АВТОМАТИЧЕСКОЕ РАЗБИЕНИЕ ТЕКСТА НА ПРЕДЛОЖЕНИЯ ДЛЯ РУССКОГО ЯЗЫКА

DETECTING SENTENCE BOUNDARIES IN RUSSIAN

Урюпина О. *(uryupina@gmail.com)*
Институт языкознания РАН,
«Ашманов и Партнеры»

В данной работе предлагается статистический алгоритм сегментации текста на предложения на материале русского языка. Алгоритм основан на контекстах знаков препинания и не требует предварительного синтаксического анализа. Сравнение с эвристическими методами показывает, что статистический подход позволяет существенно улучшить качество сегментации.

1. Введение

Большинство систем автоматической обработки языка ставят своей задачей анализ текстов, заранее разбитых на предложения. Например, парсеры определяют синтаксическую структуру предложения, системы автоматического реферирования выделяют из документа наиболее значимые предложения и так далее. В то же время, языковые данные доступны нам чаще всего в виде текстов, размеченных на абзацы, главы и другие более крупные единицы. Поэтому для их эффективного автоматического анализа необходимы соответствующие алгоритмы сегментации. К сожалению, большинство систем используют упрощенные эвристические методы разбиения на предложения. Нам не удалось найти подробных теоретических исследований данной задачи или описаний работающих алгоритмов для русского языка, за исключением очень кратких (см., например, [6]). Задача разбиения текста на предложения для английского языка описывается, в частности, в [7] и [8].

Рассмотрим простой пример, показывающий, как неправильное разбиение на предложения может породить целый ряд проблем на разных уровнях анализа документа:

(1) Но ведь Маша знает А. Б. Иванова много лет и никогда про него ничего плохого не слышала!!

Если мы ошибочно выделим здесь четыре предложения («Но ведь Маша знает А.», «Б.», «Иванова много лет и никогда про него ничего плохого не слышала!» и «!»), то дальнейшая автоматическая обработка окажется бессмысленной. Так, парсеры либо не смогут разобрать предпоследнее предложение, либо сделают это неверно. Интерпретация местоимения «него» окажется затруднительной – скорее всего, будет принято решение, что «него» кореферентно с «Б». Попытки использовать этот фрагмент в экспертной системе приведут к тому, что на вопрос «Что знает Маша?» будет выдан ответ «А». Наконец, модуль автоматического реферирования может включить одно из четырех ошибочно выделенных предложений в аннотацию документа, что приведет к потере качества.

В данной работе обсуждается алгоритм разбиения текста на предложения. Рассматриваются две связанные задачи:

1) определение, является ли терминальный знак препинания (здесь и далее под терминальными знаками мы будем понимать точку, восклицательный и вопросительный знаки) границей предложения в данном контексте,

2) определение всех границ предложений в документе.

Мы предлагаем статистический алгоритм, который можно легко адаптировать для анализа текстов разных типов (например, книг vs. веб-страниц). Важной особенностью нашего алгоритма является то, что он не опирается на синтаксический анализ. Это дает нам, во-первых, существенный выигрыш в скорости и, во-вторых, возможность анализировать практически любые тексты, независимо от их синтаксической грамотности.

В разделе 2 мы приводим примеры из Национального Корпуса Русского Языка [1], иллюстрирующие сложность задачи. В разделе 3 описываются лингвистические признаки, использованные для обучения и распознавания. В разделе 4 обсуждаются результаты экспериментов: мы сравниваем эффективность нашего алгоритма и нескольких интуитивных стратегий выделения предложений (например, «если после терминального знака идет слово с большой буквы, то это новое предложение»).

2. Примеры

В самом первом приближении можно считать, что предложение всегда начинается со слова с большой буквы и заканчивается терминальным знаком препинания. Однако и теоретические исследования (см., например, [2]), и корпусные данные говорят о том, что для наиболее точного и полного выделения предложений необходимо учитывать целый ряд дополнительных факторов.

Довольно часто точка является разделителем не предложения, а других единиц. Например, точка используется в URL веб-страниц (2, здесь и далее все примеры взяты из Национального Корпуса Русского Языка) и для обозначения даты или времени (3):

(2) Конечно, в рамках газетной статьи невозможно сделать обзор сотни докладов, поэтому мы рекомендуем посетить Интернет-страницу конференции http://www.ict.nsc.ru/ws/mol2000/, на которой размещены программа мероприятия и тезисы докладов.

(3) В 11.45 дали слово Кудрину , но он всё не шёл.

Точка используется как знак сокращения:

(4) выполнение 12 тепловозам усиленного ТР-1 с применением средств диагностики вместо ТР-2 дало экономический эффект 221,8 тыс. руб.

Стоит обратить внимание, что после сокращения в конце предложения ставится одна точка, а не две («руб.», а не «руб..»). Это затрудняет анализ: даже если мы знаем, что перед точкой сокращение, мы не можем с уверенностью сказать, что перед нами середина предложения.

Точка может быть элементом форматирования:

(5) Параметры системы питания:
линейное напряжение на проводах 81 ... 83 , В 220
фазное напряжение на проводах 81 ... 83 по отношению
к проводу 84 (нулю), В 127
частота переменного тока, Гц 50
выпрямленное напряжение на проводах, В
15—30 110
44—30 50

Нередко точка в середине предложения просто опечатка:

(6) Режиссёр Михаил. Бычков поставил в Таллине притчу о любви к невозможному и о презрении к реальности

Вопросительный и восклицательный знаки также могут употребляться в середине предложения. Чаще всего ими завершаются фрагменты в скобках (7) или кавычках (8):

(7) Дело в том, что по всяким планам «пятилеток» и заданиям ЦК советский военный комплекс создавал ядерное оружие с запасом на пять и более(!) ядерных войн.

(8) Пролетели «Тише!» Виктора Косаковского и «Фрески» Александра Гутмана.

Часто предложения заканчиваются многоточием (9) или комбинацией вопросительных и/или восклицательных знаков (10). Графически это несколько терминальных знаков, но только последний из них является концом предложения:

(9) Не в лесе и не в медицине дело...

(10) Только вот ради чего ?!

Многоточие может использоваться для передачи паузы или пропуска части текста. Две или три точки могут обозначать интервал между числами. Ни одна из точек не является концом предложения в подобных случаях:

(11) В связи с этим первый интервал пробегов был принят равным 350...700 тыс. км (середина интервала - 525 тыс. км), второй интервал — 700...1050 тыс. км (середина интервала - 875 тыс. км) и третий интервал 1050...1400 тыс. км (середина интервала — 1225 тыс. км).

Определенную проблему представляют скобки и кавычки:

(12) В случае нарушений выносится письменное предупреждение: «В вашей деятельности допускаются вот такие недочёты ...»

В подобных случаях система автоматического разбиения на предложения должна понять, что границей является не терминальный символ, а следующие за ним скобка или кавычки. К сожалению, во многих документах не соблюдаются правила постановки знаков препинания в предложениях со скобками и кавычками, что затрудняет анализ. Далеко не во всех текстах на русском языке используются кавычки-елочки. Анализировать же кавычки-лапки значительно сложнее.

Дополнительные проблемы возникают при работе с веб-данными. Так, первое слово в предложении может начинаться с маленькой буквы. Точка может отсутствовать, особенно в конце абзаца. Вопросительный знак может появиться из-за проблем с кодировкой (например, в него могут превратиться кавычки-елочки). Наконец, пробел после терминального знака ставится далеко не регулярно.

Как показывают все эти примеры, сегментация текста на предложения – это довольно сложная задача, требующая учета самых разных факторов. Ниже мы предлагаем статистический алгоритм, основанный на контекстах терминальных знаков препинания.

3. Методология

Наш алгоритм работает следующим образом: сначала из документа извлекаются *контексты* потенциальных границ предложений, потом они описываются в виде *векторов признаков*, которыми оперируют системы автоматического обучения. В нашем первом эксперименте потенциальными границами предложений являются терминальные знаки, во втором – вообще вся пунктуация. Мы не рассматриваем предложения, не заканчивающиеся никаким знаком препинания.[1]

В контекст границы мы включаем четыре элемента: сам знак препинания (punct), ближайшее псевдослово слева (left), ближайшее псевдослово справа (right) и ближайшее собственно слово справа (wright). Под псевдословом здесь и далее понимается любая последовательность символов, не включающая пробел или конец абзаца. Под словом – псевдослово, содержащее хотя бы одну букву или цифру. Также запоминается количество псевдослов слева (dleft) и справа (dright) до ближайшей потенциальной границы или конца абзаца. В Таблице 1 приводятся контексты для всех потенциальных границ в примере (1).

punct	.	.	!	!
left	А	Б	слышала	!
right	Б	Иванова	!	-отсутствует-
wright	Б	Иванова	-отсутствует-	-отсутствует-
dleft	4	1	11	0
dright	1	11	0	0

Таблица 1. Контексты для всех потенциальных границ примера (1) в предположении, что в абзаце нет других предложений.

Составленные таким образом контексты являются основой для векторов признаков. Как показывают примеры из раздела 2, для эффективного разбиения на предложения необходимо учитывать целый ряд факторов: сокращения, вид псевдослов (пунктуация, цифры и т.д.). Эта информация была добавлена в виде дополнительных признаков. В Таблице 2 приведены полные вектора признаков для (1).

Прежде всего, мы составили словарь сокращений. Для этого из размеченной части Национального Корпуса Русского Языка (около 6 миллионов слов) были извлечены все триграммы вида «псевдослово точка слово_со_строчной_буквы». Сокращения, встретившиеся только один раз, были отброшены. Данный словарь был составлен полностью автоматически и не подвергался никакой пост-обработке. Отметим, что ручная разметка корпусных данных никак не использовалась. Таким образом, для составления подобного словаря нужна только большая коллекция текстов. Каждый контекст проверяется на сокращения: если псевдослова left или right нашлись в словаре, то контекст получает дополнительные признаки abbleft и abbright соответственно.

Мы также провели классификацию псевдослов. В отдельные группы были выделены пунктуация и числа. Остальные псевдослова были разбиты на классы в зависимости от используемых символов (кириллица, латиница, кириллица+латиница, кириллица+цифры и так далее) и регистра первого символа (строчная буква, прописная буква, цифра, пунктуация). Класс псевдослов описывается признаками cleft и cright.

Ближайшее собственно слово справа, wright, описывается одним дополнительным признаком – регистр первого символа (cwright).

Наконец, мы ввели дополнительные признаки isfirst и islast для обозначения контекстов, приходящихся на самое начало или конец абзаца.

[1] В наших данных не встретилось предложений, не заканчивающихся знаками препинания, но при этом не приходящихся на конец абзаца или заголовка. Тем не менее, наш алгоритм (после необходимого переобучения и тестирования) применим и для распознавания подобных случаев.

punct	.	.	!	!
left	А	Б	слышала	!
right	Б	Иванова	!	-отсутствует-
wright	Б	Иванова	-отсутствует-	-отсутствует-
dleft	4	1	11	0
dright	1	11	0	0
abbleft	1	1	0	0
abbright	1	0	0	0
cleft	-кириллица--прописная-	-кириллица--прописная-	-кириллица--строчная-	
cright	-кириллица--прописная-	-кириллица--прописная-	-пунктуация-	-отсутствует-
cwright	-прописная-	-прописная-	-отсутствует-	-отсутствует-
isfirst	0	0	0	0
islast	0	0	0	1

Таблица 2. Вектора признаков для всех потенциальных границ примера (1) в предположении, что в абзаце нет других предложений

Вектора признаков используются для обучения и распознавания, как описывается в следующем разделе.

4. Эксперименты

Ниже описываются языковые данные и системы машинного обучения, использованные для проверки эффективности нашего метода.

Мы проверили вручную 33 документа из Национального Корпуса Русского Языка и исправили все неточности в определении границ предложений. Документы включали в себя газетные и журнальные статьи общего (политика, культура и так далее) и технического (ремонт локомотивов) содержания. Было получено 1639 предложений, 1414 из них заканчивались терминальным знаком. Из этих предложений было выделено 5230 контекстов: 2048 с терминальным знаком и 3182 с другой пунктуацией. Контексты были преобразованы в вектора признаков в соответствии с описанием, изложенным выше. Мы отобрали случайным образом 1000 векторов для тестирования, остальные были зарезервированы для обучения. Были протестировали три алгоритма машинного обучения: C4.5 [3], Ripper [4] и SVM-light [5].

Для оценки эффективности нашего алгоритма мы разработали несколько упрощенных эвристических подходов к определению границ предложений. Прежде всего, конец абзаца всегда считается концом предложения. Наша первая эвристика, termpunct, классифицирует каждый терминальный знак как конец предложения. При анализе примера (1) такой метод приведет к выделению четырех предложений.

Вторая эвристика, termpunct_cap, аналогична первой, но запрещает предложения, не начинающиеся с заглавной буквы. При анализе примера (1) такой метод приведет к выделению трех предложений: «Но ведь Маша знает А.», «Б.», «Иванова много лет и никогда про него ничего плохого не слышала!!».

Наконец, третья эвристика, advanced, дополнительно запрещает предложения, заканчивающиеся сокращением и точкой. При анализе примера (1) такой метод приведет к выделению одного предложения. Именно такой эвристикой или ее незначительными модификациями, насколько нам известно, руководствуются большинство систем автоматического анализа текста для подготовки исходных данных.

В таблице 3 приводятся результаты наших экспериментов. В первом эксперименте рассматриваются только контексты, содержащие терминальные знаки препинания. Мы выделили их в отдельную группу, поскольку большинство предложений заканчиваются именно одним из терминальных знаков. Кроме того, многие алгоритмы, описанные в литературе (см., например, [7]), не ставят своей целью анализ других знаков препинания. Во втором эксперименте рассматриваются все контексты.

Как показывают результаты, с помощью простейшей эвристики termpunct невозможно добиться удовлетворительной точности распознавания: каждая третья поставленная граница будет ошибочной. Остальные эвристики дают более приемлемое качество. Тем не менее, потеря либо полноты (при учете сокращений), либо точности (без их учета) составляет около 10%.

Статистический подход, основанный на контекстных векторах, позволяет существенно повысить качество. При использовании любого из трех протестированных модулей машинного обучения полнота и точность дости-

гают 96-99%. Тест χ^2 показывает, что рост полноты статистически значим. Во втором эксперименте также статистически значимо и увеличение точности для всех программ машинного обучения. В первом эксперименте статистически значимый рост точности удалось достичь только с помощью SVM, но, по крайней мере, ни для C4.5, ни для Ripper не засвидетельствовано падения.

	Эксперимент 1		Эксперимент 2	
	точность, %	полнота, %	точность, %	полнота, %
termpunct	67.2	100**	66.9	98.9**
termpunct_cap	90.7	97.0**	89.6	96.0**
advanced	96.4	90.4	95.0	89.6
C4.5	97.8	98.5**	98.5*	97.5**
Ripper	98.5	98.5**	98.9**	96.0**
SVM-light	99.6**	98.5**	99.6**	97.5**

Таблица 3. Качество разбиения текста на предложения: результаты систем машинного обучения и контрольных эвристик. Показатели, значительно превосходящие полноту и точность эвристики `advanced`*, отмечены * (χ^2, $p<0.05$) и ** (χ^2, $p<0.01$)*

В таблице 4 приведены два примера работы нашего алгоритма (классификатор Ripper) в сравнении с эвристикой advanced. Как видно, статистический подход дает меньше ошибок. Особенно заметна разница при анализе узкоспециальных документов (нижняя половина таблицы). Слова «1» и «2» справедливо попали в список сокращений, в результате чего были пропущены две границы (с точки зрения эвристики advanced, вторая точка в «Рис. 1.» не отличима от первой). Этот пример демонстрирует главное преимущество статистического подхода к задаче сегментации на предложения по сравнению с эвристическими методами: для анализа узкоспециальных текстов достаточно просто добавить несколько соответствующих документов в обучающую выборку.

В то же время, полностью автоматический подход имеет и свом недостатки. Как показывает пример в верхней половине таблицы 4, наш автоматически составленный список сокращений (см. раздел 3) содержит достаточно много мусора. Например, слово «есть» было сочтено сокращением. Необходима лингвистическая экспертиза для ручной пост-обработки нашего списка.

Наконец, отметим, что и статистический алгоритм допускает ошибки. Наиболее проблематичными оказались сочетания точки и кавычек-лапок: несмотря на то, что часть таких контекстов проанализирована правильно, некоторые границы оказались пропущены, как в примере в верхней половине таблицы 4. В данный момент мы проводим дополнительные эксперименты по улучшению качества сегментации текстов с кавычками-лапками.

5. Заключение

В данной работе был предложен статистический подход к задаче определения границ предложений в произвольном тексте на русском языке. Наш алгоритм основан на контекстах знаков препинания и не требует синтаксического анализа, что позволяет обрабатывать документы с высокой скоростью.

Наши эксперименты показывают, что статистический подход позволяет добиться существенно более точного и полного выделения границ, чем наиболее распространенные эвристики.

Данная работа была проведена на материале газетных статей, отобранных случайным образом из Национального Корпуса Русского Языка. В будущем мы планируем изучить применимость нашего метода для обработки менее качественных текстов, прежде всего, веб-страниц. Наши предварительные исследования выявили целый ряд дополнительных задач, возникающих при анализе веб-документов.

Список литературы

1 http://ruscorpora.ru

2 Ровинская М. Точка как Проблема. Материалы Международной Конференции Диалог. 2000.

3 Quinlan J.R. C4.5: Programs for Machine Learning. Morgan Kaufman. 1993.

4 Cohen W.W. Fast Effective Rule Induction. Proceedings of ICML. 1995.

5 Burges C.J. A Tutorial on Support Vector Machines for Pattern Recognition. Data Mining and Knowledge Discovery, 2(2). 1998.

6 http://aot.ru

7 Reynar J.C. and Ratnaparkhi A. A Maximum Entropy Approach to Identifying Sentence Boundaries. Proceedings of ANLP, pp. 16-19. 1997.

8 Stevenson M. and Gaizauskas R. Experiments on Sentence Boundary Detection. Proceedings of ANLP-NAACL. 2000.

advanced	**статистический алгоритм (Ripper)**
Статья общего содержания (культура)	
<s> Был на церемонии момент , когда прозвучала пронзительно высокая и чистая нота .___" Ника " за " Честь и Достоинство "-- вот так , всё с заглавной буквы -- вручалась Петру Ефимовичу Тодоровскому .</s> *<s> Петру Тодоровскому -- оператору и режиссёру , композитору и музыканту , солдату и просто замечательному человеку .</s>* *<s> Он молодой , ошалевший от победной весны 45-го , смотрел на нас с экрана в хуциевском фильме " Был месяц май ".</s>* *<s> Он вышел на сцену под гром аплодисментов и " Рио-риту ".</s>* *<s> Для своих ровесников и друзей так и оставшийся в его - то годы Петей Тодоровским .</s>* *<s> Он прошёл через зал ," по главной улице с оркестром ", держа в руках гитару .</s>* *<s> Спасибо вам , дорогой Петр Ефимович !</s>* *<s> За веру , верность и " Верность ", за всё ваше кино , за то , что вы сделали для нас , за вашу нескончаемую любовь , за то , что вы есть .___За то , что " и всё-таки , и всё-таки , и всё-таки мы победили "!</s>* *<s> Той весной .</s>* *<s> За то , что у нас есть эта весна .</s>* *<s> И это ее семнадцатое мгновение .</s>*	*<s> Был на церемонии момент , когда прозвучала пронзительно высокая и чистая нота .___" Ника " за " Честь и Достоинство "-- вот так , всё с заглавной буквы -- вручалась Петру Ефимовичу Тодоровскому .</s>* *<s> Петру Тодоровскому -- оператору и режиссёру , композитору и музыканту , солдату и просто замечательному человеку .</s>* *<s> Он молодой , ошалевший от победной весны 45-го , смотрел на нас с экрана в хуциевском фильме " Был месяц май ".</s>* *<s> Он вышел на сцену под гром аплодисментов и " Рио-риту ".</s>* *<s> Для своих ровесников и друзей так и оставшийся в его - то годы Петей Тодоровским .</s>* *<s> Он прошёл через зал ," по главной улице с оркестром ", держа в руках гитару .</s>* *<s> Спасибо вам , дорогой Петр Ефимович !</s>* *<s> За веру , верность и " Верность ", за всё ваше кино , за то , что вы сделали для нас , за вашу нескончаемую любовь , за то , что вы есть .</s>* *<s> За то , что " и всё-таки , и всё-таки , и всё-таки мы победили "!</s>* *<s> Той весной .</s>* *<s> За то , что у нас есть эта весна .</s>* *<s> И это ее семнадцатое мгновение .</s>*
Статья технического содержания	
<s> Чтобы объективно оценивать качество изоляции , используя явление абсорбции , в Нижегородском филиале РГОТУПСа было разработано оригинальное устройство , принципиальная схема которого приведена на рис . 1 .___Оно включает в себя : высоковольтный стабилизированный источник питания ВИП с выходным напряжением 1000 или 2500 В , измерители тока И 1 и напряжения И 2 , два высоковольтных реле Р 1 и Р 2 .___Последними управляет микроЭВМ или система автоматики , построенная на интегральных микросхемах с применением программируемых запоминающих устройств .</s> *<s> Схема замещения неоднородной изоляции тягового двигателя представлена в виде двух конденсаторов С 1 и С 2 , зашунтированных резисторами R1 и R2.</s>*	*<s> Чтобы объективно оценивать качество изоляции , используя явление абсорбции , в Нижегородском филиале РГОТУПСа было разработано оригинальное устройство , принципиальная схема которого приведена на рис . 1 .</s>* *<s> Оно включает в себя : высоковольтный стабилизированный источник питания ВИП с выходным напряжением 1000 или 2500 В , измерители тока И 1 и напряжения И 2 , два высоковольтных реле Р 1 и Р 2 .</s>* *<s> Последними управляет микроЭВМ или система автоматики , построенная на интегральных микросхемах с применением программируемых запоминающих устройств .</s>* *<s> Схема замещения неоднородной изоляции тягового двигателя представлена в виде двух конденсаторов С 1 и С 2 , зашунтированных резисторами R1 и R2.</s>*

Таблица 4. Сегментация на предложения с помощью эвристики advanced (левая колонка) и статистического алгоритма (правая колонка). Границы предложений обозначены тэгом <s>..</s>. Ошибки подчеркнуты и выделены серым.