Синсеты верхнего уровня в тезаурусе RuWordNet

В.Д. Соловьев[1], С.С. Ахтямова[2]

[1]*Казанский федеральный университет, Кремлевская 18, Казань, Россия, 420101*
[2]*Казанский национальный исследовательский технологический университет, Карла Маркса 68, Казань, Россия, 420015*

Аннотация
В статье описывается один метод верификации тезауруса русского языка RuWordNet. Рассматриваются синсеты верхнего уровня (без гиперонимов). Показано, что многие из них не представляют общие понятия и, следовательно, структура RuWordNet должна быть уточнена. Описаны общие стратегии введения дополнительных семантических отношений для устранения обнаруженных пробелов, приведены конкретные примеры рекомендуемых пополнений RuWordNet.

Ключевые слова
тезаурус, синонимы, гиперонимы, русский язык

1. Введение

В последнее время в целом ряде задач автоматической обработки текстов все чаще используются онтологии и тезаурусы. Для английского языка хорошо известна созданная более 30 лет назад лексическая онтология WordNet. Для русского языка к настоящему времени завершенной и эксплуатируемой является только одна версия – тезаурус RuWordNet [1]. Тезаурус RuWordNet содержит более 110 тыс. слов и словосочетаний русского языка, объединенных в синонимические ряды, сокращенно именуемые синетами. Между синсетами установлен следующие семантические отношения: гипо-гиперонимия, отношение к домену, часть-целое (для существительных), антонимия (для прилагательных), причинно-следственные отношения (для глаголов) и другие [1]. Высококачественные языковые ресурсы подобного рода создаются вручную, имеют сложную структуру и большой объем. Поэтому существует потребность в средствах их верификации.

В последние годы для решения этой задачи применялось несколько подходов. В статьях [2, 3] в корпусе текстов выбирались слова, близкие по дистрибутивной семантике в корпусе, но далекие в смысле расстояния в тезаурусе. Затем проводился экспертный анализ причин этого несоответствия. В статье [4] рассматривались расстояния в тезаурусе между словами квазисинонимами (по словарю Ю.Д. Апресяна). Если расстояние оказывалось велико, больше 4 шагов по семантическим отношениям, то выявлялись причины этого. В статье [5] структура синсетов в RuWordNet сопоставлялась с данными психолингвистических экспериментов, в которых носителям языка предлагалось перечислить синонимы заданных слов. В этой работе, как и в предыдущих, обнаруженные несоответствия использовались для поиска пробелов тезаурусе и предлагались пути их устранения.

В данном исследовании мы предлагаем еще один подход к верификации тезаурусов, ранее не рассматривавшийся. Он основан на анализе вершин тезауруса самого верхнего уровня.

2. Методы и результаты

В данной статье мы анализируем синсеты RuWordNet верхнего уровня. В онтологиях общего характера элементы верхних уровней репрезентируют наиболее общие сущности [6] и их обычно бывает немного. В случае, если синсет верхнего уровня представляет частные понятия, скорее всего пропущена его связь, как гипонима, с некоторым более общим синсетом. В RuWordNet обнаружено 1273 синсета без гиперонимов. Распределение по частям речи следующее. Существительные – 9, глаголы – 166, прилагательные – 1098. Как видим,

значительное большинство случаев связано с прилагательными, что объясняется отсутствием хороших классификаций прилагательных. Этот вопрос требует отдельного исследования и здесь затрагиваться не будет, остановимся на синсетах существительных и глаголов. Таким образом, целью исследования является разработка методов анализа синсетов верхнего уровня и формулировка рекомендаций по введению пропущенных семантических отношений в RuWordNet.

Следующие синсеты существительных не имеют гиперонимов:

1. {ОТНОШЕНИЕ, СВЯЗЬ МЕЖДУ СУЩНОСТЯМИ, СВЯЗЬ, ОТНОШЕНИЕ МЕЖДУ СУЩНОСТЯМИ}
2. {ПОСТОЯННАЯ СУЩНОСТЬ}
3. {ДЛЯЩАЯСЯ СУЩНОСТЬ, ПРОИСХОДЯЩАЯ СУЩНОСТЬ}
4. {ПОЛОЖЕНИЕ, РОЛЬ, МЕСТО}
5. {СОСТАВ, ЦЕЛОЕ, СОВОКУПНОСТЬ}
6. {ЧАСТЬ, СОСТАВНАЯ ЧАСТЬ}
7. {ВАРИАЦИЯ, РАЗНОВИДНОСТЬ, ВАРИАНТ}
8. {НОВИНКА, НОВОСТЬ}
9. {НОВШЕСТВО, НОВАЦИЯ}

Первые 5 синсетов отражают предельно общие понятия, которые обычно и занимают верхнее положение в онтологиях [6]. Синсеты № 6 и 7 также отражают весьма общие понятия и, в принципе, могут быть оставлены на верхнем уровне. Однако, они не занимают топовое положение в других тезаурусах и онтологиях верхнего уровня. Рассмотрим их по отдельности.

Сравним синсет {ЧАСТЬ, СОСТАВНАЯ ЧАСТЬ} с аналогичным синсетом в тезаурусе WordNet {*part, component part*}, используя основное значение при переводе слов с одного языка на другой. В WordNet у этого синсета указан гипероним {*relation*}. В RuWordNet соответствующий синсет {ОТНОШЕНИЕ 1 (отношение между сущностями)} имеет в качестве гипонимов такие синсеты, как {АССОЦИАТИВНАЯ СВЯЗЬ}, {ЧЛЕНСТВО В ОРГАНИЗАЦИИ}, {ПРИНАДЛЕЖНОСТЬ}. На этом же уровне можно включить и {ЧАСТЬ, СОСТАВНАЯ ЧАСТЬ} как гипоним к {ОТНОШЕНИЕ 1}.

Синсет №7 {ВАРИАЦИЯ, РАЗНОВИДНОСТЬ, ВАРИАНТ} близок по значению к синсету {ВАРЬИРОВАНИЕ}, однако никак с ним не связан. Возможное решение состоит в том, чтобы считать первый из них гипонимом по отношению к {ВАРЬИРОВАНИЕ}, имеющему цепочку гиперонивом, восходящую к синсету №1.

Синсеты 8 и 9 соответствуют уже весьма специфическим понятиям. Синсет {НОВИНКА, НОВОСТЬ} не встроен в структуру RuWordNet – у него нет связей с другими синсетами. В WordNet ему соответствует синсет {*novelty, freshness*} с гиперонимом {*originality*}. Предлагается и в RuWordNet привязать {НОВИНКА, НОВОСТЬ} как гипоним к {ОРИГИНАЛЬНОСТЬ 1}. Далее, любое новшество является новинкой и потому к синсету {НОВИНКА, НОВОСТЬ} следует привязать как гипоним синсет {НОВШЕСТВО, НОВАЦИЯ}. Такое решение принято, в частности, в Викисловаре, где у слова *новшество* указан гипероним *новость*.

Перейдем к глаголам. Список 166 синсетов глаголов (глагольных словосочетаний) без гиперонимов приведен на сайте проекта https://kpfu.ru/kompleksnyj-analiz-struktury-i-soderzhaniya-366287.html. В данной статье не может быть приведено исчерпывающее исследование всех 166 глаголов. Приведем в качестве примера одну стратегии связывания этих синсетов с некоторыми гиперонимами.

Если в синсет S входят словосочетания со структурой Veb – Noun, то в качестве гиперонима нужно взять синсет S’, в состав которого входит глагол Veb. Пример. В текущей версии RuWordNet есть синсет без гиперонимов {ИСПОВЕДОВАТЬ БУДДИЗМ}, состоящий из одного словосочетания со структурой Veb – Noun. Привязываем его как гипоним к синсету {ИСПОВЕДОВАТЬ}. Отметим, что этот синсет не имеет гипонимов. Попутно отметим, что в тезаурусе отсутствуют словосочетания *исповедовать ислам*, *исповедовать христианство* и др., т.е. на этом пути удается обнаружить и пробелы тезауруса другого рода.

3. Благодарность

Работа выполнена при поддержке РФФИ, грант № 18-00-01238.

4. Литература

[1] Loukachevitch, N. Comparing Two Thesaurus Representations for Russian / N. Loukachevitch, G. Lashevich, V. Dobrov // Proceedings of Global WordNet Conference. – 2018. – P. 35-44.

[2] Loukachevitch, N. Corpus-based Check-up for Thesaurus / N. Loukachevitch // Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. – Florence: Association for Computational Linguistics, 2019. – P. 5773-5779.

[3] Bayrasheva, V.R. Verification of lexic ontologies by the method of using semantic proximity of words calculated by large corpus of texts / V.R. Bayrasheva, N.V. Loukachevitch // J. Phys.: Conf. Ser. – 2020. – Vol. 1680. – P. 012004.

[4] Solovyev, V. Distribution of Quasi-Synonyms in Thesaurus for Natural Language Processing / V. Solovyev, N. Loukachevitch, V. Bochkarev // Proceedings of the Linguistic Forum: Language and Artificial Intelligence. – 2021.

[5] Solovyev, V. Semantic Similarity of Words in RuWordNet Thesaurus and in Psychosemantic Experiment / V. Solovyev, N. Loukachevitch // Advances in Intelligent Systems and Computing series. – 2021. – Vol. 1358.

[6] Bottazzi, E. Preliminaries to a DOLCE ontology of organisations / E. Bottazzi, R. Ferrario // International Journal of Business Process Integration and Management. – 2009. – Vol. 4(4). – P. 225-238. DOI: 10.1504/IJBPIM.2009.03228.