Федюкова Екатерина Алексеевна

Санкт-Петербургский государственный университет (СПбГУ),
Санкт-Петербург, Россия
katefedyukova@gmail.com

ПРОБЛЕМЫ РАЗМЕТКИ НЕОДНОСЛОВНЫХ ВЫРАЖЕНИЙ ДЛЯ ЭЛЕКТРОННОГО ТЕЗАУРУСА YARN[1]

Ключевые слова: тезаурус, неоднословная единица, словарь, WordNet.

В статье рассматриваются проблемы, возникшие при обработке неоднословных единиц при добавлении их в корпус YARN (Yet Another RussNet). Для ряда проблем предлагаются неалгоритмические решения.

Fedyukova Ekaterina

Saint Petersburg State University (SPbSU),
St. Petersburg, Russia

ISSUES OF MULTIWORD UNIT TAGGING IN COMPUTER THESAURUS YARN

Keywords: thesaurus, multiword unit, dictionary, WordNet.

The article addresses the difficulties raised in the course of adding multiword units to YARN corpus (Yet Another RussNet). Non-algorithmic ways to resolve some of the listed problems are proposed as well.

Введение

В статье пойдёт речь об исследовании, проведённом в процессе добавления в тезаурус YARN (Yet Another RussNet) неоднословных единиц. В ходе исследования рассматриваются сложности, связанные с пригодностью неоднословий для добавления, а также предлагаются неалгоритмические пути их разрешения. Исследование направлено на выявление проблем и поиск их решения, оно должно стать фундаментом для дальнейшей работы над этой темой, в ходе которой планируется создать нетривиальные, т. е. алгоритмические, способы решения проблем.

[1] Исследование ведется при поддержке гранта РГНФ номер 16-04-12019.

Тезаурус YARN и его значение для лингвистики

Тезаурус — это словарь, в котором материал упорядочен по смыслу, а не по алфавиту. Он может отражать множество смысловых отношений, но наиболее существенными являются отношения синонимии, антонимии, гипо-гиперонимии и меронимии [Мухин, 2016]. Особенности тезауруса по сравнению с любым другим словарём очевидны: в нём легче (и вообще возможно) обнаружить отношения между словами, входящими в одну смысловую группу [Мухин, 2016]. В обычном словаре «весна» окажется довольно далеко от «лета», а в тезаурусе они будут находиться рядом. Это делает тезаурус необходимым для решения задач компьютерной лингвистики, информационного поиска, машинного перевода, автоматической обработки текста и других сфер прикладной лингвистики.

YARN (Yet Another RussNet) — один из активно разрабатывающихся в данный момент электронных тезаурусов на русском языке [YARN, 2015]. Он построен по принципу WordNet (WN) — первого электронного тезауруса на английском языке, который был создан в Принстонском университете (США) в середине 1980-х гг. и совершенствуется до сих пор. Такие системы, важные в научном и практическом плане, сегодня есть на многих языках, но на русском ни одна разработка пока так и не была завершена, что определяет актуальность разработки корпуса YARN.

Синсет и его структура

Механизм добавления синсетов в тезаурус достаточно прост, что позволяет делать это любому пользователю, знакомому со структурой словарей. На сайте проекта (russianword.net) есть удобный интерфейс для создания синсета, которым можно начать пользоваться, авторизовавшись при помощи учётной записи в одной из социальных сетей. Также на данный момент разработано приложение YARN для платформы Android. В нём (на момент 27 апреля 2016) проходит тестирование по созданию синсетов (пока только однословных) по методике лингвистических замещений. Работа ведется на основе краудсорсинга, который позволяет разработчикам привлекать большое количество людей к разметке синсетов, а проекту, в свою очередь, развиваться [Braslavski et al., 2014].

В синсет входят: главное слово или словосочетание всего ряда, его синонимы и аббревиатуры, а также определение, которое должно быть достаточно ёмким и обобщать значения всех синонимов.

Необходимо учитывать, что слова, являющиеся синонимами главного слова, не могут вступать с ним в родовидовые отношения — «январь» и «месяц» ни в коем случае не могут быть в одном синсете.

Ход исследования

В ходе исследования был обработан список, сформированный на материалах Википедии и включающий более 500 словосочетаний. Благодаря этому удалось выявить пять проблем, связанных с добавлением и разметкой неоднословных выражений в YARN. Также были предложены и пути их решения.

В основном проблемы связаны с определением пригодности сочетаний для добавления их в тезаурус.

Стоит отметить, что для добавления запрещены имена собственные, нечастотные выражения и свободные сочетания слов.

Описание проблем и путей их решения

Многовариантные сочетания

Первая проблема: некоторые сочетания вида «прилагательное + существительное» допускают очень много равноправных вариантов замены прилагательного, обозначающего признак главного слова. Например, *русский народ* и *русский язык*. Они являются достаточно частотными в употреблении и кажутся носителю вполне допустимыми. Их можно в определённом контексте заменить на однословия *русские* и *русский* соответственно, но всё же для тезауруса эти сочетания — свободные, поэтому они не могут быть добавлены в словарь. Языков, как и народов, очень много, и добавлять в тезаурус все такие сочетания не имеет смысла, ведь следует помнить, что тезаурус — это не корпус текстов, и такие действия, даже если предположить, что они возможны, окажутся бесполезными.

Проблема поиска неоднословия по словарям

Также, как было сказано выше, для добавления однословных синсетов используют поиск по словарям. При этом очевидно, что в большинстве стандартных словарей неоднословия мы не найдём. А нам это, разумеется, нужно, чтобы подобрать к ним разнообразные синонимы и аббревиатуры. Эта проблема легко решается подбором однословного синонима: *образ жизни — уклад*.

Имена собственные

Иногда могут возникнуть сомнения, является ли сочетание только именем собственным. Например, *Чёрный квадрат* — это название полотна Малевича. В то же время, можно легко предложить варианты употребления этого словосочетания в речи — в качестве метафоры, в прямом значении и т. д. Но в тезаурус такое сочетание включать всё равно нельзя, поскольку в текстах, не связанных с искусством, оно используется крайне редко, что можно проверить, используя корпус: зададим в поиске слово *чёрный* на расстоянии, равное одному слову, от *квадрат*, и получим 156 вхождений этой фразы в 107-ми документах, из которых только 23 вхождения не являются названием полотна Малевича.

Количество слов в синсете

Одним из вопросов является ограничение по словам в синсете — понятие «неоднословие» лишь указывает на то, что слов больше, чем одно, а верхняя граница формально не установлена. Таким образом, мы можем добавить в число синонимов обороты, содержащие больше слов, и при этом включающие в себя изначальное неоднословие: *рубки ухода — рубки ухода за лесом*. Однако сочетание, которое добавляется в тезаурус, должно быть употребительным, устойчивым, а не просто подходить по смыслу, поэтому едва ли можно найти примеры адекватных синонимов из шести или восьми слов.

Замена опорным словом

Многие неоднословные выражения допускают замену опорным словом, например, *войсковой атаман* — очевидно, что в синонимах появится *атаман*, так как он не вступает в родовидовые отношения с неоднословием. В этом случае неоднословие легко может оказаться случайным сочетанием, и проверить это предлагается при помощи корпуса — *войсковой атаман* в разных падежных формах имеет 132 вхождения в НКРЯ, так что это сочетание можно считать употребительным, и поэтому оно включается в тезаурус.

Классификация рассмотренных проблем:

(1) определение целесообразности добавления некоторого сочетания, имеющего множество вариантов;
(2) проверка существования такого неоднословия в языке;
(3) различение имени собственного и аналогичного неоднословия;

(4) целесообразность добавления сочетания типа «прилагательное + существительное» в случае, если его можно заменить опорным словом;
(5) ограничение слов в синсете.

Итоги

Проведённая работа позволяет увидеть, что некоторые проблемы с лексикографической разметкой неоднословий могут быть решены достаточно тривиальными методами. Решение обозначенных проблем могло бы помочь в деле лексикографической разметки неоднословных выражений, что актуально как непосредственно для разработки тезауруса, так и для программ по автоматической обработке текста.

Стоит отметить, что размеченные в ходе поиска проблем синсеты прошли вторичную обработку и вскоре будут добавлены в тезаурус YARN.

ЛИТЕРАТУРА

Мухин М. Ю. Инструкция по работе с открытым электронным тезаурусом YARN (Yet Another RussNet) // NLPub. URL: https://nlpub.ru/YARN/Инструкция (дата обращения: 15.02.2016).

YARN. Итоги третьего года работы по проекту «Новый открытый электронный тезаурус русского языка» // YARN, 2015. URL: https://russianword.net/posts/third-year-results (дата обращенияЖ 15.02.2016).

Braslavski P., Ustalov D., Mukhin M. A Spinning Wheel for YARN: User Interface for a Crowdsourced Thesaurus // Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics. Gothenburg, Sweden: Association for Computational Linguistics, 2014. 106 p. — P. 101–104.