Автоматическая классификация терминов в русскоязычном корпусе текстов по корпусной лингвистике♣

© Н.В. Виноградова
Санкт-Петербургский государственный университет
natalia_vinograd@mail.ru

О.А. Митрофанова
Санкт-Петербургский государственный университет
alkonost@mail.ru

П.В. Паничева
Санкт-Петербургский государственный университет
ppolin@mail.ru

Аннотация

В статье рассматриваются результаты компьютерной обработки неразмеченных русскоязычных научных текстов. Основное содержание эксперимента – автоматическая классификация терминов-дескрипторов в текстах из корпуса по корпусной лингвистике, осуществляемая с помощью специализированного инструмента АКЛ.

1 Постановка проблемы

Совершенствование инструментов лингвистических исследований и развитие методов автоматической обработки языковых данных стимулирует решение задач, связанных с извлечением семантической информации из естественноязыковых текстов. Одной из таких задач является осуществление автоматической классификации лексики (далее АКЛ) – процедуры, результаты которой востребованы во многих областях знаний о языке. АКЛ предоставляет лингвистам возможность использовать объективные данные об иерархической структуре лексикона, собранные при анализе представительных корпусов, и строить на основе этих данных формальные онтологии и лексикографические модули, применимые в процедурах автоматической обработки текстов и допускающие пополнение из корпусов [1,12,19,20,22]. Использование инструментов АКЛ представляет интерес и в другом отношении: результаты кластеризации лексики позволяют решать вопросы автоматического индексирования текстов, тематического упорядочения документов в корпусах, способствует повышению качества информационного поиска в больших массивах текстов и пр. [2,3,7,16,24].

Автоматическая обработка научных текстов – одна из актуальных прикладных задач, решение которой может потребовать осуществления процедуры АКЛ. Примером задач такого типа является формирование и обработка массивов текстов, представляющих молодые и активно развивающиеся области знаний, логико-понятийные системы которых находятся на этапе становления. Создание специальных корпусов текстов может идти и тогда, когда границы новой области знаний ещё нечётко обрисованы, когда тематические направления внутри дисциплины недостаточно дифференцированы, когда терминология предметной области неустойчива. Очевидно, что при таких условиях сам корпус текстов может оказаться нестабильным, и лингвистическая разметка такого корпуса (прежде всего, морфологическая) вряд ли целесообразна. Однако лингвисты даже в этом случае должны иметь в своём распоряжении автоматизированные средства для предварительной обработки корпусных данных, которые обеспечивают поиск и анализ необходимой лингвистической информации. Тогда на помощь может прийти инструмент АКЛ, рассчитанный на работу с неразмеченным корпусом.

2 Цели исследования, экспериментальный материал

Обсуждаемое в настоящей статье исследование направлено на осуществление АКЛ в неразмеченных русскоязычных текстах по корпусной лингвистике. С 2002 г. на кафедре математической лингвистики СПбГУ ведутся работы по созданию корпуса русскоязычных текстов по корпусной лингвистике (руководитель проекта – В.П. Захаров). В основу корпуса легли материалы международных конференций «Корпусная лингвистика и лингвистические базы данных – 2002» (Санкт-Петербург, 5 – 7 марта 2002 г.), «Корпусная лингвистика – 2004» (Санкт-Петербург, 11 – 14 октября 2004 г.), «MegaLing–2005: Прикладная лингвистика в поиске новых путей» (Крым, м. Меганом, 27 июня – 2 июля 2005 г.), «Корпусная лингвистика – 2006» (Санкт-Петербург, 10–14 октября 2006 г.) [5,13,14,15].

Корпус включает в себя тексты различной тематики, отражающие широкий спектр проблем корпусной лингвистики: определение корпусной лингвистики как особой области научной деятельности, противопоставление её другим

Труды 9ой Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» - RCDL'2007, Переславль-Залесский, Россия, 2007.

направлениям лингвистики и языковой инженерии; определение корпуса в соотнесённости с другими типами лингвистических данных; различные аспекты создания и использования корпусов; процедуры, выполняемые при работе с корпусом (разметка, типы разметки, поиск в корпусе); типология корпусов; корпусы текстов с позиций разработчиков и пользователей; взаимодействие корпусов и корпус-ориентированных лингвистических ресурсов и пр.

В настоящее время в состав корпуса входит 105 статей на русском языке объёмом около 175 тыс. словоупотреблений. В корпусе также представлены русскоязычные тезисы докладов объёмом около 25 тыс. словоупотреблений. Материалы корпуса хранятся в текстовом формате, наряду с этим у разработчиков корпуса существует доступ к файлам с оригинал-макетами.

В ходе подготовки текстов статей к размещению в корпусе производится их ручная обработка (графематический анализ), направленная на выделение нетекстовых элементов (таблиц, рисунков, формул, гиперссылок, числовых данных и пр.) и иноязычных вкраплений, а также метаразметка, которая предполагает фиксацию основных параметров каждой статьи в её паспорте [4,6]. Наряду с библиографическим описанием эксперты включают в число параметров статьи и наборы из 10 релевантных терминов-дескрипторов, позволяющих диагностировать тематическую принадлежность текста. При формировании наборов терминов-дескрипторов учитывались не только частотность терминов в тексте, но и их содержательный вес.

Термины-дескрипторы представлены в нормализованном виде: в наборе присутствует лемма, которая противопоставляется входящим в текст словоформам, например: **корпус** (*корпус, корпуса, корпусу, корпусом, корпусе, корпусы, корпусов, корпусам, корпусами, корпусах*) и пр. Например, для текста T_2006_5_15 набор терминов-дескрипторов выглядит следующим образом: (*биграмма, запрос, интернет, корпус, поиск, пользователь, сервис, слово, текст, частота*).

3 Используемые методы и инструменты обработки данных

Поставленная в рамках данного исследования цель предполагает использование компьютерного инструмента АКЛ, позволяющего производить процедуры латентного семантического анализа и кластеризации.

В АКЛ возможно использование целого ряда методов кластеризации: иерархических (агломеративных, дивизимных), неиерархических (например, итеративных – *K*-средних, *K*-медианы), гибридных методов (анализ различных подходов к кластеризации лингвистических объектов см., например: [18,25]). Выбор того или иного метода кластеризации определяется условиями эксперимента (умеренный или значительный объём корпуса; наличие или отсутствие ограничений на число итоговых кластеров и пр.). На нынешнем этапе реализации проекта были задействованы иерархический (агломеративный) метод кластеризации и неиерерархический метод (*K*-средних).

Выделение кластеров лексем в тексте производится на основе процедуры латентного семантического анализа (ЛСА). Идеология и практика ЛСА обсуждаются во многих научных и энциклопедических описаниях (см., например, такие ресурсы, как http://lsi.research.telcordia.com/lsi/LSIpapers.html, http://en.wikipedia.org/wiki/Latent_semantic_analysis и пр). Говоря об особенностях практического использования ЛСА, следует отметить существование нескольких разновидностей данного метода, учитывающих возможную сложность обрабатываемого массива документов и взаимосвязи их фрагментов (см., например, [11,21]).

С лингвистической точки зрения, суть ЛСА заключается в определении содержательной близости лексических единиц при сопоставлении их синтагматических свойств (иначе говоря, их сочетаемости с другими элементами контекста, дистрибуции) [17,23]. С инженерной точки зрения, ЛСА предполагает представление множества контекстов употребления исследуемых лексем как точек или векторов дистрибуций в *N*-мерном пространстве. Вычислив расстояния *d* между точками или сравнив вектора дистрибуций, можно получить количественную оценку тесноты семантических связей слов. При вычислении расстояний применяются различные меры близости: мера Евклида, мера Хэмминга и пр., производится вычисление значения косинуса угла между векторами дистрибуций и пр. (о преимуществах и недостатках мер см. [8,9]). Результаты измерений, сохраняемые в матрице расстояний, используются при кластеризации: чем ближе синтагматические свойства лексем (а стало быть, чем ближе их значения), тем меньше расстояние между векторами их дистрибуций и тем больше вероятность их объединения в один кластер. Сформированные таким образом кластеры лексем допускают дальнейшую лингвистическую интерпретацию.

В обсуждаемом проекте применяется инструмент АКЛ, разрабатываемый на кафедре математической лингвистики СПбГУ [10]. Поскольку технические аспекты разработки инструмента АКЛ заслуживают самостоятельного рассмотрения, в данной статье приводится лишь краткое его описание.

Программа АКЛ, созданная на языке Python, предусматривает три блока: блок предварительной обработки текста и вычисления расстояний между исследуемыми лексемами с помощью меры Евклида, блок иерархического кластерного анализа и блок кластерного анализа методом *K*-средних. При активизации программы определяются следующие параметры:

- имя файла, содержащего анализируемый текст (*text.txt*);
- имя файла, содержащего лексемы, отношения между которыми требуется исследовать (*words.txt*);
- ширина контекстного окна (± *s*);
- наличие/отсутствие весовых значений для ближних/удалённых элементов контекстов (*yes*/*no*);
- метод кластеризации (иерархический или *K*-средних);
- количество кластеров, которое необходимо получить (*C*).

Результаты кластеризации выводятся в виде многоуровневого списка слов с помощью скобочной записи. Наряду с этим пользователь получает данные о частотности исследуемых лексем в обрабатываемом тексте и значения расстояний во всевозможных парах лексем из анализируемого набора.

4 Обработка экспериментальных данных

В ходе эксперимента по АКЛ на материале текстов из русскоязычного корпуса по корпусной лингвистике производилась автоматическая обработка текстов и соответствующих им наборов терминов-дескрипторов. Были выполнены следующие процедуры:

- определение частоты встречаемости каждого термина-дескриптора в тексте;
- вычисление расстояний *d* между парами терминов-дескрипторов в наборах при ширине контекстного окна [-5…+5] и с учётом весов элементов контекстов;
- осуществление кластеризации терминов-дескрипторов для каждого текста иерархическим методом при заданной глубине иерархии $C = 5$.

В качестве примера приведём результаты обработки текста T_2002_72_79, описываемого терминами-дескрипторами (*архив, банк, данные, корпус, массив, поиск, разметка, текст, формат, чешский*).

Установлена частота употребления терминов-дескрипторов в данном тексте и определена тройка наиболее частотных элементов в наборе:

***корпус* (*f* = 43),**
***текст* (*f* = 25),**
***данные* (*f* =13),**
поиск (*f* = 8),
чешский (*f* = 6),
разметка (*f* = 4),
массив (*f* = 2),
формат (*f* = 2),
архив (*f* = 1),
банк (*f* = 1).

Определены расстояния между парами терминов-дескрипторов и выявлены наиболее тесно связанные друг с другом элементы (см. фрагмент ниже):

d(*корпус, корпус*) = 0,000
***d*(*корпус, текст*) = 0,344**
***d*(*корпус, данные*) = 0,509**
***d*(*корпус, поиск*) = 0,6739**
d(*корпус, чешский*) = 0,737
d(*корпус, разметка*) = 0,984
d(*корпус, массив*) = 1,477
d(*корпус, формат*) = 1,492
d(*корпус, архив*) = 1,848
d(*корпус, банк*) = 2,088

Результаты кластеризации для рассматриваемого текста при заданных условиях эксперимента таковы: (*архив, банк, массив, формат* (*разметка* (*чешский* (*поиск* ((*текст, корпус*) *данные*))))).

Аналогичные данные о частотности терминов-дескрипторов, о расстояниях между ними в наборах получены и для остальных текстов корпуса.

Например:

T_2002_27_39 (*массив, база, данные* ((*переводческая, память*) (*система* (*текст, перевод*))) (*корпус, параллельный*));

T_2004_5_15 (*построение, компьютерный, тезаурус, понятие* (*словарь* (*валентность* (*частота* (*контекст* (*корпус, текст*))))));

T_2004_24_28 (*поиск, ИПС, программа, индоевропейский* (((*язык* (*тезаурус* (*текст, модуль*))) *корпус*) *интерфейс*));

T_2004_86_98 (*метаразметка, разметка, словарь, паспорт* ((*исторический* (*данные* ((*корпус, текст*) *параметр*))) *метаданные*));

T_2004_149_158 (*параметр, единица, категория, значение* (*фундаментальная* ((*разметка* (*падеж, корпус*)) (*лингвистика, корпусная*))));

T_2004_304_315 (*категоризация, категория, класс, инженерия (классификация (корпус ((текст (семантическая, разметка)) лингвистика*))));

T_2006_5_15 (*поиск, интернет, запрос, пользователь* (*сервис* (*частота* ((*текст, корпус*) (*слово, биграмма*))))));

T_2006_16_24 (*разметка, формат, поиск, тег* (((*фрагмент* (*слово* (*текст, житие*))) *корпус*) *цитата*));

T_2006_303_306 (*источник, поиск, словарь, картотека* (*корпус* ((*топоним* ((*ландшафт, культурный*) *топонимический*)) *данные*))).

Очевидно, последовательность формирования кластеров терминов-дескрипторов отражает естественные связи элементов исследуемых текстов, что подтверждается частотными данными и значениями расстояний между парами элементов. Особенностью полученных кластеров является то, что в них зафиксированы как синтагматические (например, (*переводческая, память*)), так и парадигматические связи терминов-дескрипторов (например, (*массив, база, данные*)). Вместе с тем, разграничение этих основополагающих типов связей на уровне текста зачастую затруднено: например, термины *текст* и *корпус*, *слово* и *биграмма* могут находиться в парадигматических отношениях, если интерпретируются как разноплановые текстовые единицы (*текст* ≠ *корпус*, *слово* ≠ *биграмма*), или в

синтагматических отношениях, если указывается, что между ними допустимы отношения включения (*текст* ⊃ *корпус*, *слово* ⊃ *биграмма*). Тем самым, в процессе создания модели предметной области «Корпусная лингвистика» обобщение выявленных связей терминов-дескрипторов до родо-видовой иерархии понятий производится на достаточно широких основаниях, а сама результирующая иерархия при этом оказывается более насыщенной.

Результаты кластеризации позволяют оценить диапазон понятийных категорий, релевантных для предметной области «Корпусная лингвистика». Вероятно, такие термины-дескрипторы, как *корпус*, *текст*, *данные*, *разметка*, *тег*, *поиск*, *слово*, *лемма*, *словоформа*, *контекст* и пр. представляют понятийное ядро указанной предметной области.

В целях уточнения характера связей между понятийными категориями, выраженными исследуемыми терминами, была проведена серия экспериментов с текстами, для которых наблюдается частичное совпадение наборов дескрипторов. В ряде случаев результаты кластеризации совпадающих терминов-дескрипторов для разных текстов оказались идентичными. Так, обнаружены пары текстов, применительно к которым группы общих для них дескрипторов упорядочиваются единообразно: (*словарь* (*корпус, текст*)), (*частота* (*корпус, текст*)), (*массив* (*данные* (*корпус, текст*))). Безусловный интерес представляют те случаи, когда кластеризация терминов-дескрипторов, разделяемых парой текстов, приводит к несовпадающим результатам. Например, отношения в пятёрке дескрипторов, общих для пары текстов, устанавливаются следующим образом: (*формат* (*разметка* (*поиск* (*текст, корпус*)))) vs. (*разметка* (((*корпус, текст*) *формат*) (*поиск*)). Применительно к другой паре текстов их общие дескрипторы также упорядочиваются по-разному: (*поиск* (*слово* (*текст, корпус*))) vs. (*поиск* (*корпус* (*слово, текст*))).

Сравнение иерархий терминов-дескрипторов, полученных для разных документов, создаёт почву для их тематической рубрикации. Если результаты экспериментов свидетельствуют о единообразии связей между дескрипторами, можно сделать предположение и о тематическом сходстве текстов. Обратное может указывать на то, что тексты не представляют одно тематическое направление или на то, что в паре тематически близких текстов по-разному расставлены акценты.

Процедуры отбора и кластеризации дескрипторов, характеризующих предметную область «Корпусная лингвистика», позволяют перейти с терминологического уровня представления знаний на онтологический и сформировать упорядоченное множество понятийных категорий, которые необходимо включить в формальную онтологию рассматриваемой предметной области.

Ниже приведена развернутая модель онтологии по корпусной лингвистике, созданной на основе текстов из русскоязычного корпуса данной тематики и реализованной в онторедакторе Protégé:

- *предметная область «Корпусная лингвистика»*
- *корпус данных*
  корпус текстов
  тип корпуса
  - *работа с корпусом*
    разработчик
    отбор данных
    оцифровка данных
    разметка
    корпус-менеджер
    пользователь
    поиск
    запрос
    терминальная цепочка символов
    регулярное выражение
    лемма
    тег
    результат
    конкорданс
    контекст
    словоуказатель
    статистика

При задании структуры онтологии и формулировке понятийных категорий разработчики руководствовались как эмпирическими данными, полученными в процессе АКЛ, так и научными описаниями сферы корпусной лингвистики (см., например, [6]). Проверка адекватности онтологии была проведена применительно к целостному корпусу текстов и принесла положительные результаты.

5 Основные результаты исследования

Автоматическая обработка текстов статей из русскоязычного корпуса по корпусной лингвистике с учётом терминов-дескрипторов способствует решению комплекса задач, среди которых:

- структурирование знаний в предметной области «Корпусная лингвистика», что предполагает упорядочение терминологии, выявление понятийных категорий, характеризующих данную предметную область, а также исследование естественных связей между категориями, проявляющихся в специальных текстах;
- подготовка данных для создания онтологии предметной области «Корпусная лингвистика» и для осуществления процедуры автоматической классификации текстов, что предполагает выявление основных тематических областей в рамках корпусной лингвистики.

Перспективные направления развития исследования связаны с разработкой инструментов для определения количественных оценок близости текстов и с проведением автоматической классификации текстов внутри тематических областей.

Литература

[1] Азарова И.В., Марина А.С. Автоматизированная классификация контекстов при подготовке данных для компьютерного тезауруса RussNet // Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции «Диалог–2006». – М., 2006. – С. 13–17.

[2] Баглей С.Г., Антонов А.В., Мешков В.С., Суханов А.В. Кластеризация документов с использованием метаинформации // Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции «Диалог–2006». – М., 2006. – С. 38–45.

[3] Браславский П. Автоматические операции с запросами к машинам поиска интернета на основе тезауруса: подходы и оценки // Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции «Диалог–2004». – М., 2004. – С. 79–84.

[4] Волков С.Св., Захаров В.П., Дмитриева Е.А. Метаразметка в историческом корпусе XIX века // Труды международной конференции «Корпусная лингвистика – 2004». – СПб., 2004. – С. 86–98.

[5] Доклады научной конференции «Корпусная лингвистика и лингвистические базы данных». – СПб., 2002.

[6] Захаров В.П. Корпусная лингвистика: Учебно-методическое пособие. – СПб., 2005.

[7] Крижановский А.А. Автоматизированное построение списков семантически близких слов на основе рейтинга текстов в корпусе с гиперссылками и категориями // Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции «Диалог–2006». – М., 2006. – С. 297–302.

[8] Митрофанова О.А. Измерение семантической информации в тексте на основе анализа латентных связей // Труды Международной конференции «MegaLing–2005»: Прикладная лингвистика в поиске новых путей. – СПб., 2005. – С. 80–89.

[9] Митрофанова О.А. Новые разработки в области измерения семантических расстояний // XXXV Международная филологическая конференция. Вып. 21. Секция математической лингвистики. Ч. 2. – СПб., 2006. – С. 3–11.

[10] Митрофанова О.А., Мухин А.С., Паничева П.В. Автоматическая классификация лексики в русскоязычных текстах на основе латентного семантического анализа // Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции «Диалог–2007». – М., 2007. – С. 413–421.

[11] Некрестьянов И.С. Тематико-ориентированные методы информационного поиска. Дис. … канд. физ.-мат. наук. – СПб., 2000.

[12] Сидорова Е.А. Технология разработки тематических словарей на основе сочетания лингвистических и статистических методов // Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции «Диалог–2005». – М., 2005. – С. 443–449.

[13] Труды международной конференции «Корпусная лингвистика – 2004». – СПб., 2004.

[14] Труды международной конференции «Корпусная лингвистика – 2006». – СПб., 2006.

[15] Труды Международной конференции «MegaLing–2005»: Прикладная лингвистика в поиске новых путей. – СПб., 2005.

[16] Buscaldi D., Rosso P., Alexandrov M., Ciscar A.J. Sense Cluster Based Categorization and Clustering of Abstracts // Computational Linguistics and Intelligent Text Processing: Proceedings of the 7th International Conference CICLing–2006. LNCS 3878. – Springer-Verlag, 2006. – P. 547–550.

[17] Gamallo P., Gasperin C., Augustini A., Lopes G. P. Syntactic-Based Methods for Measuring Word Similarity // Text, Speech and Dialogue: Fourth International Conference TSD–2001. LNAI 2166. – Springer-Verlag, 2001. – P. 116–125.

[18] Pantel P. Clustering by Committee. Ph.D. Dissertation. Department of Computing Science, University of Alberta: 2003. –
[19] http://www.isi.edu/~pantel/Content/publications.htm

[20] Pantel P., Lin D. Document Clustering with Committees // SIGIR–02. Tampere: 2002. –
[21] http://www.isi.edu/~pantel/Content/publications.htm

[22] Pekar V. Linguistic Preprocessing for Distributional Classification of Words // Proceedings of the COLING–04 Workshop on Enhancing and Using Electronic Dictionaries. – Geneva: 2004. – P. 15–21.

[23] Rohde D.L., Gonnerman L.M., Plaut D.C. An Improved Method for Deriving Word Meaning from Lexical Co-Occurrence. –
[24] http://dlt4.mit.edu/~dr/COALS/Coals.pdf

[25] Shin S.-I., Choi K.-S. Automatic Word Sense Clustering Using Collocation for Sense Adaptation // Proceedings of the Second International WordNet Conference (GWC–2004). – Brno, Czech Republic: 2004. – P. 320–325.

[26] Smrž P., Rychlý P. Finding Semantically Related Words in Large Corpora // Text, Speech and Dialogue: Fourth International Conference (TSD–2001). LNAI 2166. – Springer-Verlag, 2001. – P. 108–115.

[27] Stein B., Meyer zu Eissen S. Document Categorization with MajorClust // Proceedings of the 12th Workshop on Information Technology and Systems (WITS–02). – Barcelona, Spain: 2002. – P. 91–96.

[28] Stein B., Niggemann O. On the Nature of Structure and its Identification // P. Widmayer, G. Neyer, S. Eidenbenz (eds.). Graph-Theoretic Concepts in Computer Science. LNCS 1665. – Springer-Verlag, 1999. – P. 122–134.

Automatic Term Clustering in the Corpus of Russian Texts on Corpus Linguistics

Vinogradova N., Mitrofanova O., Panicheva P.

The article deals with the results of automatic processing of raw scientific texts in Russian. The essence of the experiment in question is automatic clustering of terms-descriptors used in the texts from the corpus on Corpus Linguistics. The procedure is fulfilled with the help of specialized research tool.

♣Отдельные этапы работы выполняются при финансовом обеспечении из средств гранта Президента РФ для поддержки молодых российских ученых № МК-9701.2006.6.
Авторы выражают благодарность В.П. Захарову за мудрые советы в области корпусной лингвистики, способствовавшие развитию проекта, и за предоставление корпусных данных, А.С. Мухину за консультации по программированию лингвистических задач, участникам семинара по корпусной и компьютерной лингвистике ИЛИ РАН за полезные обсуждения, студентам кафедры математической лингвистики СПбГУ за помощь в подготовке текстов для размещения в корпусе, а также рецензентам конференции RCDL'2007 за ценные замечания по совершенствованию текста статьи.