

Точка как проблема

Мария Ровинская,

МГУ, филологический факультет

marr@df.ru

Настоящий доклад посвящен рассмотрению одной из многочисленных проблем, возникающих при создании систем автоматического синтеза речи. Создатель такого рода систем с неизбежностью сталкивается с тем, что письменные тексты состоят не только из орфографических слов и знаков пунктуации, они часто включают в себя различного рода специальные знаки, сокращения, аббревиатуры и т.д. Вместе с тем и знаки пунктуации могут выполнять не только свою основную функцию, но и некоторые дополнительные, которые также необходимо отделять от собственно пунктуационного употребления, поскольку в зависимости от функции, которую выполняет тот или иной знак, текст может иметь разные варианты прочтения. Так, например, восклицательный знак в конце предложения и восклицательный знак, обозначающий факториал в математической записи, представляют собой, в сущности, разные знаки – первый служит для чтеца сигналом, что предложение закончено и должно получить соответствующее интонационное оформление, а второй требует произнесения слова «факториал». Подобная многозначность свойственна практически всем знакам пунктуации.

Мы рассмотрим нетривиальную проблему, возникающую при транскрибировании текста, содержащего сокращения с точкой.

Примеры^[1]:

(1) Так что дискуссия не закрыта, и она, безусловно, будет продолжаться до самого 1999 г., когда июльским решениям предстоит воплотиться в жизнь.

(2) Деятельность св. князя Владимира по реформированию языческого Пантеона известна даже меньше, чем его злодейские подвиги юности.

(3) Иначе как объяснить запланированное в проекте падение реального объема ВВП на 6 процентов по сравнению с планом 1996 г.?

(4) 14 марта 1995 г. Служба получила свой основополагающий закон.

(5) С 1988 года публиковался в литературных журналах и альманахах "Зеркало", "Личное дело" и др. Книги стихотворений...

(6) Кто-то, памятуя о советских временах, возможно, испытывает генетический страх перед самими словами "идея", "идеология" и пр. А кому-то в принципе ничего этого не нужно.

(7) Из Гусь-Хрустального выезжают в 2 часа ночи, с пересадкой на ст. Черусти, в 8 утра - уже в Москве, в 18.

(8) Согласно ст.137 Конституции изменения в ст. 65 (то есть в составе Федерации) могут производиться на основании федерального конституционного закона.

(9) Генеральным директором Центра назначается проф. Сухих.

Задача интерпретации сокращений с точкой разбивается на две сравнительно независимые части. Во-первых, необходимо уметь автоматически отделять пунктуационное употребление точки от «паразитического», когда точка в тексте не обозначает конец повествовательного предложения и, следовательно, не должна служить для системы синтеза маркером для присвоения тексту соответствующих просодических характеристик. Можно сказать, что точка в последнем случае является частью сокращенного слова, которая показывает, что перед прочтением необходимо восстановить полную форму сокращения, что, собственно, и составляет вторую часть исследуемой проблемы. Трудности возникают уже на уровне установления исходного слова (или сочетания слов), которые выступают в тексте в сокращенном виде, даже если оставить за пределами рассмотрения проблему определения грамматической формы сокращенного слова.

1. I. Определение функции точки.

Прежде всего следует отметить, что точка может употребляться не только в качестве знака пунктуации и показателя сокращенного слова, но и в функции своего рода разделителя в обозначении номеров пунктов и подпунктов (2.3.1.), дат (*Итоги, 18.04.1997*), времени (21.40), больших и дробных чисел (2.600.000 тонн металла / 2.6 млн. тонн металла^[2]). Для каждого из приведенных примеров существуют отдельные правила чтения. Так, перед номером подпункта должно прозвучать слово «*подпункт*», названия месяцев в датах произносятся полностью, а не обозначаются цифрами, причем в конце добавляется слово «*года*» («*восемнадцатое апреля тысяча девятьсот девяносто седьмого года*»); если цифры в записи обозначают время, то они читаются по отдельности до и после точки («*двадцать один сорок*»), а если цифрами записано дробное число, то должно быть озвучено разделение на дробную и целую часть числа («*две целых шесть десятых*» или, в упрощенном варианте, «*два и шесть*»). Таким образом, употребление точки в качестве разделителя составляет отдельную проблему для автоматического синтеза речи.

Рассмотрим существующие возможности для автоматической интерпретации точки как знака пунктуации и как показателя сокращенного слова. Первый тест, которому необходимо подвергнуть знаки, окружающие встреченную в тексте точку, помогает в ряде случаев решить поставленную задачу независимо от значения конкретных слов, которые мы исследуем. Согласно правилам русского языка каждое новое предложение должно начинаться с *прописной буквы*. В случае, когда после точки следует строчная буква (а также запятая, точка с запятой, тире, двоеточие) мы имеем дело с сокращенным словом в середине предложения и нам остается правильно его восстановить и затранскрибировать (примеры 1 и 2). После сокращения могут также следовать вопросительный или восклицательный знаки, но и в этих случаях точка не является знаком пунктуации, конец предложения обозначен отдельно (пример 3). Главную трудность составляют примеры, когда за точкой следует прописная буква. В этом случае точка может служить либо только маркером сокращения, либо одновременно и показателем сокращенного слова, и знаком пунктуации. Так, пример 4 допускает два различных варианта прочтения. В зависимости от интерпретации функции точки после сокращенного слова, можно считать, что в примере приводится одно предложение, в котором за сокращением следует имя собственное, или два отдельных предложения. Видимо, на этом этапе анализа следует рассматривать каждый случай сокращения отдельно, используя заранее подготовленный электронный словарь сокращений, о котором несколько подробнее будет сказано ниже. Каждое сокращенное слово допускает некоторые варианты контекста, часто ограниченные, что может позволить

системе автоматического синтеза речи правильно интерпретировать функцию точки в каждом конкретном случае с приемлемой долей погрешности. Скажем, хотя за сокращением *проф.* (*профессор*) чаще всего следует прописная буква (с которой начинается фамилия профессора), очевидно, что предложение не может заканчиваться подобным сокращенным словом, следовательно, точка в этом случае не является знаком пунктуации (пример 9).

Существует также ряд сокращений, тяготеющих к употреблению в конце предложения. В таких случаях точка совмещает выполнение двух функций – пунктуационной и показателя сокращения. К подобным словам относятся *т.д.*, *т.п.*, *др.*, *пр.* [3] (см. примеры 5, 6). Учитывая тенденцию некоторых сокращений к употреблению в конце предложения, при автоматическом чтении текста можно считать, что прописная буква после них всегда начинает новое предложение. Список сокращений, которые обычно употребляются в конце предложения, чрезвычайно ограничен. По общему правилу сокращенные слова встречаются в середине предложения, даже если позиция в конце предложения формально для них возможна, а употребления, подобные примеру 4, допускающие неоднозначное прочтение, крайне редки и свидетельствуют о стилистической небрежности. Подобную закономерность можно объяснить важностью позиции конца предложения и стремлением избежать трудностей при чтении. В ряде случаев для определения функции точки необходимо восстановить полный вариант сокращения. Например, сокращение *г.* в значениях *‘город’*, *‘гора’* или *‘господин’* никогда не встречается в конце предложения, а в значении *год* эта позиция является самой частотной (на 384 употребления сокращения *г.* в значении *‘год’* 320 приходится на конец предложения).

1. II. Восстановление сокращенного слова.

Сокращения с точкой читаются в своей полной форме, но в отличие от обычных слов и даже аббревиатур графическая форма сокращения сама по себе практически ничего не говорит читателю о полном виде слова; чтобы прочесть сокращение, мы используем дополнительную экстралингвистическую информацию. Построить единые алгоритмы, которые позволили бы по произвольному сокращенному слову восстановить его полную форму, хотя бы для групп сокращений не представляется возможным. Но, с другой стороны, список сокращений в языке фиксирован, следовательно, его можно занести в базу данных. Такой электронный словарь сокращений для русского языка создан, он содержит порядка 1200 сокращений и некоторую дополнительную информацию о релевантных для автоматического синтеза особенностях употребления сокращенных слов.

Следует отметить, что некоторые тексты изначально содержат списки используемых в них сокращений. Чаще всего такие списки состоят из специальных, редко употребляемых сокращений. Наличие такого списка упрощает задачу автоматического чтения текста, тем более, что сокращения в таком случае, как правило, не допускают двусмысленных толкований.

При отсутствии предназначенного специально для данного текста списка сокращенных слов появляется проблема чтения омонимичных сокращений. Представляется, что при решении этой проблемы в первую очередь было бы полезно учитывать специфику текста. Конечно, идеальная система автоматического синтеза речи может озвучить любой, произвольный текст, но умение определить его жанр и специфику помогли бы сразу решить большую часть сложностей в исследуемой проблеме. К примеру, достаточно большую часть сокращенных слов составляют топонимы, и дополнительная информация о тексте позволит правильно интерпретировать сокращение *алб.*, которое может обозначать *‘алебастровый завод’* и *‘албанский’* (скажем, в словарях), и другие подобные случаи. Учет специфики текста часто избавляет создателей систем автоматического синтеза речи от необходимости алгоритмизировать сложный семантический контекстный анализ употребления сокращений, позволяя без труда разделять такие сокращенные слова как *‘рота’* и *‘родительный’* (падеж) – *р.*, *‘край’* и *‘критический’* (в физических или химических текстах) – *кр.* и другие.

В случаях, когда один из двух вариантов значений омонимичного сокращения достаточно редко встречается, можно ориентироваться на *частотные характеристики* употребления. Скажем, сокращение *стр.* гораздо чаще обозначает страницу, чем строку, а анализ специфики текста в этом случае не помогает разделить эти два значения.

На следующем этапе процесса восстановления полной формы сокращенного слова добавляется *анализ контекстного окружения* сокращения. Например, после сокращения *ст.* в значении *статья* следуют цифры, обозначающие номер статьи, а после того сокращения в значении *станция* – название станции, т.е. имя собственное, написанное с прописной буквы (см. примеры 7 и 8).

Проиллюстрируем предлагаемый алгоритм интерпретации сокращений с точкой на примере одного из самых частых сокращенных слов, встречающихся в русских текстах – *г.*^[4], которое может иметь значения '*год*', '*город*', '*господин/госпожа*', '*гора*' (в порядке уменьшения частоты употребления).

Примеры:

(10) Лужков Юрий Михайлович. Родился 21 сентября 1936 г. в Москве в семье плотника.

(11) Рукоположен в священники в 1958 г. Вдовец, имеет двух дочерей.

(12) Если "курсы" наших двух стран сохранятся, к 2010 г. Вьетнам по объемам душевого производства ВВП обгонит Россию.

(13) Наверное, надо оговориться: г. Южа попал в особо тяжкое положение из-за чудовищно неблагоприятного демографического состава населения.

(14) Автор этих картин - г. Семейко - работает в жанре гиперреализма.

(15) Запретить ввоз, разведение и продажу на территории г. Москвы собак породы пит-бультерьер, а также американский и стаффордширский терьер.

Чтобы определить, обозначает ли точка после сокращения конец повествовательного предложения, необходимо отделить употребление сокращения в значении '*год*', которое может употребляться и часто встречается в позиции конца предложения, от остальных значений, для которых эта возможность исключена. Такое разделение сравнительно легко можно произвести, учитывая особенности употребления сокращения в разных значениях – '*год*' встречается после цифр, в предложении часто встречаются временные интервалы («в конце 1999 г.»), названия месяцев («в сентябре 1980 г.»), а после сокращения в остальных значениях следует имя собственное. В примере 3 перед сокращением стоят цифры, обозначающие год, а после – имя собственное. В подобных спорных случаях приоритетным должно быть прочтение «*год*»: во-первых, оно гораздо чаще встречается (в публицистических текстах на 421 употреблений исследуемого сокращения 384 приходится на значение '*год*'), и во-вторых, оно не так легко может быть опущено, как остальные (ср. пример 1 и вариант *Родился 21 сентября 1936 в г. Москве*).

Далее, если за сокращением следует строчная буква (запятая, тире, двоеточие, точка с запятой), то, очевидно, точка не завершает предложения, а выступает лишь показателем сокращения (пример 10). В случае, когда за сокращением в значении '*год*' следует прописная буква, необходимо установить, является ли следующее слово нарицательным или собственным именем. Как показывает анализ публицистических текстов, в первом случае точка обычно обозначает конец предложения, а во втором – нет (примеры 11 и 12).

Перейдем к более сложной задаче разделения значений ‘*город*’, ‘*господин/госпожа*’ и ‘*гора*’. Чтобы правильно читать сокращенное слово, можно хранить в памяти компьютера названия городов и гор, но часто встречаются, например, названия городов, неотличимые от фамилий (*В окне показался г. Николаев* – город? господин?). В некоторых случаях может помочь анализ контекста употребления сокращения, но такой анализ обычно проводится в пределах одного предложения и не всегда полезен; так, в примере 13 и город Южа, и господин Южа могли попасть в особо тяжкое положение по упомянутым причинам, а анализ примера 14 не дает нам информации о поле г. Семейко (господин? госпожа?). В качестве одного из вариантов решения поставленной задачи можно предложить опускать сокращения во всех употреблениях, вызывающих затруднения. Такое решение, как правило, не приводит к искажению смысла читаемого текста (ср. вариант чтения примера 15: *Запретить ввоз, разведение и продажу на территории Москвы собак породы пит-бультерьер, а также американский и стаффордширский терьер*) и позволяет экономить компьютерные и временные ресурсы[5].

Предварительная схема интерпретации сокращения «г.»

«г.»

анализ контекстного окружения

‘год’		‘город, господин/госпожа, гора’
+ строчная буква		ъ
(, - ; :)	+ прописная буква	середина предложения
		сокращение может быть опущено
ъ		
середина предложения	имя собственное	имя нарицательное
	ъ	ъ
		середина предложения конец предложения

Литература:

Алексеев Д.И., Гозман И.Г., Сахаров Г.В. Словарь сокращений русского языка. М., 1984 г.

Розенталь Д.Э. Пунктуация и управление в русском языке. М., 1988 г.

997. Dutoit. An Introduction to Text-to-Speech Synthesis. Dordrecht-Boston-London. 1997.

[1] Примеры взяты из корпуса публицистических текстов.

[2] Существуют два способа записи дробных чисел, в одном случае в качестве разделителя используется точка («европейская традиция»), а во втором случае дробная часть числа отделяется от целой при помощи запятой («русская традиция»). В настоящее время в различных текстах используются оба варианта записи, что еще более осложняет задачу автоматического чтения текста.

[3] Скажем, на 56 употреблений сокращения *др.* на конец предложения приходятся 39 из них, 12 употреблений встречаются в скобках, и лишь 5 раз это сокращение употреблено в середине предложения, причем в 3 случаях после сокращения следует тире, а в остальных – запятая.

[4] Приведенные выше пути решения проблемы автоматического чтения сокращенных слов, конечно, не охватывают абсолютно все случаи употребления сокращений, что хорошо видно на нашей фразе, где сокращение *г.* должно читаться так, как оно написано - «*г точка*». Очевидно, что иногда для правильного чтения требуется дополнительный синтаксический, семантический и прагматический анализ.

[5] Подобное решение проблемы чтения сокращений применимо и к другим словам. Например, чтобы правильно прочитать сокращение *д. ф. н.* (*‘доктор филологических/ философских наук’*), можно не проводить анализ контекста, а читать «доктор наук».