Что может помочь компьютеру понять, кто стоял на балконе

What can help the computer to learn who was on the balcony

Юдина М. В. (maria_yu@abbyy.com)
ABBYY

В статье перечислены некоторые факторы, влияющие на разрешение синтаксической неоднозначности, с точки зрения возможности их использования в автоматическом анализе текста, а также показаны результаты опыта семантического подхода к разрешению синтаксической неоднозначности.

Одну из наибольших проблем для автоматической обработки текста составляет синтаксическая омонимия, или синтаксическая неоднозначность, т. е. возможность построить несколько синтаксических структур на основе одной и той же последовательности языковых знаков. В подавляющем большинстве случаев омонимию может разрешить только человек. В ряде случаев омонимия принципиально неразрешима без дополнительной информации (например, в предложении (1): *Маша читала и писала письма*, см. [Юдина, Янович, Фёдорова 2007]). Тем не менее, в реальной жизни мы очень редко замечаем синтаксическую омонимию, в силу способности нашего синтаксического парсера мгновенно анализировать не только синтаксическую структуру, но и ситуацию, контекст, делать логические выводы о смысле предложения. Научить этому машинный анализатор представляется практически невозможным.

Изучение синтаксической омонимии сводится, как правило, к исчислению всех потенциально возможных конструкций (см. например, [Иорданская 1967]). В системах автоматического анализа текста такая омонимия (если набор эвристик конкретного парсера в принципе может распознать некоторые типы омонимии), разрешается в пользу одного из вариантов случайным образом или с помощью статистики.

Рассмотрим один из наиболее изученных и популярных типов синтаксической неоднозначности — «раннее-позднее закрытие», или омонимию относительного придаточного предложения (см., к примеру, обзор в [Фёдорова, Янович 2004]). Напомним, что данная омонимия заключается в наличии двух интерпретаций для предложения (2) *Преступник застрелил служанку актрисы, которая стояла на балконе*: первое прочтение, называемое «ранним закрытием» (далее РЗ), соответствует пониманию «служанка стояла на балконе», а второе прочтение — «позднее закрытие» (далее ПЗ) — пониманию «актриса стояла на балконе».

Феномен раннего-позднего закрытия широко исследован на разных языках, однако на данный момент все полученные данные носят лишь теоретический характер, не находя применения в системах автоматической обработки текста и машинного перевода. Остановимся на некоторых факторах, влияние которых на выбор той или иной интерпретации в рамках разрешения омонимии раннего-позднего закрытия представляется доказанной. Данные факторы, как кажется, в перспективе могли бы быть использованы и в целях автоматического разрешения неоднозначности.

1. Длина придаточного предложения

Влияние длины придаточного предложения на выбор РЗ или ПЗ было проведено в работе [Fodor 1998]. Как оказалось, английский язык, в котором предыдущие эксперименты выявили предпочтение ПЗ, следует данному принципу далеко не всегда. Например, если придаточное предложения было длинным, ПЗ встречалось редко или не наблюдалось вовсе; иная картина наблюдалась в случае придаточных, состоящих из одного просодического слова: ПЗ было преобладающим. Данный факт был

проверен также на материале других языков: арабского, хорватского, французского, немецкого и испанского. Результаты оказались схожими: короткие придаточные гораздо чаще модифицировали второе существительное именной группы (далее ИГ). Для объяснения данной закономерности Дж. Фодор выдвинула Закон антигравитации. Данный закон гласит, что если присоединяемая составляющая имеет просодически «легкий» статус, то она, скорее всего, присоединяется к зависимому существительному ИГ; в то время, как присоединение более «тяжелых» составляющих зависит от двух факторов: от соотношения «просодической тяжести» придаточного предложения и ИГ, которую оно призвано модифицировать, и от просодических особенностей данного языка. Результаты были проверены и на русском материале в работе [Фёдорова, Янович 2004], подтвердившей результаты Фодор.

Фактор длины придаточного наиболее легко реализуем в системах автоматического синтаксического анализа.

2. Лингвистическая настройка

Необходимо также упомянуть о Гипотезе лингвистической настройки (*Linguistic Tuning*), рассматриваемой в работе [Mitchell et al. 1995] (см. также [Драгой 2006]). Данная гипотеза предполагает, что одним из факторов, влияющих на разрешение синтаксической неоднозначности, является предыдущий лингвистический опыт человека в разрешении подобной неоднозначности. В частности, это означает, что выбор интерпретации предложения основан на той стратегии, которая является наиболее частотной в конкретном языке. Эта гипотеза была подтверждена на французском и английском материале.

Гипотеза лингвистической настройки может служить опорой для статистического подхода к разрешению синтаксической неоднозначности.

3. Одушевленность существительных, входящих в ИГ

Исследование [Brysbaert, Mitchell 1996], проведенное для проверки гипотезы лингвистической настройки на материале нидерландского языка, показало прямо противоположные результаты. В работе [Desmet et al. 2002] приводится объяснение этому факту. Было доказано, что при наличии в ИГ одушевленного и неодушевленного существительных одушевленное существительное выбирается чаще, чем неодушевленное в любой позиции, при наличии в ИГ двух одушевленных существительных преобладает РЗ, а при наличии двух неодушевленных — ПЗ; таким образом, исследование [Brysbaert, Mitchell 1996], в экспериментальном материале которого предложения с одушевленными именами составляли меньшинство, не может служить опровержением Гипотезы лингвистической настройки.

Фактор одушевленности/неодушевленности также достаточно просто учесть при автоматическом разрешении неоднозначности.

4. Контекст

Первым экспериментом, исследовавшим влияние контекста на присоединение придаточных предложений и проведенным по методике заканчивания предложений, стала работа Т. Десмета и коллег ([Desmet, De Baecke et al. 2002]). Ими был проведен эксперимент на базе нидерландского языка. Экспериментальный материал состоял из 30 предложений, к каждому было придумано по три контекста. Контекст, склоняющий к первому (главному) существительному ИГ, вводил двух возможных референтов для главного существительного и одного возможного референта для зависимого существительного ИГ; контекст, склоняющий ко второму, зависимому, имени, наоборот, вводил двух возможных референтов для зависимого имени и одного возможного референта для главного имени; и, наконец, в нейтральном контексте референты либо не вводились вовсе, либо вводился лишь один возможный референт для обоих существительных. Гипотеза авторов состояла в том, что, поскольку нидерландский язык относится к языкам с тенденцией к РЗ, нейтральный контекст будет способствовать присоединению относительного придаточного к вершине ИГ. При контексте, склоняющем к вершине именной группы, РЗ будет преобладать, а при контексте, склоняющем к зависимому существительному ИГ, будут возможны случаи ПЗ. Результаты эксперимента полностью подтвердили гипотезу.

Серия подобных экспериментов на русском материале была проведена в работе [Юдина 2006]. Всего в исследовании приняли участие 165 человек, было обработано 2970 экспериментальных предложений. Общие результаты экспериментов таковы: 62 % РЗ после нейтрального контекста, 91 % РЗ после контекста, склоняющего к первому имени ИГ, 60 % РЗ после контекста, склоняющего ко второму имени ИГ.

Фактор контекста на данном этапе его изученности наиболее сложен для формализации с целью использования его для автоматического разрешения неоднозначности, так как помимо референциального аспекта введения участников неминуемо содержит также и семантическую информацию о ситуации в целом, что, несомненно, также влияет на тип закрытия.

5. Прайминг

Явление синтаксического прайминга заключается в следующем: при ответной реакции на какой-либо стимул говорящий склонен использовать те синтаксические конструкции, которые он в недавнем прошлом каким-либо образом обработал (услышал, прочитал, сказал). Одним из типичных проявлений синтаксического прайминга является синтаксическая координация участников диалога. Высказывание, осуществляющее преднастройку, называется «праймом», а высказывание, на порождение или понимание которого, как предполагается, окажет влияние прайм, называют «целью». Первый подобный эксперимент был проведен в [Scheepers 2003] на немецком материале. Экспериментальный блок состоял из четырех предложений: прайм с ранним закрытием (далее РЗ-прайм), прайм с поздним закрытием (далее ПЗ-прайм), базовый прайм (блокировал присоединение относительного придаточного) и целевое предложение (далее цель). Суть эксперимента заключалась в следующем: РЗ- и ПЗ-праймы могли быть продолжены испытуемыми только одним способом, базовый прайм не предполагал конструкции с относительным придаточным, а цель была составлена так, что она могла быть закончена испытуемыми двояко. Предполагалось, что вынужденное использование испытуемым РЗ или ПЗ в прайме вызовет использование соответствующей структуры в синтаксически неоднозначном предложении — цели. Базовые праймы включались в экспериментальные наборы для того, чтобы проверить, каково будет предпочтение закрытия в цели в отсутствие прайминга. Результаты эксперимента показали значительный прайминг-эффект.

Мы провели серию аналогичных экспериментов на русском материале ([Юдина, Фёдорова 2009]). Всего было проведено 3 эксперимента, в которых приняли участие 131 человек, было обработано 2430 экспериментальных предложений. Основной эксперимент серии показал значительный эффект синтаксического прайминга: 57 % РЗ после РЗ-прайма и 46 % РЗ после ПЗ-прайма. Количество случаев РЗ после базового прайма составило 60 %, что подтвердило результат предыдущих экспериментов.

Влияние эффекта прайминга на разрешение неоднозначности имеет больше теоретическую значимость, нежели практическую; однако, наличие не только глобальной, но и локальной настройки на разрешение неоднозначности определенным способом также может помочь более тонкому и правильному статистическому подходу к разрешению неоднозначности.

Таким образом, после проведения экспериментов, исследовавших влияние прайминга и контекста, в нашем распоряжении оказалась огромная база экспериментальных предложений с разрешенной неоднозначностью раннего-позднего закрытия. При обобщении результатов всех проведенных экспериментов можно выявить среднее для русского языка распределение типов закрытия в отсутствие какого-либо склонения к одному из типов закрытия, будь то контекст или прайминг. Мы считаем, что это распределение равно 60/40 %, что вполне согласуется с другими экспериментами с ранним-поздним закрытием на русском материале (см., например, [Sekerina 2003]). Принятие этих цифр за некий базовый коэффициент дает возможность более подробно исследовать предложения, явно выбивающиеся из этих показателей, или же наоборот, выявить наиболее «стандартные» с точки зрения закрытия предложения.

6. Тип глагольной вершины (предикат главного предложения)

Все проведенные нами эксперименты были основаны на приблизительно одном и том же экспериментальном материале. Некоторые предложения, казавшиеся нам неудачными, мы заменяли другими, менялся экспериментальный дизайн, но большая часть предложений оставалась неизменной. В результате мы оказались обладателями объемной базы экспериментальных предложений, при этом все они имели однотипную структуру: *Subject Verb* NN_{gen} *[Relative Clause]*.

Оказалось, что некоторые предложения показывают схожие результаты: например, большее по сравнению с нормальным количество случаев РЗ или ПЗ. В части случаев виной тому, несомненно, был общий контекст предложения, обрисовывавший ситуацию, заведомо так или иначе связанную с одним из существительных ИГ в качестве участника. Так, большинство контекстов, оказавшихся склоняющими к ПЗ, содержат концепт того же семантического поля, что и второе существительное ИГ (например, (3) *Редакция поручила Косте написать статью о шоу-бизнесе, но он совершенно не знал, с чего начать. Коля попросил помощи у менеджера певицы,…*: «певица» относится к семантическому полю «шоу-бизнес»), или концепт, тесно связанный со вторым существительным тематически (например, (4) *Вся страна со страхом наблюдала за развитием событий в захваченном здании. Власти обещали освободить заложницу террориста,…*).

Однако если не брать в расчет контекст, а оперировать лишь непосредственно предложениями, оказывается, что одинаковым поведением обладают предложения с одинаковым типом глаголов-вершин. Например, для закрытия экспериментального предложения (5) *В парке друзья встретили ассистентку профессора,…* испытуемые предпочитали присоединение придаточного предложения к первому имени, что было подтверждено результатами двух экс-

периментов. Мы предположили, что ИГ, состоящая из одушевленных существительных разного рода, неравноправна с точки зрения закрытия, а именно, что существительные женского рода обладают свойством «перетягивать» закрытие. Для проверки этого факта мы провели дополнительный эксперимент, в котором поменяли род у существительных ИГ (например, «мама однокурсника» вместо «папа однокурсницы»), однако, этот эксперимент показал слабое влияние женского рода (в основном свойством притягивания закрытия обладали маркированные слова вроде «директриса»). Большинство предложений, несмотря на изменение рода существительных, показали результаты, полностью идентичные основному эксперименту.

Поэтому мы предположили, что существует и еще один фактор, влияющий на разрешение синтаксической неоднозначности, а именно, тип глагольной вершины. Из всех отобранных нами экспериментальных предложений мы составили список глагольных вершин, встречающихся более чем в одном экспериментальном предложении каждого эксперимента в отдельности или по всем экспериментам в целом. Всего были обработаны материалы четырех проведенных нами в разные годы экспериментов. Поскольку наши эксперименты имели разный дизайн и разные задачи (в некоторых экспериментах предложение функционировали со своими контекстами, в некоторых — без), можно считать, что одинаковые результаты, показанные предложением, не связаны с контекстом или другим влиянием.

Таким образом были выделены следующие глаголы:

Заметить
Увидеть
Смотреть
Слушать
Встретить
Навестить
Столкнуться
Разругаться
Поссориться
Договориться
Познакомиться
Узнать

Мы составили таблицу, иллюстрирующую количество РЗ и ПЗ в предложениях, содержащих данный глагол. Критерием оценки стало количество РЗ при нейтральном контексте (или после базового прайма) и при позднем контексте (или после прайма, склоняющего к ПЗ). Мы считали, что предикат способствует РЗ, если число РЗ в вышеуказанных условиях составляло 100–80 %, и к ПЗ — если меньше 50 %.

Далее, каждый предикат был отнесен к определенному типу согласно тезаурусу Роже ([Roget's Thesaurus 2000]). Мы выделили несколько групп однотипных предикатов; некоторые предикаты, такие, как «нагрубить» или «разговориться», пришлось не учитывать в исследовании, так как в рамках выборки они оказались единственными предикатами своего типа.

Результаты можно увидеть в таблице 1.

Наибольшую сложность вызвал предикат «увидеть». Дело в том, что во всех экспериментальных предложениях, кроме одного, этот глагол показывал большой процент ПЗ. Оказалось, практически во всех экспериментальных предложениях глагол «увидеть» означал скорее «увидеть и узнать», например (6) *Выходя на улицу, Катя вдруг увидела ученицу доктора, …* (увидела и узнала, что это именно ученица доктора). Лишь в предложении (7) *В конце концов Алексей увидел посетительницу директора, …* «увидеть» функционирует в своем прямом значении глагола зрения. Обозначим его «увидеть1», а другое

Таблица 1

предикат	тип закрытия	класс предиката по Роже
заметить	Р	matter → organic matter → vision
смотреть	Р	matter → organic matter → vision
слушать	Р	matter → organic matter → hearing
увидеть1	Р	matter → organic matter → vision
встретить	Р	space → motion → arrival
навестить	Р	space → motion → arrival
столкнуться	Р	space → motion → arrival
разругаться	П	volition → individual volition → antagonism → dissention
поссориться	Р\П	volition → individual volition → antagonism → dissention
договориться	Р	volition → individual volition → antagonism → concord
познакомиться	П\Р	intellect → formation of ideas → results of reasoning → knowledge
узнать	П	intellect → formation of ideas → results of reasoning → knowledge
увидеть2	*П*	*intellect → formation of ideas → results of reasoning → knowledge*

значение будем маркировать как «увидеть2» (выделено в таблице курсивом). Однако в словарях не фиксируется подобное различие смыслов; результаты, показанные предложением «Алексей увидел…» могут быть связаны с особенностями ИГ «посетительница директора»: например, причина может быть в разных референтных статусах ИГ «посетительница директора» и «ученица доктора». Данный вопрос подлежит дальнейшему изучению.

Как видно из таблицы, глаголы одного класса показывают удивительное единство в том, как разрешается синтаксическая неоднозначность в предложении, где вершиной является данный предикат. А именно: предложения с глаголами чувственного восприятия и глаголы движения имеют большой процент РЗ, глаголы, связанные с мышлением и интеллектом — ПЗ. Класс глаголов отношений неоднороден: глагол «договориться» показывает практически 100 % РЗ. Как нам кажется, это может быть связано со структурой ИГ: экспериментальное предложение (8) *Я решил договориться с женой строителя, …* почти все испытуемые заканчивали описанием того, что может сделать с (нерадивым) строителем жена (запретит пить, будет бить каждый вечер и т. п.)

Также не совсем ясна ситуация с глаголом «познакомиться»: обычно этот глагол ведет себя по варианту ПЗ, но одно предложение в двух экспериментах показало почти 100 % РЗ: (9) *Вчера Татьяна наконец познакомилась с секретаршей начальника, ….* Видимо, это также связано с особенностью самого этого предложения, точнее, с неудачной структурой ИГ с точки зрения семантики. Вероятно, «секретарша начальника» — это настолько устойчивый концепт, что опознаётся скорее как одно целое, чем как два независимых объекта.

Не вошедшие в нашу выборку по причине своей «единичности» в рамках экспериментальной выборки или в рамках классификации по Роже глаголы (например, *нагрубить, попрощаться*) также, тем не менее, показывают довольно однородные результаты. Нам кажется, что имеет смысл продолжить исследования в этом направлении.

Конечно, данные результаты нельзя считать окончательными и доказанными. Но они, на наш взгляд, иллюстрируют возможность семантического подхода к синтаксической неоднозначности. Класс предиката хорошо формализуем, поэтому не составит труда встроить семантическую классификацию предикатов в синтаксический анализатор.

Литература

1. *Юдина М. В., Федорова О. В., Янович И. С.* Синтаксическая неоднозначность в эксперименте и в жизни // Диалог. М.: 2007
2. *Иорданская Л. Н.* Синтаксическая омонимия в русском языке (с точки зрения автоматического анализа и синтеза) // НТИ. 1967. № 5.
3. *Фёдорова О. В., Янович И. С.* Об одном типе синтаксической многозначности, или Кто стоял на балконе. // МГУ, 2004
4. *Fodor J. D.* Learning to parse? // Journal of Psycholinguistic Research, 27, 2, 1998.
5. *Mitchell D. C., Cuetos F., Corley M. M. B. & Brysbaert M.* Exposure-based models of human parsing: Evidence for the use of coarse-grained (nonlexical) statistical records. // Journal of Psycholinguistic Research, 24, 6, 1995.
6. *Драгой О. В.* Разрешение синтаксической неоднозначности: правила и вероятности. // Вопросы языкознания, № 6, 2006
7. *Brysbaert M., Mitchell D. C.* Modifier attachment in sentence parsing: Evidence from Dutch. // Quarterly Journal of Experimental Psychology, 49A, 3, 1996.
8. *Desmet T., Brysbaert M. & De Baecke C.* The correspondence between sentence production and corpus frequencies in modifier attachment. // Quarterly Journal of Experimental Psychology, 55A (3), 2002.
9. *Desmet T., De Baecke C., Brysbaert M.* The influence of referential discourse context on modifier attachment in Dutch. // Memory & Cognition 2002, 30 (1), 2002
10. *Юдина М. В.* Понимание и порождение высказываний с синтаксической неоднозначностью (на примере относительных придаточных в русском языке) // Диалог. М.: 2006
11. *Scheepers C.* Syntactic priming of relative clause attachments: persistence of structural configuration in sentence production. // Dundee: University of Dundee, 2003
12. *Юдина М. В., Фёдорова О. В.* Разрешение синтаксической неоднозначности: эффекты прайминга и самопрайминга. // Диалог. М.:2009
13. *Sekerina I.* The Late Closure Principle in Processing of Ambiguous Russian Sentences. // The Proceedings of the Second European Conference on Formal Description of Slavic Languages. Universität Potsdam, Germany. 2003
14. *Roget's* Thesaurus Of English Words And Phrases // Penguin Books, London, 2000.