

УДК 81'367

Анализ некоторых проблем синтаксических анализаторов русского языка

Пантелеев М.Ф., студент

*Россия, 105005, г. Москва, МГТУ им. Н.Э. Баумана,
кафедра «Компьютерные системы и сети»*

*Научный руководитель: Самарев Р.С., к.т.н, доцент
Россия, 105005, г. Москва, МГТУ им. Н.Э. Баумана
bauman@bmstu.ru*

Введение

В настоящее время существует очень важная проблема извлечения информации из текста на естественном языке для его дальнейшей обработки. Автоматизация этого процесса представляется очень важной задачей в основном потому, что каждый час в СМИ появляется огромное количество информации, которое человек обработать просто не под силу. Один из возможных и наиболее действенных способов извлечения информации из текста базируется на использовании синтаксических деревьев предложений.

Проблемы синтаксических анализаторов

Можно выделить три подгруппы причин, способствующих появлению синтаксической неоднозначности, создающей основную проблему синтаксических анализаторов:

1. Неоднозначность определения входящих в предложение единиц.
2. Неоднозначность определения связей между единицами.
3. Особенности самой синтаксической структуры.

Рассмотрим каждую из этих подгрупп по отдельности. В качестве примера будет приводиться дерево синтаксическому разбору предложения. Используются анализатор АОТ, как один из наиболее известных свободного распространяемых анализаторов, а также анализатор Treevial, разрабатываемый на факультете ВМиК МГУ им. М.В. Ломоносова.

Неоднозначность определения входящих в предложение единиц.

Сюда можно отнести следующие явления:

1. Омонимия и полисемия.
2. Грамматическая конверсия.
3. Неоднозначность интерпретации проформы.

Омонимия и полисемия. Омонимы – разные по значению, но одинаковые по звучанию и написанию слова, морфемы и другие единицы языка. Пример: «Звериная клетка»/«клетка в тетради», «Косил косой косой косой». Будем считать, что омонимия – случайное совпадение слов, в то время, как полисемия – наличие у слова разных исторических связанных значений. Например, “бор” в значении “сосновый лес” и “бор” в значении “химический элемент” являются омонимами, в то время, как “эфир” в смысле “радиовещание и телевидение” и “эфир” в смысле органического вещества являются полисемией, так как оба происходят от древнегреческого слова αἰθήρ, которое в переводе означает “горный воздух”.

Пример: “Косил косой косой косой”. Видно, что существует несколько вариантов интерпретации предложения в зависимости от положения в нём главного действующего лица – косого.

Результаты разбора показаны на рис. 1,2.

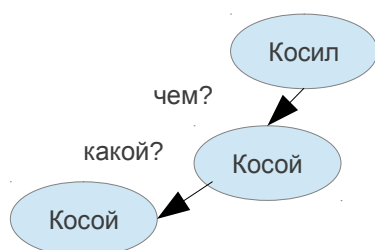


Рис. 1. Пример разбора анализатором АОТ

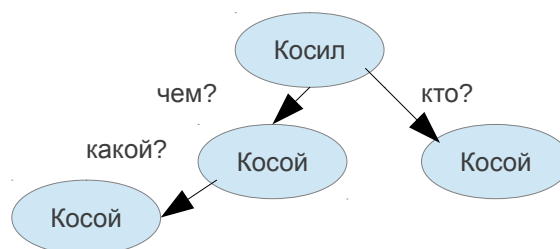


Рис. 2. Пример разбора анализатором Treeview

Можно сказать, что синтаксический анализатор АОТ плохо справился с разбором данного предложения, так ему не удалось правильно определить главное действующее лицо. Также он посчитал, что два последних слова являются прилагательными, которые характеризуют один предмет – “коса”. Анализатор Treeview предоставил несколько вариантов разбора, в числе которых был верный.

Грамматическая конверсия. Грамматическая конверсия — это выражение одного и того же действия в разных направлениях.

Пример: “Раскапывают погребённых в земле слепых исполинов”. Ожидается один из следующих вариантов интерпретации предложения анализатором: “[Раскапывают

погребённых] [в земле слепых исполинов]” или “[Раскапывайте погребённых в земле] [слепых исполинов]”. Результаты разбора показаны на рис. 3,4.

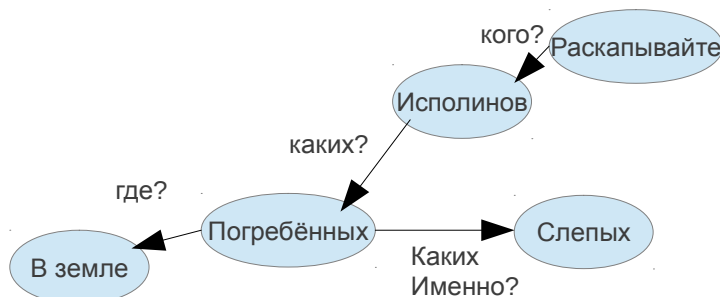


Рис. 3. Пример разбора анализатором AOT

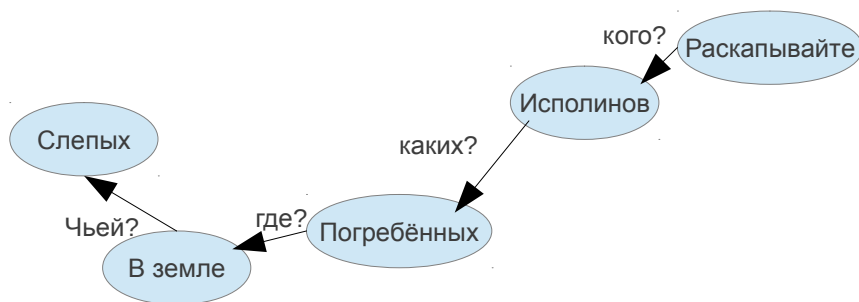


Рис. 4. Пример разбора анализатором Treevial

Синтаксический анализатор AOT правильно разобрал предложение, отдав предпочтение второму варианту разбора.

Синтаксический анализатор системы Treevial предложил несколько вариантов разбора предложения, в том числе предложенный анализатором AOT, как минимум два были верны.

Неоднозначность интерпретации проформы.

Пример: “Доклад ученого, о котором я вам говорил”. Здесь неопределенность заключается в том, что непонятно, о ком говорит действующее лицо - о докладе или об учёном. Результат разбора приведён на рис. 5.

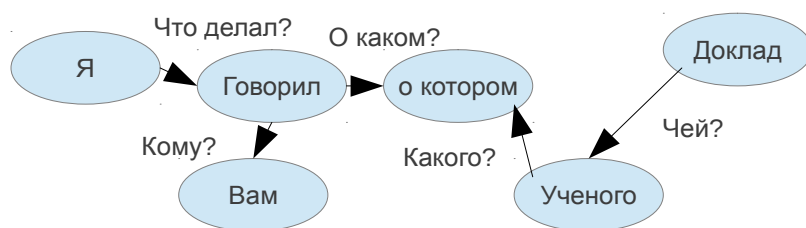


Рис. 5. Пример разбора анализатором АОТ

Синтаксический анализатор АОТ правильно разобрал предложение, посчитав, что ранее говорилось об учёном.

Синтаксический анализатор системы Treevial предложил один вариант разбора, отличающийся только тем, что проформой был определён доклад.

Неоднозначность определения синтаксических связей между этими единицами.

Сюда можно отнести следующие явления:

1. Факультативность сирконстантов.
2. Вариативность актантов.
3. Факультативность актантов.

Факультативность сирконстантов. Сирконстант – слово, отражающее, при каких обстоятельствах происходит действие. Например: “вчерашнее пятно”, “завтрашний день”. Сирконстанты не являются обязательными элементами и имеют способность легко находить в предложении потенциальных “хозяев”.

Пример: “Новость о преступлении в полицейском участке”. Неоднозначность заключается в том, что непонятно, произошло ли преступление непосредственно в полицейском участке или о нем там узнали.

Результаты разборов приведены на рис. 6, 7, 8:

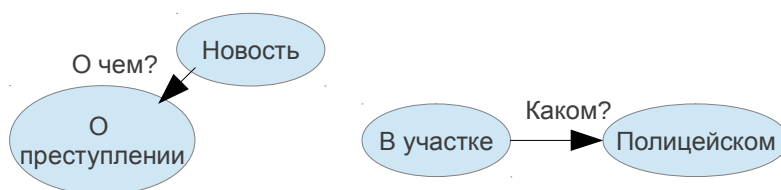


Рис. 6. Пример разбора анализатором АОТ

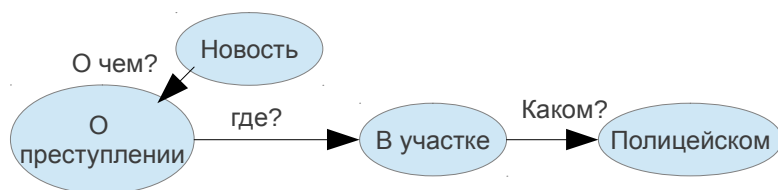


Рис. 7. Пример разбора анализатором Treevial

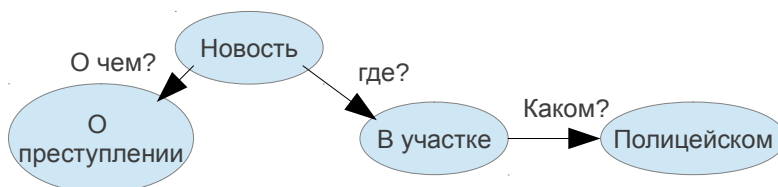


Рис. 8. Пример разбора анализатором Treevial

Синтаксический анализатор АОТ не смог полностью правильно построить дерево, так как в первом варианте не хватает отношения “преступление (где?) в участке”, а во втором – отношения “новость (где?) в участке”.

Синтаксический анализатор системы Treevial предложил несколько вариантов разбора, в том числе два ожидаемых и полностью верных.

Вариативность актантов. Актант – активный, действующий участник ситуации. Вариативность актантов означает способность слова заполнять валентности актантами разных типов, а также подчинять себе разные формы.

Пример: “Имеется возможность просьбы начальника избежать”. Неоднозначность в том, что неизвестно, имеется ли возможность избежать просьбы или имеется возможность не просить. Имеются и другие интерпретации. Результаты разборов показаны на рис. 9,10.

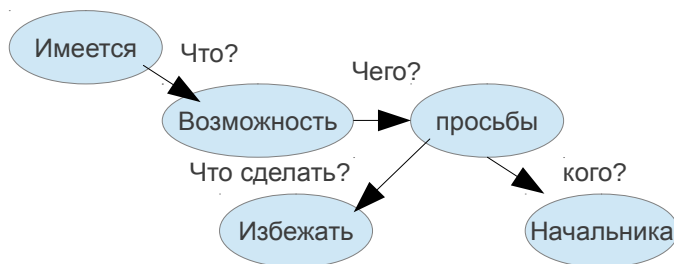


Рис. 9. Пример разбора анализатором АОТ

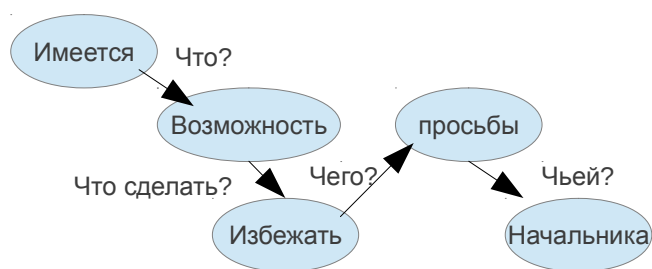


Рис. 10. Пример разбора анализатором Treevial

Анализатор АОТ продемонстрировал построение, близкое к ожидаемому при второй интерпретации. Для полноты не хватает отношения “возможность (что сделать?) избежать”.

Синтаксический анализатор системы Treevial продемонстрировал верное построение дерева, совпадающее с первым вариантом.

Факультативность актантов, то есть возможность относить зависимый элемент как к одному, так и к другому хозяину.

Пример: “Учитель пения не слышит”. Неоднозначность в следующем: Учитель по пению не слышит что-либо или учитель не слышит пения. Результат разбора приведён на рис. 11.

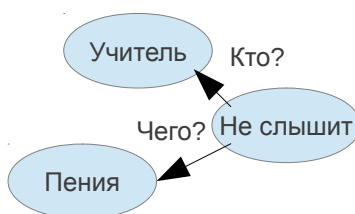


Рис. 11. Пример разбора анализатором АОТ

Видно, что синтаксический анализатор АОТ правильно произвёл разбор предложения, который соответствует второму варианту интерпретации.

Синтаксический анализатор системы Treevial не предоставил ни одного варианта построения дерева предложения.

Неоднозначности, вызванные особенностью синтаксической структуры

Сюда можно отнести следующие явления:

1. Наличие зависимого элемента, который может быть отнесён к одному или нескольким членам предложения.
2. Наличие элемента, который может состоять в отношении с несколькими

однородными членами предложения или только с одним из них.

3. Случаи перераспределения однородности.

Здесь неоднозначность связана со свойствами самой синтаксической структуры.

Наличие зависимого элемента, который может быть отнесён к одному или нескольким членам предложения.

Пример: “Век атомной энергетики, автоматизации”. Здесь неоднозначность в неопределённости: характеризует ли прилагательное “атомной” только существительное “энергетики”, или ещё и существительное “автоматизации”?

Результат разбора показан на рис.12.

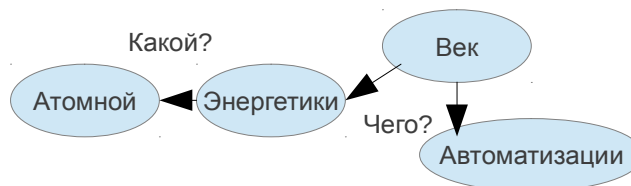


Рис. 12. Пример разбора анализаторами AOT, Treevial

Анализатор AOT правильно произвёл разбор, что соответствует первой интерпретации.

Синтаксический анализатор системы Treevial произвел разбор идентично.

Наличие элемента, который может состоять в отношении с несколькими однородными членами предложения или только с одним из них.

Рассмотрим предложение “Доверять дочери и отцу запрещали”. Здесь союз “и” может выступать как в роли соединительной – доверять запрещали (дочери и отцу), так и в роли усилительной конструкции – доверять дочери запрещали даже отцу. Результаты разбора приведены на рис. 13, 14.

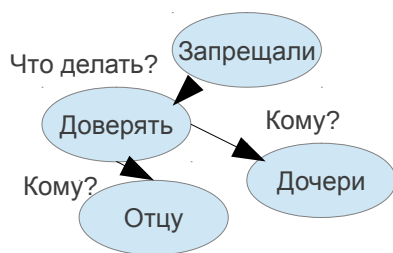


Рис. 13. Пример разбора анализатором АОТ

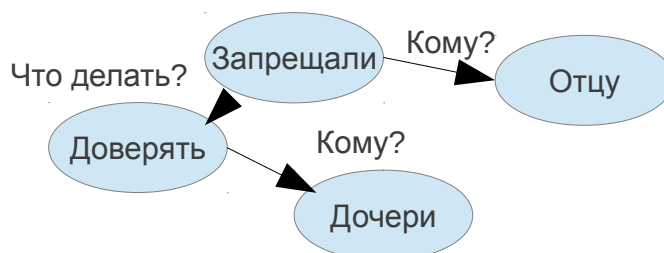


Рис. 14. Пример разбора анализатором Treeeval

Синтаксический анализатор АОТ правильно произвёл разбор предложения, посчитав, что союз “и” соединяет однородные члены.

Синтаксический анализатор Treeeval правильно произвёл разбор предложение, отдав предпочтение второму варианту, однако этот вариант разбора имел далеко не самый высокий приоритет среди предложенных.

Случаи перераспределения однородности.

Можно рассмотреть следующий пример: “Мой брат или Петя и Миша останутся”. Союзы “и” и “или” могут по-разному объединять однородные члены предложения. Так, в одном случае останутся брат или Петя с Мишей, а в другом – Миша – либо с братом, либо с Петей. Результат разбора приведен на рис. 15:

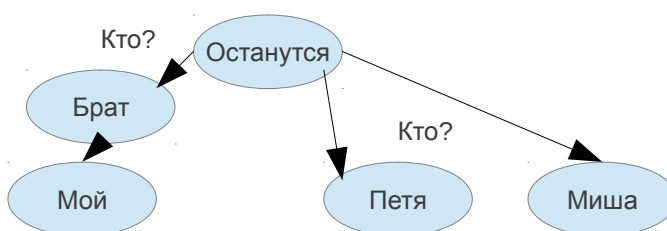


Рис. 15. Пример разбора анализатором АОТ

Можно заметить, что средствам АОТ не удалось удовлетворительно разобрать приведённый пример, потому он не подходит ни под один из возможных вариантов разбора.

Синтаксический анализатор Treeeval также не смог предложить верный вариант разбора.

Заключение

Проведенные исследования показали, что АОТ смог предоставить лишь один не всегда правильный вариант разбора, а Treevial практически всегда предлагал несколько вариантов построения синтаксических деревьев, но эти правильные варианты, как правило, приходилось отыскивать среди множества неправильных (для предложения “Косил косой косой косой” было порождено более 100 вариантов). Используемые примеры продемонстрировали следующее: при проектировании синтаксического анализатора необходимо учитывать, что для правильного построения дерева разбора предложения недостаточно просто использовать правила, описывающие отношения между словами, необходимы результаты семантического анализа. Сегодня решена только задача поверхностного семантического анализа: например, средства АОТ производят поверхностный семантический анализ только для одного разбираемого предложения. Однако может возникнуть такая ситуация, когда для снятия неоднозначности в каком-либо предложении может потребоваться информация, содержащаяся в другом. Нетрудно привести такие примеры, в которых использование контекста невозможно и из нескольких вариантов разбора надо выбрать наиболее правдоподобный, то есть подходящий по смыслу.

Список литературы

1. Программа mystem. Режим доступа: <http://company.yandex.ru/technologies/mystem/> (дата обращения 15.03.14).
2. Программа solarix. Режим доступа: <http://www.solarix.ru/> (дата обращения 20.03.14).
3. Программа Treeton. Режим доступа: <http://starling.rinet.ru/treeton/> (дата обращения: 15.03.14).
4. Программа АОТ. Режим доступа: <http://www.aot.ru/> (дата обращения 15.03.14).
5. Митренина О.В. Проблемы неоднозначности синтаксического анализа: дис. ... канд. фил. наук. Санкт-Петербург 2005. 133с. (дата обращения 09.02.14).
6. Стеммер Snowball. Режим доступа: <http://snowball.tartarus.org/> (дата обращения 12.03.14).