АВТОМАТИЧЕСКОЕ РАЗРЕШЕНИЕ АНАФОРЫ: СЛУЧАЙ НЕОДНОЗНАЧНОСТИ АНТЕЦЕДЕНТА МЕСТОИМЕНИЯ *КОТОРЫЙ*

Анна Маракасова
ИППИ РАН, НИУ ВШЭ
Москва, Россия

Неоднозначность антецедента местоимения *который* возникает крайне редко. В современных системах автоматической обработки текста практически не ставится вопрос о её разрешении. Более того, среди неоднозначных случаев довольно много таких, когда и человеку трудно однозначно выбрать одного из кандидатов в качестве антецедента. Был проведен эксперимент, который показывает, что с помощью грамматических (в том числе синтаксических) и семантических признаков можно достичь точности автоматического выбора кандидата в более чем 90%.

Ключевые слова: автоматическое разрешение анафоры, местоимение *который*, синтаксический анализ.

1. Относительные предложения с местоимением *который*

В большинстве относительных предложений с местоимением *который* отсутствует неоднозначность его антецедента в силу семантических ограничений и ограничений на согласование по категориям рода, числа, одушевленности. Анализ относительных предложений на материале НКРЯ это подтверждает.

Статистический анализ релятивных конструкций корпуса позволяет учитывать только случаи грамматических ограничений (с некоторой погрешностью). Так, в основном корпусе со снятой омонимией имеется 23585 сложных предложений с местоимением *который*, среди них грамматическая неоднозначность антецедента возможна по крайней мере в 6555 случаях, что составляет около 27% предложений. В действительности доля предложений с неоднозначностью антецедента местоимения *который* заметно меньше (около 1,2%), потому что в большинстве случаев при выборе антецедента действуют семантические ограничения. Так в предложении

> *Расстрелян* ***вице-мэра*** ***Пекина*** ***Лю*** ***Чжихуа****, который семь лет возглавлял управление китайской «Силиконовой долиной» - наукоградом Чжунгуанцунь в северо-западном университетском предместье Пекина.*

местоимение *который* соотносится с именной группой *вице-мэр Лю Чжихуа*, а не с городом Пекин[1]. Забегая несколько вперед, отметим, что наш алгоритм

[1] В этом примере, как и в последующих, мы выделяем жирным начертанием все слова, являющиеся кандидатами в антецеденты, а подчеркиванием – действительный антецедент местоимения *который*.

автоматического разрешения анафоры основывается на синтаксической структуре предложения, поэтому в случаях, когда антецедентом является не одна словоформа, а именная группа, релятивная клауза подчиняется вершине этой именной группы.

Для проведения эксперимента был составлен корпус относительных предложений с грамматической неоднозначностью. При его составлении использовался глубоко аннотированный корпус русских текстов СинТагРус [Дьяченко et al., 2015], содержащий помимо морфологической разметки синтаксическую, в виде деревьев зависимостей. Всего СинТагРус содержит 5582 предложения с местоимением *который*, их которых 723 (около 13%) имеют грамматическую неоднозначность антецедента[2].

2. Разрешение анафоры местоимения *который* в системе ЭТАП-3

Поиск антецедента относительного местоимения *который*, в отличие от поиска антецедента других местоимений, не представляет сложности для семантического модуля лингвистического процессора ЭТАП-3 [Iomdin et al., 2012]. Дело в том, что на вход семантического модуля поступает синтаксическая структура того или иного предложения в виде дерева зависимостей. Благодаря этому антецедент местоимения *который* легко идентифицируется: антецедентом является существительное главной клаузы, управляющее релятивной (иными словами вершина относительной клаузы подчиняется искомому антецеденту по релятивному отношению).

Однако в некоторых случаях различного рода грамматических ограничений и ограничения на проективность не достаточно, чтобы верно установить релятивную связь, и выбор верного антецедента местоимения *который* либо не возможен, например, как в

> *Кроме того, НС справляются с* ***проблемой размерности****, которая не позволяет моделировать линейные зависимости в случае большого числа переменных.*

либо осуществляется при помощи семантической и/или экстралингвистической информации. Интересно заметить, что то, насколько хорошо система способна автоматически разрешать анафору (второй случай), может служить некоторой оценкой, насколько система может “понимать” текст. В частности, схемы[3] Терри Винограда включают немало предложений, где для ответа на вопрос требуется установить антецедент местоимения *который*.

[2] В данном случае предложения были отобраны вручную после предварительной фильтрации результатов поиска, которая дала 1307 предложений.

[3] Схемы Винограда представляют собой специальным образом подобранные предложения с вопросами. Вопросы составлены так, что для ответа на них требуется понимание значения предложения (для человека) или умение производить логический вывод (для машины).

Чтобы уменьшить число случаев ошибочного установления релятивной связи, было написано правило, которое, используя в том числе и онтологическую информацию, проверяет корректность выбора вершины релятивной клаузы и, при необходимости, изменяет ее. Данное правило основывается на данных психолингвистических экспериментов [Юдина, Федорова & Янович 2007, Sekerina 2003], а также на статистике, собранной по корпусу. Размер нашего корпуса совсем небольшой. Тем не менее статистические подсчеты по корпусу согласуются с данными, собранными по более представительным корпусам. Так, распределение случаев раннего и позднего закрытия (повсеместно используемый в психолингвистике термин для обозначения выбора первого или второго кандидата в антецеденты местоимения *который* в случае его неоднозначности) в собранном корпусе составляет 59/41%, что соответствует данным экспериментов И.Секериной и М.Юдиной (60/40 %)[4].

Выбор в пользу того или иного кандидата вершины относительного предложения делается на основании следующих выделенных нами признаков (признаки указаны в том порядке, в котором они применяются):

1) одушевленность/неодушевленность предполагаемых кандидатов[5];
2) наличие аппозитивной связи между двумя кандидатами при их непосредственном подчинении;
3) наличие местоимения *тот, такой*, подчиняющегося одному из кандидатов[6];
4) наличие согласования по модели управления предиката - вершины релятивной клаузы с одним из кандидатов (онтологические дескрипторы);
5) наличие у одного из кандидатов предиката - вершины релятивной клаузы в качестве значения той или иной лексической функции;
6) определенный тип синтаксической структуры, связывающей кандидаты;
7) расстояние между кандидатами.

---

[4] В нашем случае к раннему закрытию относятся не только случаи, когда антецедентом является предпоследнее существительное перед началом придаточной клаузы, но и предложения типа *Основным удостоверяющим документом выступает **сертификат ключа подписи**, который центр выдаёт участникам электронного документооборота*, где между антецедентом и началом релятивной клаузы может быть 2 и более словоформы.

[5] Психолингвистические эксперименты на материале нидерландского языка выявили такие закономерности: при наличии одушевленного и неодушевленного существительных одушевленное существительное выбирается чаще, чем неодушевленное в любой позиции, при наличии двух одушевленных существительных преобладает раннее закрытие, а при наличии двух неодушевленных -- позднее закрытие [Brysbaert & Mitchell, 1996]. На нашем корпусе подтвердились две выше описанные закономерности (при наличии двух одушевленных существительных, в 87% случаев наблюдается раннее закрытие; при наличии одушевленного и неодушевленного в 67% случаев антецедентом является одушевленное существительное), которые и были включены в правило.

[6] Анализ размеченного корпуса показал, что кандидат, имеющий в качестве зависимого местоимение *тот* или *такой*, с большей вероятностью является антецедентом, чем кандидат, не имеющий такого зависимого: *Он составлен по библиографическим бюллетеням и по тем **спискам трудов**, которые были предоставлены некоторыми из авторов*.

3. Результаты эксперимента

480 (из 723) предложений собранного корпуса использовались для тестирования правила, остальные – для его отладки. Точность работы правила составила 0,92.

В силу того, что правило основывается на синтаксической структуре зависимостей, оно успешно справляется с отождествлением вершины именной группы среди цепочки кандидатов:

*Для "Единой России" это стало очевидным после отстранения от дел бывшего **председателя** **генсовета** партии **Александра Беспалова**, который пытался сохранить "монополию на власть".*

*Другим вариантом **приёма** **смешения** стилей является зевгема (от греч. "связь, **объединение разнопланового**"), при котором происходит намеренное нарушение законов сочетаемости слов.*

Для статистических систем разрешения анафоры такие конструкции представляют особенную сложность.

Использование лексических функций и онтологических дескрипторов модели управления позволяет делать выбор кандидата с учётом семантики. Так, в предложении

*Он понимал всю **нелепость** **жизни**, которую ведут люди*,

благодаря тому, что глагол *вести* является значение лексической функции OPER1 существительного *жизнь* (что неверно для существительного *нелепость*), правило выбирает верное опорное слово релятивной клаузы.

На основании наличия дескриптора (Agent) соответствующей строки модели управления вершины придаточного среди дескрипторов одного из кандидатов (Agent, PhysicalObject, Organization, Collection), успешно решается неоднозначность в предложении

*Мы сообщаем контактные **телефоны** **организаций**, которые проводят с учащимися и учителями экскурсии, организуют экологическую практику*.

Семантическая информация помогает выбрать верного антецедента в зависимости от контекста предложения, ср.:

*Совершенно не допускается применять циклование, особенно на облицованной поверхности, так как при этом уменьшается **толщина** **облицовки**, которая при реставрации в будущем может разрушиться.*

*Прочное приклеивание набора к основе достигается лишь при равномерной **толщине** **облицовки**, которую получают выравниванием мозаики.*

Ошибки правила объясняются, во-первых, ошибками в построении исходной синтаксической структуры предложения (из-за большой длины предложения, инверсного порядка слов, прямой речи и других стилистических особенностей).
Во-вторых, отсутствием необходимой словарной информации для использования семантических признаков. Так, при анализе предложения

*На **фоне кризиса**, который настиг некоторых её соседей, Магадан выглядит островком спокойствия и благополучия.*

правило ошибается. Модель управления глагола *настигать* не содержит набора дескрипторов, допустимых в позиции его предикативного зависимого; подходящих лексических функций также не имеется. Таким образом, получается, что нет признаков, с помощью которых можно было бы предпочесть один кандидат другому (*фон* vs. *кризис*), поэтому правило выбирает первого из кандидатов, как было бы верно в большинстве подобных случаев.

В-третьих, ошибки встречаются в тех предложениях, где отсутствует даже потенциальная возможность сделать верный выбор, как, например, в

*К шести вечера обычно возвращалась обладательница самой просторной во всей **квартире комнаты**, которую она добровольно отдавала под чью-нибудь свадьбу, под именины.*

4. Литература

1. *Дяченко П.В., Иомдин Л.Л., Лазурский А.В., Митюшин Л.Г., Подлесская О.Ю., Сизов В.Г., Фролова Т.И., Цинман Л.Л.* Современное состояние глубоко аннотированного корпуса текстов русского языка (СинТагРус) // Сборник «Национальный корпус русского языка: 10 лет проекту». Труды Института русского языка им. В.В. Виноградова. М., 2015. Вып. 6. С. 272-299.
2. *Юдина М.В., Федорова О.В., Янович И.С.* Синтаксическая неоднозначность в эксперименте и в жизни // Компьютерная лингвистика и интеллектуальные технологии. Труды международной конференции “Диалог-2007”. М., 2007. С. 202-208.
3. *Brysbaert M., Mitchell D.C.* Modifier attachment in sentence parsing: Evidence from Dutch. // Quarterly Journal of Experimental Psychology, 49A, 3, 1996.
4. *Iomdin L.L., Petrochenkov V.V., Sizov V.G., Tsinman L. L.* ETAP parser: state of the art. // Computational Linguistics and Intellectual Technologies. International Conference (Dialog’2012). Moscow: RGGU Publishers, Issue 11(18). 2012. Pp. 830–843.
5. *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. Pp. 376-384.