АЛГОРИТМ СЕГМЕНТАЦИИ ТЕКСТА НА СИНТАКСИЧЕСКИЕ СИНТАГМЫ ДЛЯ СИНТЕЗА РЕЧИ

AN ALGORITHM OF TEXT SEGMENTATION ON SYNTACTIC SYNTAGMAS FOR TTS SYNTHESIS

Лобанов Б.М. *(lobanov@newman.bas-net.by)*
Объединенный институт проблем информатики НАН Беларуси, Минск, Беларусь

Предлагается алгоритм сегментации текста на синтаксические синтагмы, основанный на анализе устойчивых фразеологических и грамматико-смысловых словосочетаний, составляющих предложение. Основной смысл выделения в предложении рассматриваемых словосочетаний заключается в том, что теперь свобода его разделения на синтагмы ограничивается, а именно: граница синтагмы может находиться только за пределами словосочетаний, но не внутри их.

Введение

Просодическая разметка текста заключается в его членении на синтагмы, разметке синтагм на акцентные единицы и маркировке интонационного типа синтагм [1].

Под синтагмой понимается самостоятельная в интонационном смысле часть предложения или всё предложение. Установка границ синтагм влияет на передачу интонационных характеристик при синтезе речи, а также на передачу смыслового содержания. При разбиении текста на синтагмы важно не поставить границу синтагмы там, где она может нарушить смысловое восприятие речи (или передачу смыслового содержания текста), например, между предметом и его признаком. Для установки границ синтагм при синтезе речи по тексту используются определённые правила синтагматического членения, базирующиеся на синтаксическом анализе текста, на учёте фактора речевого дыхания [2], а также на статистическом анализе особенностей синтагматического членения естественной речи конкретного диктора [3LobTs].

Первым этапом просодической разметки текста является его членение на пунктуационно-лексические синтагмы [4Present]. Пунктуационно-лексической синтагмой (ПЛС) считается всё предложение (при отсутствии в нём знаков препинания) или часть предложения, ограниченная любым знаком препинания или каким-либо из лексических маркеров. Следует заметить, что даже после разбиения предложения на ПЛС их длина может оказаться всё же слишком большой.

Пример:

«*Но молодая жена упорно продолжала отстирывать белую в кровавых пятнах рубаху мужа посиневшими от холода руками в железном тазике с ледяной водой*».

Очевидно, что при отсутствии механизма дальнейшего членения таких предложений на более мелкие синтаксические синтагмы (СС) неизбежно возникнут затруднения в понимании смысла синтезированной речи. Идеальным решением проблемы дальнейшего членения такого рода ПЛС на СС было бы использование комплекса правил их глубинного синтаксического разбора [5Bogus]. Однако, в виду сложности и недостаточной разработанности таких правил, в данной работе предлагается использование процедуры поверхностного синтаксического анализа, опирающейся на доступную морфосинтаксическую информацию о словосочетаниях, составляющих ПЛС.

1. Общая структура алгоритма сегментации на синтаксические синтагмы

Словосочетание рассматривается в грамматике как пара по смыслу и грамматически связанных слов, выделяемая из предложения [6Gramm]. Являясь наряду со словом элементом построения предложения, словосочетание выступает в качестве одной из основных синтаксических единиц. Непосредственной целью

рассматриваемой процедуры поверхностного синтаксического анализа является предварительное разбиение ПЛС на последовательность словосочетаний 2-х типов: устойчивые фразеологические словосочетания (ФЛС) и грамматико-смысловые словосочетания (ГСС). Основной смысл выделения в ПЛС словосочетаний типа ФЛС и ГСС заключается в том, что теперь свобода разделения ПЛС на СС ограничивается. Граница синтагмы может находиться только за пределами ФЛС или ГСС, но не внутри их.

Предлагаемая общая структура процедуры просодической разметки на синтаксические синтагмы для синтеза речи по тексту, основанная на анализе словосочетаний, представлена на рис. 1.

Рис. 1. Общая структура алгоритма сегментации на синтаксические синтагмы

Рассмотрим подробно функционирование каждого из блоков алгоритма, представленного на рис.1.

2. Выделение фразеологических словосочетаний

В анализируемой синтагме отмечаются ФЛС, найденные в словаре устойчивых словосочетаний. К фразеологическим словосочетаниям относятся [6]:

– фразеологические сращения – «*попасть впросак», «бить баклуши», «ничтоже сумняшеся», «собаку съесть»* и др.

– фразеологические единства – «*зайти в тупик», «бить ключом», «плыть по течению», «брать в свои руки», «прикусить язык»* и др.

– фразеологические сочетания – «*потупить взор»,* «*щекотливый вопрос»,* «*бархатный сезон*», «*поголовные аресты»* и др.

Выделяются следующие типы компонентного состава фразеологизмов:

– сочетание прилагательного с существительным: *краеугольный камень, заколдованный круг, лебединая песня*;

– сочетание существительного в именительном падеже с существительным в родительном падеже: *точка*

зрения, камень преткновения, бразды правления, яблоко раздора;

– сочетание имени существительного в именительном падеже с существительными в косвенных падежах с предлогом: *кровь с молоком, душа в душу, дело в шляпе*;

– сочетание предложно-падежной формы существительного с прилагательным: *на живую нитку, по старой памяти, на короткой ноге*;

– сочетание глагола с существительным (с предлогом и без предлога): *окинуть взором, посеять сомнения, взять в руки, взяться за ум, водить за нос*;

– сочетание глагола с наречием: *попасть впросак, ходить босяком, видеть насквозь*;

– сочетание деепричастия с именем существительным: *спустя рукава, скрепя сердце, сломя голову*.

Позиции слабых и сильных словесных ударений в устойчивых сочетаниях могут быть определены в словаре сочетаний, при этом одно из слов обязательно несёт сильное ударение. При отсутствии помет слабых и сильных ударений вполне допустима установка сильного ударения на каждом из слов устойчивого словосочетания.

3. Объединение слов в грамматико-смысловые словосочетания

В зависимости от того, какое слово является первым в словосочетании, различаются основные лексико-грамматические группы словосочетаний. Классификация словосочетаний по признаку первого слова может быть представлена следующей схемой.

Группа прилагательных словосочетаний. Эта группа включает прилагательные, местоимения-прилагательные, порядковые числительные и причастия, которые сочетаются:

• С существительным (*полезная книга, зелёную листву, свою находку, унесённые ветром)*.

Признаки словосочетания: прилагательное + существительное (в одном падеже).

• С инфинитивом (*способный работать, готовый учиться)*.

Признаки: прилагательное + инфинитив.

Группа наречных словосочетаний. Эта группа сочетается:

• С инфинитивом (*безнаказанно игнорировать, хорошо петь)*.

Признаки: наречие + инфинитив.

• С наречием (*очень удачно, по-прежнему хорошо)*.

Признаки: наречие + наречие

• С существительным *(далеко от дома, наедине с сыном, незадолго до экзаменов)*.

Признаки: наречие + существительное (в косвенном падеже с предлогом).

• С местоимением-существительным (*недалеко от них, наедине с ней, незадолго до неё)*.

Признаки: наречие + местоимение-существительное (в косвенном падеже с предлогом).

Группа глагольных словосочетаний. Глагольная группа сочетается:

• С инфинитивом (*предложил выучить, просит взять)*.

Признаки: глагол (в любой форме) + глагол (инфинитив).

• С деепричастием (*идёт оглядываясь, говорить улыбаясь)*.

Признаки: глагол (в любой форме) + деепричастие.

• С наречием (*поступал справедливо, заниматься вдвоем)*.

Признаки: глагол (в любой форме) + наречие.

• С существительным (*искать покоя, писал брату, стоять у дороги, подъехал к дому, встретиться с друзьями)*.

Признаки: глагол (в любой форме) + существительное в косвенном падеже.

• С местоимением-существительным **(***искать их, писал ему, стоять около неё, подъехал к нему, встретился с ними)*.

Признаки: глагол (в любой форме) + местоимение-существительное в косвенном падеже.

Группа числительных словосочетаний. Эта группа сочетается:

• С существительным (*две книги, оба друга, трое в шинелях, сто рублей)*.

Признаки: количественное или собирательное числительное + существительное, в одном падеже.

Группа существительных словосочетаний. Эта группа включает существительные и местоимения-существительные и сочетается:

• С существительным (*письмо родителям, его доклада, оценку выступления, входом в театр)*.

Признаки: существительное + существительное (в отличающихся падежах без предлога или с предлогом).

• С наречием (*прогулка верхом, судак по-польски)*.

Признаки: существительное + наречие.

Перечисленные правила объединения слов в ГСС представлены в таблице 1, где по горизонтали расположены типы групп словосочетаний в порядке степени (силы) его связности со вторыми в паре словами. В таблице указано также место предпочтительной установки полного (+) и частичного (=) ударений.

Тип словосочетания / Второе слово в паре		Прила-гательное	Наречное	Глаголь-ное	Числи-тель-ное	Существи-тельнное
		1	**2**	**3**	**4**	**5**
Инфинитив	**1**	(=) (+)	(=) (+)	(=) (+)	–	–
Деепричастие	**2**	–	–	(=) (+)	–	–
Наречие	**3**	–	(+) (=)	(=) (+)	–	(=) (+)
Существительн	**4**	(=) (+)	(+) (=)	(+) (=)	(+) (=)	(+) (=)
Местоимение	**5**	–	(+) (=)	(+) (=)	–	–

Таблица 1. Правила объединения слов в словосочетания

Из таблицы видно, что в соответствии с правилами русской грамматики [3], допустимыми и наиболее частотными являются 14 различных типов ГСС. С учётом этого предлагается следующая последовательность действий по разметке синтагм на словосочетания.

1. В синтагме отыскиваются пары слов - прилагательные словосочетания, состоящие из слова группы прилагательных и стоящего справа от него существительного либо инфинитива глагола. Эти пары слов объединяются в словосочетания. Если такой пары не находится, то слово из группы прилагательных остаётся «одиноким».

2. Затем в синтагме рассматриваются оставшиеся слова, т.е. не объединённые в словосочетания по п. 1, и отыскиваются пары слов - наречные словосочетания, состоящие из двух наречий или наречия и стоящего справа от него инфинитива глагола, либо существительного или местоимения-существительного с предлогом. Если таковые находятся, то они объединяются в словосочетания, если нет, то наречие остаётся «одиноким».

3. Далее в синтагме рассматриваются оставшиеся слова, и отыскиваются пары слов - глагольные словосочетания, т.е. глагол в любой форме и стоящие справа от него наречие, инфинитив или деепричастие, которые объединяются в одно словосочетание. Если таковых не обнаружено, то глагол может быть объединён с существительным или с местоимением-существительным в косвенном падеже, стоящим справа, Если их нет, то глагол остаётся «одиноким».

4. В оставшихся необъединённых словах ищутся пары слов - числительные словосочетания, состоящие из количественного или собирательного числительного и стоящего справа от него существительного, согласованного с числительным по падежу, которые объединяются в словосочетания. Если нет, то числительное остаётся «одиноким».

5. Наконец, в оставшихся необъединённых словах ищутся соседние пары слов - существительные словосочетания, состоящие из слова группы существительных и стоящего справа от него наречия либо существительного или местоимения-существительного, которые объединяются в словосочетания. Если таких слов не находится, то существительное остаётся «одиноким».

Ниже в примере (3) показана разметка 4-х различных ПС на словосочетания в соответствии с предложенной последовательностью действий (см. п.п. 1 – 5). Словосочетания отмечены квадратными скобками, в круглых скобках указан тип словосочетания (см. таблицу 1), буквой *Ъ* обозначено объединение знаменательных и служебных слов в фонетическое слово.

(3) Пример:

Если Вам [необходимо активировать(2)] [услугу передачи(5)] даных дляЪвашего [мобильного номера(1)].

Но благодаря [разумному сочетанию(1)] лекарств он [смог остановить(3)] [развитие болезни(5)] [вЪбольшинстве случаев(5)].

Тогда тарификация [Ваших звонков(1)] [начинается сЪмомента(3)] соединения [сЪ телефоном абонента(5)].

[Идеальным решением(1)] [проблемы членения(5)] [такого рода(1)] предложений наЪсинтагмы [былоЪбы использование(3)] [комплекса правил(5)] разбора [наЪсинтаксические компоненты(1)].

4. Расширение двухсловных ГСС до трёх- и более словных сочетаний

Если ПЛС, обработанная в соответствии с указанной выше в п.п. 1 – 5 последовательностью действий по разметке синтагм на словосочетания, содержит слова, не вошедшие в созданные двухсловные сочетания, то рассматривается возможность расширения до трёх- и более словных сочетаний по следующей схеме.

1. Рассматриваются полученные «прилагательные словосочетания».

а) если перед прилагательным словосочетанием стоит слово из группы прилагательных, то оно дополнительно включается в это словосочетание.

б) если после прилагательного словосочетания стоит слово из группы существительных в родительном падеже, то оно дополнительно включается в это словосочетание.

2. Если после наречного или глагольного словосочетания стоит слово из группы существительных в родительном падеже, то оно дополнительно включается в это словосочетание.

3. Если после числительного или существительного словосочетания стоит слово из группы существительных в родительном падеже, то оно дополнительно включается в это словосочетание.

В примере (3.1) показан результат расширения двухсловных сочетаний ПС примера (3) в соответствии с изложенными выше правилами. Здесь в круглых скобках указан тип словосочетания, в квадратных – двусловные сочетания, в фигурных – расширенные по п.п..1 – 3 трёхсловные сочетания

(3.1) Пример:

Если Вам [необходимо активировать(2)] {[услугу передачи(5)] даных} {дляЪвашего [мобильного номера(1)]}.

Но благодаря {[разумному сочетанию(1)] лекарств} он [смог остановить(3)] [развитие болезни(5)] [вЪбольшинстве случаев(5)].

Тогда тарификация [Ваших звонков(1)] {[начинается сЪмомента(3)] соединения} [сЪтелефоном абонента(5)].

[Идеальным решением(1)] [проблемы членения(5)] {[такого рода(1)] предложений} наЪсинтагмы [былоЪбы использование(3)] {[комплекса правил(5)] разбора} [наЪсинтаксические компоненты(1)].

. В приведенном выше примере, несмотря на реализацию полной последовательности действий по разметке синтагм на словосочетания и по их расширению, остаются ещё отдельные слова, не вошедшие ни в одно из словосочетаний. Следующим, последним шагом, является дополнительное расширение словосочетаний путём включения в их состав слабоударных слов, таких, как многосложные предлоги, союзы и местоимения. Если же и после такого расширения словосочетаний остаются отдельные слова, то они определяются как частный случай однословных сочетаний.

Результат применения этого правила иллюстрируется примером (3.2).

(3.2) Пример:

{Если Вам [необходимо активировать(2)]} {[услугу передачи(5)] данных} {дляЪвашего [мобильного номера(1)]}.

{Но благодаря [разумному сочетанию(1)] лекарств} {он [смог остановить(3)]} [развитие болезни(5)] [вЪбольшинстве случаев(5)].

{Тогда [тарификация]} [Ваших звонков(1)] {[начинается сЪмомента(3)] соединения} [сЪтелефоном абонента(5)].

[Идеальным решением(1)] [проблемы членения(5)] {[такого рода(1)] предложений} [наЪсинтагмы] [былоЪбы использование(3)] {[комплекса правил(5)] разбора} [наЪсинтаксические компоненты(1)].

5. Расстановка слабых и сильных словесных ударений

Для двухсловных сочетаний в таблице 4.2 указано место предпочтительной установки сильного и слабого ударений, т.е. на первом или втором слове словосочетания. Такое распределение позиций слабого и сильного ударений не претендует, конечно, на универсальность. Оно характеризует среднестатистическую тенденцию для достаточно широкого набора различных текстов. При определённых условиях (индивидуальная манера чтения, стремление к определённой ритмической структуре и др.) знаки (+) (=) для данного словосочетания могут меняться местами, либо оба знака индицировать сильное ударение – (+) (+). Важную роль может играть также наличие некоторых индикаторов потенциальной «слабости» или «силы» какого-либо из слов в словосочетании. В частности, к индикатору «слабости» может быть отнесена принадлежность слова к группе потенциально слабоударных слов, таких, как многосложные предлоги и частицы, союзы и местоимения. К индикатору «силы» может быть отнесено наличие перед словом усилительной или отрицательной частицы.

После применения указанных правил расстановки сильных и слабых ударений пример (3.2) перепишется в следующем виде.

(3.3) Пример:

{Е=сли Ва=м [необходи=мо активи+ровать(2)]} {[услу+гу переда=чи(5)] да=нных} {дляЪва=шего [моби=льного но+мера(1)]}.

{Но= благодаря= [разу=мному сочета+нию(1)] лека=рств} {о=н [смо=г останови+ть(3)]} [разви+тие боле=зни(5)] [вЪбольшинстве+ слу=чаев(5)].

{Тогда= [тарифика+ция]} [Ва=ших звонко+в(1)] {[начина+ется сЪмоме+нта(3)] соедине=ния} [сЪтелефо+ном абоне=нта(5)].

[Идеа=льным реше+нием(1)] [пробле+мы члене=ния(5)] {[тако=го ро+да(1)] предложе=ний} [наЪсинта+гмы] [бы+лоЪбы испо=льзование(3)] {[ко+мплекса пра=вил(5)] разбо=ра} [наЪсинтакси=ческие компоне+нты(1)].

На заключительном этапе целесообразно провести окончательную корректировку позиций сильных и слабых ударений с точки зрения приближения к оптимальной организации ритмической структуры синтагм. При этом уточняются ситуации, когда в ГрС имеется более одного слова со слабым ударением. Окончательная корректировка осуществляется, исходя из необходимости соблюдения следующих условий:

– в ГрС не должно быть двух следующих подряд слов со слабым ударением. В этом случае в одном из этих слов, например во втором, слабое ударение заменяется на сильное.

– в ГрС количество слов со слабым ударением не должно быть больше количества слов с сильным ударением. Например, последовательность (=) (+) (=) заменяется на последовательность (=) (+) (+).

Следует заметить, что приведенные здесь правила отражают лишь среднестатистические закономерности. Окончательные условия особой выделенности того или иного слова могли бы быть адекватно определены только в результате глубокого синтаксического и семантического анализа предложений, что в настоящий момент пока недостижимо.

После применения указанных правил корректировки пример (3.3) перепишется в следующем виде.

(3.4) Пример:

{Е=сли Ва+м [необходи=мо активи+ровать(2)]} {[услу+гу переда=чи(5)] да+нных} {дляЪва=шего [моби+льного но+мера(1)]}.

{Но= благодаря+ [разу=мному сочета+нию(1)] лека+рств} {о=н [смо+г останови+ть(3)]} [разви+тие боле=зни(5)] [вЪбольшинстве+ слу=чаев(5)].

{Тогда= [тарифика+ция]} [Ва=ших звонко+в(1)] {[начина+ется сЪмоме+нта(3)] соедине=ния} [сЪтелефо+ном абоне=нта(5)].

[Идеа=льным реше+нием(1)] [пробле+мы члене=ния(5)] {[тако=го ро+да(1)] предложе+ний} [наЪсинта+гмы] [бы+лоЪбы испо=льзование(3)] {[ко+мплекса пра=вил(5)] разбо+ра} [наЪсинтакси=ческие компоне+нты(1)].

6. Разметка ПЛС на акцентные единицы (АЕ) и синтаксические синтагмы

Разметка полученной в примере (3.4) последовательности слов на акцентные единицы осуществляется по следующим правилам:

1. Разметка на АЕ осуществляется раздельно для каждой ГСС.

2. Если в ГСС имеются слова со слабым ударением, то каждое из них объединяется в одну АЕ с сильноударным словом, стоящим слева или справа от него.

3. Оставшиеся слова с сильным ударением отмечаются как отдельные АЕ.

После применения указанных правил разметки на АЕ пример (3.4.) перепишется в следующем виде.

(3.5) Пример:

{(Е=сли Ва+м) (необходи=мо активи+ровать)} *[2]*
{(услу+гу переда=чи) (да+нных)} *[2]*
{(дляЪва=шего моби+льного) (но+мера)}. *[2]*

{(Но= благодаря+) (разу=мному сочета+нию) (лека+рств)} *[3]*
{(о=н смо+г) (останови+ть)} *[2]*
{(разви+тие боле=зни)} *[1]*
{(вЪбольшинстве+ слу=чаев)}. *[1]*

{(Тогда= тарифика+ция)} *[1]*
{(Ва=ших звонко+в)} *[1]*
{(начина+ется) (сЪмоме+нта соедине=ния)} *[2]*

{(сЪтелефо+ном абоне=нта)}. *[1]*
{(Идеа=льным реше+нием)} *[1]*
{(пробле+мы члене=ния)} *[1]*
{(тако=го ро+да) (предложе+ний)} *[2]*
{(наЪсинта+гмы)} *[1]*
{(бы+лоЪбы испо=льзование)} *[1]*
{(ко+мплекса пра=вил) (разбо+ра)} *[2]*
{(наЪсинтакси=ческие компоне+нты)}. *[1]*

В примере (3.5) круглыми скобками отмечены полученные АЕ в каждом из ГрС, которые ограничены фигурными скобками и помещены на отдельных строках., причём справа от каждой строки указано количество АЕ в данной ГСС.

Как уже указывалось, основной смысл предварительного разбиения ПЛС на ФЛС и ГСС заключается в том, что теперь свобода разделения ПЛС на СС ограничивается, т.к. граница между СС не может находиться внутри ФЛС или ГСС. В простейшем случае границей каждой СС могут служить границы ГСС. В этом случае, как видно из примера (3.5), каждая СС будет включать различное количество АЕ: от 1-й до 3-х.

Если же требуемый стиль чтения предполагает, что СС должна включать по возможности не менее 2- х АЕ, то в этом случае получим схему членения, показанную на примере (3.6), где справа от каждой строки указано количество АЕ в данной СС.

(3.6) Пример:
{(Е=сли Ва+м) (необходи=мо активи+ровать)} *[2]*
{(услу+гу переда=чи) (да+нных)} *[2]*
{(дляЪва=шего моби+льного) (но+мера)} *[2]*
{(Но= благодаря+) (разу=мному сочета+нию) (лека+рств)} *[3]*
{(о=н смо+г) (останови+ть)} *[2]*
{(разви+тие боле=зни)} {(вЪбольшинстве+ слу=чаев)} *[2]*
{(Тогда= тарифика+ция)} {(Ва=ших звонко+в)} *[2]*
{(начина+ется) (сЪмоме+нта соедине=ния)} {(сЪтелефо+ном абоне=нта)} *[3]*
{(Идеа=льным реше+нием)} {(пробле+мы члене=ния)} *[2]*
{(тако=го ро+да) (предложе+ний)} {(наЪсинта+гмы)} *[3]*
{(бы+лоЪбы испо=льзование)} {(ко+мплекса пра=вил) (разбо+ра)} *[3]*
{(наЪсинтакси=ческие компоне+нты)} *[1]*

Заключение

Описанный алгоритм сегментации на синтагмы используется в составе системы синтеза речи по тексту «МультиФон» [7]. Образцы синтезированной речи будут продемонстрированы во время доклада.

Список литературы

1. Лобанов Б.М. и др. Алгоритмы синтеза просодических характеристик речи по тексту в системе ”Мультифон” // Компьютерная лингвистика и интеллектуальные технологии: труды международной конференции Диалог’2007, М.: Издательский центр РГГУ, 2007. – С. 550-558.

2. Кривнова О.Ф. Фактор речевого дыхания в интонационно-паузальном членении речи // В кн: Лингвистическая полифония / Изд. «Языки славянских культур»– Москва, 2007 – С. 424-443.

3. Lobanov, B., Tsirulnik, L. Statistical study of speaker’s peculiarities of utterances into phrases segmentation // Speech Prosody: proceedings of the 3-rd International conference, Dresden, Germany, May 2–5, 2006. – Dresden, 2006. – V. 2. – P. 557–560.

4. Лобанов Б.М., Сизонов О.Г., Цирульник Л.И. Алгоритм интонационной разметки повествовательных предложений для синтеза речи по тексту / в наст. Сб. трудов Диалог’2008

5. Boguslavsky I., Karnevskaya E., Lobanov B. Generation of Intonation and Accentuation on the of Synthetic Speech on the Basis of Morpho-Syntactic Knowledge // Proc.of International Workshop «Integration of Language and Speech» – Moskow, 1995, pp. 11-28.

6. Валгина Н.С. Современный русский язык / http://www.hi-edu.ru/

7. Лобанов, Б.М. «МУЛЬТИФОН» - система персонализированного синтеза речи по тексту на славянских языках // В кн: Лингвистическая полифония / Изд. «Языки славянских культур»– Москва, 2007 – С. 849-866.