Междисциплинарные науки

УДК 81.33

РАССТАНОВКА ЗНАКОВ ПРЕПИНАНИЯ МЕЖДУ ОДНОРОДНЫМИ ЧЛЕНАМИ В ИНТЕРНЕТ-ЗАПРОСАХ[3]

И. Р. Васильева. Ульяновский государственный университет (Россия),
e-mail: i_vasy@inbox.ru

Резюме. Используя языковые анализаторы текста aot.ru, в статье предлагается алгоритм расстановки недостающих запятых между однородными членами предложения в Интернет запросах, выполненных на русском языке.

Ключевые слова: информационный запрос, синтаксический анализатор, дерево, семантический граф, однородные члены предложения.

В данной статье предложен алгоритм расстановки знаков препинания между однородными определениями, сказуемыми и подлежащими в Интернет запросах, выполненных на русском языке.

В предложенной работе автор опирается на достижения группы разработчиков Рабочая группа Aot.ru.[1] Предложенные ими алгоритмы и правила анализа предложения позволяют строить адекватные семантические графы для многих предложений русского языка. Более подробно см обзор.[2]

Одновременно с наличием множества программ анализа художественного текста, источник[3] утверждает, что « в интернете нет такого ресурса, которым выполняется проверка пунктуации». Автором данной статьи было введено предложение «Белый снег пушистый падает кружится и на землю тихо падает ложится» в некоторые программы, производящие синтаксический анализ. В результате, были обнаружены ошибки со словом «кружится» (программа Орфограммка[4], редактор WORD). Они предложили поставить «ь» после «т» в слове кружится, но ни одна из программ не увидела однородных сказуемых. Что же говорить об информационных Интернет запросах, которые представляют собой фразы естественного языка (ЕЯ), построенные с нарушениями ЕЯ грамматики, и, в частности, пунктуации. Машинный лингвистический анализ таких фраз является новым и актуальным в компьютерной лингвистике.

Алгоритм.

Перед началом работы алгоритма прогоняем каждый Интернет-запрос через морфологический, синтаксический и семантический анализаторы программы Aot.ru[1]. В результате мы получаем или текст с разметкой или семантические графы, соответствующие данной фразе. Заметим, что графов может быть несколько из-за грамматических ошибок в Интернет запросах.. В дальнейшем, фразу семантический граф заменим одним словом – дерево.

[3] *Рецензент: Крашенинникова Наталья Алексеевна, канд. т. наук, доцент кафедры английского языка для профессиональной деятельности УлГУ (Ульяновск, Россия).*

В целом, алгоритм базируется на методе прохода по глубине по дереву. Сначала выполняется проход по все узлам, имеющим глубину X, и только потом выполняется переход к узлам с глубиной X + 1[5]. Чтобы обеспечить условие выхода и универсальность данного алгоритма автор предлагает любое дерево, построенное по ЕЯ запросу, начинать с фиктивной корневой вершины с любой символьной пометкой. Такой ход позволяет избежать ввода дополнительных условий проверки и выход из возможного цикла.

В результате прохода по дереву мы применяем правила – форматки[1] для определения однородных членов предложений к новой вершине и уже рассмотренным вершинам этого уровня. В случае, если по правило срабатывает, то в параметре новой вершины как в тексте, так и в дереве появится разметка ОДНР ПРИЛ, в противном случае такой разметки не будет..

После окончания прохода по дереву у вершин появятся разметки ОДНР ПРИЛ, ОДНОР_ИНФ, ОДНОР_НАР, ОДНОР_ИГ, по которым на выходе алгоритма будут ставятся запятые между однородными членами предложения.

Разметки ОДНР ПРИЛ, ОДНОР_ИНФ, ОДНОР_НАР, ОДНОР_ИГ будут появляться два раза. Первый раз перед началом прохода по дереву, а второй раз – при проверке форматки для вершин одного уровня, своего рода верификация. Это связано с грамматическими ошибками в Интернет-запросах, что приводит к появлению нескольких семантических графов для одного и того же запроса. Так же возможно, что оба графа будут вполне адекватными, но с разной расстановкой запятых. Это связано с некоторой языковой двусмысленностью. Например, фраза «Белый снег искрящийся иней» может соответствовать двум фразам: «Белый снег, искрящийся иней» и «Белый снег искрящийся, иней».

Математически возможность реализации данного алгоритма на ЕЯ запросах автор доказывает индукционно. Предполагается, что на входе либо текст с синтаксической и морфологической разметкой, либо дерево.

УТВЕРЖДЕНИЕ. В любом ЕЯ-запросе можно правильно расставить запятые между однородными членами предложения.

ДОКАЗАТЕЛЬСТВО.

Доказательство проводим индукцией по числу вершин, зависящих от рассматриваемой корневой.

База индукции:

Количество вершин, зависящих от корневой N = 1.

Естественно, такому дереву соответствует фраза без однородных членов, и расставлять запятые не нужно.

Индукционное предположение:

Количество вершин, зависящих от корневой N>1.

Считаем, что для такого дерева и соответствующего запрос проведена успешная развертка и расставлены все знаки препинания между однородными членами.

Шаг индукции:

Количество вершин, относящихся к корневой, равно N+1 вершины. Причем для первых N вершин по индукционному предположению развертка проведена успешно и все знаки препинания уже расставлены. А для (N+1)-й развертка не была проведена.

Смотрим на синтаксические характеристики (N+1)-й вершины. Если такие вершины есть, то мы имеем однородные члены предложения, нам нужно ставить знак препинания перед этой вершиной. В вектор вершины мы пишем «**,**».

Если нет вершин с одинаковыми синтаксическими и морфологическими разметками, то нет и однородных членов для ра**сс**матриваемой вершины. Соответственно, никаких знаков препинания ставить не нужно и поле «знак препинания» в векторе вершины оставляем пустым.

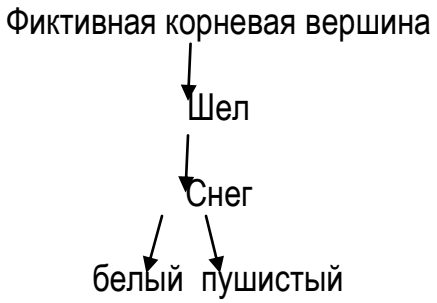
Затем переходим к следующей вершине.

ДОКАЗАТЕЛЬСТВО закончено.

Замечание. Процесс, описанный в доказательстве, конечен, так как число возможных уровней не может превышать количества слов в ЕЯ запросе.

Пример 1: Шел белый пушистый снег.

Дерево после синтаксического анализатора.



1 шаг. Корневая вершина – фиктивная. От нее зависит слово Шел. Это первая вершина из зависящих, до нее не было вершин, поэтому нам не с чем сравнивать синтаксические параметры. Переходим к следующей зависящей вершине этого уровня. Таких нет, поэтому переходим на уровень ниже.

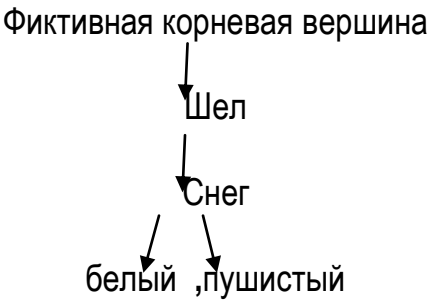
2 шаг. Корневая вершина – Шел. От нее зависит одно слово Снег. Это первая вершина из зависящих, до нее не было вершин, поэтому нам не с чем сравнивать синтаксические параметры. Переходим к следующей зависящей вершине этого уровня. Таких нет, поэтому переходим на уровень ниже.

3 шаг. Корневая вершина – Снег. От нее зависит слово Белый. Это первая вершина из зависящих, до нее не было вершин, поэтому нам не с чем сравнивать синтаксические параметры. Переходим к следующей зависящей вершине этого уровня. Это слово - Пушистый. До него есть слово Белый с теми же синтаксическими функциями, поэтому они однородные, и в вектор характеристик слова Пушистый мы вносим знак «,». Больше вершин зависящих от слова Снег нет. Идем к следующему шагу.

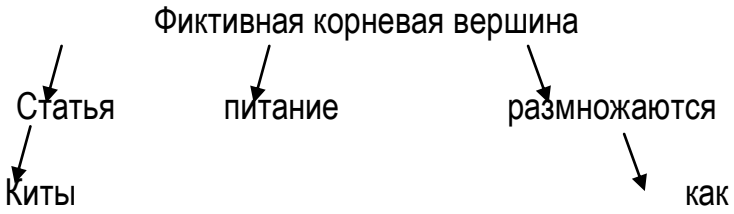
4 шаг. Берем первую вершину из рассмотренных зависящих вершин. Это вершина Белый. Она станет корневой. Это лист, поэтому мы берем следующую зависящую вершину Пушистый. Это тоже лист. Больше на этом уровне вершин нет. Мы поднимаемся на слово Снег. На этом уровне тоже нет других вершин. Мы откатываемся назад на слово Шел. На этом уровне тоже нет других вершин. Мы откатываемся назад на фиктивную вершину – корневую вершину всего дерева.

Алгоритм завершен.

В результате наше дерево имеет следующий вид:



Пример 2: Статья про китов питание как размножаются
Дерево после синтаксического анализатора.



1 шаг. Корневой вершиной будет фиктивная корневая вершина. Первая зависящая вершина – Статья. Так как она – первая, то переходим ко второй зависящей вершине – Питание. У нее те же синтаксические характеристики, как у слова Статья, поэтому в ее параметры вносим знак «,». Переходим к следующей зависящей вершине Размножаются. У нее те же синтаксические характеристики, как у Статья, поэтому в ее параметры тоже вносим знак «,». Переходим к следующей зависящей вершине. Таких больше нет.

2 шаг. Из рассмотренных выше зависящих вершин берем первую вершину– Статья. Она будет корневой. От нее зависит слово Киты. У этого слова нет однородных, так как оно первое, зависящее от слова статья. Переходим к следующей зависящей вершине этого уровня. Таких больше нет. Спускаемся вниз.

3 шаг. Из рассмотренных выше зависящих вершин берем вершину– Киты. Она будет корневой. Это лист. Поэтому откатываемся на слово Статья.

4 шаг. Берем следующую вершину, зависящую от слова Статья. Таких больше нет. Откатываемся назад к фиктивной корневой вершине.

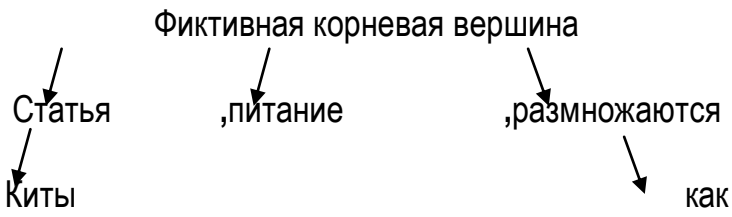
5 шаг. Берем следующую корневую вершину – Питание. Это лист, поэтому рекурсивно откатываемся назад к фиктивной корневой вершине.

6 шаг. Берем следующую корневую вершину – Размножаются. От нее зависит слово Как. У этого слова нет однородных, так как оно первое, зависящее от слова Размножаются. Переходим к следующей зависящей вершине этого уровня. Таких больше нет. Поэтому откатываемся назад к фиктивной корневой вершине.

7 шаг. Переходим к следующей зависящей вершине. Таких больше нет. А так как корневая вершина – фиктивная, алгоритм останавливается.

Алгоритм завершен.

В результате наше дерево имеет следующий вид:



Проходим по ЕЯ запросу и одновременно смотрим метки соответствующих вершин. Имеем запрос:

Статья про китов, питание, как размножаются

Литература:

1. http://www.aot.ru/
2. http://zlat.spb.ru/CatalogImages/File/pdf/comp_progr.pdf
3. http://www.internet-technologies.ru/articles/article_1629.html
4. http://orfogrammka.ru
5. http://study-and-dev.com/blog/sda_theory_trees/

Vasil'eva I.R. Rasstanovka znakov prepinanija mezhdu odnorodnymi chlenami v Internet zaprosah / I.R. Vasil'eva // Nauka. Mysl'. - № 3. – 2015.

© И.Р. Васильева, 2015.
© «Наука. Мысль», 2015.

– ● –

Abstract. Using language text analyzers aot.ru, in the article the algorithm of placing missing commas between homogeneous members of the proposals to the online requests made in Russian is suggested.

Keywords: information request, syntactic analyzer, tree, semantic graph, homogeneous members of sentence.

– ● –

Сведения об авторе

Ирина Романовна **Васильева**, к. ф.-м. н., доцент кафедры Английского языка для профессиональной деятельности Ульяновского государственного университета (Ульяновск, Россия).

– ● –

Подписано в печать 02.12.2015.
© Наука. Мысль, 2015.