Для оценки точности критерия сходства синсетов был проведён следующий эксперимент.

Из тезауруса YARN были получены все синсеты[2] и оставлены те из них, для которых есть эквивалентные. Согласно нашим расчётам, в YARN (по состоянию на 10 июля 2015г.) содержалось 44 тыс. синсетов, связывающих 54 тыс. слов. Из них почти 28 тыс. пар эквивалентных синсетов, среди которых 15,7 тыс. различных синсетов. Таким образом, YARN содержит до 35 % синсетов с одинаковым смыслом.

Случайным образом было выбрано 100 пар эквивалентных синсетов s и s_e, таких что:

$$s \in S,\ s_e \in \mathrm{S_e},\ |s| < |s_e|,$$

где || обозначает мощность множества и соответствует количеству слов в синсете. Затем из большего синсета s_e каждой пары выбирались слова *words*, отсутствующие в меньшем s. Согласно критерию эквивалентности синсетов, эти слова являются кандидатами на включение в меньший синсет. Все отобранные синсеты содержали 3–7 слов включительно (нижняя граница гарантировала, что синсет формировался не «случайно», и он определяет некоторую концепцию; верхняя граница отфильтровывала в достаточной степени полные синсеты).

Затем был проведён опрос, в котором участникам (из числа носителей русского языка) предлагалось выбрать слова из *words*, добавление которых в синсет s не искажает смысл. Очевидно, что если участник выбирал какие-то слова, это означало, что синсеты действительно имеют общий смысл. Из-за того, что участники опроса не являлись экспертами, мы собрали 3 оценки для каждого синсета, т.е. всего было получено 300 оценок. Считалось, что слово необходимо добавить в синсет, если его выбрало более одного человека.

[2] Исходные данные доступны по адресу http://russianword.net/yarn-synsets.csv.

Хотя бы одно слово было добавлено в 73 синсета, при этом всего было добавлено 159 слов. Это говорит о том, что каждый синсет, который был пополнен, увеличился в среднем более, чем на 2 слова (159 / 73 ≈ 2,2). Это свидетельствует о том, что в 73 случаях из 100 синсеты, которые мы считаем эквивалентными, действительно описывают одинаковые понятия. В таких случаях синсет, из которого выбирались слова на добавление, может быть удалён из ресурса, как дубликат модифицированного синсета.

Заключение

Проведённый опрос носителей русского языка позволил сделать вывод о качестве предлагаемого метода определения синсетов, отражающих одинаковые концепции. Оценка точности метода получена на основе точности критерия эквивалентности синсетов, который формулируется следующим образом: синсеты, содержащие хотя бы пару одинаковых слов, разделяют общий смысл. Точность этого критерия составляет 73 %. На основании этой оценки и количества эквивалентных синсетов в тезаурусе YARN можно сказать, что до 25 % его синсетов описывают понятия, представленные в ресурсе неуникальным образом.

Важно отметить, что проведённый анализ качества был осуществлён без привлечения специалистов из числа профессиональных лексикографов. За счёт получения трёх оценок на один синсет и их последующего мажорирования оказалось достаточно участия одних лишь носителей языка.

Проведённое исследование выявило необходимость в очистке синсетов тезауруса YARN от синсетов-дубликатов. В этой связи метод повышения качества синсетов, созданных с помощью краудсорсинга, является предметом дальнейших исследований.

Благодарности

Исследование выполняется при финансовой поддержке РГНФ (проект № 13-04-12020 «Новый открытый электронный тезаурус русского языка») и научной группы «Разработка методов анализа, обработки, визуализации и прогнозирования многомерных данных для современных информационных систем» Уральского федерального университета им. первого Президента России Б.Н. Ельцина.

Литература

1. Jarmasz M. Roget's Thesaurus and Semantic Similarity. Proc. of Conf. on Recent Advances in Natural Language Processing. Borovets: John Benjamins Publishing Company, 2003. 212–219 pp.
2. Fellbaum C. WordNet: An Electronic Lexical Database. Cambridge: 1998. 447 p.
3. Викисловарь URL: wiktionary.org.
4. Meyer C. M., Gurevych I. Wiktionary: A new rival for expert-built lexicons? Exploring the possibilities of collaborative lexicography. Electronic Lexicography. Oxford: Oxford University Press, 2012. 259–291 pp.
5. Kiselev Y., Krizhanovsky A., Braslavski P., et al. Russian Lexicographic Landscape: a Tale of 12 Dictionaries. Computational Linguistics and Intellectual Technologies: papers from the Annual conference "Dialogue". Moscow: RGGU, 2015. 254–271 pp.
6. Yet Another RussNet URL: russianword.net.
7. Braslavski P., Ustalov D., Mukhin M. A Spinning Wheel for YARN: User Interface for a Crowdsourced Thesaurus. Proc. of the Demonstrations at the 14th Conf. of the EACL. Gothenburg: ACL, 2014. 101–104 pp.
8. Киселёв Ю.А. Перспективы использования жанровой классификации Веб документов в поисковых системах. Инженерный вестник Дона. 2012. №4–2 URL: ivdon.ru/ru/magazine/archive/n4p2y2012/1425.

9. Красников И.А., Никуличев Н.Н. Гибридный алгоритм классификации текстовых документов на основе анализа внутренней связности текста. Инженерный вестник Дона. 2013. №3 URL: ivdon.ru/ru/magazine/archive/n3y2013/1773.

10. WordNet URL: wordnet.princeton.edu.

References

1. Jarmasz M. Roget's Thesaurus and Semantic Similarity. Proc. of Conf. on Recent Advances in Natural Language Processing. Borovets: John Benjamins Publishing Company, 2003. 212–219 pp.

2. Fellbaum C. WordNet: An Electronic Lexical Database. Cambridge: 1998. 447 p.

3. Vikislovar' [Wiktionary] URL: wiktionary.org.

4. Meyer C. M., Gurevych I. Wiktionary: A new rival for expert-built lexicons? Exploring the possibilities of collaborative lexicography. Electronic Lexicography. Oxford: Oxford University Press, 2012. 259–291 pp.

5. Kiselev Y., Krizhanovsky A., Braslavski P., et al. Russian Lexicographic Landscape: a Tale of 12 Dictionaries. Computational Linguistics and Intellectual Technologies: papers from the Annual conference "Dialogue". Moscow: RGGU, 2015. 254–271 pp.

6. Yet Another RussNet URL: russianword.net.

7. Braslavski P., Ustalov D., Mukhin M. A Spinning Wheel for YARN: User Interface for a Crowdsourced Thesaurus. Proc. of the Demonstrations at the 14th Conf. of the EACL. Gothenburg: ACL, 2014. 101–104 pp.

8. Kiselev Yu.A. Inženernyj vestnik Dona (Rus), 2012, №4–2 URL: ivdon.ru/ru/magazine/archive/n4p2y2012/1425.

9. Krasnikov I.A., Nikulichev N.N. Inženernyj vestnik Dona (Rus), 2013, №3 URL: ivdon.ru/ru/magazine/archive/n3y2013/1773.

10. WordNet URL: wordnet.princeton.edu.