Задача автоматической расстановки знаков пунктуации в распознанной спонтанной русской речи

Дмитрий Анатольевич Бирин,
генеральный директор филиала ФГУП «НИИ «Квант» , Санкт-Петербург

Александр Евгеньевич Булашевич,
кандидат технических наук, научный сотрудник филиала ФГУП «НИИ «Квант», Санкт-Петербург

Марианна Юрьевна Грекис,
инженер ФГУП «НИИ «Квант», Санкт-Петербург

Аннотация:

Основная цель процесса распознавания речи — получение на выходе удобочитаемого, ясного текста. В русском языке это практически невозможно без знаков препинания. Проблема в том, что существующая система правил пунктуации была разработана для письменного языка. В спонтанной речи эти правила часто не соблюдаются и даже нарушаются. Кроме того, для спонтанной речи характерны такие явления, которые не описаны в правилах, сформулированных для литературного (письменного) языка, поскольку эти явления там практически отсутствуют (например, хезитационный поиск, самоисправления и т.д.). Таким образом, задача заключается в том, чтобы адаптировать классические правила для спонтанной речи и разработать систему автоматической пунктуации, которая сможет превратить последовательность распознанных слов спонтанной речи в понятный письменный текст. На данном этапе наша система позволяет в большинстве случаев однозначно определять границы предложения и с определённой точностью ставить внутренние знаки препинания.

Ключевые слова: распознавание спонтанной речи, пунктуация в спонтанной речи, автоматическая расстановка знаков препинания.

ВВЕДЕНИЕ

Русский язык, как в устной, так и в письменной форме, представляет собой весьма сложную систему, функционирующую на основе набора неких принятых стандартов, так называемых правил, в соответствии с которыми в письменной речи употребляется та или иная буква (графема) или проставляется тот или иной знак препинания. Эти стандарты — результат труда многих квалифицированных специалистов, и они изложены в многочисленных авторитетных изданиях. Однако в

реальной жизни человек использует выработанные каноны лишь в определённой мере, которые зависят прежде всего от его так называемого социолингвистического статуса, т.е. уровня образования, культурного уровня, окружающего социума и т.п. Не все русские владеют правилами русского языка в совершенстве. «Среднестатистический» гражданин на практике достигает лишь определённого уровня грамотности, то есть использует лишь тот объём знаний, который получен им в процессе обучения, например в объёме средней школы. Это касается владения как устной, так и письменной речью.

Значительную сложность в письменной речи представляет собой система пунктуации. Правила пунктуации весьма сложны, простановка знаков препинания часто неоднозначна и зависит от определённых исходных условий. Совершенно очевидно, что разработка алгоритмов для автоматизации процесса расстановки знаков препинания даже в «нормативном» тексте представляется весьма сложной задачей. Еще более трудной оказывается задача расстановки знаков препинания в спонтанной, т.е. заранее не продуманной, не «запрограммированной» речи, когда фразы составляются «на ходу», выражаясь простым языком, когда слово опережает мысль. В этом случае говорящий не будет выстраивать правильные синтаксические конструкции, подбирать соответствующую лексику и т.п. Даже без какого-либо научного анализа понятно, что в этом случае языковые нормы «перестают работать» — разрушаются синтаксические связи, предложения часто неполные, слова могут не согласоваться друг с другом и т.п.

Расстановка знаков препинания в тексте, представляющем собой задокументированную спонтанную речь, является непростой задачей даже для лингвиста, что позволяет представить всю сложность попытки автоматизации решения данной задачи.

1. ПУНКТУАЦИЯ В СПОНТАННОЙ РЕЧИ: ПРОБЛЕМАТИКА

Очевидно, что системы пунктуации письменной речи недостаточно для транскрибирования устной речи. Например, невозможно средствами пунктуации передать выделенность слова, невозможно различить фразы «***Петр*** *приехал?*» и «*Петр* ***приехал****?*», выражающие различные по смыслу вопросы.

Возникающие здесь трудности объясняются существенным различием между двумя системами национального языка — разговорной речью и кодифицированным литературным языком, на который преимущественно рассчитаны действующие правила пунктуации [1].

Многие явления, которые характерны для спонтанной речи, в письменной речи практически не встречаются. Например, хезитация, самоперебивы, повторы, неоформленность синтаксических и семантических связей в предложении (исключение — фрагменты художественной литературы, передающие устную речь персонажей). Проблема в том, что правила пунктуации, рассчитанные на литературный язык, эти явления никак не описывают.

Трудности возникают уже на этапе определения границ предложения.

Очень просто выделить предложение в письменном тексте. Знаки препинания — точка, вопросительный, восклицательный знаки, многоточие — почти однозначно отмечают конец предложения. И дальше можно определять его тип: сложное или про-

стое, выявлять связи между членами предложения и т.д. Казалось бы, те же процедуры легко проделает любой человек, а тем более лингвист, и со спонтанной звучащей речью. Однако, как показывает практика и как пишут исследователи, занимающиеся спонтанной (живой) речью, выделить в ней предложение — задача весьма сложная, а может быть, даже невыполнимая [2].

На филологическом факультете СПбГУ был проведен перцептивный эксперимент по членению спонтанных монологов на предложения. Предполагалось, что наличие звука позволит экспертам-аудиторам точно и единодушно определить границы предложений. Но это не так. Мнения экспертов во многом разошлись. Как описано в работе А.И. Рыко — С.Б. Степановой [2], аудиторы реализуют различные стратегии: «максималистскую» и «минималистскую». При реализации «максималистской» стратегии в качестве предложений выделяются длинные, многосинтагменные структуры. В этом варианте членения текста в большом количестве представлены бессоюзные предложения, сложносочиненные и сложноподчиненные предложения. «Минималистская» стратегия позволяет провести границу везде, где только это можно сделать: два и более простых предложений вместо одного бессоюзного сложного и т.п. Фактически любая завершающая пауза или затянувшаяся межсинтагменная пауза в рамках такой стратегии — повод для проведения границы между предложениями. Кроме этого «минималистская» стратегия предполагает ориентацию на интонацию текста. Но самое интересное, что в зависимости от темпа речи, качества звучания, ораторского мастерства диктора аудиторы не последовательны и придерживается то одной, то другой стратегии.

Суть в том, что единственным надежным пограничным сигналом, разбивающим речевой поток на отрезки с ограниченным лексическим составом являются паузы разной природы (вдохи и вздохи, молчание, хезитация, ларингализация). Образующиеся при этом интервалы непрерывного «говорения» существенно различаются по длительности и структуре [3].

В частности, в интервале между двумя паузами может быть произнесено только одно слово из длинной фразы, а несколько слов, разделенных запятыми, — без каких-либо перерывов.

В то же время отрезок речевого акустического сигнала между двумя последовательными паузами может содержать несколько синтагм в их общепринятом значении, причем на их стыках не обнаруживается никаких разрывов в мелодическом контуре, которые предположительно могли бы маркировать границы между синтагмами и одновременно являлись бы маркерами границ между словами. Более того, на стыке двух синтагм может происходить стяжение гласных с полной утратой какой-либо физической границы между ними.

Таким образом, оказывается, что анализ мелодической составляющей просодических признаков (аудиторский или инструментальный) не обеспечивает адекватного синтагматического членения. Поэтому А.В. Венцов приходит к выводу, что в акустическом сигнале, представляющем

спонтанную речь, не существует никаких физических признаков границ между лексическими единицами, кроме пауз [4].

Большинство исследователей сегментируют речь на элементарные дискурсивные единицы (ЭДЕ) [5], соответствующие клаузе (элементарному предложению) или синтагме. Такое членение удобно для исследования речи и процессов речепорождения, но не совсем подходит для сохранения смысловой целостности распознанного текста.

2. ОПИСАНИЕ РАЗРАБАТЫВАЕМОЙ СИСТЕМЫ ПУНКТУАЦИИ

В нашей лобаратории ведётся разработка системы пунктуации (СП), которая должна расставить знаки препинания в потоке слов, являющимся результатом автоматического распознавания звучащей речи. Заметим, задача автоматической расстановки знаков препинания осмысленна только в связке с автоматическим распознаванием речи, так как в противном случае текст уже появляется со знаками препинания. При этом точность расстановки знаков препинания напрямую зависит от точности результатов распознавания.

Задача автоматической пунктуации естественно распадается на три подзадачи: определять границы предложения, выбирать завершающий знак препинания и ставить внутренние знаки препинания. При этом внутренние знаки можно ставить в соответствии с правилами русского языка чисто по тексту, но определение границ предложений и классификация завершающего знака чисто по тексту невозможно (собственно, именно вследствие этой невозможности и возникли различные завершающие знаки).

Соответственно, работа была разделена на несколько направлений:
- Создание речевых корпусов;
- Акустическое направление: детектор границы фраз и классификатор завершающего знака;
- Лингвистическое направление: формализация и программная реализация грамматических норм русского языка с целью расстановки внутренних знаков препинания;
- Разработка методики оценки качества пунктуации.

2.1. Создание речевых корпусов

В рамках этого направления был подготовлен речевой корпус «Пунктуация». Общий объём корпуса — около 27 часов. Объём эталонной части — 8,5 часа.

На данном этапе не исследовались и не анализировались спонтанные монологи. Нашей задачей была разработка системы пунктуации для спонтанной диалогической речи.

Эталонная база проаннотирована в двух видах:
- по акустике — знаки проставлены только там, где они выражены диктором акустически. Обычно в таких местах присутствует пауза (различного характера);
- по правилам русского языка — в соответствии с нормами письменной речи со ссылкой на правило, по которому поставлен тот или иной знак.

2.2.1. Детектор границы

Задача детектора границы — определить места постановки завершающих знаков препинания, т.е. границы предложений распознанного текста. Именно текста, а не речи.

Как уже говорилось ранее, среди лингвистов существует точка зрения, что в спонтанной речи вообще отсутствует достаточно четкое членение на фразы. В случае спонтанного диалога в отличие от спонтанного монолога задача детектора границ существенно отличается наличием длинных пауз на месте речи собеседника.

Длительность и структура пауз прежде всего связаны с уровнем «ораторского мастерства» диктора, умением формулировать мысль и чётко ее излагать. Так называемые «публичные» люди или люди, часто говорящие вслух, разделяют свои мысли более короткими паузами ввиду того, что им не требуется длительного обдумывания. Длительные паузы чаще всего говорят о смене диктора или завершении высказывания с ожиданием ответной реакции. Как правило, такие отрезки речи заканчиваются «конечным» знаком препинания или обрывом высказывания (перебив собеседником). Короткие паузы чаще говорят о наличии «внутреннего» знака препинания (запятой, тире, двоеточия и т.п.), но вполне могут быть и «завершающими».

В текущей версии СП используется только один признак — длительность паузы, понимаемой как интервал от конца слова до начала следующего. Это позволяет с приемлемой точностью определять границы предложений. Чисто по длительности паузы в спонтанной диалогической речи (из нашего РК) удается обнаружить границу со спутыванием (SER) порядка 15%. Длительность пороговой паузы — обучаемый параметр. Интересным является факт довольно слабой эластичности погрешности обнаружения границы от величины порога. Улучшить этот результат вспомогательными мерами (адаптация порога по длительности к темпу речи и т.п.) не удалось. Причина в том, что пауза — необходимый, но нестабильный признак границы фразы. В потоке речи достаточно часто встречаются внутрифразовые паузы, более длинные, чем межфразовые паузы в этой же аудиозаписи. На *рисунке 1* приведен пример пауз в конкретном аудиофайле.

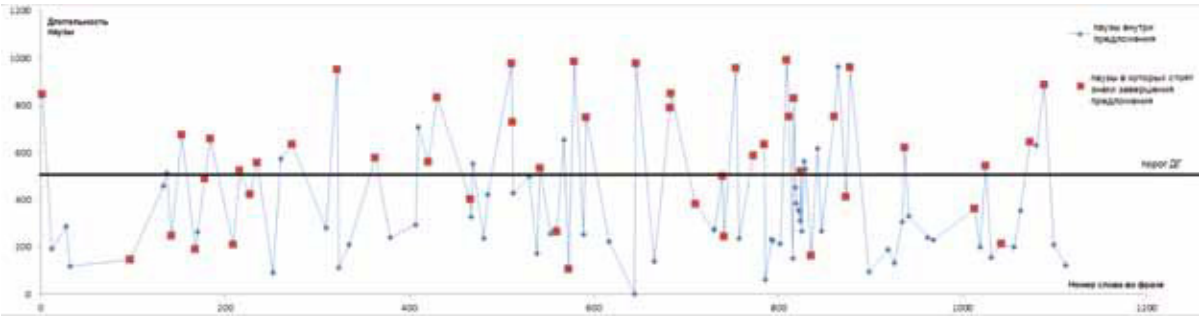

Рис. 1. Длительность межфразовых и внутренних пауз

Большинство фраз завершается понижением основного тона. Поэтому для обнаружения границ фраз был использован признак «скачок основного тона на паузе» в предположении, что большой положительный скачок коррелирует с наличием границы фразы. Однако количественное улучшение точности обнаружения границ составило менее 1%, что неожиданно. Добавление этого признака почти не улучшило разделения (рис. 2).

Быть может, надо использовать более сложные признаки, основанные на основном тоне.

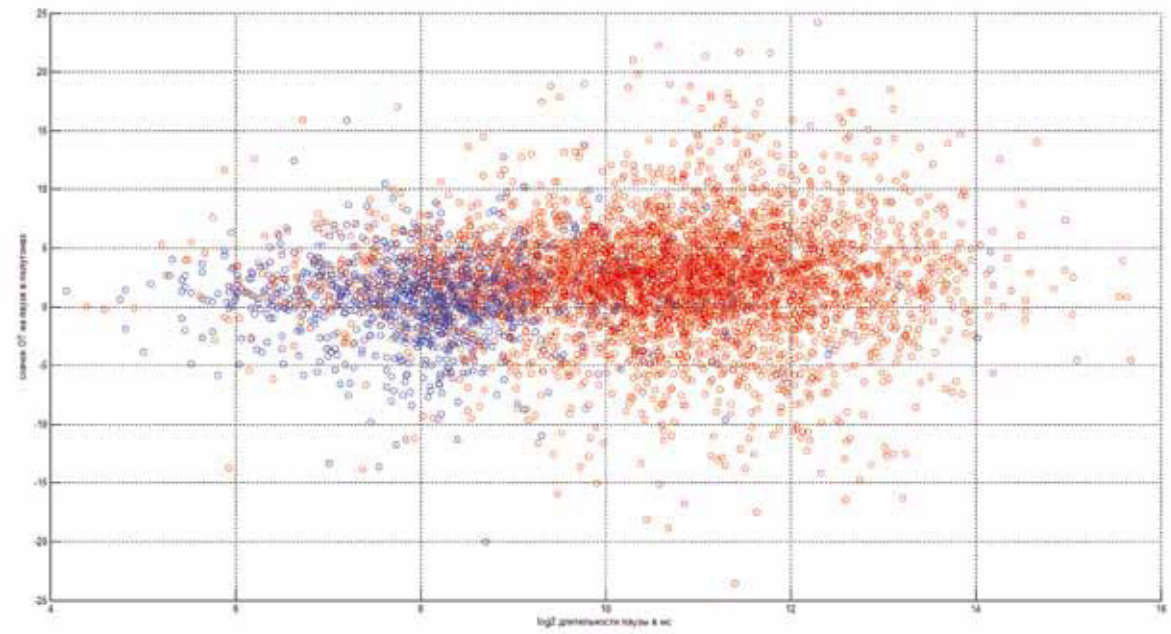

Рис. 2. Длительность пауз и скачки основного тона

2.2.2. Классификатор завершающего знака

Классификатор знака нуждается в серьезном совершенствовании. Текущая версия классификатора использует признаки, характеризующие форму трека основного тона. Для коротких фраз (одно-два слова) используются признаки, вычисленные относительно конца фразы. Для длинных (более трёх слов) фраз используются признаки, характеризующие форму трека основного тона около участка с наиболее быстрым изменением основного тона — предполагается, что там расположен интонационный центр фразы. Основная нерешенная проблема — правильное определение интонационного центра, необходимое для отделения релевантных изменений основного тона от нерелевантных. SER по завершающим знакам в настоящий момент 42%, что явно недостаточно для практического использования.

2.3. Лингвистическое направление

На этом направлении разработан программный механизм расстановки знаков препинания, реализующий правила, записанные на формальном языке YAML.

При разработке этого направления авторы опирались на правила русской пунктуации, изложенные в работах Д.Э. Розенталя [1], и справочник, подготовленный Институтом русского языка им. В.В. Виноградова РАН и Орфографической комиссией при Отделении историко-филологических наук Российской академии наук под редакцией В.В. Лопатина [6].

Для нашей задачи классические правила были формализованы и структурированы. Из таблицы правил номера правил были введены в эталонную базу «Пунктуация». Была подсчитана статистика частотности правил пунктуации, на основании которой ведутся работы по реализации правил. Кроме статистики при последовательности реализации правил учитывается не только частотность, но и возможность реализации данного правила без привлечения дополнительных программных средств.

Для тестирования лингвистического блока СП созданы отдельные тесты, полученные из эталонных аннотаций путём удаления внутренних знаков препинания. В настоящий момент реализованы основные правила русского литературного языка. Например, такие, как:

- Запятые в сложноподчинённых предложениях;
- Обособление вводных слов и выражений;
- Выделение междометий;
- Запятые при частицах и др.

В ходе разработки каждое правило проходит тестирование на примерах из эталонной части корпуса «Пунктуация».

Реализация каждого правила, которое кажется однозначным в письменной речи, сталкивается с большим количеством исключений в спонтанной речи.

Основная сложность возникает с теми явлениями, которые отсутствуют в письменном языке. Однако при этом они очень часто встречаются в спонтанной речи. Это такие явления, как:

- самоисправления;
- вербальная хезитация;
- неоформленность синтаксических и семантических связей;
- нарушение привычных связей в предложении.

Опора на интонацию и паузы различного рода (вдохи и вздохи, молчание, хезитация, ларингализация), как установлено, не может считаться надёжным критерием, поскольку между интонацией и пунктуацией нет полного соответствия (возможно: есть пауза, но нет знака препинания; или: есть знак, но нет паузы).

В некоторых случаях представляется возможным находить какие-то соотношения между конструкциями разговорной речи и конструкциями речи письменной (кодифицированного литературного языка), проводить аналогию между теми и другими. Но иногда такое сопоставление невозможно и приходится искать новые решения.

Например, как ставить запятые при маркерах хезитационного поиска? Это явление часто встречается в спонтанной речи, но никак не описано в правилах пунктуации, составленных для литературного языка, поскольку оно там отсутствует.

По сути, это выражения, которые диктор подставляет, когда не может сразу подобрать какое-то слово или словосочетание (это, это самое):

А вы вагончик как, на колёсах привезли или на этой, на шаланде?

Если диктор подобрал нужное слово, то оно появляется после вербального хезитатива, раскрывая смысл предложения, поэтому представляется уместным поставить между ними запятую. Если никаких знаков не ставить, то смысл изменится или будет потерян. Ср.: *телефон этого, моряка; телефон этого моряка*.

Если поиск оказывается неудачным и искомое слово не найдено, можно говорить о появлении своеобразного словосочетания с использованием вербального хезитатива:

Потому что дорога очень это самое, в темноту нельзя ездить.

Из данного примера видно, что запятая перед вербальным хезитативом была бы лишней, нарушая структуру синтагмы. А поскольку первая часть высказывания (до появления вербального хезитатива и включая его) строится независимо от того, найдено ли впоследствии искомое слово, представляется логичным и в случаях с удачным поиском ставить запятые аналогичным образом, то есть без запятой перед вербальным хезитативом, но с запятой между ним и искомым словом.

Другая особенность спонтанной устной речи — самоисправления говорящего.

*Да *чи(чистим), снег чистим.*
*Я холодильник *ща(сейчас), холодец *щас(сейчас) поставлю.*

По сути, дискурс прерывается и синтагма выстраивается заново. Очевидно, что при расстановке знаков препинания игнорировать подобные речевые сбои нельзя. Поэтому было принято решение в таких случаях ставить запятую.

К сожалению, не всегда всё бывает столь очевидно. При спонтанном порождении речи человек думает и говорит одновременно, поэтому связи между членами предложения часто размыты. Случается так, что фраза выстраивается до некого слова или словосочетания, а потом уже произнесённое слово или словосочетание становятся началом новой синтагмы — формируются две равнозначные части с общим центром. Яркий пример — дублирование подлежащего или сказуемого:

Ну ты Людке сама-то ты скажи...
Есть опять же в любой ситуации есть но.

По нормам литературного языка, с одной стороны, в предложении не может быть два сказуемых, не разделённых запятой, а с другой, нельзя разделять запятой непосредственно подлежащее и сказуемое (если между подлежащим и сказуемым стоит вводное слово или другой член предложения, требующий обособления, то это, по сути, не запятые, разделяющие подлежащее сказуемое, а запятые, выделяющие оборот). Кроме того, в случае дублирования подлежащего или сказуемого не

ясно, куда ставить запятую, поскольку фраза произносится как интонационно целостная, оба продублированных слова равноценны и равновесны, и получается, нет оснований отделять одно из них запятой.

Более простой случай — нарушение привычных связей в предложении (инверсия, перестановка слов). В литературном языке инверсия тоже возможна, но не в таких масштабах. Например, придаточное предложение со смещённым союзным словом:

Вот это вот, я чем переболела, группу дают людям.
*Ходить там, *ничё(ничего) нету когда.*

В данном случае вопрос о месте постановки запятой не стоит, поскольку выделяется придаточная часть сложноподчинённого предложения, независимо от позиции союзного слова, однако подобная перестановка слов серьёзно усложняет задачу автоматической расстановки знаков препинания.

2.4. Разработка методики оценки качества пунктуации

При разработке методики оценки качества автоматической пунктуации участники проекта столкнулись с явлением, которое было не существенно при оценке качества распознавания: трудность создания ручного эталона вследствие низкой степени согласия разных аудиторов.

Данное явление обусловлено качественным отличием письменной речи от спонтанной устной речи и проявляется как на уровне членения потока слов на предложения, так и на уровне классификации знаков.

Упомянутый выше эксперимент с членением потока слов на предложения профессиональными лингвистами показал:

- При членении потока слов аудитор в значительной мере полагается на смысл сообщения (законченность мысли);
- Степень согласия аудиторов достаточно низка.

Исходя из этого, сейчас проводится работа по введению в эталоны вариативных знаков для маркировки мест, где имеется существенная неоднозначность. Для этого надо как уточнить сами эталоны, так и изменить программу оценки качества пунктуации для поддержки вариативных знаков.

ЗАКЛЮЧЕНИЕ

На настоящий момент достигнуты следующие результаты: SER по завершающим знакам 42%, по внутренним знакам — 42%, по всем знакам препинания — 39%. Парадокса тут нет, так как при раздельном учете внутренних и завершающих знаков ошибка замены точки на запятую (весьма частая ошибка) учитывается как две ошибки (удаление точки и вставка запятой), а при общем учете — как одна ошибка замены.

Основная проблема — сложность самой задачи. Расстановка знаков пунктуации в потоке распознанных слов есть не задача распознавания (пунктуации в звучащей речи нет, и «распознать» её там невозможно), а задача перевода — перевода с устного на письменный. Главное отличие задачи перевода от задачи распознавания — различие выразительных средств входного и выходного языков. Так как устная речь, вообще говоря, богаче письменной, выразить знаками препинания на письме то, что выражается в устной речи просодически, — очень нелегко, а иногда и принципиально невозможно.

ЛИТЕРАТУРА

1. *Розенталь Д.Э.* Справочник по пунктуации. — М.: АСТ, 1997.
2. Стратегии членения спонтанной речи на синтаксические единицы // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог 2009» (Бекасово, 27-31 мая 2009 г.).— Вып. 8 (15). — М.: РГГУ, 2009. — С. 438–443.
3. *Комовкина Е.П., Слепокурова Н.А.* Анализ межпаузальных интервалов в спонтанном тексте: предварительные результаты // Череповецкие научные чтения-2009: Материалы Всероссийской научно-практической конференции, посвященной Дню города Череповца (2–3 ноября 2009 г.) / Ч. 1. Литературоведческие и лингвистические науки в начале XXI в. — Череповец: ГОУ ВПО ЧГУ, 2010. — С. 47–51.
4. *Венцов А.В.* Спонтанная речь: проблемы сегментации // IX выездная школа-семинар «Проблемы порождения и восприятия речи»: Материалы. — Череповец: ГОУ ВПО «Череповецкий государственный университет», 2010. — С. 96–101.
5. *Кибрик А.А., Подлесская В.И.* (ред.) Рассказы о сновидениях: Корпусное исследование устного русского дискурса. — М.: ЯСК, 2009.
6. *Лопатин В.В.* (ред.) Правила русской орфографии и пунктуации. Полный академический справочник. — М: АСТ, 2009.

A TASK OF AUTOMATIC PUNCTUATION IN RECOGNIZED RUSSIAN SPONTANEOUS SPEECH

Dmitry A. Birin,
the General Director of branch FGUP "scientific research Institute "Kvant" , Saint-Petersburg

Alexander E. Bulashevich,
candidate of technical Sciences, researcher of the branch of FSUE "scientific research, Institute "Kvant", Saint-Petersburg

Marianna Y. Grekis,
engineer of FGUP "scientific research Institute "Kvant», Saint-Petersburg

Abstract

The main purpose of speech recognition process is to produce readable, understandable text at output. In the Russian language it is hardly possible without punctuation marks. There is a very complicated system of punctuation rules for the Russian language. The problem is that these rules were developed for written language. Most of them are

not observed or are even broken in spontaneous speech. There are also some phenomena in spontaneous speech which are not described in the rules for literary (written) language simply because these phenomena do not meet in the written text (hesitation search, self-repairs etc.). Thus, the task is to adopt classic rules for spontaneous speech and to develop an automatic punctuation system that would be able to transform a sequence of recognized words received from spontaneous speech into a comprehensible written text. At this stage our system allows to detect sentence boundaries in most cases and placing some internal punctuation marks with a certain accuracy.

Keywords: spontaneous speech recognition, punctuation of recognized speech, automatic punctuation.