АВТОМАТИЗИРОВАННАЯ КЛАССИФИКАЦИЯ КОНТЕКСТОВ ПРИ ПОДГОТОВКЕ ДАННЫХ ДЛЯ КОМПЬЮТЕРНОГО ТЕЗАУРУСА RUSSNET

AUTOMATIC CONTEXT CLUSTERING FOR COMPUTER THESAURUS RUSSNET

И. В. Азарова *(azic@bsr.spb.ru)*
Санкт-Петербургский государственный университет

А. С. Марина (*a_s_marina@rambler.ru)*
Санкт-Петербургский государственный университет

Методика построения компьютерного лексикона RussNet включает задание рамок валентностей для единиц словаря. Параметры рамок позволяют более четко разграничивать «синсеты» тезауруса и разрешать неоднозначность при автоматическом анализе текста. Спецификация рамок валентностей определяется по статистически устойчивым контекстным маркерам значения в корпусе текстов: морфо-синтаксическим и семантическим.

В докладе обсуждается процедура автоматизированной классификации морфологически размеченных контекстов корпуса текстов, которая будет использоваться как предварительное распределение контекстов по лексическим группам – семантическим деревьям RussNet. Процедура включает вычисление дистрибуции тегов леммы в заданном окне анализа, определение метрики сходства дистрибуций, задание центров кластеров для дистрибуций основных семантических деревьев. Обсуждаются предварительные результаты классификации для контекстов глаголов.

1. Введение

Компьютерный тезаурус RussNet[1], который разрабатывается на кафедре математической лингвистики Санкт-Петербургского государственного университета, строится на основании принципов, общих для wordnet-словарей (Fellbaum 1997). Словарь опирается на данные корпуса современных тексов объемом 21 млн. словоупотреблений, базовую часть (60%) которого составляют газетные и журнальные статьи. Тематика статей довольно разнообразна, поэтому эта часть корпуса позволяет задать общеупотребительную лексику. Кроме того, в корпус включены деловые тексты (10%), научные (15%) и литературно-художественные (15%), что позволяет выделить употребительную экономическую, социальную и научную терминологию. Словарь RussNet является оригинальным ресурсом (Азарова, Синопальникова, Яворская, 2003) в том смысле, что он не был переведен с Принстонского прототипа WordNet.

В процессе разработки в тезаурус были внесены дополнительные структуры, которые были связаны как с методикой построения словаря, так и ориентированы на использование словаря в процедурах автоматической обработки текста (Азарова, Иванов, Овчинникова, 2005).

Единицей тезауруса RussNet является синонимический ряд (синсет), в который входят как отдельные слова, так и устойчивые выражения (MWE – multiword expressions), используемые в корпусе для передачи соответствующего значения. Элементы синсета упорядочены в соответствии с частотой употребления лексической единицы (слова, выражения) в данном значении в корпусе. Единицей измерения частоты в корпусе является ipm (items per million) – частота вхождений на 1 миллион словоупотреблений в корпусе.

Для задания частотного упорядочения значений полисемантичного слова используется разметка выборочной совокупности контекстов корпуса. Эта процедура производится вручную, что приводит к большим затратам времени. Было исследовано варьирование объема выборки и выявлено, что 100-150 случайно выбранных контекстов в корпусе дают такое же распределение долей значений, что и 1000, что подтверждается и другими работами (Leacock & Chodorow, 1998), в которых использовалась разметка значений в контекстах.

Распределение долей значения в корпусе дает довольно характерную картину: около 80% слов имеют четко выраженное первое частотное значение, которое представлено в 50-70% контекстах корпуса, низкочастотные значения представлены 1-3%.

[1] http://www.phil.pu.ru/depts/12/RN

Методика разметки значений в выборочной совокупности потребовала введения описания частоты встречаемости грамматических и семантических параметров контекстных маркеров, которые позволяют объективировать принятие решения относительно сходства или различия значений. Для описания выделенных контекстных маркеров задается особая структура – рамка валентностей, которая «расширяет» структуру wordnet-словаря, поскольку в Принстонском тезаурусе WordNet не было предусмотрено описание контекстной информации. Рамка валентностей в RussNet имеет подробную спецификацию.

2. Спецификация валентных рамок

Рамки валентностей включают описание статистически устойчивых характеристик размеченных контекстов (Azarova et al, 2005). Среди общих параметров рамки указываются

(а) ее активность или пассивность, которая определяет, в каком синтаксическом качестве – главного или зависимого слова – выступают описываемые элементы синсета;

(б) в какой синтаксической структуре – предикативной или атрибутивной – реализуется валентность.

В рамку может входить несколько валентностей. Параметрами валентности являются

(в) ее обязательность-факультивность, которая вычисляется как частотность реализации данной валентности в корпусе: низкочастотные (<35%) валентности не вносятся в рамку, границей устойчивости валентности является пороговое значение (65%);

(г) семантическая функция валентности, которая понимается как позиция в аргументной структуре пропозициональных значений семантических деревьев RussNet, объединяющих синсеты родовидовыми отношениями.

Валентность может иметь несколько частотных вариантов реализации, при этом каждый вариант описывается

(д) семантическими ограничениями, накладываемыми на реализацию валентной позиции, которые задаются в терминах структур и подструктур семантических деревьев RussNet (целые деревья, поддеревья, отдельные синсеты), а также группировок деревьев;

(е) морфо-синтаксическими параметрами, т. е. частотными заполнениями позиции словами определенной ЧР с уточнением других параметров поверхностной структуры (предложно-падежная форма существительных, видовое значение инфинитива, лексико-грамматическая характеристика наречий и т. п.).

Описанная структура рамок валентностей используется в процедуре разрешения неоднозначности (Azarova et al, 2005), в которой сопоставляется полученный грамматический разбор фразы/предложения, и рамки валентности для лемм, выделенных в процессе анализа. Случаи соответствия рамок разбору позволяют либо однозначно выбрать вариант анализа, либо выбрать предпочтительные варианты разбора.

Соотношение параметров рамок валентностей в гипонимическом дереве используется в сложных случаях принятия решений о семантических связях между значениями (Азарова, Иванов, Овчинникова, 2006). Там же рассматривается сложная схема наследования рамок валентностей, которая показывает, что статистически устойчивые контекстные признаки охватывают большую часть вхождений значений из семантического дерева в текст. Все это позволяет предположить, что возможна формулировка процедуры автоматической классификации контекстов как реализации значений из некоторого (или некоторых) деревьев тезауруса RussNet.

3. Использование дистрибуции морфологически размеченных контекстов

При разработке процедуры автоматической классификации контекстов мы использовали идеи, которые были сформулированы в ряде работ по вычислению контекстных признаков для разграничения значений слов на материале английского языка.

3.1. Дистрибуции ЧР тегов

Работа К. Ликок и М. Ходорова (Leacock & Chodorow) была посвящена разграничению значений полисемантичного слова *serve* на основе выборки обучающих контекстов, которые описывались тремя дистрибуциями. Первая дистрибуция вычислялась по ЧР разметке обучающих контекстов: в каждой позиции окна анализа[2] накапливалась частота тегов. Во вторую дистрибуцию включались слова, не получившие ЧР разметки, в первую очередь, знаки препинания, предлоги, артикли и числа. Третья дистрибуция включала леммы основных ЧР, которые были объединены в группу левого и правого контекста без учета удаленности от рассматриваемого слова. Для каждого из четырех значений глагола были получены такие дистрибуции.

При тестировании размеченный контекст сравнивался с полученными дистрибуциями, мера сходства определялась через произведение вероятностей появления тега в i-ой позиции окна анализа. Поскольку вероятность в произведении могла быть нулевой, к ней прибавлялась единица. Оценки, полученные по трем дистрибуциям, перемножались, самое большое произведение указывало на нужное значение.

В результате данного исследования было показано, что использование дистрибуции ЧР тегов и дополнительных признаков контекста позволяют разграничивать значения полисемантичного слова довольно устойчиво (80-83%). Результаты анализа зависят от окна анализа и того, насколько хорошо представлены обучающие контексты для значений.

[2] Первоначально в качестве окна анализа было выбрано ±2 позиции от рассматриваемого слова.

Сопоставляя результаты классификации с другими исследованиями авторы пришли к следующим выводам (1) объем первичной обработки текста (например, выделение синтаксически связанных фрагментов) не влияет существенно на результат;
(2) невозможно достичь устойчивого разграничения низкочастотных значений, поскольку сложно создать удовлетворительную обучающую выборку;
(3) омонимы (т. е. контрастные значения) разграничиваются гораздо надежнее, чем сходные значения;
(4) объемные обучающие совокупности улучшают результат, но не в такой степени, в какой возрастает время вычислений и подготовки таких совокупностей.

3.2. Исследование закономерностей распределения ЧР тегов в локальных контекстах лексико-семантических групп глаголов

Приведенный выше метод исследования дает хорошие результаты применительно к языкам с фиксированным порядком слов, но, возможно, мало применим к языкам типа русского со свободным порядком слов. Для предварительной проверки было проведено «пилотное» исследование, в ходе которого сравнивались дистрибуции двух групп: глаголов говорения (*сказать, говорить, спросить, ответить, просить*) и движения (*идти, пойти, выйти, вернуться, ходить*). Для каждого глагола из корпуса было отобрано 200 контекстов в прямом значении (без переносных или фразеологически связанных употреблений). Окном анализа были [-6…+6] позиций от рассматриваемого глагола, при этом одну позицию могло занимать слово или знак препинания.

Сравнение производилось по трем распределениям: (1) знаки препинания; (2) продуктивные грамматические классы (существительные собственные или нарицательные; прилагательные; глаголы; наречия; деепричастия; причастия); (3) непродуктивные грамматические классы (союзы; междометия; числительные; частицы; предлоги; местоимения). После подсчета дистрибуции, из рассмотрения были исключены теги, частота которых была меньше 5% (10 вхождений). В силу низкой частоты они не могли служить критерием различения, в эту группу попали причастия, деепричастия, междометия, числительные и многие знаки препинания.

Для каждого тега был построен график, на котором указывались его частота в позициях окна анализа для каждого из глаголов группы, а также средняя частота для группы (см. распределение предлогов для глаголов перемещения на Рис. 1; прилагательных для глаголов говорения на Рис. 2).

По совпадению направления кривых на графиках вычислялось, есть ли закономерность в распределении данного тега для рассматриваемой группы (ср. на Рис. 1 закономерность есть, на Рис. 2 – нет). Для глаголов *говорения* следующие теги имели характерное распределение на протяжении окна анализа: существительные, глаголы, местоимения, запятые, кавычки, двоеточие, тире, а для глаголов движения – существительные, наречия, местоимения, союзы, предлоги, запятые, точки. Некоторые теги распределялись специфическим образом только в правом (например, союзы для глаголов говорения) или левом контексте (например, частицы и предлоги для глаголов говорения; частицы для глаголов движения).

Далее были построены графики разницы частоты для групп глаголов в позициях окна анализа по которым были определены наиболее характерные позиции для групп глаголов.

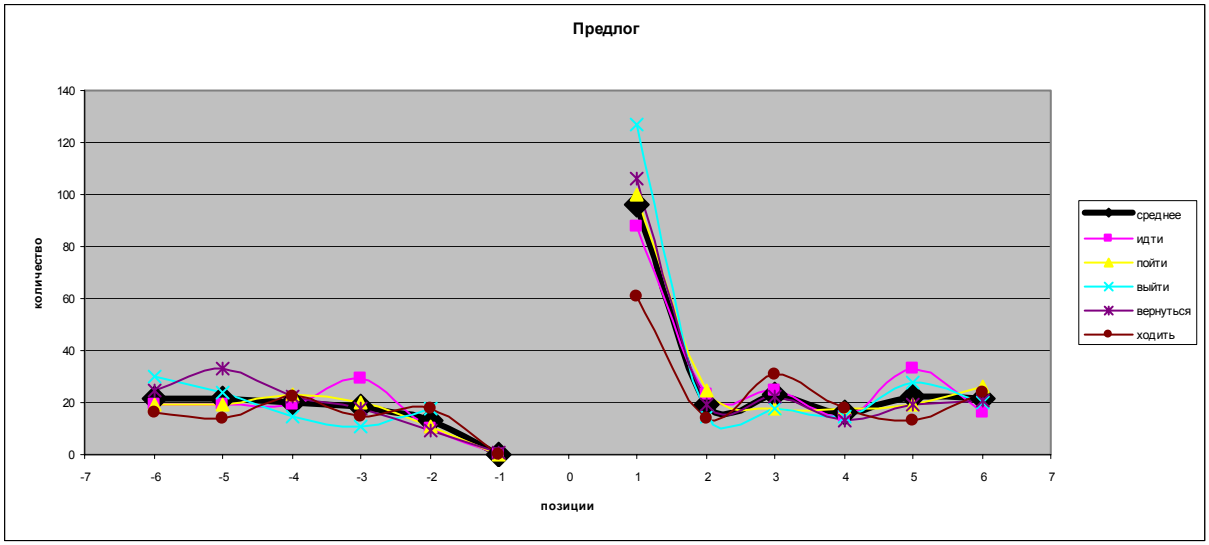

Рис. 1. Распределение частоты тега предлога по позициям локального контекста для глаголов движения

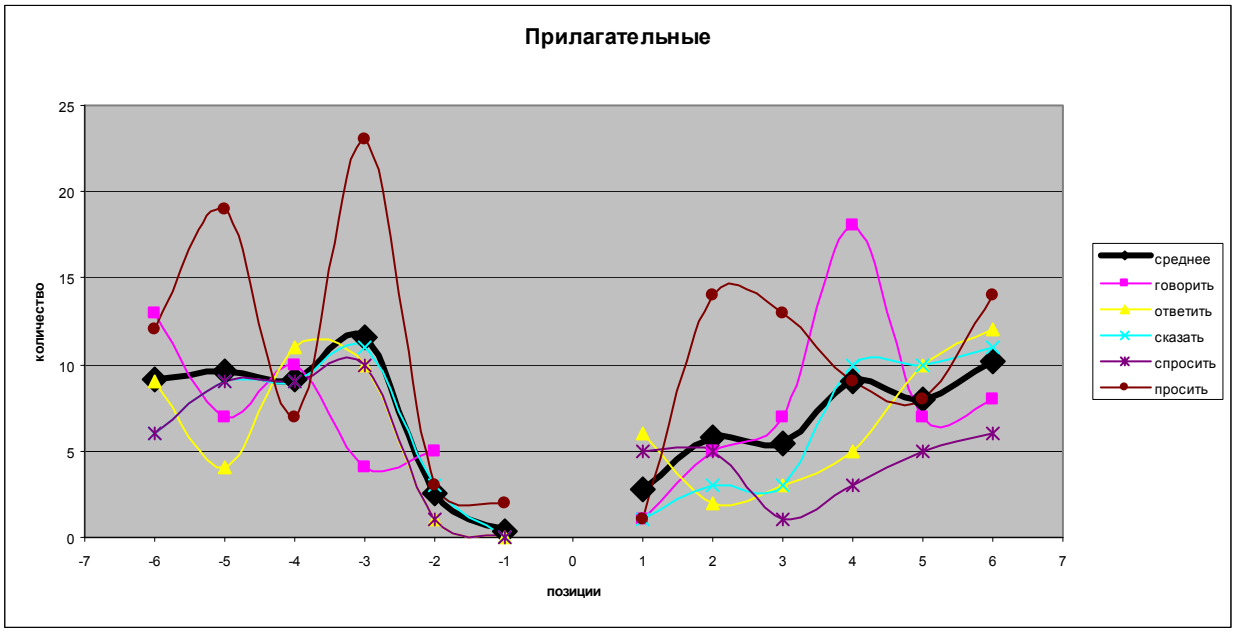

Рис. 2. Распределение частоты тега прилагательных по позициям локального контекста для глаголов говорения

Для глаголов говорения характерными позициями являются:
+1 (запятая – 38,8/19,8[3]; двоеточие – 43,2/43,2; местоимения – 37,4/31);
+2 (кавычки – 36,6/34,6; союзы – 31,6/20,4);
-1 (тире – 18/16,4).

Для глаголов движения:
+1 (наречия – 27/20; предлоги – 96,4/80,2);
+2 (прилагательные – 13,2/7,4; нарицательные существительные – 73/53,6);
-1 (глаголы – 15,8/15,6).

Полученные данные позволили сделать вывод о том, что подсчет дистрибуции тегов для русских контекстов можно использовать в процедуре автоматической классификации для некоторых глагольных групп. Несмотря на то, что для рассмотренных групп наиболее важными позициями оказались те, которые непосредственно следуют или предшествуют глаголу, в общем случае вопрос о размере окна анализа должен быть рассмотрен более подробно.

4. Автоматизированная классификация морфологически размеченных контекстов

В настоящем исследовании была выбрана 21 лексико-семантическая группа глаголов, для каждой группы отобраны частотные представители, общее количество глаголов равно 51. Анализируемый глагол представлен случайной выборкой из 200 морфологически размеченных контекстов. Поскольку в первую очередь необходимо уточнить параметры окна анализа, первоначальными границами являются [-10…+10] позиций. Морфологические теги указывают не только ЧР, но и некоторые значения грамматических категорий (для имен задается падеж, для глагола – личная или неличная форма, видовое значение). Знаки препинания отмечались единым тегом PM.

Совокупность контекстов глагола использовалась для подсчета дистрибуции тегов в окне анализа. Полученные дистрибуции сравнивались между собой, сходство в i-ой позиции окна анализа между словами *a* и *b* вычислялось как косинус угла между векторами частоты встречаемости ЧР тегов в i-ой позиции:

$$sim(a_i, b_i) = \frac{\sum_N a_{ij} \times b_{ij}}{\sqrt{\sum_N a_{ij}^2 \times \sum_N b_{ij}^2}}$$

Фрагмент общей схемы распределения параметров сходства по позициям окна приведен на Рис. 3. Исходя из данных кажется весьма правдоподобным, что для различения глаголов вполне достаточным окном анализа является [-1…+2], хотя отдельные флуктуации наблюдаются и в других позициях (-9, -7, -6, -2, +3, +10). Распределение параметров сходства указывает на четкое отличие дистрибуции глагола *успеть* от всех остальных, причем наиболее «далекими» от него являются в порядке убывания расхождений глаголы *стоять*, *идти*, (*брать-взять*). И наоборот, наиболее «близкими» (не противопоставленными) являются дистрибуции *брать-взять* и *идти-стоять*. Если для первой пары однотипность контекстов выглядит естественной, то во втором случае очевидно, что необходимо несколько изменить характеристики дистрибуции.

[3] Первое число – средняя частота тега в позиции, второе – разница между средними частотами для групп глаголов.

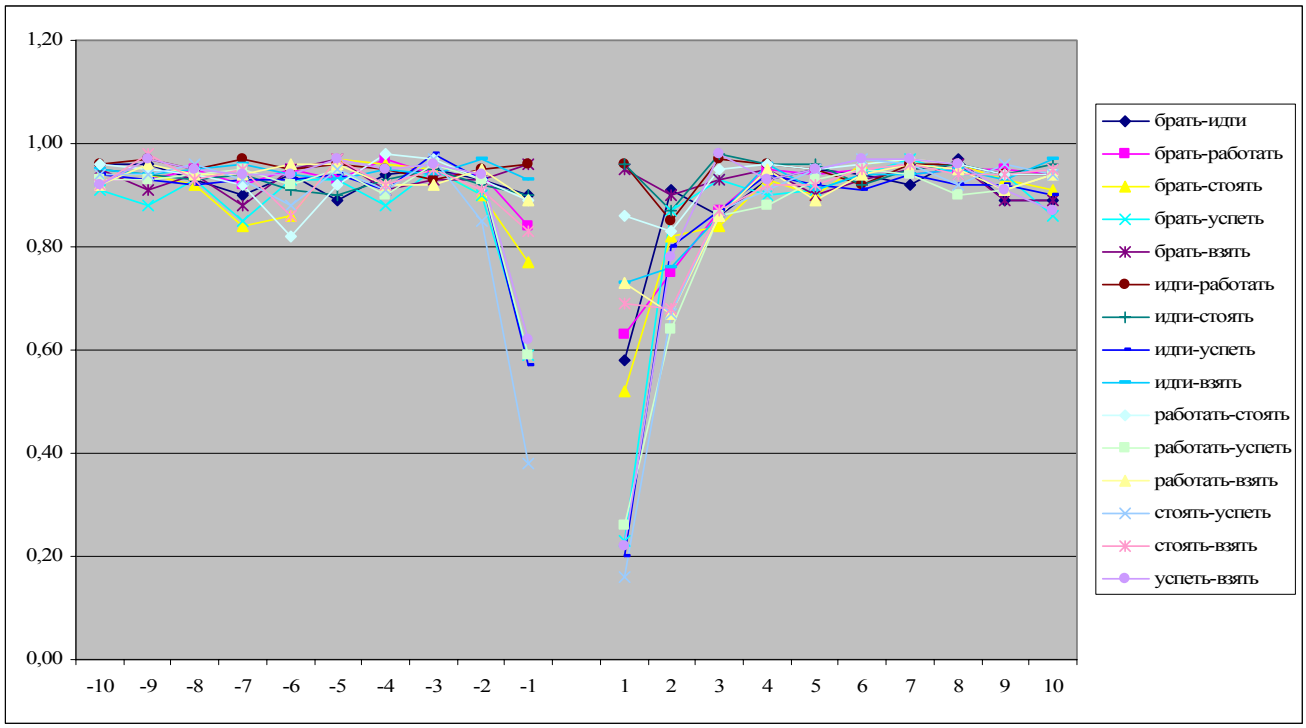

Рис. 3. Фрагмент схемы распределения параметров сходства между глагольными дистрибуциями в позициях окна анализа

Возможно, более четкие параметры дистрибуции контекста будут получены при объединении всех имен под одним тегом, если оставить в качестве дополнительного параметра характеристику падежного значения. Кажется естественным использование конкретных предлогов для описания дистрибуции.

5. Заключение

Проведенное исследование показывает, что с высокой долей вероятности можно использовать дистрибуцию тегов для разграничения типов контекстов лексико-семантических групп глаголов. Параметры окна анализа, а также характер тегов для описания контекстов следует дополнительно уточнить.

В случае четкого противопоставления хотя бы некоторых групп глаголов, можно будет задать параметры дистрибуции для центров кластеров. При автоматической классификации тестовые контексты будут относится к одной или нескольким группам, что позволит сократить первичную ручную обработку материала и получать более последовательные результаты, возможно, прототипические описания для рамок валентностей.

Список литературы

1. *Azarova I.V., Ivanov V.L., Ovchinnikova E.A., Sinopalnikova A.A. RussNet as a Semantic Component of the Text Analyser for Russian. // GWC 2006: Third International WordNet Conference. Jeju Island, Korea, January 22-26 2006: Proceedings. Brno: Masaryk University, 2005. P. 19–27.*
2. *WordNet: An Electronic Lexical Database / C. Fellbaum (ed.) MIT Press. 1997.*
3. *Leacock C., Chodorow M. Combining Local Context and WordNet Similarity for Word Sense Identification. In "WordNet: An Electronic Lexical Database". C. Fellbaum (ed.) MIT Press, 1998. P. 265–283.*
4. *Pantel P., Lin D. Word-for-Word Glossing with Contextually Similar Words // Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics, May 27 – June 1, Edmonton, 2003.*
5. *Voorhees E.M. Using WordNet for Text Retrieval // WordNet: an Electronic Lexical Database. Ch. Fellbaum (ed.), MIT Press, 1998. P. 285–303.*
6. *Азарова И.В., Иванов В.Л., Овчинникова Е.А. Семантическая структура пропозиции при извлечении фактов из текстов на русском языке // Компьютерная лингвистика и интеллектуальные технологии. Труды Международной конференции Диалог'2005 (Звенигород, 1-7 июня 2005 г.) М., 2005. С. 6–11.*
7. *Азарова И.В., Иванов В.Л., Овчинникова Е.А. Использование схемы наследования рамок валентностей в тезаурусе RussNet для автоматического анализа текста // Труды Международной конференции Диалог'2006.*
8. *Азарова И.В., Синопальникова А.А., Яворская М.В. Принципы построения wordnet-тезауруса RussNet // Компьютерная лингвистика и интеллектуальные технологии. Труды Международной конференции Диалог'2004 ("Верхневолжский", 2–7 июня 2004 г .) М., 2004. С. 542–547.*