АВТОМАТИЧЕСКАЯ КЛАССИФИКАЦИЯ ЛЕКСИКИ В РУССКОЯЗЫЧНЫХ ТЕКСТАХ НА ОСНОВЕ ЛАТЕНТНОГО СЕМАНТИЧЕСКОГО АНАЛИЗА[1]

AUTOMATIC WORD CLUSTERING IN RUSSIAN TEXTS BASED ON LATENT SEMANTIC ANALYSIS

Митрофанова О.А*. (alkonost-om@yandex.ru)*
Мухин А.С. *(antonmuhin@gmail.com)*
Паничева П.В. *(ppolin@yandex.ru)*
Санкт-Петербургский государственный университет

Исследование посвящено созданию и использованию инструмента автоматической классификации лексики, применимого при анализе неразмеченных русскоязычных текстов. Обсуждаются результаты экспериментов по кластеризации со сменой параметров, проведённых на материале текстов различных типов.

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 [Азарова, Марина 2006; Сидорова 2005; Pantel, Lin 2002; Pekar 2004; Shin, Choi 2004]. 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. [Баглей, Антонов, Мешков, Суханов 2006; Браславский 2004; Крижановский 2006; Buscaldi, Rosso, Alexandrov, Ciscar 2006; Stein, Meyer zu Eissen, 2002].

Процедуры АКЛ широко применяются в прикладных лингвистических разработках. Для исследователей открыт доступ к ряду специализированных ресурсов, позволяющих выделять из корпусов текстов кластеры близких по значению слов: см., например, COALS (**http://dlt4.mit.edu/~dr/COALS/**), InfoMap (**http://infomap.stanford.edu**), Google-Sets (**http://labs.google.com/sets**), SenseClusters (**http://senseclusters.sourceforge.net/),** LexSem (**http://www.isi.edu/~pantel/Content/Demos/LexSem/cbc.htm**), Word Clusters (**http://www.cs.ualberta.ca/~lindek/demos/wordcluster.htm**), DSM (**http://clg.wlv.ac.uk/demos/similarity/**). Надо признать, что число ресурсов, которые выполняют процедуру АКЛ и смежные с ней операции на основе русскоязычных текстов и лексикографических баз, не столь велико. Очевидно, что создание открытых модулей АКЛ для русского языка стало необходимостью.

2. Цели исследования, используемые методы и процедуры

Обсуждаемый в настоящей статье проект направлен на построение русскоязычного ресурса АКЛ, который позволял бы качественно выделять лексико-семантические классы слов из текстов разных объёмов и разных типов, допускал бы классификацию лексики с различными условиями, открывал бы возможность использования результатов классификации в других системах автоматической обработки текста.

[1] Исследование выполняется при финансовой поддержке гранта Президента РФ для поддержки молодых российских ученых № МК-9701.2006.6. Авторы благодарят И.В. Азарову, М.А. Александрова, В.П. Захарова, А.С. Марину, В.С. Савицкого за помощь в реализации проекта.

Реализация проекта проводится в несколько этапов, первым из которых является создание инструмента АКЛ для работы с неразмеченными текстами. В дальнейшем планируется усовершенствование инструмента АКЛ для обработки размеченных текстов и корпусов параллельных текстов.

Поставленная цель предполагает компьютерную реализацию алгоритма кластерного анализа, или автоматической классификации объектов без учителя. При АКЛ возможно использование целого ряда методов кластеризации: иерархических (аггломеративных, дивизимных), неиерархических (например, итеративных – *K*-средних, *K*-медианы), гибридных методов (анализ различных подходов к кластеризации лингвистических объектов см., например: [Pantel 2003; Stein, Niggemann 1999]). 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 первом этапе реализации проекта были задействованы иерархический (аггломеративный) метод кластеризации и неиерерархический метод (*K*-средних).

Выделение кластеров лексем в тексте производится на основе процедуры латентного семантического анализа (ЛСА). Идеология и практика ЛСА обсуждается во многих исследованиях (см., например, **http://lsi.research.telcordia.com/lsi/LSIpapers.html**). С лингвистической точки зрения, суть ЛСА заключается в возможности определения содержательной близости лексических единиц при сопоставлении их синтагматических свойств (иначе говоря, их сочетаемости с другими элементами контекста, дистрибуции) [Gamallo, Gasperin, Augustini, Lopes 2001; Smrž, Rychlý 2001]. 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. [Митрофанова 2005; Митрофанова 2006]). 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.

3. Компьютерная реализация инструмента автоматической классификации лексики

В ходе подготовки экспериментов по АКЛ велась разработка соответствующего программного обеспечения. Программа АКЛ, созданная на языке Python, предусматривает три блока: блок предварительной обработки текста и вычисления расстояний между исследуемыми лексемами, 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*).

Первый блок программы обеспечивает обработку входного текста. Прежде всего, обнаруживаются все вхождения исследуемых лексем в текст, затем производится автоматическое выделение границ контекстов в соответствии с заданной шириной контекстного окна. Далее возможно автоматическое определение весов элементов контекста: чем ближе позиция элемента контекста к исследуемой лексической единице, тем выше его вес, и наоборот. В дальнейшем для каждой лексемы *l* формируется множество контекстов её употребления, которое представляется в виде вектора дистрибуции в *N*-мерном пространстве. Измерения пространства задаются элементами контекстов k_i ($i = 1 \ldots N$) для исследуемой лексемы, а значения координат вектора соответствуют коэффициенту взаимной встречаемости *l* и k_i. Затем производится операция сравнения векторов дистрибуций всех исследуемых лексем применительно к обрабатываемому тексту. Вычисление расстояний *d* осуществляется с использованием меры Евклида. Особенностью данной меры является её нелинейный характер, что следует учитывать при интерпретации количественных результатов. Итог работы программы на данном этапе – матрица расстояний между векторами дистрибуций для каждой пары исследуемых лексем. Данные о близости дистрибуций лексем в обрабатываемом тексте используются при кластеризации.

Второй и третий блоки программы обеспечивают кластеризацию лексем анализируемого текста. По выбору пользователя производится иерархический кластерный анализ или кластерный анализ методом *K*-средних.

При осуществлении иерархического кластерного анализа в тексте осуществляется пошаговое формиро-

вание совокупностей двух и более лексем, имеющих близкую дистрибуцию и образующих кластеры. Процедура повторяется до тех пор, пока все лексемы не объединятся последовательно в один кластер, или пока количество промежуточных кластеров (фактически, глубина иерархии) не достигнет числа, указанного пользователем. Таким образом происходит построение иерархической структуры, отражающей сходства и различия дистрибуций лексем: лексемы со сходной дистрибуцией на начальном шаге попадают в наиболее глубокий кластер, а слова с различной дистрибуцией на последнем шаге присоединяются к наиболее обширному кластеру или формируют собственные кластеры.

Альтернативный метод кластерного анализа – метод *K*-средних, при выборе которого пользователь может заранее указать желаемое число кластеров. В этом случае элементы кластеров назначаются случайным образом, затем вычисляются центры кластеров. На каждом шаге итерации для каждого элемента заново происходит поиск ближайшего кластера, а также заново вычисляются центры кластеров. Процедура повторяется до тех пор, пока элементы не перестанут изменять своё местоположение в структуре, т.е. пока центры кластеров не стабилизируются.

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. Эксперименты по автоматической классификации лексики

Для оценки эффективности работы и определения исследовательских возможностей разрабатываемого инструмента АКЛ была проведена серия экспериментов по извлечению и обработке данных из неразмеченных текстов:

- автоматическая классификация терминов-дескрипторов в научных текстах (на материале статей из русскоязычного корпуса по корпусной лингвистике);
- автоматическая классификация глагольной лексики в экспериментальном корпусе (на материале базовых глаголов русского языка и корпуса глагольных контекстов);
- автоматическая классификация лексики в параллельных текстах (на материале тематической группы существительных – обозначений живых существ, используемых в тексте оригинала и перевода повести-притчи Дж. Оруэлла «Скотный двор»).

4.1. Автоматическая классификация терминов-дескрипторов в научных текстах

Автоматическая обработка научных текстов – одна из актуальных прикладных задач, решение которой может потребовать осуществления процедуры АКЛ на основе неразмеченного корпуса. 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.

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 трёх 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.), «Корпусная лингвистика – 2006» (Санкт-Петербург, 10–14 октября 2006 г.) [КЛ и ЛБД 2002; КЛ 2004; КЛ 2006]. При подготовке текстов к размещению в корпусе производится их метаразметка, которая предполагает фиксацию основных параметров каждой статьи в её паспорте [Волков, Захаров, Дмитриева 2004]. Наряду с библиографическим описанием в число параметров статьи предлагается включить наборы релевантных терминов-дескрипторов, уточняющих содержательные характеристики текста.

В качестве иллюстрации работы обсуждаемого инструмента АКЛ произвольным образом были отобраны 10 статей из корпуса (далее T1–T10) – это тексты различной тематики, отражающие широкий спектр проблем корпусной лингвистики: определение корпусной лингвистики как особой области научной деятельности, проти-

вопоставление её другим 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.

Для каждой из статей были выявлены 10 терминов-дескрипторов, позволяющих диагностировать тематическую принадлежность текста. Термины-дескрипторы представлены в нормализованном виде: в набор включается лемма, которая противопоставляется входящим в текст словоформам, например: **корпус** (*корпус, корпусов, корпусе, корпуса, корпусы, корпусах, корпусом, корпусу*) и пр.

В ходе эксперимента производилась автоматическая обработка исследуемых 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 учетом весов элементов контекстов);
- осуществление кластеризации терминов-дескрипторов для каждого текста иерархическим методом и методом *K*-средних с различными параметрами (при глубине иерархии / конечном числе кластеров $C = 3, 5, 7, 9$).

В качестве примера приведём результаты обработки текста T1, описываемого терминами-дескрипторами (*архив, банк, данные, корпус, массив, поиск, разметка, текст, формат, чешский*).

Установлена частота употребления терминов-дескрипторов в тексте T1 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*) = 0,0
***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

Получены результаты кластеризации с использованием иерархического метода и метода *K*-средних (глубина иерархии / конечное число кластеров $C = 3, 5, 7, 9$) (см. табл. 1). Очевидно, последовательность формирования кластеров терминов-дескрипторов отражает естественные связи элементов исследуемого текста, что подтверждается частотными данными и значениями расстояний между парами элементов. Что касается параметров кластеризации, то наиболее удачные результаты получены при значении $C = 5$ (иерархический метод), $C = 5, 7$ (метод *K*-средних).

C	***Иерархический метод***
3	(архив, банк, массив, разметка, формат, чешский, (поиск ((текст, корпус) данные)))
5	**(архив, банк, массив, формат (разметка (чешский (поиск ((текст, корпус) данные)))))**
7	(архив, банк, (массив ((разметка (чешский (поиск ((текст, корпус) данные)))) формат)))
9	(банк (архив (массив ((разметка (чешский (поиск ((текст, корпус) данные)))) формат)))
C	***Метод K-средних***
3	((архив) (банк) (данные, корпус, массив, поиск, разметка, текст, формат, чешский))
5	**((архив) (банк) (разметка) (данные, корпус, поиск, текст, формат, чешский) (массив))**
7	((архив) (банк) (данные, корпус, текст) (формат, чешский) (массив) (поиск) (разметка)
9	((архив) (банк) (данные) (чешский) (массив) (поиск) (разметка) (корпус, текст) (формат))

Таблица 1. Результаты кластеризации терминов-дескрипторов для текста T1

Аналогичные данные о частотности терминов-дескрипторов, о расстояниях между ними в наборах для соответствующих текстов, о вариантах кластеризации при заданных условиях были получены для остальных исследуемых текстов. Результаты кластеризации текстов T1–T10 (см. табл. 2) позволяют оценить диапазон категорий, в большей или меньшей степени релевантных для предметной области «Корпусная лингвистика». Вероятно, такие термины-дескрипторы, как *корпус*, *текст*, *данные*, *разметка*, *поиск*, представляют понятийное ядро указанной предметной области.

Текст	**Иерархический метод**
T1	(архив, банк, массив, формат (разметка (чешский (поиск ((текст, корпус) данные)))))
T2	(массив, база, данные ((переводческая, память) (система (текст, перевод))) (корпус, параллельный))
T3	(поиск, ИПС, программа, индоевропейский (((язык (тезаурус (текст, модуль))) корпус) интерфейс))
T4	(построение, компьютерный, тезаурус, понятие (словарь (валентность (частота (контекст (корпус, текст))))))
T5	(метаразметка, разметка, словарь, паспорт ((исторический (данные ((корпус, текст) параметр))) метаданные))
T6	(параметр, единица, категория, значение (фундаментальная ((разметка (падеж, корпус)) (лингвистика, корпусная))))
T7	(источник, поиск, словарь, картотека (корпус ((топоним ((ландшафт, культурный) топонимический)) данные)))
T8	(категоризация, категория, класс, инженерия (классификация (корпус ((текст (семантическая, разметка)) лингвистика))))
T9	(поиск, интернет, запрос, пользователь (сервис (частота ((текст, корпус) (слово, биграмма)))))))
T10	(разметка, формат, поиск, тег (((фрагмент (слово (текст, житие))) корпус) цитата))
Текст	**Метод К-средних**
T1	((архив) (банк) (разметка) (данные, корпус, поиск, текст, формат, чешский) (массив))
T2	((данные) (корпус) (параллельный) (перевод, текст, переводческая, память, массив, база) (система))
T3	((корпус, текст) (программа) (тезаурус, ИПС, модуль, индоевропейский, язык) (поиск) (интерфейс))
T4	((корпус, текст, словарь, контекст, понятие, валентность) (частота) (построение)) (компьютерный) (тезаурус))
T5	((данные, паспорт, параметр) (исторический) (разметка) (метаразметка) (корпус, текст, словарь, метаданные))
T6	((параметр) (лингвистика, корпусная, фундаментальная) (единица, категория) (корпус, значение, падеж) (разметка))
T7	((картотека) (словарь) (источник) (корпус, данные, топоним, топонимический, ландшафт, культурный) (поиск))
T8	((семантическая, категория, классификация, класс, лингвистика, инженерия) (разметка) (корпус) (текст) (категоризация))
T9	((интернет) (запрос, пользователь) (сервис, поиск, биграмма, слово) (частота) (корпус, текст))
T10	((корпус) (текст, слово, житие, цитата, фрагмент, тег) (разметка) (формат) (поиск))

Таблица 2. Результаты кластеризации терминов-дескрипторов для текстов T1–T10 (C = 5)

В целях уточнения характера связей между понятийными категориями, выраженными исследуемыми терминами, была проведена серия экспериментов с текстами, для которых наблюдается частичное совпадение наборов дескрипторов. Кластеризация совпадающих элементов в наборах терминов-дескрипторов производилась с помощью иерархического метода.

В ряде случаев результаты кластеризации совпадающих терминов-дескрипторов для разных текстов оказались идентичными. Так, применительно к текстам T1 и T2 четвёрка общих дескрипторов упорядочивается единообразно – (*массив* (*данные* (*корпус, текст*))); то же самое можно отметить и в отношении троек дескрипторов для текстов T4 и T5 – (*словарь* (*корпус, текст*)), а также T4 и T9 – (*частота* (*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 текстов T1 и T10, устанавливаются следующим образом: T1 – (*формат* (*разметка* (*поиск* (*текст, корпус*)))); T10 – (*разметка* (((*корпус, текст*) *формат*) (*поиск*)). Применительно к текстам T9 и T10 общие дескрипторы также упорядочиваются по-разному: T9 – (*поиск* (*слово* (*текст, корпус*))); T10 – (*поиск* (*корпус* (*слово, текст*))). Не совпали результаты кластеризации тройки дескрипторов, одновременно характеризующих тексты T6 и T8: T6 – (*лингвистика* (*разметка, корпус*)); T8 – (*корпус* (*лингвистика, разметка*)).

Сравнение иерархий терминов-дескрипторов, полученных для разных документов, даёт основания для оценки степени близости самих текстов. 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.

Итак, автоматическая обработка текстов статей из 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 и классификацию текстов внутри этих областей, а также разработку инструментов для определения количественных оценок близости текстов.

4.2. Автоматическая классификация глагольной лексики в экспериментальном корпусе

Среди используемых в настоящее время инструментов работы с корпусами довольно значительную роль играют те компьютерные средства, которые ориентированы на автоматическую обработку данных о глагольной лексике. Процедуры АКЛ позволяют извлекать и анализировать информацию о разграничении отдельных значений глаголов, об их сочетаемости, о семантических и синтаксических валентностях непосредственно из корпусов текстов (см., например, [Азарова, Марина 2006; Chklovski, Pantel 2004; Goetz, Hogue 2004; Resnik, Diab 2000; Wunsch, Hinrichs 2006] и пр.). В свою очередь, формализованные описания глагольной лексики, основанные на эмпирических данных, необходимы для эффективного функционирования систем автоматического понимания текстов (см., например, [Giuglea, Moschitti 2004; Chen, Palmer 2005; Palmer, Rosenzweig, Cotton 2001; Lopatková, Bojar, Semecký, Benesová, Zabokrtský 2005] и пр.). Поэтому в ходе реализации обсуждаемого проекта значительное внимание было уделено проверке эффективности работы инструмента АКЛ на материале корпуса глагольных контекстов.

Для проведения эксперимента были отобраны 14 высокочастотных глаголов русского языка, относящихся к основным лексико-семантическим классам (интеллектуальная деятельность, восприятие, владение, созидательная деятельность, социальная деятельность, физическое воздействие, перемещение/местоположение в пространстве и пр.) – *думать, понимать, видеть, смотреть, брать, дать, делать, работать, держать, бросать, идти, ехать, стоять, лежать* [ТСРГ 1999]. Эксперименты по кластеризации проводились на основе корпуса глагольных контекстов, включающего свыше 100 тыс. с/у [Азарова, Марина 2006]. Данный корпус был сформирован на основе случайных выборок контекстов употребления глаголов русского языка из корпуса Бокрёнок, применяемого на кафедре математической лингвистики СПбГУ. В ходе эксперимента использовалась версия корпуса глагольных контекстов без морфологической разметки.

Исследуемые лексемы, представляющие ядерную часть глагольной лексики русского языка, являются многозначными и одновременно соотносятся с несколькими лексико-семантическим классами. Вместе с тем, валентные свойства данных глаголов, предопределяющие их поведение в контексте и задающие круг их синтагматических соседей, всё же позволяют их дифференцировать. Об этом свидетельствуют результаты экспериментов по иерархической кластеризации четвёрок глаголов при ширине контекстного окна [-5…+5] и с учётом весов элементов контекстов. Например, удалось корректно сформировать кластеры в следующих наборах глагольных лексем:

(*идти, ехать* (*видеть, смотреть*))
(*идти, ехать* (*делать, работать*))
(*брать, дать* (*видеть, смотреть*))
(*держать, бросать* (*думать, понимать*))
(*стоять, лежать* (*думать, понимать*))

Данные о расстояниях между глаголами, полученные на основе сравнения их дистрибуций в корпусе, также выглядят вполне убедительно и допускают содержательную интерпретацию. Оценка разброса значений расстояний между глаголами показала, что расхождения являются значимыми.

Глаголы, относящиеся к одному лексико-семантическому классу, проявляют высокое сходство дистрибуций (расстояние между ними мало); при кластеризации они формируют целостные кластеры:

d(*думать, понимать*) = 0,107 d(*делать, работать*) = 0,117

Глаголы, представляющие разные лексико-семантические классы, отличаются по своим сочетаемостным способностям (что подтверждается высоким значением расстояния):

d(*понимать, лежать*) = 0,152 d(*видеть, идти*) = 0,151

Любопытно заметить, что различие дистрибуций наблюдается также у глаголов, хотя и принадлежащих к одному лексико-семантическому классу, но находящихся в отношениях контраста, например, являющихся конверсивами: d(*брать, дать)* = 0,131. Тем не менее, это не препятствует их корректной кластеризации. Следовательно, расхождение дистрибуций данных глаголов менее существенно, чем расхождение дистрибуций глаголов из разных лексико-семантических классов.

Итак, в ходе экспериментов с четвёрками глаголов из экспериментальной группы удалось получить положительные результаты. Хотя исследования свидетельствуют о том, что при решении задач АКЛ применительно к глагольной лексике предпочтительна работа с размеченными корпусами значительного объёма и осуществление кластеризации по тегам, есть аргументы в пользу того, что желаемая цель может быть достигнута и при обращении к неразмеченному корпусу текстов.

4.3. Автоматическая классификация лексики в параллельных текстах

Инструменты АКЛ открывают весьма привлекательные возможности при исследовании параллельных текстов. Сравнительный анализ количественных данных об употреблении слов, о степени их семантической близости помогает устанавливать распределение лексических единиц разных языков внутри лексико-семантических и тематических групп. Информация о соотношении элементов кластеров, полученная при параллельной обработке текстов оригинала и перевода, имеет высокую ценность в определении адекватности перевода, при проведении контрастивных исследований [Андреева 2006; Беляева 2004; Гарабик, Захаров 2006]. Очевидно, что применение модулей классификации повышает эффективность поиска в параллельных корпусах, позволяет извлекать данные для пополнения и корректировки многоязычных словарей, для проверки качества работы систем машинного перевода [Widdows, Dorow, Chan 2002].

Эксперименты по классификации лексики в параллельных текстах производились на основе текстов оригинала и перевода повести-притчи Дж. Оруэлла «Скотный двор» (G. Orwell, «Animal Farm»). Используемые тексты представлены в электронной библиотеке М. Мошкова. Объем текстов составляет примерно 24 тыс. с/у (русский текст), 30 тыс. с/у (английский текст). Исследовалась тематическая группа «Живые существа», объединяющая существительные, которые обозначают человека, животных и птиц. В экспериментальный набор вошли более 50 лексем, присутствующих в оригинальном и переводном текстах и релевантных с точки зрения сюжета. Эксперименты по иерархической кластеризации лексем проводились при ширине контекстного окна [-5…+5], с учётом весов контекстных элементов.

Обработка текстов с помощью обсуждаемого инструмента АКЛ позволила поучить данные о частотности лексических единиц, о расстояниях между исследуемыми словами в пределах текстов оригинала и перевода, а также о вариантах их кластеризации.

Используемый инструмент АКЛ успешно справляется с задачей разграничения микрогрупп в пределах заданной тематической группы. Так, кластеризация позволяет противопоставлять названия животных и птиц:

(*ворон* (*овца, животное*))	(*raven* (*animal, sheep*))
(*цыплята* (*животное, кошка*))	(*chickens* (*cat, animal*))
(*осел* (*утка, птица*))	(*donkey* (*duck, bird*))
(*коза* (*утята, птица*))	(*goat* (*ducklings, bird*))

Следует заметить, что и для русских, и для английских видовых наименований родовое имя определяется корректно:

((*голубь, утка*) *птица*) (*bird* (*duck, pigeon*)).

Кластеризация существительных, являющихся названиями представителей одного вида, также была успешной. В некоторых случаях иерархия имён для русского и английского текстов является идентичной:

(*цыплята* (*курица, петух*)) (*chickens* (*hen, cockerel*)).

Несовпадение результатов кластеризации можно объяснить асимметрией переводческих соответствий и/или различной частотой употребления элементов пары «лексема языка оригинала – лексема языка перевода» (*лошадь* (*f* = 20); *кобыла* (*f* = 2), *кобылка* (*f* = 2); *жеребята* (*f* = 2), *жеребец* (*f* = 2); *horse* (*f* = 25); *mare* (*f* = 4); *foal* (*f* = 2)):

(*кобыла* (*жеребята, лошадь*)) (*foal* (*horse, mare*)).

Отдельного обсуждения заслуживают результаты кластеризации в следующих четвёрках:

(*птица* (*человек, животное*) *боров*) (*bird* (*boar* (*man, animal*))).

В рассматриваемом случае наиболее вероятной причиной отступления от логического соотношения родовых имён и видового наименования является специфика употребления существительных *боров* и *boar* в тексте повести-притчи, а также особенности её сюжетной линии. Это подтверждается данными о расстояниях между парами лексем:

d(*animal, boar*) = 0,783	*d*(*животное, боров*) = 0,817
d(*animal, human*) = 0,206	*d*(*животное, человек*) = 0,224
d(*animal, bird*) = 0,950	*d*(*животное, птица*) = 0,838

Проведённые эксперименты подтверждают перспективность совершенствования инструмента АКЛ для дальнейшей работы с корпусами параллельных текстов.

5. Перспективы развития исследования

Осуществление первого этапа проекта по созданию и практическому применению инструмента АКЛ, рассчитанного на работу с русскоязычными текстами, привело к желаемым результатам.

В дальнейшем планируется:

- работа по техническому совершенствованию инструмента: введение дополнительных возможностей при кластеризации (расширение круга используемых методов кластеризации), при измерении расстояний между лексемами (добавление метрик); модернизация пользовательского интерфейса (добавление режима визуализации результатов);
- проведение лингвистических экспериментов с более сложными параметрами: обработка размеченных текстов, текстов различной тематической принадлежности, разнообъёмных текстов, параллельных текстов; кластеризация в наборах лексем при различных условиях; использование инструмента АКЛ в решении практических задач по извлечению семантической информации из корпусов текстов.

Список литературы

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. Андреева Е.Г. Анализ переводческих соответствий на материале параллельного корпуса тестов // Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции «Диалог–2006». М.: 2006. С. 26–30.
3. 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.
4. Беляева Л.Н. Лексикографический потенциал параллельного корпуса текстов // Труды международной конференции «Корпусная лингвистика – 2004». СПб.: 2004. С. 55–64.
5. 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.
6. 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.
7. Гарабик Р., Захаров В.П. Параллельный русско-словацкий корпус // Труды международной конференции «Корпусная лингвистика – 2006». СПб.: 2006. С. 81–87.
8. КЛ и ЛБД 2002 – Доклады научной конференции «Корпусная лингвистика и лингвистические базы данных». СПб.: 2002.
9. КЛ 2004 – Труды международной конференции «Корпусная лингвистика – 2004». СПб.: 2004.
10. КЛ 2006 – Труды международной конференции «Корпусная лингвистика – 2006». СПб.: 2006.
11. 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.
12. 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.
13. 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.
14. 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.
15. ТСРГ 1999 – Толковый словарь русских глаголов: Идеографическое описание / Под ред. Л.Г. Бабенко. М.: 1999.
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. Chen J., Palmer M. Clustering-based Feature Selection for Verb Sense Disambiguation // Proceedings of the 2005 IEEE International Conference on Natural Language Processing and Knowledge Engineering (IEEE NLP–KE 2005). Wuhan, China: 2005. P. 36–41.
18. Chklovski T., Pantel P. VerbOcean: Mining the Web for Fine-Grained Semantic Verb Relations // Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP–04). Barcelona, Spain: 2004. P. 33–40.
19. 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.
20. Giuglea A.-M., Moschitti A. Knowledge Discovering Using FrameNet, VerbNet and PropBank // Proceedings of the Workshop on Ontology and Knowledge Discovering (ECML–2004). Pisa, Italy: 2004. P. 929–936.
21. Goetz M., Hogue A. Parallel Clustering of English Verbs into Levin Classes // MIT 6.338/6.863 Joint Final Project Report, May 2004. URL: **http://secondthought.org/notes/6.338-final.html**
22. Lopatková M, Bojar O., Semecký J., Benesová V., Zabokrtský Z. Valency Lexicon of Czech Verbs VALLEX: Recent Experiments with Frame Disambiguation // Text, Speech and Dialogue: Eighth International Conference TSD–2005. LNAI 3658. Springer-Verlag, 2005. P. 99–106.
23. Palmer M., Rosenzweig J., Cotton Sc. Automatic Predicate Argument Analysis of the Penn TreeBank // Proceedings of HLT–2001: First International Conference on Human Language Technology Research. San Francisco: 2001. URL: **http://acl.ldc.upenn.edu/H/H01/H01-1010.pdf**
24. Pantel P. Clustering by Committee. Ph.D. Dissertation. Department of Computing Science, University of Alberta: 2003. URL: **http://www.isi.edu/~pantel/Content/publications.htm**
25. Pantel P., Lin D. Document Clustering with Committees // SIGIR–02. Tampere: 2002. URL: **http://www.isi.edu/-~pantel/Content/publications.htm**
26. 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.
27. Resnik P., Diab M. Measuring Verb Similarity // Proceedings of the 22nd Annual Meeting of the Cognitive Science Society (CogSci). Philadelphia: 2000. P. 399–404.
28. 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.
29. 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.
30. 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.
31. 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.
32. Widdows D., Dorow B., Chan Ch.-K. Using Parallel Corpora to Enrich Multilingual Lexical Resources // Third International Conference on Language Resources and Evaluation (LREC–2002). Las Palmas: 2002. P. 240–245.
33. Wunsch H., Hinrichs E.W. Latent Semantic Clustering of German Verbs with Treebank Data // Proceedings of the Fifth Workshop on Treebanks and Linguistic Theories (TLT–2006). Prague, Czech Republic: 2006. P. 151–162.