3. TEI CONSORTIUM, eds.: TEI P5: Guidelines for Electronic Text Encoding and Interchange. http://www.tei-c.org/Guidelines/P5/

4. Hu Zheng,: Textual Dictionary File Format. StarDict project on GitHub, https://github.com/huzheng001/stardict-3/blob/master/dict/doc/TextualDictionaryFileFormat, accessed 8/6/2016

5. Гүлзура Жумакунова: Түркчө-Кыргызча сөздүк, 50 000 сөз. Кыргыз-Түрк «Манас» Университетинин басылмалары, Бишкек, 2005.

6. Гүлзура Жумакунова: Түркчө-Кыргызча сөздүк, 50 000 сөз. Кыргыз-Түрк «Манас» Университетинин басылмалары, Бишкек, 2005. pages 28-29

УДК 510.5, 519.768.2

МЕТОДОЛОГИЯ АВТОМАТИЗИРОВАННОГО ПОПОЛНЕНИЯ СЛОВАРЯ СИСТЕМЫ МАШИННОГО ПЕРЕВОДА ДЛЯ КАЗАХСКО-РУССКОЙ И КАЗАХСКО-АНГЛИЙСКОЙ ЯЗЫКОВОЙ ПАРЫ

У.А. Тукеев, *Казахский Национальный Университет имени аль Фараби, Алматы, Казахстан. ualsher.tukeyev@gmail.com*
Д.Р. Рахимова, *Казахский Национальный Университет имени аль Фараби, Алматы, Казахстан.di.diva@mail.ru*
Ж.М. Жуманов, *Казахский Национальный Университет имени аль Фараби, Алматы, Казахстан. z.zhake@gmail.com*

Аннотация: В статье описывается методология автоматизированного пополнения словаря системы машинного перевода Apertium для казахско-русской и казахско-английской языковой пары. Цель данной методологии состоит в оказании помощи пользователю обнаружить лучшую морфологическую парадигму в одноязычном морфологическом словаре Apertium. Приведены практические результаты.

Ключевые слова: заполнение словарей, Apertium, казахский, русский и английский язык.

METHODOLOGY OF THE AUTOMATED ENRICHMENT OF MACHINE TRANSLATION SYSTEM DICTIONARIES FOR KAZAKH-RUSSIAN AND KAZAKH-ENGLISH LANGUAGE PAIR

U.A. Tukeyev, *Al Farabi Kazakh National University, Almaty, Kazakhstan. ualsher.tukeyev@gmail.com*
D.R.Rakhimova, *Al Farabi Kazakh National University, Almaty, Kazakhstan. di.diva@mail.ru*
Zh.M. Zhumanov *Al Farabi Kazakh National University, Almaty, Kazakhstan. z.zhake@gmail.com*

Abstract: This paper describes the methodology of the automated enrichment dictionary of the machine translation system Apertium for the Kazakh-Russian and Kazakh-English language pair. The purpose of this methodology consists in assistance to the user to find the best morphological paradigm in the monolingual morphological Apertium dictionary. Practical results are presented.

Keywords: filling of dictionaries, Apertium, Kazakh, Russian and English.

Введение

На данный момент существует много различных словарей, как печатные, так и электронные. Словари, используемые в машинном переводе (МП) может содержать переводы на различные языки сотен тысяч слов и фраз, а также предоставить пользователям

дополнительные возможности. Такие, как давая пользователю возможность выбрать языки и направление перевода, обеспечить быстрый поиск слов, возможность ввода фразы и т.д.

На сегодняшний день существует множество методов расширения словарей. Мы применили метод, реализованный Микелям Эспла-Гомис(Esplà-Gomis M) и др. [2], и адаптировали для казахского языка. Мы использовали данный инструмент для заполнения словарей казахско-русского и казахско-английского языка в свободной / с открытым исходным кодом системы Apertium машинного перевода.

1.Казахско-русский и казахско-английский словарь

Apertium представляет собой систему машинного перевода поверхностно-трансферного типа. Следовательно, в основном он имеет дело со словарями и правилами поверхностного трансфера. На практике поверхностный трансфер отличается от глубокого тем, что при нём не выполняется полный синтаксический разбор предложений [4]. А правила, в отличии от операций на дереве синтаксического разбора, представляют собой операции с группами лексических единиц. В Apertium используются три типа словарей: два монолингва словарей и один двуязычный словарь для каждой языковой пары. Так мы имеем для казахско-русского и казахско-английского языковых пар следущие:

apertium-rus.rus.dix: морфологический словарь для русского языка: он, в свою очередь, содержит информацию о словоизменении на русском языке.

apertium-eng-kaz.eng.dix: одноязычный словарь английского языка.

apertium-kaz.kaz.lexc: морфологический словарь для казахского языка: он содержит информацию о словоизменении на казахском языке.

apertium-eng-kaz.eng-kaz.dix: двуязычный словарь для англо-казахской пары.

apertium-kaz-rus.kaz-rus.dix: двуязычный словарь, содержит переводные соответствия слов и символов двух языков. В языковой паре любой из языков, составляющих эту пару, может быть как входным, так и выходным языком, т. е. эти термины употребляются условно.

Словари системы Apertium имеют формат XML, каждое слово имеет тег, показывающие ее лексические свойства (часть речи, число, склонение и др.). В контексте системы Apertium парадигма является примером склонения/спряжения определённой группы слов. В морфологическом словаре леммы ссылаются на парадигмы, что позволяет нам показать все словоформы этих лемм без необходимости записи всех возможных окончаний.

Примером использования парадигмы может служить следующее. Допустим, мы хотим добавить в словарь существительные свойство и старшинство. Вместо записи одинаковых окончаний, мы можем записать окончания форм слова свойство, а потом сказать "старшинство изменяется как свойство". Или рассмотрим пример на английском языке, мы хотим добавить в словарь прилагательные happy и lazy. Вместо записи одинаковых окончаний:

happy, happ (y, ier, iest)

lazy, laz (y, ier, iest)

Мы можем записать окончания форм слова happy, а потом сказать "lazy изменяется как happy", или "shy изменяется как happy", "naughty изменяется как happy", "friendly изменяется как happy" и т. д. В этом примере happy и свойство и будет парадигмой, моделью изменения всех остальных. Точное описание определения парадигм будет дано позже. Парадигмы определяются в тэгах <pardef> и используются в тэгах <par> [3].

Наполнение двуязычных словарей. Таким образом, теперь у нас есть два морфологических словаря и далее мы перейдём к двуязычному словарю. Двуязычный словарь описывает соответствия слов. Все словари имеют один и тот же формат (которой описан в DTD, dix.dtd). В этом словаре парадигмы создаются для каждой части речи отдельно, например, для глаголов и для прилагательных есть различные парадигмы. В дальнейшем в словаре будут описаны следующие:

```
<!-- numerals -->

<e><p><l>нөл<s n="num"/></l><r>ноль<s n="num"/></r></p><par n="__num_gender"/></e>
<e><p><l>бір<s n="num"/></l><r>один<s n="num"/></r></p><par n="__num_gender"/></e>
<e><p><l>екі<s n="num"/></l><r>два<s n="num"/></r></p><par n="__num_gender"/></e>
<e><p><l>үш<s n="num"/></l><r>три<s n="num"/></r></p><par n="__num_gender"/></e>
<e><p><l>төрт<s n="num"/></l><r>четыре<s n="num"/></r></p><par n="__num_gender"/></e>
<e><p><l>бес<s n="num"/></l><r>пять<s n="num"/></r></p><par n="__num_gender"/></e>
<e><p><l>алты<s n="num"/></l><r>шесть<s n="num"/></r></p><par n="__num_gender"/></e>
<e><p><l>жеті<s n="num"/></l><r>семь<s n="num"/></r></p><par n="__num_gender"/></e>
<e><p><l>сегіз<s n="num"/></l><r>восемь<s n="num"/></r></p><par n="__num_gender"/></e>
<e><p><l>тоғыз<s n="num"/></l><r>девять<s n="num"/></r></p><par n="__num_gender"/></e>
<e><p><l>он<s n="num"/></l><r>десять<s n="num"/></r></p><par n="__num_gender"/></e>
```

Рисунок 1. - Пример заполнения имени числительного для двуязычного казахско-русского словаря.

Как уже видно, заполнение самого словаря требует определенные знания и затрачивается не мало времени. Связи с этим появилось востребованность в автоматизированной системе заполнения словарей.

2. Описание метода автоматизированного пополнения словаря системы машинного

Этот метод состоит из двух этапов:

На первом этапе, обобщенное дерево суффиксов [1], в котором все суффиксы всех парадигм склонения и спряжения в словаре хранятся эффективным способом с комментариями, какие парадигмы генерируют каждый из этих суффиксов. Это дерево позже используется для анализа неизвестного слова и определения способов, которыми оно может быть разбито на пары основа/суффикс.

На втором этапе, метод использует различные стратегии для того, чтобы показать пользователю различные поверхности форм в результате комбинации корней и парадигм, полученных на первом этапе. Затем пользователь может решить, какие из этих словоформ правильны или нет, помогая системе в конечном счете найти наиболее подходящую пару основа/парадигма для слова, которая вставляется в словарь. Были опробованы две стратегии определения слов, которые спрашиваются у пользователя, с целью как можно большего уменьшения взаимодействий с ним: один - эвристический, второй - основанный на использовании деревьев решений ID3 [2].

В этом разделе мы описываем эксперименты, проводимые, чтобы попытаться адаптировать методику, ограничения наших подходов и возможные шаги улучшения этого метода для языков.

2.1 Описание проблемы

Эта методология, описанная в [2], основана на ряде предположений, которые определяют возможность использования метода. Эти предположения:

Изменения слов в языке происходит с использованием суффиксов, т.е. способ, которым язык производит формы слов, заключается в добавлении суффиксов к статической основе (это условие выполняется для казахского и русского языков);

все суффиксы, генерируемые парадигмами в словаре могут быть получены заранее;

две парадигмы не должны генерировать один и тот же набор суффиксов.

Предыдущие работы [2] показывают, что третье условие не всегда выполняется (Например, парадигмы для топонимов и антропонимов, как правило, создают те же самые суффиксы (в случае английского языка, это был бы пустой суффикс)), и в этих случаях метод полезен только, чтобы уменьшить количество возможных основ/парадигм. Оставив этот факт в стороне, эта методология была успешно использована для каталонского, испанского, английского, баскского, мальтийского и хорватского языков. Обширные эксперименты, проведенные на хорватском языке, позволяют хорошо оценить возможность использования

метода для русского языка, учитывая, что оба они - славянские языки с очень похожим системами склонения и спряжения. Метод был протестирован со словарем русского языка в Apertium (apertium-rus.rus.dix) и неизвестными русскими словами. Он работает, как ожидалось, но обширные эксперименты не проводились. В следующем разделе содержится набор использованных неизвестных русских слов и команды, использованные для тестирования. Однако, казахский язык является гораздо более сложной проблемой. Основные трудности адаптации метода для казахском языка связаны с:

форматом словаря: в то время как английский и русский языки используют, основанный на XML формат dix, изначально используемый Apertium для хранения одноязычных словарей, казахский язык основан на формализме HFST для конечных автоматов.

двухуровневыми морфологическими правилами: важной проблемой, с которой сталкиваются при работе с системой казахского языка на Apertium, является то, что он использует двухуровневые морфологические правила, которые изменяют результат конкатенации основ и суффиксов. Это означает, что поверхностная форма не производится непосредственно словарем и, таким образом, невозможно извлечь прямым путем основы и суффиксы из формы. Это делает невозможным использование метода, описанного в [2] напрямую.

Таким образом, должно быть определено лучшее адаптированное решение для того, чтобы приспособить данный метод для казахского языка, заменив первый шаг процесса (обнаружение пар кандидатов основа/парадигма с помощью обобщенного дерева суффикса) более специфичным решением.

2.2 Подход решения проблемы адаптации к казахскому языку

Была изучена возможность параллельного подхода, который многократно использует как можно больше идей, воплощенных в методологии, используемой для русского языка.

Обратите внимание, что, как и в других тюркских языках, когда основа, такая как «мектеп» (школа), получает дополнительные морфемы в свои суффиксы, например, для генерации формы в 3-ем лице притяжательной формы и в творительном падеже, последние буквы основы иногда изменяются, в результате добавления суффикса, но суффикс также может изменяться, в зависимости от последних букв основы: «мектебінен» (с его школы), «мектеп» изменился на «мектеб», но окончание «інен» отличается от суффикса («сыдан»), который относится к «қалта» (карман): қалтасынан (с его кармана). Мы подсчитали, что, в худшем случае, могут измениться три буквы основы.

В этих условиях, чтобы быть в состоянии помочь пользователю в добавлении неизвестных слов в словари, мы разработали пакет скриптов, которые позволяют создавать и компилировать два модифицированных словаря из существующих apertium–kaz.kaz.lexc и одного вспомогательного файла.

Был разработан новый класс в нашем первоначальном пакете java, который читает выходные данные этого процесса и использует тот же интерфейс, чтобы спросить пользователя о правильности поверхностной формы слова и вставляет полученных кандидатов в словарь. Стоит отметить, что первые два действия не нужно повторять до тех пор, пока определение парадигм в словаре не изменится каким-либо образом. Поэтому, все ресурсы, необходимые для применения этого метода, могут быть получены только один раз и затем применены на лету к новым словам. Вывод всего этого процесса затем может быть подан в DictionaryAnalyser, написанный на Java, который задает вопрос.

3.Практические результаты

В этом конечном результате мы проверили возможность использования нашего подхода для помощи неспециалистам в задаче добавления новых слов в морфологические словари казахского и русского языка в Apertium. Мы сделали предварительные испытания, чтобы подтвердить, что наш оригинальный подход подходит для русского языка без изменений, но не подходит для казахского языка, который использует совершенно другой

формат и стратегию для построения морфологических одноязычных словарей. Мы разработали коллекцию скриптов, которые позволяет нам адаптировать оригинальные идеи нашего метода для казахского языка. С помощью этих сценариев, можно создать файл, содержащий все возможные основы/парадигмы и коллекцию поверхностных форм из списка неизвестных слов, которые затем предлагаются пользователю. В заключении, мы адаптировали оригинальный интерфейс для того, чтобы адаптировать его для чтения этого файла, что позволяет использовать метод двоичных вопросов для вставки новых слов в словарь казахского языка.

После запуска инструмента пользователь может выбрать один из различной комбинации кандидатских стеблями и парадигм, отвечая на вопросы, заданные системой. Когда пользователь подтверждает, что слова были обнаружены правильно, они перемещаются в соответствующий словарь раздел. В случае, когда система обнаруживает более одного решения для слова, все возможные варианты записываются в словарь наряду с количеством найденных возможных вариантов.

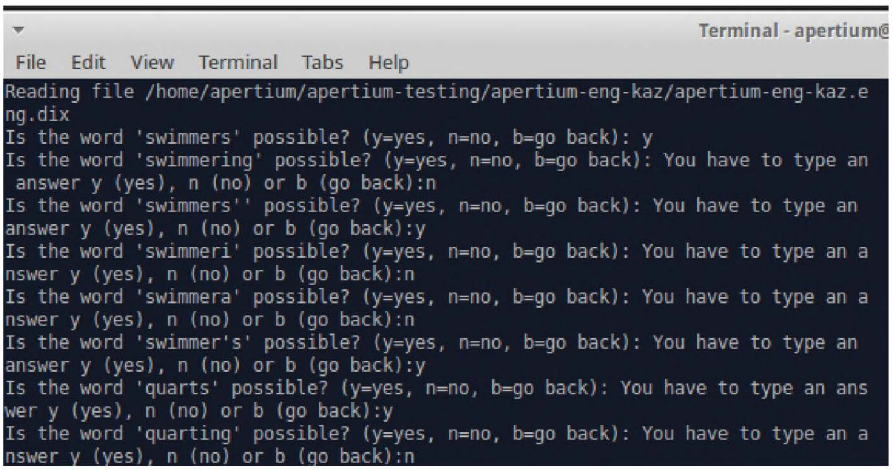

Рисунок 2.- Пример применения метода в заполнении словаря.

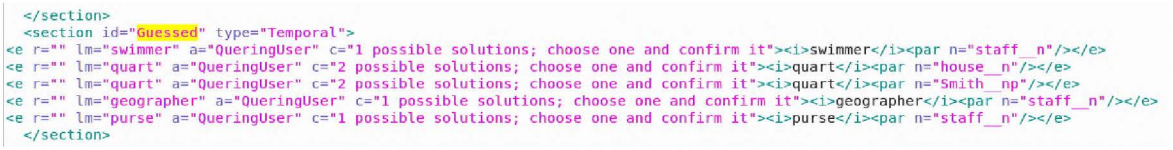

Рисунок 3.- Генерация словарных статей

Заключение

Данный метод был впервые применен для сложных языковых пар, как казахско-русский и казахско-английский. Система была адаптирована для казахского языка: был создан словарь apertium-kaz-kaz-suffixgen.lexc, в котором поддельная, зависимая от парадигмы, основа был связана с каждой парадигмой открытого класса и разработано 113 лексических форм. Мы используем эту методологию, чтобы расширить количество слов в словарях, которые в основном эффекты к качеству машинного перевода. Технология позволяет пользователям, которые не обладают глубокими знаниями в области вычислительной представления морфологии, но понимают язык и активно участвуют в построении словарей. Это означает, что все больше людей могут добавлять словарные статьи, создающие большие словари в более короткие сроки.

Список литературы

1. McCreight E.M. A Space-Economical Suffix Tree Construction Algorithm. //Journal of the ACM. – 1976. - 23 (2). – P.262–272.

2. Esplà-Gomis M., Sánchez-Cartagena V.M., Pérez-Ortiz J.A., Sánchez-Martínez F., Forcada M.L., Carrasco R.C. An efficient method to assist non-expert users in extending dictionaries by assigning stems and inflectional paradigms to unknown words. //Proceedings of EAMT 2014, The Seventeenth Annual Conference of the European Association for Machine Translation (Dubrovnik, June 16–18, 2014). – 2014. – P.19–26.

3. Abduali B., Sundetova A., Zhanbussunov N., Musabekova Zh., Study of the problem of creating structural transfer rules for the Kazakh-English and Kazakh-Russian machine translation systems on Apertium platform. //Вестник Казну им. Аль-Фараби, том 20 (Messenger of Al-Farabi KazNU, vol. 20) by proceedings of the International Conference "Computational and Informational Technologies in Science, Engineering and Education" (CiTech-2015, 24-27 September, 2015). - Almaty: Al-Farabi Kazakh National University Press, 2015. - P.77-82

4. Рахимова Д.Р., Қалдашбеков Е.Е., Мұсабекова Ж.Ғ., Абақан М., Кызырканова С., Жамалиева А., Абдуали Б. Apertium платформасы негізінде орыс тілінен қазақ тіліне аудару жүйесін құру. //Материалы международной научной конференции студентов и молодых ученых "Фараби Әлемі», Алматы, Казахстан, 13-16 апреля, 2015 г. - Алматы, 2015. - С.175.

УДК 621.3

ОНТОЛОГИЧЕСКАЯ МОДЕЛЬ ИМЕНИ СУЩЕСТВИТЕЛЬНОГО ДЛЯ СИСТЕМЫ КАЗАХСКО- КИТАЙСКОГО МАШИННОГО ПЕРЕВОДА

Унзила Каманур *Евразийский Национальный Университет им.Л.Н.Гумилева, Астана, 010000, Қазақстан Unzila.88@mail.ru*
Алтынбек Шарипбай *Евразийский Национальный Университет им.Л.Н.Гумилева, Астана, 010000, Қазақстан*
Гүлмира Бекманова *Евразийский Национальный Университет им.Л.Н.Гумилева, Астана, 010000, Қазақстан*
Лена Жеткенбай *Евразийский Национальный Университет им.Л.Н.Гумилева, Астана, 010000, Қазақстан*

Цель статьи - В этой работе будет рассмотрен создание онтологической модели имен существительных казахского, русского языков, предназначенных для системы машинного перевода. Эта модель даст нам возможность сравнить сходства и различие двух языков. Также можно применить для разработки информационного поиска, машинного перевода и автореферирования, диалоговых и других систем.

Ключевые слова: агглютинативные языки, морфология, онтологии.

ONTOLOGICAL MODEL OF NOUNS SYSTEM FOR KAZAKH-CHINESE MACHINE TRANSLATION

Unzila Kamanur *L.N. Gumilyov Eurasian National University, Astana, 010000, Kazakhstan Unzila.88@mail.ru*
Altynbek Sharipbay *L.N. Gumilyov Eurasian National University, Astana, 010000, Kazakhstan*
Gulmira Bekmanova *L.N. Gumilyov Eurasian National University, Astana, 010000, Kazakhstan*
Lena Zhetkenbay *L.N. Gumilyov Eurasian National University, Astana, 010000, Kazakhstan*

Abstract. In this work, the machine translation for jüyecine Kazakh, Chinese nouns establishment of the ontological model. This model allows us to compare the similarities and differences between the two languages. Information retrieval, machine translation and avtoreferattaw, and other systems can be used to create a dialogue.