М.Р. Шаехов, г.Казань

ПРИНЦИПЫ ПОСТРОЕНИЯ И ВОЗМОЖНОСТИ ИСПОЛЬЗОВАНИЯ WORDNET ДЛЯ ТЮРКСКИХ ЯЗЫКОВ

Аннотация: В статье представлен обзор имеющихся на данный момент WordNet-подобных ресурсов для тюркских языков, проанализированы структура, принципы и порядок разработки. Также представлены первые шаги по построению WordNet для татарского языка (TatarNet) на основе перевода данных тезауруса русского языка РуТез и последующих обработки и дополнений. На примере достигнутых результатов обсуждается общая методология разработки ворднетов и рассматриваются ключевые проблемы, возникающие при построении синсетов для татарского языка.

Ключевые слова: wordnet, ворднет, татарский язык, тюркские языки, синсет.

Во многих языках мира имеется доступ к комплексным лексическим ресурсам. Традиционные ресурсы, такие как двуязычные и одноязычные словари, тезаурусы и лексиконы составляются лексикографами. Но в эпоху развития компьютерных технологий и масштабного распространения большого объема информации приоритетными становятся цифровые и автоматические системы сбора, систематизации, обработки и хранения лексических данных. Один из таких ресурсов – WordNet, впервые созданный для английского языка в Принстонском университете (США), поэтому получивший название Принстонского WordNet (далее – PWN) [Miller G.A., 1995]. Со времени публикации в 1995 году данный ресурс много раз дополнялся и совершенствовался и стал основой для создания многочисленных подобных ему ворднетов для других языков мира, но его сущность и цель остались неизменными.

По своей сущности WordNet представляет собой лексико-семантический ресурс, объединяющий информацию классического словаря и дополнительные

данные по смысловым связям. В нем значения слов объединяются под общим концептами, которые называют синсетами (от английского synonym sets – «набор синонимов»). В результате получается комплексный словарь, доступный для использования компьютерными технологиями, удобный для анализа текста и других исследований в области языка [Лукашевич Н.В., 2011. – С.68.].

WordNet-подобные ресурсы для тюркских языков

На данный момент составление WordNet-подобных ресурсов является популярным способом систематизации и автоматизации лексического инструментария во многих языках мира, в том числе и в тюркских, на которых мы остановимся поподробнее.

Турецкий WordNet (https://bitbucket.org/ozlemc/twn/downloads/) [Özlem Çetinoğlu etc., 2018; Bilgin Orhan etc., 2004] был разработан в Sabancı University (Стамбул) в рамках проекта BalkaNet, целью которого было построение ворднетов среднего размера для шести языков, территориально расположенных на Балканском полуострове: болгарского, чешского, греческого, румынского и турецкого [Tufis D. etc., 2004]. BalkaNet был построен на основе комбинации обоих подходов создания ворднетов. Сначала из Принстонского WordNet было отобрано множество синсетов и осуществлен их перевод. Затем, на основе подхода merge, было разработано множество синсетов по балканской тематике, общие для всех ворднетов BalkaNet. И, наконец, были разработаны синсеты, специфичные для каждого из языков BalkaNet. В частности, размер турецкого WordNet внутри данного многоязычного ресурса составляет 14,795 синсетов.

Помимо перечисленных выше методов использования, разработчики Турецкого WordNet видят первоочередную пользу данного ресурса в возможности экспортирования смысловых отношений в другие ворднеты и получения искомого слова по его описанию.

KeNet (http://haydut.isikun.edu.tr/kenet.html) – еще один турецкий WordNet, построенный на базе современных словарей турецкого языка, содержащий 113 217 синсетов [Razieh Ehsani, 2018; Razieh Ehsani. KeNet…,

2018]. В отличие от BalkaNet, данный ресурс был построен в результате ручной обработки снизу-вверх на основе Турецкого современного словаря, составленного Турецким институтом языка [http://sozluk.gov.tr (Дата доступа: 01.07.2019)]. Из данного общедоступного онлайн-словаря были отобраны синонимы, которые две группы аннотаторов соотносили между собой по общим значениям с помощью специально созданного приложения. Затем для извлечения смысловых значений были использованы автоматические методы извлечения и обработка текста из Турецкой Википедии и Викисловаря.

Extended Open Multilingual Wordnet (http://compling.hss.ntu.edu.sg/omw/summx.html). Цель работы над данным ресурсом – создать условия исследователю лексической семантики одного или нескольких языков для использования ворднетов этих языков без правовых и технических преград через единый интерфейс [Francis Bond etc., 2013]. Результатом данного исследования стал Открытый многоязычный Wordnet, который содержит всего более 2 миллионов значений для 117,659 концептов из более 1000 языков мира. Ресурс был построен путем комбинирования открытых ворднетов с данными, автоматически извлечёнными из Wiktionary и Unicode Common Locale Data Repository (CLDR).

Открытый многоязычный Wordnet содержит данные и по русскому языку (20,138 синсетов, которые покрывают 64% ядерных концептов PWN), и по семи тюркским языкам: азербайджанскому (1923 синсетов, покрытие 35%), казахскому (1124 синсетов, покрытие 8%), киргизскому (793 синсетов, покрытие 7%), татарскому (550 синсетов, покрытие 5%), туркменскому (680 синсетов, покрытие 7%), турецкому (7953 синсетов, покрытие 35%) узбекскому (889 синсетов, покрытие 8%).

BabelNet (http://babelnet.org/) – еще один многоязычный, сверх-большой широкоохватный лексико-семантический ресурс [Roberto Navigli etc., 2012]. Изначально, в 2013 году, он был построен при помощи комбинирования самой большой многоязычной открытой интернет-энциклопедии – Википедии – с самым популярным компьютезированным лексиконом – WordNet. Синсеты

WordNet и вики-страницы Википедии были интегрированы в единые концепты (в т.н. Babel синсеты) через автоматическое преобразование, а затем лексические пустоты в ограниченных ресурсами языках были заполнены при помощи машинного перевода.

Самое главное преимущество BabelNet перед другими ворднетами заключается в том, что он постоянно пополняется за счет расширения своих источников. На данный момент (последняя дата обновления – февраль 2018 года) BabelNet объединяет более 15 миллионов синсетов для 284 языков мира на основе обработки данных, полученных из 47 источников.

В данном ресурсе также представлены почти все тюркские языки. В частности, в нем содержится 69,756 определений и 609,628 текстовых входов татарских слов, которые относятся к 1,862,223 концептам.

WordNet для татарского языка

Изучив богатый опыт создания ворднетов по всему миру, и понимая, что ограниченность ресурсов и недостаточная разработанность лексической базы на данный момент не позволит создать полноценный WordNet для татарского языка (TatarNet) на основе монологических словарей, мы решили взять за основу Тезаурус русского языка РуТез [Лукашевич Н.В., 2014]. В результате автоматической обработки данного ресурса, в том числе автоматического перевода, была сформирована таблица с текстовыми входами на русском языке с переводом на татарский язык. Основной задачей была проверка соответствия этого перевода на основе Русско-татарского словаря под редакцией Ф.А. Ганиева, редактирование и добавление возможных вариантов. Основанием для этого служили данные колонок с гипонимами, гиперонимами и глоссарием, так как приоритетной задачей была не оценка правильности перевода отдельных слов, а передача понятий оригинала на язык перевода.

При анализе и редактировании текстовых входов на язык перевода можно наблюдать интересные моменты передачи синсетов русского языка на татарский язык.

1. Синсеты из русского языка представляют собой существительные и отыменные глаголы. Почти все они на татарский язык переводится в форме существительного (исем) или отыменного глагола (исем фигыль). Например: *величие – бөеклек, олылык; вескость – авыр булу, саллы буллу.*

В некоторых случаях, при переводе отыменных глаголов употребляются обе части речи. Например: *бездействие – бер нәрсә дә эшләмәү; чара күрмәү, гамьсезлек; гегемония – гегемония, җитәкчелек итү, өстенлек.*

2. В анализируемой таблице, как и во всех русско-татарских словарях, много слов, передаваемых на татарский язык при помощи описательной конструкции. Всего на конец файла насчитывается 375 примеров такой формы передачи текстовых входов, не употребляемых на татарском языке. Здесь решающим фактором являлось не количество слов, а наличие в словарях только отдельных компонентов словосочетания или описательного оборота. Например: *коренник – төпкә җигелгән ат; выскочка – сикергәк, ялагай, сәнәктән көрәк булган кеше.*

Эти описательные обороты можно поделить, в зависимости от лексического значения и состава слова источника, на 4 группы:

А) Коренные слова, не имеющие соответствующего варианта на татарском языке по причине того, что эти понятия не свойственны культуре этого народа. Например: *именинник – исем бәйрәмен үткәрүче; клюка – кәкре башлы таяк; конура – эт оясы.*

Б) Термины и понятия, не имеющие перевода на татарском языке, передаваемые заимствованиями и/или описательным оборотом. Например: *дротик – дротик, кыска саплы сөңге; горизонталь – ятма сызык, горизонталь, горизонталь сызык; котельная – пар казаннары бинасы; микрокосм – кечкенә зурлыклар дөньясы, микрокосм.*

В) Сложные слова, не имеющие идентичных по строению слов на языке перевода. Например: *водосток – су агып төшә торган торба; двустволка – ике көпшәле мылтык; естествоиспытатель – табигать фәннәре белгече.*

Г) Слова с повтором вида передаваемого понятия. Это или названия месяцев, или наименования растений и деревьев. Например: *январь – гыйнвар, гыйнвар ае; вяз – карама, карама агачы; липа – юкә, юкә агачы.*

Повторяться могут и слова, описывающие национальность или принадлежность к чему-либо другому. Например: *японец – япон, япон кешесе; девочка – кыз бала; иноходец – юрга, юрга ат.*

3. В обработанных данных часто встречаются текстовые входы, обозначающие названия профессий, национальностей и рода занятий женского рода. Из-за того, что в татарском языке отсутствует морфологическая категория рода, для их передачи применяются лексические средства, что приводит к увеличению описательных конструкций. Кроме того, в текстовых входах добавляется еще конкретизация возраста: кыз (девушка) или хатын (женщина). Например: *активистка – активистка, актив хатын, актив кыз; караимка – караим хатыны, караим кызы; купальщица – су коенучы хатын, су коенучы кыз; манекенщица – манекенчы хатын, манекенчы кыз.*

4. Проблемную область составляют синсеты на русском языке, для которых в татарском языке нет своих соответствующих понятий. Это слова религиозной тематики – понятия православия, не свойственные религии татар – исламу. Статистически их немного – всего 32. Например: *ересь – ересь; молебен – молебен; миропомазание – миро белән майлап чукындыру.*

Как видно из примеров, они передаются 3 путями:

1) без изменений (ересь, канон, коляда, священник);

2) при помощи описательной конструкции (чиркәү манарасы, чиркәүдә чаң кагучы);

3) заменой близкого понятия из мусульманской терминологии: *исповедник – тәүбә иттерүче, тәүбәче, тәүбә-истигъфарчы.*

Общие выводы

По итогам анализа и обработки синсетов на русском языке с их передачей на татарский язык можно сделать несколько выводов:

1. Разработка WordNet представляет собой интересное и продуктивное направление для изучения, систематизации лексического богатства отельных языков, в том числе и татарского языка. В процессе работы не только подтверждаются общепринятые особенности в переводе отдельных лексических единиц из одного языка на другой, но и выявляются не заметные до этого тонкости передачи культуры и своеобразия народного колорита языковыми средствами.

2. Работа над синсетами также четко показывает западающие моменты в лексикографии отдельных языков. К примеру, процесс работы был бы более продуктивным, а итоги более достоверными при наличии проработанных электронных версий словарей на обоих языках.

3. В процессе работы также подтверждается необходимость обновления и обогащения словарей татарского языка терминологией, современными и широко употребляемыми просторечными словами для более точной передачи заимствований средствами самого языка.

На данный момент завершена только первичная обработка автоматически полученных данных. На наш взгляд, в дальнейшем работу над TatarNet можно продолжить в следующей последовательности:

- проверка качества и репрезентативности полученных данных через сопоставление с частотным словарем, созданным на основе Татарского национального корпуса «Туган тел» [ТНК], добавление пропущенных значений;

- сопоставление полученных данных с ядром Принстонского WordNet, добавление пропущенных значений;

- формирование полноценных синсетов из полученных текстовых входов;

- выстраивание смысловых связей между полученными синсетами и их заполнение пустых синсетов;

- дальнейшее дополнение полученной системы другими частями речи и т.д.

Литература

Лукашевич, Н.В. РуТез-Lite, опубликованная версия тезауруса русского языка РуТез / Н.В. Лукашевич, Б.В. Добров, И.И. Четверкин // Международная конференция по компьютерной лингвистике Диалог-2014. – 2014. – С. 340-349.

Лукашевич Н.В. Тезаурусы в задачах информационного поиска. – М.: Изд-во Моск. ун-та, 2011. – 511 с.

Bilgin O., Çetinoglu Ő., Oflazer K. Building a Wordnet for Turkish // Romanian Journal of Information Science and Technology. – 2004. – V. 7, No 1–2. – P. 163–172.

Francis Bond, Ryan Foster. Linking and Extending an Open Multilingual Wordnet // Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL 2013). Pp. 1352-1362

Miller G.A. “Wordnet: a lexical database for english,” Communications of the ACM, vol. 38, no. 11, pp. 39–41, 1995.

Özlem Çetinoğlu, Orhan Bilgin, Kemal Oflazer. Turkish Wordnet // Kemal Oflazer, Murat Saraçlar (eds). Turkish Natural Language Processing. Springer, 2018. doi:10.1007/978-3-319-90165-7_15

Razieh Ehsani, Ercan Solak, Olcay Taner Yildiz. Constructing a WordNet for Turkish Using Manual and Automatic Annotation // ACM Transactions on Asian and Low-Resource Language Information Processing, Volume 17 Issue 3, May 2018. Article No. 24. doi:10.1145/3185664

Razieh Ehsani. KeNet: A Comprehensive Turkish Wordnet and Using It in Text Clustering. PhD Thesis. Işık University, 2018

Roberto Navigli, Simone Paolo Ponzetto. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network // Artificial Intelligence, Volume 193, December 2012. Pp. 217–250

Tufis D., Cristea D., Stamou S. BalkaNet: Aims, Methods, Results and Perspectives. A General Overview // Romanian Journal of Information Science and Technology, Volume 7, Numbers 1–2, 2004. Pp. 9–43