ЭКСПЕРИМЕНТАЛЬНАЯ РЕАЛИЗАЦИЯ СЕГМЕНТАЦИОННОГО АНАЛИЗА РУССКОГО ПРЕДЛОЖЕНИЯ[1]

EXPERIMENTAL IMPLEMENTATION OF RUSSIAN SENTENCE SEGMENTATION ANALYSIS

Баталина А.М. *(batalina_anna@rambler.ru),* ***Епифанов М.Е.*** *(xeme@rambler.ru),*
Кобзарева Т.Ю. *(stamstam@mtu-net.ru),* ***Кушнарёва Е.В.*** *(likart@mail.ru),*
Лахути Д.Г. *(delir1@yandex.ru)*
Российский государственный гуманитарный университет, Москва

В статье описывается моделирование и отладка сегментационного анализа русского предложения в инструментальной среде для экспериментов с алгоритмами поверхностно-синтаксического анализа.

Введение

В докладе описаны результаты очередного этапа разработки создаваемой в Институте лингвистики РГГУ [1-3] объектной среды для моделирования системы правил поверхностно-синтаксического анализа (ПСА) [4-6] с целью их проверки и совершенствования их организации. На данном этапе объектом экспериментальной реализации были правила сегментационного анализа русского предложения.

Правила ПСА автор-лингвист организует в виде алгоритмов. Каждый такой алгоритм представляет собой совокупность правил, содержащих проверки некоторых лингвистических ситуаций в предложении. Правила соединены переходами, причем переход к следующему правилу осуществляется в зависимости от того, какие условия предыдущего правила выполнены, а какие нет. Алгоритмы записываются автором в документах Word похожим на блок-схему образом.

ПСА в рамках рассматриваемого подхода представляет собой последовательное применение алгоритмов к предложению в заданном порядке. ПСА состоит из нескольких стадий, и сегментационный анализ (СА) – одна из них. Ранее была осуществлена пробная реализация, причем в сжатые сроки, других, отличных от СА, этапов ПСА: предсинтаксического, предсегментационного и внутрисегментного анализа. Этот опыт охарактеризован в работе [7]. Для того чтобы охарактеризованные в работе [7] алгоритмы могли анализировать не только простые, но и синтаксически сложные предложения, необходимо выполнить СА.

Характеристика сегментационного анализа русского предложения в рамках рассматриваемого подхода приведена в работе [4]. Коротко говоря, на этом этапе предложение разделяется на «зоны анализа» (сегменты), которые далее обрабатываются одна за другой. То есть внутрисегментный анализ (ВА) последовательно применяется к каждому сегменту, в результате чего строится граф его синтаксических связей. После этого должен работать межсегментный анализ, задача которого – соединить обработанные сегменты синтаксическими связями специального типа.

В отличие от рассматриваемого в данной работе подхода, в других системах при синтаксическом анализе русского предложения сегментация обычно или не выделяется в отдельную задачу [8-10], или осуществляется перебором заранее заданных комбинаций сегментов [11, 12]. В предлагаемой же системе [6] задачу сегментации решает отдельный модуль сегментации [4, 5], представляющий собой комплекс рекурсивных алгоритмов, позволяющих анализировать линейные структуры любой конечной длины и «громоздкости» [13].

Ранее уже проводилась реализация рассматриваемого подхода к СА с использованием компилируемого языка программирования Delphi Pascal [5, 14]. Она предполагалась как часть возможной системы (реализация выполнялась в рамках работы над диссертацией), ориентированной на «конечный результат», т.е. обработку большого массива предложений. Её автор не ставил перед собой задачу долговременной поддержки усовершенствования самой системы алгоритмов СА, а именно исправления ошибок в алгоритмах, их пополнения, оптимизации их структуры. Конечно, в процессе этой реализации алгоритмы СА были существенным образом доработаны, однако в итоге была запрограммирована их конкретная, актуальная на тот момент времени версия. При этом модификации кода программы вследствие изменения алгоритмов и последующая ее отладка требовали все большего времени.

[1] Доклад подготовлен при частичной поддержке РФФИ (грант 06-06-80434).

В данной работе мы акцентируем внимание на экспериментальном моделировании и отладке алгоритмов СА. Изменения объектной модели алгоритмов СА и проверка их правильности на тестовых примерах в применяемой для этого инструментальной среде не требуют слишком большого времени в связи с предметной ориентированностью языка входного описания модели и возможностью пошаговой интерпретации правил в процессе отладки.

Краткое описание алгоритмов сегментационного анализа

Данный вариант СА рассчитан на сегментацию любых грамматически правильных неэллиптичных предложений литературного письменного языка, не являющихся записью или имитацией устной речи. Он представляет собой блок алгоритмов, работающих в заданном порядке. В результате применения СА к предложению-примеру предложение будет поделено на «зоны анализа» (сегменты). Внутри них будет производиться ВА с учетом проективности уже построенных подчинительных и сочинительных связей.

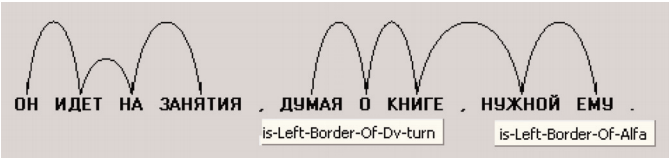
СА был построен, исходя из следующих представлений:

- костяком русского предложения (S) является цепочка простых («главных») предложений (таких предложений будет как минимум одно) – **β-сегментов**;
- каждое из них может быть разорвано вложением в него придаточных предложений, обособленных согласованных определений, деепричастных, предложных, вводных и сравнительных оборотов – **α-сегментов**;
- каждый α-сегмент, в свою очередь, может быть разорван такими же вложениями, причем количество вложений теоретически не ограничено.

Для реализации СА, покрывающего большинство наиболее распространенных случаев, была создана его *экспресс-версия*. Она ориентирована на анализ предложений с синтаксически сложной линейной структурой. Отличием экспресс-СА от полной версии СА является то, что в нем не предполагается возможности разрывающих вложений α-сегментов в α-сегменты.

Экспресс-СА включает в себя определение границ простых-главных предложений, причастных, деепричастных и других обособленных оборотов и придаточных предложений на базе анализа сочинения и предикативной структуры предложения, включая анализ неморфологического предиката. В ходе экспресс-СА границам сегментов приписываются свойства «быть границей простого-главного/обособленного оборота/придаточного такого-то типа». Эти границы учитываются при осуществлении ВА и межсегментного анализа.

Работа экспресс-СА уже промоделирована в объектной среде для экспериментов с алгоритмами ПСА, отлажена и используется в системе. Приведем примеры работы экспресс-СА:



(1) Первой запятой приписано свойство «левая граница деепричастного оборота» и второй запятой приписано свойство «левая граница α-сегмента» (т.е. здесь – причастного оборота).

(2) Запятой приписано свойство «левая граница придаточного».

(3) Первой и третьей запятой приписаны свойства «левая граница α-сегмента» и «левая граница сочиненного β-сегмента» соответственно. Вторая запятая однозначно определяется как правая граница причастного оборота «думающий о книгах». В случае присутствия какого-либо вложенного сегмента это могло бы быть неверно.

Лингвистический базис полной версии СА, еще не включавшей анализа тире, двоеточия и точки с запятой, был охарактеризован в работе [4]. В объектной среде для экспериментов с алгоритмами ПСА в данный момент ведется моделирование и отладка алгоритмов полного СА, дополненного алгоритмами анализа тире, двоеточия и точки с запятой. Собственно полный сегментационный анализ S включает в себя два этапа:

1. определение границ α-сегментов с одновременным восстановлением целостности α-сегментов, разорванных вложениями;
2. определение границ β-сегментов с одновременным восстановлением их целостности.

Эти задачи решают одиннадцать алгоритмов, использующие стандартные подпрограммы проверки согласования и управления:

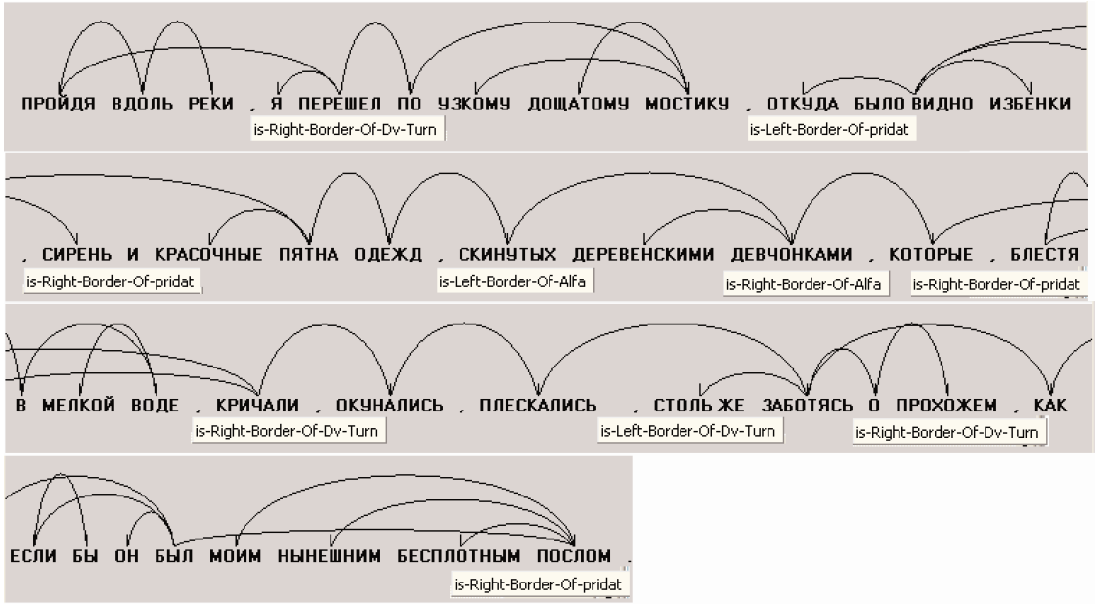
1. Блок алгоритмов определения левых границ α-сегментов осуществляет поиск **α-отрезков** – ограниченных знаками препинания безусловных левых составляющих α-сегментов. Содержит алгоритмы:
- поиск левой границы α-сегмента – запятой,
- анализ тире,
- анализ двоеточия с подпрограммой поиска синтаксически выраженных родовидовых отношений,
- анализ точки с запятой,
- алгоритм определения левой границы и хозяина обособленного согласованного определения с вершиной причастием или его синтаксическим аналогом (А-оборота).

2. Алгоритм построения α-сегментов представляет собой итеративное удлинение каждого очередного найденного в 1-м блоке α-отрезка грамматически допустимым присоединением ближайшего к нему справа **β-отрезка** – отрезка, еще не идентифицированного как составляющая α-сегмента. Алгоритм использует как подпрограмму алгоритм анализа сочинения.

3. Блок алгоритмов построения **β-сегментов**: поиск границ β-сегментов. Содержит алгоритмы:
- элиминирование α-сегментов с анализом места разрыва,
- поиск не найденных при построении α-сегментов сказуемых («неморфологических предикатов»),
- поиск границ зон влияния сказуемых,
- анализ сочинения предикатов и поиск границ β-сегментов.

Рассмотрим работу перечисленных алгоритмов полной сегментации на двух примерах из прозы Набокова.

(4) Результат применения алгоритмов полной сегментации к предложению «Пройдя вдоль реки, я перешел по узкому дощатому мостику, откуда было видно избенки, сирень и красочные пятна одежд, скинутых деревенскими девчонками, которые, блестя в мелкой воде, кричали, окунались, плескались, столь же заботясь о прохожем, как если бы он был моим нынешним бесплотным послом.»:



Первый этап – построение α-сегментов (придаточных предложений, деепричастных и других обособленных оборотов) – любых сегментов, осложняющих линейную структуру β-сегментов (простых-главных).

1. Алгоритм поиска левых границ α-сегментов:

a. S разбивается на отрезки – части предложения между ближайшими знаками препинания. Знаки препинания внутри уже построенных именных и предложных групп, как и все остальные компоненты фрагментов, ограниченных этими связями, из рассмотрения исключаются.

b. Ищутся отрезки, в которых есть подчинительный союз, деепричастие или причастие/прилагательное, для последних специальная подпрограмма (алгоритм анализа А-оборота) ищет хозяина в отрезке слева. Такие отрезки являются безусловными левыми компонентами α-сегментов:
[***Пройдя*** *вдоль реки*] –деепричастный оборот (Д-об), [*я перешел по узкому дощатому мостику*], [***откуда*** *было видно избенки*] – придаточное, [*сирень и красочные пятна одежд*], [***скинутых*** *деревенскими девчонками*] – обособленный А-оборот, [***которые***] – придаточное, [***блестя в мелкой воде***] – Д-об, [*кричали*], [*окунались*], [*плескались*], [***столь же заботясь о прохожем***] – Д-об, [***как если бы он был моим нынешним бесплотным послом***] –придаточное.
На этом шаге в предложении найдено семь минимальных составляющих α-сегментов (см. иллюстрацию к примеру (4)).

2. Алгоритм поиска правых границ α-сегментов с одновременной ликвидацией их разрывов вложениями α-сегментов:

a. Первый справа налево отрезок α-сегмента (**α-отрезок**) – [*как если бы он был моим нынешним бесплотным послом*]. Он последний в S, его правая граница – точка. Объявляем его α-сегментом придаточным и исключаем из рассмотрения.

b. Берем следующий α-отрезок [***столь же заботясь о прохожем***]. Правее него только исключенный α-сегмент, поэтому правая граница отрезка и есть правая граница сегмента. Исключаем его из рассмотрения.

c. Берем следующий α-отрезок [***блестя в мелкой воде***]. В отрезке непосредственно справа есть глагол в личной форме *кричали*, который не может быть внутри Д-об: этот отрезок присоединить нельзя. В силу проективности сегментов отрезки правее уже рассмотрению не подлежат. Отрезок объявляется сегментом и исключается из рассмотрения.

d. Следующий α-отрезок [***которые***]. Отрезок непосредственно справа от него исключен из рассмотрения. Ближайший справа претендент на присоединение – отрезок [*кричали*]. Он может быть или не быть частью анализируемого. В силу тривиальной неполноты отрезка [**которые**] и согласования существительного – подчинительного союза в Им.п. с глаголом *кричали* соединяем отрезки и получаем новый отрезок придаточного: [**которые**]+ [*кричали*]=[***которые кричали***]. Справа еще не опознанный [*окунались*], содержащий единственное слово – глагол, согласующийся с *кричали*, соответственно присоединяем его: [***которые кричали***]+ [*окунались*]=[***которые кричали, окунались***]. Аналогично [***которые кричали, окунались***]+ [*плескались*]=[***которые кричали, окунались, плескались***]. Правее только исключенные сегменты, поэтому построенный отрезок объявляется сегментом.

e. Правее α-отрезка [***скинутых деревенскими девчонками***] нет не исключенных из рассмотрения отрезков, поэтому он объявляется α-сегментом.

f. Следующий налево α-отрезок [***откуда было видно избенки***]. Справа от него только отрезок [*сирень и красочные пятна одежд*], в котором нет сказуемых, поэтому запятая между отрезками может не быть границей придаточного, только если она в этих отрезках сочиняет слова – не сказуемые. Алгоритм анализа сочинения находит сочинение существительных *избенки* и *сирень*: [***откуда было видно избенки***]+[*сирень и красочные пятна одежд*]=[***откуда было видно избенки, сирень и красочные пятна одежд***].

g. [***Пройдя вдоль реки***] объявляется сегментом, так как в единственном сохранившемся справа от него отрезке есть глагол в личной форме (ср. 2.с.).

Второй этап – построение простых-главных (β-сегментов).

Элиминируем все α-сегменты. Остается единственный отрезок [*я перешел по узкому дощатому мостику*], который и объявляется β-сегментом.

Для иллюстрации более сложного варианта работы второго этапа рассмотрим пример.

(5) Но взрывом веселья мгновенно разлучая нас, в сумраке началась снежная свалка, и кто-то, спасаясь, падая, хрустя, хохоча с запышкой, влез на сугроб, побежал, охнул сугроб, произвел ампутацию валенка.

1. После исключения шести построенных аналогично первому примеру α-сегментов остаются β-отрезки, в цепочке которых [*и кто-то*] и [*влез на сугроб*] объединяются, так как запятая между ними (в месте разрыва) не сочиняет, и *кто-то* и *влез* согласованы.
2. Определяются сказуемые: [*в сумраке* ***началась*** *снежная свалка*], [*и кто-то* ***влез*** *на сугроб*], [***побежал***],[***охнул*** *сугроб*], [***произвел*** *ампутацию валенка*].
3. Ищем границы зон влияния сказуемых. Это – запятые между отрезками, так как между каждыми двумя сказуемыми ровно одна запятая.
4. Все сказуемые – личные глаголы. Проверяется согласование сказуемых: отрезок [*в сумраке началась снежная свалка*] объявляется сегментом, так как его сказуемое *началась* и ближайший справа от него глагол *влез* не согласуются.
5. Остальные сказуемые согласуются. Чтобы определить, где кончается цепочка сочиненных сказуемых с подлежащим *кто-то*, нужно найти сказуемое, в отрезке которого есть собственное подлежащее. Подлежащее появляется в отрезке [***охнул*** *сугроб*], дальше претендентов на эту роль нет. Соответственно, строим два β-сегмента: [*и кто-то* ***влез*** *на сугроб,* ***побежал***], [***охнул*** *сугроб,* ***произвел*** *ампутацию валенка*].

В целом полный и экспресс-СА в системе дополняют друг друга. Существует возможность проверить, корректно ли применение экспресс-СА к анализируемому предложению, и в случае отрицательного ответа перейти к полному СА.

Изменения в инструментальной среде

Объектное моделирование СА потребовало некоторой доработки инструментальной среды для экспериментов с алгоритмами ПСА. В частности, было завершено описанное в [1] объектное представление иерархии сегментов предложения. (Алгоритмы, с которыми велась работа ранее, не требовали такого представления, т.к. либо они применяются внутри одного сегмента, и мы использовали синтаксически простые предложения для их отладки, либо для их применения сегментация вообще не актуальна.) В объектной модели предложения, учитывающей его деление на сегменты, последние представлены объектами, сохраняющими:
- грамматические свойства сегмента в целом,
- ссылки на его дочерние объекты, представляющие либо слова предложения, либо вложенные непосредственно в него сегменты,
- ссылку на родительский объект.

Корневой объект соответствует предложению «в целом». Следует заметить, что упомянутые ссылки представляют дуги дерева сегментов и не имеют отношения к синтаксическим связям внутри предложения.

Однако в данной модели сегмент, как и слово предложения, может быть синтаксически связан с другим словом или сегментом.

Поведение объектов, представляющих члены и части предложения поддерживает следующую функциональность:
- последовательный ВА сегментов в соответствии с их иерархией,
- возможность в процессе ВА и межсегментного анализа рассматривать сегмент как единое целое или как совокупность его составляющих с нужной степенью детализации,
- проведение синтаксических связей между ними.

Заключение

Итогом данной работы стала пробная реализация модуля сегментационного анализа русского предложения в двух вариантах. Имплементированный экспресс-СА ориентирован на сложносочиненные и сложноподчиненные предложения без разрывающих вложений α-сегментов в α-сегменты. Моделируемый и отлаживаемый модуль полного СА ориентирован на полный сегментационный анализ предложения оговоренного выше типа с любой степенью вложенности сегментов. В [7] была охарактеризована совокупность промоделированных алгоритмов, которая отлаживалась на простых предложениях. После окончания моделирования и отладки полного СА она будет работать в составе полной системы алгоритмов ПСА на сегментах любого типа. Реализация остается открытой для пополнений и модификаций, которые в рассматриваемой среде требуют сравнительно малых человеко-временных ресурсов.

Список литературы

1. Баталина А.М., Епифанов М.Е., Ивличева О.О., Кобзарева Т.Ю., Лахути Д.Г. Инструментальная среда для экспериментов с алгоритмами поверхностно-синтаксического анализа // Компьютерная лингвистика и интеллектуальные технологии: Труды Международной конференции Диалог'2004 («Верхневолжский», 2-7 июня 2004 г.). М.: Наука. С. 32-38.
2. Баталина А.М. Объектное моделирование поверхностно-синтаксического анализа // Девятая национальная конференция по искусственному интеллекту с международным участием КИИ-2004: Труды конференции. М.: Физматлит, 2004. Т.2, с. 462-471.
3. Баталина А.М., Айриян Г.Ю., Епифанов М.Е., Кобзарева Т.Ю., Лахути Д.Г. Автоматизация отладки алгоритмов поверхностно-синтаксического анализа // Компьютерная лингвистика и интеллектуальные технологии: Труды Международной конференции Диалог'2005 (Звенигород, 1-6 июня 2005 г.). С. 45 - 50.
4. Кобзарева Т.Ю. Принципы сегментационного анализа русского предложения // Московский лингвистический журнал. М.: РГГУ, 2004. Т.8, №1, с. 31-80.
5. Кобзарева Т.Ю., Лахути Д.Г., Ножов И.М. Сегментация русского предложения // Труды конференции. Седьмая национальная конференция по искусственному интеллекту с международным участием. КИИ' 2000 Москва. Издательство Физико-математической литературы, 2000. С. 879-880.
6. Кобзарева Т.Ю. Иерархия задач поверхностно-синтаксического анализа русского предложения // НТИ. 2007. Сер.2, № 1, с. 23 - 35.
7. Баталина А.М., Епифанов М.Е., Кобзарева Т.Ю., Кушнарёва Е.В., Лахути Д.Г. Опыт экспериментальной реализации алгоритмов поверхностно-синтаксического анализа // Компьютерная лингвистика и интеллектуальные технологии: Труды Международной конференции Диалог'2006 (Бекасово, 31 мая - 4 июня 2006 г.). М.: Наука. С. 51-56.
8. Апресян Ю.Д., Богуславский И.М., Иомдин Д.Л., Лазурский А.В., Перцов Н.В., Санников В.З., Цинман Л.Л. Лингвистическое обеспечение системы Этап-2 // М.: Наука, 1989.
9. Иорданская Л.Н. Автоматический синтаксический анализ. 1961. Том 2. Межсегментный синтаксический анализ // Новосибирск: Изд. «Наука», 1967.
10. Мельчук И.А. Автоматический синтаксический анализ. 1964. Том 1. Общие принципы. Внутрисегментный синтаксический анализ // Новосибирск: 1964.
11. Агранат Т.Б., Кулагина О.С. Об электронном словаре сочетаемости сложносочиненных и сложноподчиненных предложений // Труды Международного семинара Диалог'2001, Аксаково 2001. Т.2, с.13-15.
12. Кулагина О.С. Об одном подходе к установлению отношений между простыми предложениями в составе сложного при автоматическом анализе текстов // Математические вопросы кибернетики. 2001. №10.
13. Севбо И.П. О громоздкости синтаксических структур. // НТИ. 1971. Сер.2., №2.
14. Ножов И.М. Процессор синтаксической сегментации русского предложения // НТИ. 2003. Сер. 2, № 11, с. 26-37.