ПОВЕРХНОСТНЫЕ ФИЛЬТРЫ ДЛЯ РАЗРЕШЕНИЯ СЕМАНТИЧЕСКОЙ ОМОНИМИИ В ТЕКСТОВОМ КОРПУСЕ

Кобрицов Б.П., Ляшевская О.Н., Шеманаева О.Ю.

ВИНИТИ РАН, Москва

neuralman (at) yandex.ru, olesar (at) mail.ru, shemanaeva (at) yandex.ru

SHALLOW RULES FOR WORD-SENSE DISAMBIGUATION IN TEXT CORPORA

Kobritsov B.P., Lashevskaja O.N., Shemanaeva O.Yu.

VINITI RAN, Moscow

В текстах новостей и газетно-журнальных статей особенно велика доля языковых штампов типа *вступить в силу*. Использование поверхностных фильтров, основанных на частотных устойчивых сочетаниях слов, обеспечивает большую точность разрешения лексико-семантической омонимии при семантической разметке текстовых корпусов.

Лексико-семантическая информация в Национальном корпусе русского языка

Идея настоящего проекта[*] родилась в ходе работ по "ручному" снятию морфологической омонимии в Корпусе современного русского языка (www.ruscorpora.ru). Когда разметчики, накопив опыт работы по снятию омонимии в художественной литературе, перешли к обработке газетных и других нехудожественных текстов, выяснилось, что в них из статьи в статью или, например, в мемуарах одного автора довольно часто повторяются одни и те же обороты: *в самом деле*; *несмотря на то, что*... и др. Д.Н. Шмелев еще в 1964 году отмечал растущую роль многословных оборотов в современном русском языке: "Некоторые из слов… все активнее употребляются для выражения разного рода отношений между предметами и явлениями, обозначенными другими словами. В связи с этим их функция в предложении становится близкой функции предлогов, союзов. Ср. такое употребление слов *сфера, лицо, дух, мера* и т. п.: *перед лицом военной угрозы, в сфере распределения материальных благ, со стороны общества, в расчете на массового потребителя, в духе взаимопонимания, по мере того как, в связи с этим, что* и т.п." [Шмелев 2002:87-88]. Очевидно, что учет подобных конструкций - устойчивых коллокаций (см., например, обсуждавшиеся на "Диалоге" работы [Большаков, Галисия-Аро 2003; Борисова 1995; Добровольский 2003; Копотев 2004 и др.]) может быть полезен не только при снятии морфологической омонимии, но также в "малом" синтаксисе (shallow parsing) и при разрешении лексической многозначности в семантической разметке. Кроме того, информация о вхождении слова в некоторый устойчивый оборот представляет самостоятельную ценность для пользователя корпуса.

Семантическая разметка Корпуса современного русского языка содержит информацию о принадлежности лексемы к одному или нескольким традиционным лексико-семантическим классам, таким как "глаголы движения", "каузативные глаголы", "части тела", "имена деятеля" и т. п. В настоящее время семантический разбор получают имена существительные, прилагательные, числительные, местоимения, глаголы и наречия (подробнее о разметке см. [Кустова et al. 2004; Кустова et al. (в печати)], а также документацию на сайте http://www.ruscorpora.ru/corpora-sem.html). Процедура аннотации корпуса основана на семантическом словаре, в котором каждое словарное значение слова имеет свой словарный вход[1] и представляется в виде набора параметров, ср.:

ПРИПЛЫТЬ – "глагол движения", "каузативный глагол", "приставочный глагол".

Фасетность, т. е. параметризация словарного значения по нескольким основаниям, не ведет к семантической омонимии: последняя возникает, если разные значения слова относятся к разным лексико-семантическим классам:

ЗОЛОТОЙ 1 (*золотое кольцо*) – "относительное прилагательное";

[*] Работа подготовлена при поддержке ООО "ЯНДЕКС" (www.yandex.ru) и фонда РФФИ (грант № 05-06-80396).

[1] С точки зрения представления семантической информации в корпусе традиционная омонимия (*жить в МИРЕ и согласии ~ МИР животных*) и многозначность (*ВСПЫШКА гнева ~ фотографическая ВСПЫШКА*) считаются явлениями одного порядка; соответственно, термин "разрешение семантической омонимии" (word-sense disambiguation) охватывает оба явления.

ЗОЛОТОЙ 2 (*золотые кудри*) – "прилагательное цвета", "относительное прилагательное";

ЗОЛОТОЙ 3 (*золотой ребенок*) – "прилагательное оценки", "качественное прилагательное".

Программа первичного семантического парсинга работает без учета контекста и приписывает лемме семантические признаки, относящиеся ко всем ее значениям. Естественно, это создает много шума при поиске. Во-первых, некоторые слова вообще не получает семантического разбора, если они отсутствуют в семантическом словаре. Во-вторых, часть слов получает несколько альтернативных разборов: это слова с семантической омонимией. В третьих, слова могут иметь ошибочный разбор (при разборе "вручную" эти слова получили бы разбор, отличный от словарного). Не будем также забывать, что в случае, если семантическая разметка накладывается на корпус с неснятой морфологической омонимией, процент ошибок многократно умножается.

В нашей предыдущей публикации [Кобрицов, Ляшевская 2004] обсуждалась проблема снятия семантической омонимии с помощью глубинных фильтров - на основе глобальных правил сочетаемости семантических классов, например, "названия одежды не могут употребляться в роли субъекта при глаголах эмоции", ср. *амазонка* <'человек', ~~'одежда'~~> *рассмеялась*. Как показывает практика, точность глобальных правил далека от 100 процентов [Кобрицов 2004].

Поверхностные фильтры, которые на входе содержат комбинацию из 2-х, 3-х и т. д. лексем (или даже словоформ), ср.:

w1 *до*
w2 *сих*
  сей (A-PRO) <"указательное мест.">
w3 *пор*
  пóра (S) <"предметное имя", "простр.:отверстие">
  пóра (S) <"предметное имя", "простр.:пустота">
  порá (S) <"непредметное имя", "период времени">
  порá (PRAEDIC)
→ порá (S) <"непредметное имя", "период времени">,

напротив, почти всегда позволяют предсказать единственно правильный разбор. Проблема лишь в том, что для того чтобы существенно уменьшить семантическую омонимию в корпусе, требуются тысячи таких фильтров. Таким образом, перед нами стояла задача максимально автоматизировать работу по подготовке исходного материала для экспертов и выделить самые частотные коллокации для создания наиболее эффективных поверхностных фильтров.

Частотные устойчивые коллокации

Списки устойчивых коллокаций были получены на базе корпуса публицистики[2], включающего материалы московских и региональных газет за 1998-2004 гг., новостей, радиоинтервью, а также мемуарную литературу. Объем обработанного на настоящий момент корпуса составляет около 15 млн слов. Были получены реестры двусловных и трехсловных "жестких" коллокаций, в которых составляющие непосредственно примыкают друг к другу (т. е. расстояние между словами не превышает единицы). Наш алгоритм не учитывал пары и тройки словоформ, разделенные границами предложения, а также любыми знаками препинания. Таким образом, сюда попали сочетания типа *Московский комсомолец* и не попали сочетания типа *газета "Московский комсомолец"*. Словосочетания с переменой мест составляющих считались разными коллокациями (ср. *российское государство* и *государство российское*).

Реестры коллокаций были обработаны с помощью лингвистических и статистических методов. На основании информации о частях речи из списков были исключены коллокации, не образующие синтаксического единства, типа "и в" (CONJ + PR), "в самом" (PR + APRO) и др. (метод Джастесона и Каца [Justeson, Katz 1995]). Был также использован стоп-лист малоинформативных слов, например *или*/*этот*/*мой* + S.

Как известно, ни один из существующих методов статистического ранжирования коллокаций (см. их обзор в [Manning, Schütze 1999; Pearce 2002; Jiangsheng Yu et al. 2003]) не позволяет с уверенностью различить "хорошие" и "плохие", т. е. случайные, коллокации. В список представляемых на суд эксперта оборотов были включены все коллокации, абсолютная частотность которых превышала 100 употреблений. Оставшиеся коллокации были ранжированы с помощью формулы

$$MI = \frac{\text{frequency}(w1,w2)^2}{\text{frequency}(w1) \bullet \text{frequency}(w2)},$$

использовавшейся ранее при обработке Кембриджского корпуса английского языка. Верхняя часть полученного списка была также включена в short-list.

Наконец, была учтена информация о дискурсивных сдвигах значения односложных элементов (ср. существительное в функции предлога *типа*) и о более чем 3-словных оборотах, полученная из работ [Русская грамматика 1980; Зализняк 1977/2003; Рогожникова 2003; Шведова 1960/2003].

Характеристика поверхностных фильтров

Поверхностные фильтры, работающие на базе устойчивых коллокаций, включают данные (1) о лемме, (2) о частеречных, (3) словоклассифицирующих и (4) словоизменительных признаках составляющих, (5) об их исходной семантической разметке, а также (6) о некоторых грамматических и лексико-семантических характеристиках ближайшего

[2] Входит в состав Корпуса современного русского языка.

контекста (например, "родительный падеж" для оборота *типа кого-чего-л.*).

Для увеличения скорости обработки больших массивов оборотов эксперт может сортировать список по любому из параметров, например, обрабатывать весь все обороты с вариантом семантического разбора "период времени". Кроме того, эксперт может проследить дерево вложенных коллокаций, ср.:

НЕ ГОВОРЯ О <предл. пад.>

НЕ ГОВОРЯ О том, что...

НЕ ГОВОРЯ О том, как...

я НЕ ГОВОРЮ О <предл. пад.>

мы НЕ ГОВОРИМ уже О том что...

Говоря о частотных коллокациях, обычно имеют в виду, что частотность употребления двух слов рядом друг с другом ведет к тому, что в их значении появляются нетривиальные компоненты, нарушающие строгий принцип композициональности значения [Manning, Schütze 1999][3]. В связи с этим на выходе поверхностных фильтров может находиться не только один семантический разбор из нескольких данных, но и новый, несловарный разбор. (Заметим, что на этапе подготовки short-листа коллокаций в список были включены обороты, у которых межлексемная и семантическая омонимия изначально отсутствовала).

С помощью поверхностных фильтров в семантическую разметку также добавляется особая информация о функционировании слова в составе оборота. В частности, таким образом размечаются:

– сложные служебные лексические единицы: составные предлоги, союзы, наречия, частицы, вводные слова (*в связи с, в случае если, на самом деле* и др.);

– топонимы (*Нижний Новгород*);

– неоднословные обозначения лиц (*Владимир Путин, В. Путин, президент Путин*) и др.

Поверхностные фильтры позволяют зафиксировать дискурсивные сдвиги частеречной принадлежности: существительное в функции предлога (*типа, вида*); междометие или наречие в функции существительного (*пройти на ура; Никакого суетливого послезавтра!* (Э. Радзинский) и т. п.).

Отмечаются следующие случаи выветривания значения:

– употребление глагола в качестве лексической функции (ср. *вступить в силу*);

– употребление прилагательного в качестве лексической функции Magn (*круглый дурак*);

– включение предлога в модель управления глаголов и имен (*делить на…, проблема с…, министр по…, один из…*);

– идиоматическое употребление одной или нескольких составляющих (ср. *круглый стол*: весь оборот относится к классу "мероприятие").

Оценка эффективности работы фильтров

Подведем итоги, достигнутые после создания фильтров на основе первой тысячи самых частотных устойчивых коллокаций.

Оценка эффективности проводилась в три этапа. На первом этапе было подсчитано количество словоупотреблений в исходном корпусе и число нераспознанных словоупотреблений. Уровень распознавания текста (лемматизации) составил приблизительно 98%. Некоторые такие слова (*госдума*, *СМИ* и др.) были обнаружены в списке высокочастотных коллокаций. После определения их исходной формы и частеречной принадлежности и применения фильтров к текстам точность разметки была повышена на 0,26%.

На втором этапе было подсчитано число словоформ с межлексемной омонимией (в среднем 1,53 разборов на каждое словоупотребление). Разрешение межлексемной омонимии с помощью поверхностных фильтров позволяет достичь точности разметки 1,13 разборов на одно словоупотребление.

На последнем этапе оценивалось качество собственно семантической разметки. Число слов, размеченных по лексико-семантическим параметрам, составило в настоящей версии корпуса 67% всех словоупотреблений. Однако точность семантической разметки представляется разумным рассчитывать без учета слов, относящихся к предлогам, союзам, частицам и другим частям речи, на которые семантическая разметка не распространяется. По данным подкорпуса со снятой омонимией, к знаменательным частям речи должно принадлежать порядка 75,9% всех словоупотреблений. Таким образом, относительно них доля семантически размеченных слов составила 88,3% (13,5 млн из 15,3 млн) .

Мы лишены возможности сравнить наш корпус с некоторым "золотым стандартом", ибо русских корпусов, размеченных "вручную" описанным выше методом, не существует. Это, безусловно, затрудняет оценку правильности семантической разметки.

Предлагается два показателя оценки относительного улучшения качества разметки. Первый – показатель полного снятия омонимии,

$$WSD = \frac{x_f - x_0}{N} \cdot 100\%,$$

где x_f – число слов с единственным правильным разбором в корпусе после применения фильтров, x_0 – аналогичное число в исходном корпусе, N = 13 538 782 – база для оценки точности

[3] Классические идиомы, впрочем, оказались большей частью вне зоны нашего внимания – их частотность в текстах слишком мала. В то же время были обработаны высокочастотные коллокации, в которых не отмечается диффузии значения, ср. *министр финансов, президент Путин*.

разметки, общее число слов знаменательных частей речи.

Коэффициент WSD учитывает те случаи, когда фильтр однозначно снимает семантическую омонимию:

$w\{s_1, \dots s_n\} \rightarrow w\{s_i\}$.

Для тех случаев, когда фильтр снимает только часть омонимии, например,

$w\{s_1, \dots s_n\} \rightarrow w\{s_1,s_2\} -$

используется коэффициент

$$WSR = \left(\frac{s(N\text{-}x_0)}{N\text{-}x_0} - \frac{s(N\text{-}x_f)}{N\text{-}x_f}\right) \bullet 100\%,$$

где $N\text{-}x_0$, $N\text{-}x_f$ – число слов с неснятой омонимией, а $s(N\text{-}x_0)$, $s(N\text{-}x_f)$ – общее количество разборов у слов с неснятой омонимией.

Фильтры, построенные на первой тысяче самых частотных коллокаций, позволяют разрешить семантическую неоднозначность для 800 тыс. словоупотреблений полностью (WSD ≈ 6%) и для 100 тыс. словоупотреблений частично (WSR ≈ 1,5%).

Список литературы:

1) Большаков И.А., Галисия-Аро С.Н. Сколько страниц на данном языке содержит Интернет? // Труды международной конференции Диалог'2003. М., 2003.
2) Борисова Е.Г. Коллокации. Что это такое и как их изучать? М., 1995.
3) Зализняк А.А. Грамматический словарь русского языка. М., 1977. 4-е изд.: М., 2003.
4) Добровольский Д.О. Корпус параллельных текстов как инструмент анализа литературного перевода. Труды международной конференции Диалог'2003. М., 2003.
5) Кобрицов Б.П. Модели многозначности русской предметной лексики: глобальные и локальные правила разрешения омонимии. Автореф… канд. филол. наук. М.: РГГУ, 2004.
6) Кобрицов Б.П., Ляшевская О.Н. Автоматическое разрешение семантической неоднозначности в Национальном корпусе русского языка // Компьютерная лингвистика и интеллектуальные технологии: труды международной конференции Диалог'2004. Под ред. И.М.Кобозевой, А.С.Нариньяни, В.П.Селегея. М., 2004.
7) Копотев М. «Несмотря на» «потому что», или Многокомпонентные единицы в аннотированном корпусе русских текстов. Диалог'2004. М., 2004.
8) Кустова Г.И., Ляшевская О.Н., Падучева Е.В., Рахилина Е.В. Национальный корпус русского языка как инструмент семантико-грамматического исследования лексики // Международная конференция "Корпусная лингвистика - 2004". Тезисы докладов. СПб.: СПбГУ, 2004. С. 50-51.
9) Кустова Г.И., Ляшевская О.Н., Падучева Е.В., Рахилина Е.В. Опыт семантического расширения морфологической разметки: таксономическая классификация лексики в Национальном корпусе русского языка // Научная и техническая информация, сер. 2. Информационные процессы и системы (в печати).
10) Русская грамматика. М., 1980.
11) Рогожникова Р. П. Словарь эквивалентов слова. М., 2003.
12) Шведова Н.Ю. Очерки по синтаксису русской разговорной речи. М., 1960. 2-е изд.: М., 2003.
13) Шмелев Д.Н. О семантических изменениях в современном русском языке // Шмелев Д.Н. Избранные труды по русскому языку. М., 2002.
14) Jiangsheng Yu, Zhihui Jin, Zhenshan Wen. Automatic Detection of Collocation // The 4th Chinese lexical semantics workshop, Hong-Cong, 2003. http://icl.pku.edu.cn/yujs/papers/pdf/col.pdf.
15) Justeson J.S., Katz S.M. Technical terminology: some linguistic properties and an algorithm for identification in text // Natural Language Engineering, 1995, 1(1). P. 9-27.
16) Manning C.D., Schütze H. Foundations of Statistical Natural Language Processing // Cambridge, Massachusetts: The MIT Press, 1999. Ch. 5. Collocations. http://nlp.stanford.edu/fsnlp/promo/colloc.pdf.
17) Pearce D. A comparative evaluation of collocation Extraction Techniques // Third International Conference on Language Resources and Evaluation. May, 2002. Las Palmas, Canary Islands, Spain. 2002. http://www.informatics.susx.ac.uk/users/darrenp/academic/dphil/publications/data/Conferences/lrec2002/paper.pdf.