УДК 81'322

ОБ ЭКОНОМИИ УСИЛИЙ ПРИ РАЗРАБОТКЕ МНОГОЯЗЫЧНЫХ ЛИНГВИСТИЧЕСКИХ Е-РЕСУРСОВ

С.О. Шереметьева, Н.Ю. Дюмин

TOWARDS AN ECONOMY OF EFFORT IN MULTILINGUAL E-RESOURCE DEVELOPMENT

S.O. Sheremetyeva, N.Yu. Dyumin

Описана новая методика автоматизированного построения многоязычных лексических ресурсов, позволяющая экономить усилия и время разработчиков. Экономия достигается за счет повторного использования отдельных компонентов, ранее разработанных для других языков и приложений, с их последующей адаптацией к выполнению новых задач и включением в разрабатываемую нами систему на определенных этапах обработки языкового материала. Методика иллюстрируется на примере извлечения эквивалентных именных групп для английского и русского языков.

Ключевые слова: электронные ресурсы, многоязычный лексикон, автоматическая обработка текста, экономия усилий.

The paper describes a new methodology for automatic extraction of aligned multilingual lexical resources with economy of development effort and time. The economy is achieved by reusing some of the components developed earlier for other langugages and applications and adapting them to fulfil particular tasks in the suggested extraction workflow. The methodology is illustrated on the example of aligned NP extraction from the English-Russian language pair.

Keywords: resource extraction, alignment, multilingual lexicon, natural language processing, economy of effort.

Введение

Электронные лингвистические ресурсы являются неотъемлемой частью как профессиональной, так и повседневной деятельности людей. Качество электронных словарей и автоматической обработки информации, например, информационного поиска и автоматического перевода оказывают все большее влияние на уровень научных и технических исследований.

Создание электронных лингвистические ресурсов – очень трудоемкий и времязатратный процесс, особенно если они предназначены для многоязычных систем, где необходимость определения межъязыковых эквивалентов вызывает дополнительные трудности и увеличивает время разработки. В связи с этим нужны такие методики и программный инструментарий, которые при повышении качества результата позволяли бы экономить усилия, время, а следовательно, и финансовые затраты на создание ресурсов.

Настоящая статья посвящена проблемам снижения трудоемкости и повышения оперативности разработки одноязычных и многоязычных лексиконов как лингвистического ресурса, имеющего первостепенное значение.

Одним из очевидных путей повышения оперативности разработки лингвистических ресурсов является автоматизация их создания и повторное использование методик и компонентов уже разработанных систем с адаптацией их к новым языкам и системам.

Несмотря на большое внимание, которое уделяется автоматизации процесса разработки лексиконов и существование определенного количества

Шереметьева Светлана Олеговна, доктор филологических наук, доцент, профессор кафедры лингвистики и межкультурной коммуникации, Южно-Уральский государственный университет (г. Челябинск). E-mail: linklana@yahoo.com

Дюмин Никита Юрьевич, преподаватель кафедры общей лингвистики, Южно-Уральский государственный университет (г. Челябинск). E-mail: nikita.dyumin@gmail.com

Svetlana O. Sheremetyeva, PhD (Habilitation), professor of the Linguistics and Intercultural Communication department, South Ural State University (Chelyabinsk). E-mail: linklana@yahoo.com

Nikita Yu. Dyumin, lecturer of General Linguistics department, South Ural State University (Chelyabinsk). E-mail: nikita.dyumin@gmail.com

алгоритмов и систем, направленных на решение как этой проблемы в целом, так и отдельных ее этапов, полученные результы до сих пор недостаточно эффективны. Такие системы, как правило, разрабатываются на основе корпусного анализа в рамках статистического или гибридных подходов, сочетающих статистику и лингвистические знания. В качестве примера можно привести системы автоматического выравнивания предложений[1, 2], системы автоматического выравнивания слов[3, 4], системы автоматического извлечения терминов[5] и др. Эффективность таких систем в значительной степени зависит от объема используемого корпуса: чем больше корпус, тем лучше результаты. Причем при статическом подходе точнее извлекается более частотная лексика, в то время как автоматическое извлечение низкочастотных лексических единиц является проблематичным. Пренебрежение низкочастотной лексикой ведет как к значительному ухудшению покрываемости словарей, так и потере ключевых слов, например, при информационном поиске.

В своей работе мы представляем новую гибридную методику автоматизированного построения многоязычного лексикона на основе извлечения лексических единиц различных типов из параллельных корпусов текстов. Методика предусматривает адаптацию отдельных компонентов автоматической обработки текстов (АОТ), ранее разработанных для конкретных языков и прикладных задач, и включение их в разабатываемую нами систему на определенных этапах обработки языкового материала. Особенностью методики является ее эффективность на текстах небольшого объема и достоверное извлечение как высокочастотных, так и низкочастотных многокомпонентных словосочетаний. Методика иллюстрируется на примере построения двуязычного англо-русского лексикона именных групп (ИГ).

Алгоритм автоматизированного построения двуязычного лексикона

Представленная в статье методика разработана в рамках гибридного подхода к АОТ и сочетает такие статистические методы как вычисление n-gram и частотности, а также лингвистические знания о структуре ИГ конкретного языка. Описываемая ниже методика автоматизированного извлечения многоязычных ИГ формулируется в виде алгоритма «RAlign» для английского и русского языков и может быть перенесена на другие пары языков.

Построение лексикона включает следующие этапы:

1) выравнивание параллельных (английских и русских, в нашем случае) текстов по предложениям;

2) извлечение одноязычных ИГ из выровненных предложений;

3) извлечение межъязыковых эквивалентных ИГ.

На вход системы подается файл в формате TMX. Формат TMX предназначен для обмена данными памяти переводов и поддерживается наиболее крупными представителями данной отрасли, такими как TRADOS Workbench, OmegaT, Abbyy Aligner, Google «Инструменты переводчика» и т. д. В TMX файле выровненные тексты разделены на сегменты (тег <seg>..</seg>), причем каждый сегмент текста на одном языке соответствует сегменту текста на другом языке (рис. 1).

Из рис. 1 видно, что в корпусе английскому сегменту «The results of calculations on the supercomputer “Uranus” are presented» соответствует русское «Приводятся результаты расчетов на суперкомпьютере “Уран“». Выровненные в файле сегменты могут быть равны по протяженности предложению. В том случае, когда одному предложению на языке оригинала соответствуют два или больше предложений языка перевода, в параллельный сегмент добавляются все предложения, передающие содержание входного предложения, например*: «We investigate the initial-finish value problem for the Boussinesque-Love equation by reducing it to the initial-finish value problem for the Sobolev type equation of the second order.»* соответствует русским предложениям *«Рассматривается начально-конечная задача для неоднородного уравнения Буссинеска – Лява.» и «Проводится редукция к абстрактной начально-конечной задаче для уравнения Соболевского типа второго порядка.»*. В настоящее время имеются эффективные системы автоматического выравнивания предложений, позволяющие экспортировать результат в формате TMX. В частности, в нашем исследовании мы использовали продукт Trados WinAlign. Пример автоматического выравнивания приведен на рис. 2.

На следующем этапе происходит извлечение ИГ из выровненных разноязычных сегментов по

```
<tu creationdate="20111214T150319Z" creationid="ALIGN!">
<tuv xml:lang="EN-US">
<seg>The results of calculations on the supercomputer "Uranus" are presented.</seg>
</tuv>
<tuv xml:lang="RU-RU">
<seg>Приводятся результаты расчетов на суперкомпьютере "Уран".</seg>
</tuv>
</tu>
```

Рис. 1. Фрагмент файла TMX

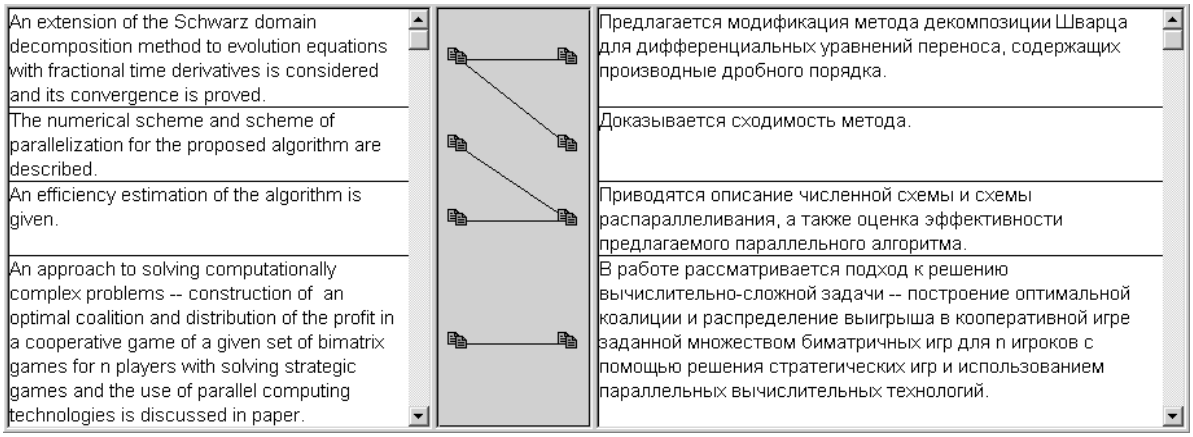

Рис. 2. WinAlign: автоматическое выравнивание предложений

методике, предложенной в работе[6] для английского языка, которая включает в себя следующие основные этапы.

1. Ввод текста любого объема и вычисление n-грам ($n = \{1, 2, 3, 4\}$).

2. Фильтрация n-грам с помощью лексикалистских правил.

3. Лемматизация.

Правила фильтрации представляют собой предписания удалять n-грамы (цепочки слов) которые не могут быть именными группами: вычисленные n-грамы обрабатываются набором стоплистов (списками лексем, относящихся к частям речи, запрещенных на определенных позициях в многокомпонентной ИГ грамматикой конкретного языка). Таким образом, автоматическое извлечение ИГ обеспечивается универсальным алгоритмом, который оперирует зависимыми от английского языка стоп-листами, последовательность применения которых определяется порядком слов в английской ИГ.

Эта методика может быть использована для извлечения ИГ из текстов на других языках[7], при этом универсальная часть алгоритма используется повторно, а стоп-листы наполняются иноязычными лексемами и меняется порядок их применения к вычисленным n-грам в соответствии с грамматикой обрабатываемого языка. Мы адаптировали описанный алгоритм для русского языка.

В нашей работе мы полностью использовали предложенную методику фильтрации (идентификации) ИГ, однако применили ее не к вычисленным n-грамам (мы не вычисляем n-грамы входных текстов), а к выровненным сегментам параллельных текстов, полученных на выходе этапа 1 построения лексикона, тем самым снимая ограничение в 4 компонента на длину извлекаемых ИГ. Лемматизация ИГ не проводилась по причинам, приведенным ниже.

Итак, на вход второго этапа поступают параллельные английские и русские сегменты, каждый из которых обрабатывается следующим образом:

• сегмент делится на отрезки, ограниченные знаками пунктуации: *«Показано**,** что новый алгоритм не приводит к появлению дуплетов Фруассара**,** в отличие от имеющихся в Maple и Mathematica процедур вычисления аппроксимаций Паде.» делится на «Показано», «что новый алгоритм не приводит к появлению дуплетов Фруассара», «в отличие*

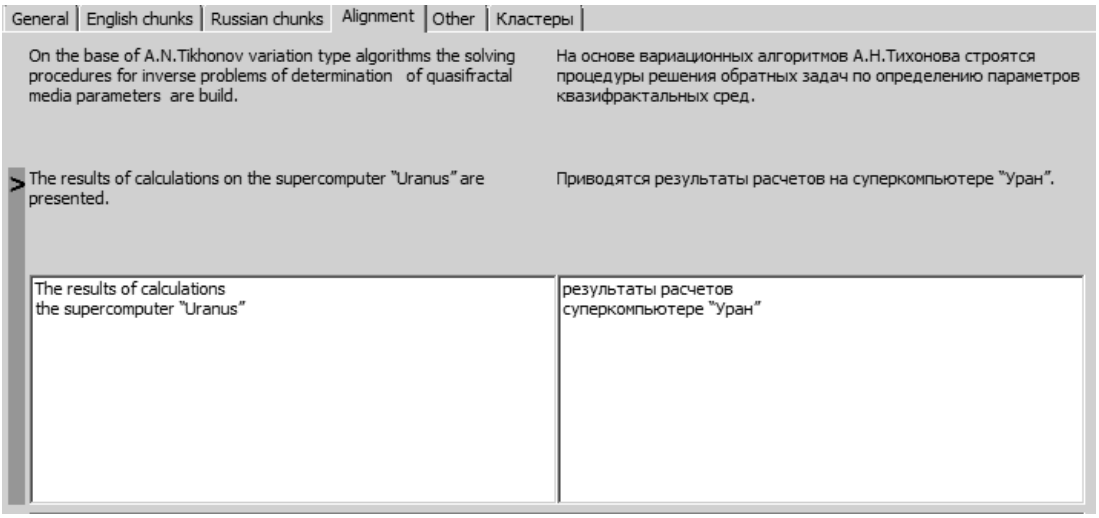

Рис. 3. Извлеченные из параллельных текстов ИГ

от имеющихся в Maple и Mathematica процедур вычисления аппроксимаций Паде»;

• каждый из полученных таким образом отрезков обрабатывается фильтрующими правилами для английского или русского языка соответственно, и в результате извлекаются английские и русские ИГ. В нашем примере мы получим «*новый алгоритм» «появлению дуплетов Фруассара», «Maple и Mathematica процедур вычисления аппроксимаций Паде».*

На третьем этапе работы системы пользователю предлагается подтвердить или отвергнуть соответствие полученных ИГ «The results of calculations on the supercomputer “Uranus”» и «результаты расчетов на суперкомпьютере “Уран”» (рис. 3). В случае подтверждения, что как правило происходит, двуязычные эквиваленты автоматически заносятся в словарь. После чего пользователю для утверждения представляется новая порция автоматически извлеченных русско-английских эквивалентных ИГ, при этом, если в новой порции текста извлекаются пары ИГ, уже занесенные в словарь, они пользователю не предъявляются. Пользователь имеет возможность редактирования ИГ перед занесением в словарь.

Как уже было отмечено, при внесении новых пар в лексикон не производится их лемматизация, в частности, потому что в системе предусмотрен инструмент накопления морфологической информации об ИГ: в ходе работы формируются кластеры, в которых собраны парадигмы ИГ одного языка и соответствующая парадигма эквивалентной ИГ другого языка. Например, если извлечены пары «the supercomputer Uranus: суперкомпьютере “Уран”», «the supercomputer Uranus: суперкомпьютеру “Уран”» и «the supercomputer Uranus: суперкомпьютером “Уран”», система устанавливает соответствия этих ИГ, формируя при этом словоизменительную парадигму ИГ (рис. 4).

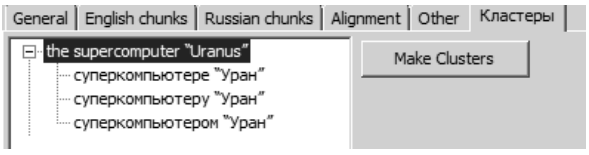

Рис. 4. Представленная в тексте парадигма русской ИГ и ее английский эквивалент

Помимо информации о морфологии ИГ в подобных кластерах могут собираться синонимичные конструкции, либо многозначные термины. Каждый случай предоставляет возможность собрать дополнительную лингвистическую информацию.

Заключение

В статье описана новая методика автоматизированного построения многоязычных лексиконов, которая иллюстрируется на примере английского и русского языков. Методика предусматривает адаптацию отдельных компонентов АОТ, ранее разработанных для конкретных языков и прикладных задач, и включение их в разрабатываемую нами систему на определенных этапах обработки языкового материала. Основным из таких компонентов является методика извлечения именных групп (ИГ) из текста на английском языке, которую авторы перенесли на материал русского языка и адаптировали для решения поставленной задачи.

Методика была проверена путем ее имплементации в виде системы «RAlign» на языке программирования Delphi и тестирования на материале параллельных английских и русских аннотаций к статьям сборника «Вестник ЮУрГУ. Серия „Математическое моделирование и программирование“».

[1] Braune F., Fraser A. Improved Unsupervised Sentence Alignment for Symmetrical and Asymmetrical Parallel Corpora // Coling, Beijing, 2010.

[2] Li P., Sun M., Xue P. Fast-Champollion: A Fast and Robust Sentence Alignment Algorithm // Coling, Bejing, 2010.

[3] Unsupervised Word Alignment with Arbitrary Feature / Ch. Dyer, J. Clark, A. Lavie, N.A. Smith // ACL Portland, 2011.

[4] Germann U. Yawat: Yet Another Word Alignment Tool // ACL Columbus, 2008.

[5] Lefever E., Macken L., Hoste V. Language-independent bilingual terminology extraction from a multilingual parallel corpus // EACL Athens, 2009.

[6] Sheremetyeva S.O. On extracting multiword NP terminology for MT // EAMT-2009, Universitat Politècnica de Catalunya, Barcelona, Spain. P. 205–212.

[7] Шереметьева С.О., Дюмин Н.Ю., Мыларщикова Т.Ю. О возможности межъязыкового переноса систем автоматической обработки текста // V МНПК «Прикладная лингвистика в науке и образовании». СПб., 2010. С. 345–350.

Поступила в редакцию 14 марта 2012 г.