Л. Н. Беляева
L. N. Beliaeva

ПАРАЛЛЕЛЬНЫЙ КОРПУС ТЕКСТОВ В ЗАДАЧАХ ЛЕКСИКОГРАФИЧЕСКОГО АНАЛИЗА

PARALLEL CORPORA IN LEXICOGRAPHY

Аннотация. В статье рассматриваются проблемы использования параллельных и псевдопараллельных корпусов текстов для решения задачи извлечения и выравнивания терминов. Оцениваются возможности извлечения терминов из одноязычных текстов на глобальном английском и русском языках, предлагаются ограничения процедуры лемматизации при анализе текстов на русском языке. Анализируются особенности критерия терминологичности.

Abstract. The paper considers the problems of parallel and comparable corpora for terms extraction and translation. Methods of terms extraction from monolingual texts on global English and Russian languages are discussed, limitations of lemmatizing procedures at Russian texts analysis are proposed. Termhood criterion is analyzed.

Активное изменение ситуации в науке и технике, появление новых направлений и, более того, новых отраслей знаний приводит к резкому отставанию специализированных лексикографических источников, предназначенных для поддержки работы переводчиков. Анализ специализированных переводных словарей, издаваемых в нашей стране и/или включенных в различные автоматизированные словарные системы, позволяет фиксировать их несоответствие как современному уровню науки и техники, так и основным направлениям развития отраслей знаний. Это связано не только с естественным отставанием лексикографии, связанным с необходимостью переработки больших массивов современной информации, но и с традиционным подходом к созданию словарей с опорой на уже опубликованные источники, а уже затем на результат анализа переведенных авторами текстов.

Идея создания автоматизированных систем извлечения терминов из корпусов параллельных текстов насчитывает уже более 20 лет и в той или иной степени реализуется в различных проектах. На этом пути кроме сложностей, связанных с отсутствием симметричности терминологических систем разных языков, особую проблему представляет отбор переводов для корпуса параллельных текстов, поскольку их качество часто является сомнительным. Поэтому обращение к псевдопараллельным (сопоставимым) корпусам текстов, при организации которых возможна экспертная оценка текстов на сопоставляемых языках, вполне естественно.

Особым источником сопоставимых текстов являются материалы конференций, посвященных одной и той же научной проблеме, но проводимых в разных странах и с разными рабочими языками. Поскольку для зарубежных научных конференций основным рабочим языком является английский, то для решения задачи создания переводных словарей целесообразно составление корпусов материалов конференций на английском и русском языках.

При сопоставительном анализе корпусов таких текстов выявляется несколько дополнительных «подводных» камней. «Английские» тексты в своем большинстве написаны на глобальном английском языке, что в реальности означает нарушение синтаксической структуры предложения, вызванное влиянием родных языков, и отсутствие гармонизации терминологии, в результате чего термины часто представляют собой переводы соответствующих лексических единиц родного языка автора, а не стандартизированные номинации. «Русские» тексты, в свою очередь, «отягощены» безумным научным канцеляритом, частотным использованием синтаксических структур с объектом в первой позиции предложения и отсутствием явных границ между именными группами, номинирующими термины и выполняющими разные роли в предложении (см., например, фрагмент предложения *построение соответствующих различным конструктивным параметрам*

семейства силовых характеристик упругопластических демпферов). В ситуации развитой падежной омонимии, характерной для именной лексики русского языка, это приводит к невозможности корректного установления границ терминов и их структуры.

При использовании корпусов сопоставимых текстов вопрос о выравнивании переходит в особую плоскость. В случае параллельных корпусов текстов основным является выравнивание по предложениям, которое опирается на формальные показатели границ и частей предложений, соответствие объемно-прагматических структур текстов. При всех возникающих технических и лингвистических сложностях этот процесс вполне реализуем. В случае текстов сопоставимых возможно только терминологическое выравнивание, опирающееся на выявление характерных для обоих массивов корпуса однословных терминологических единиц и их сопоставление в качестве кандидатов в переводные эквиваленты, а также поиск устойчивых словосочетаний с этими однословными терминами в качестве ядер. Дальнейший сопоставительный анализ требует привлечения знаний из переводных автоматизированных словарей, позволяющих верифицировать выбранные пары терминов.

Простые именные словосочетания – именные группы (ИГ), т.е. синтаксические конструкции, ядерным элементом которых является существительное, в тексте функционально равноценны слову, но по сути представляют собой свертку предложения, т.е. являясь скорее единицей синтаксиса, чем лексикона. При этом (вслед за Хомским) можно утверждать, что внутренняя структура зависимостей в именной группе отражает структуру зависимостей соответствующего предложения. Проблема заключается в том, как эту структуру распознать в свертке, и каковы те механизмы, которые позволяют нам в тексте распознать структуру, свернутую многократно (см., например, в рамках одного текста переход от четырехкомпонентной именной группы *seismic stability direct analysis*, введенной в заголовочном

комплексе, к трехкомпонентной конструкции *seismic direct analysis,* а затем к аббревиатуре *SDA*).

Поскольку ИГ представляет собой именно свертку, стяжение структуры высказывания, то это внешнее упрощение конструкции и формы вызывает ее семантическое усложнение: показатели наличия связи между конкретными компонентами и типов связи между элементами, которые в предложении вводятся с помощью коннекторов и реляторов разного уровня, в именной группе сняты.

При коммуникации на одном языке и, более того, при коммуникации носителей языка в рамках одного языка для профессиональных целей возможность адекватного распознавания номинируемых объектов поддерживается совпадением тезаурусов участников, общим культурным и профессиональным фоном и установкой. Поэтому именные группы, являющиеся способами номинации сложных объектов, даже с пропуском некоторых компонентов, понимаются однозначно.

Иная ситуация возникает при коммуникации на одном глобальном языке специалистов с разными родными языками, ее отражением являются тексты, включаемые в материалы конференций. Именно здесь возникает особая ситуация, интересная как для исследования влияния интерференции родного языка на язык английский, так и для решения проблем как ручного, так и машинного перевода.

Опора на референциальный статус терминов – именных групп в научном тексте позволяет предположить, что установка автора на передачу информации и ее понимание требует экспликации в тексте связей между элементами таких словосочетаний. Специально проведенный анализ массивов текстов разных предметных областей и языков для специальных целей показал[1], что появление в тексте ИГ длиной более 2 элементов

---

[1] *Беляева Л.Н., Камшилова О.Н., Филимонова Г.И.* Исследование семантико-синтаксической структуры английской именной группы в

(при средней длине простой именной группы, составляющей в научном тексте 4 элемента) сопровождается появлением двухкомпонентных ИГ в ближайшем окружении этого термина, в пределах 2-3 предложений или в комбинации заглавия, ключевых слов и реферата. Следовательно, при ручном переводе можно использовать эту ситуацию как ключ для диагностики структуры именной группы и выбора адекватного перевода. При автоматизации поиска терминологических единиц в корпусе текстов выделенные именные словосочетания с одним и тем же ядром или с вложенной коллокацией (см., например, именные группы *abdominal visceral and subcutaneous adipose tissue compartments, adipose tissue, abdominal and subcutaneous adipose tissue*) можно рассматривать как терминологическое гнездо и искать в сопоставимом корпусе подобные группы, опираясь на результаты лемматизации.

При этом следует иметь в виду, что в английском тексте по отношению к именным группам лемматизация дает корректный результат, то в русском языке она может быть просто опасна. При решении задачи извлечения терминов процедура лемматизации именных словосочетаний в русском языке должна опираться на позицию ядра именной группы и ее формальных границ: все согласованные определения, предшествующие ядру, и само ядро должны лемматизироваться, а несогласованные определения в постпозиции к ядру должны оставаться в текстовой форме. При таком условии установление границ ИГ оказывается процедурой, определяющей результат лемматизации.

Опора на границы групп и установление ядер позволит формировать терминологические гнезда в границах текста и сопоставлять их с выделенными кандидатами в термины на других языках на основе возможных переводов ядерных существительных.

системном и функциональном аспекте. – УДК 81-13; 81’373; 001.4, № госрегистрации 01.2.0 310077, инв. № 3Н – 13/03.

Большинство автоматизированных систем извлечения терминов используют либо статистический, либо лингвистический подход, либо их сочетание. При этом учитывается частота лексической единицы в тексте, информация о сочетаемостных предпочтениях лексических единиц, отношение правдоподобия для двухсловных и многословных терминов, и другие метрики.

Особый интерес представляют гибридные подходы, использование которых представляет собой попытку преодоления ограничений односторонних подходов к решению задачи извлечения терминов на основе как лингвистических, так и статистических элементов.

Одним из основных методов оценки степени терминологичности является метод автоматического выявления многокомпонентных терминов в тексте[2]. В качестве исходного материала для анализа при этом используется корпус текстов исходного языка, на его основе формируется список кандидатов в многокомпонентные термины. Эти термины упорядочиваются по степени терминологичности, устанавливаемой на основе сравнения: суммарных частот компонентов словосочетания, частоты самого словосочетания, включенного в более длинные структуры, частоты ядерного слова и словосочетания в национальном корпусе текстов. Получаемый в результате список оценивается экспертом в конкретной предметной области. Как правило, подходы с использованием показателей терминологичности основаны на объединении лингвистической и статистической информации. Лингвистическая информация состоит из грамматической разметки корпуса текстов по частям речи, лингвистический фильтр ограничивает тип извлекаемых терминов и использует список стоп-слов (антипризнаков).

---

[2] *Delpech E., Daille B.* Dealing with lexicon acquired from comparable corpora: validation and exchange // Proceedings, 9th Conference on Terminology and Knowledge Engineering (TKE). – Fiontar, Dublin City University, 2010. Pp. 229-223.

Лингвистическая база, необходимая для реализации этого метода, включает следующие компоненты:

1. Информацию о части речи, извлекаемую из результатов грамматической разметки корпуса текстов

2. Собственно лингвистический фильтр, применяемый к размеченному корпусу текстов, чтобы исключить те цепочки, извлечение которых не требуется по формальным признакам. К таким моделям относится неразрешенная комбинаторика частей речи. При этом возможно применение как «закрытого» фильтра, разрешающего извлечение цепочек слов только конкретных типов, так и «открытого» фильтра, в котором перечисляются только неразрешенные типы цепочек.

3. Список слов-антипризнаков.

Статистическая часть анализирует частотные характеристики цепочки лексических единиц, являющихся кандидатами в термины. Необходимость гибридного подхода определяется тем, что доступная для анализа статистическая информация без специальной лингвистической фильтрации не является достаточной для того, чтобы дать полезные результаты.

Выбор конкретного лингвистического фильтра и его наполнения зависит от того, каким образом в системе сбалансированы полнота и точность: предпочтение точности над полнотой требует использовать закрытый фильтр, в то время как предпочтение полноты определяет использование открытого фильтра.

Методы, используемые в статистических системах, варьируются от простых подсчетов частот до вычисления сложных статистических индикаторов для измерения силы связи элементов коллокаций, встретившихся в структуре кандидата на роль термина[3].

---

[3] TTC Project: Terminology Extraction, Translation Tools and Comparable Corpora //URL: **http://www.ttc-project.eu/about-ttc/concept-and-objectives**

Построение корпуса параллельных текстов для конкретного языка для специальных целей и установление комплекса лингвистических и статистических параметров представляют собой скоррелированные задачи и позволяют устанавливать термины и переводные соответствия.