УДК 519.688

В.В. Чемерилов, А.С. Фадеев

Система автоматического разрешения омографии на основе семантической связи слов смежных предложений в текстовом отрывке

Описана методика автоматического разрешения омографии в вопросительных предложениях на основе смысловой связи с ответом. Данная методика впервые рассматривает семантические связи слов за пределами одного предложения при разрешении омографии. В качестве входных данных использовались результаты анализа семантических классов слов семантико-синтаксического анализатора Semsin. На основе данной методики был разработан программный алгоритм, позволяющий автоматически маркировать омографы SSML-тегом, который меняет акустические характеристики синтезированной речи. На специальной тестовой выборке, состоящей из сверхфразовых единств, был проведен эксперимент по оценке качества работы алгоритма, давший положительные результаты. В ходе работы также были выявлены недочеты, которые связаны с ошибками анализа семантических классов слов семантико-синтаксического анализатора Semsin и невозможностью распознавания контекста в пределах одного предложения.

Ключевые слова: семантическая связь предложений, синтез русской речи по тексту, разрешение омографии, омографы, вопросительное предложение, сверхфразовое единство, анализатор Semsin.

doi: 10.21293/1818-0442-2018-21-3-42-48

Одной из важнейших задач, связанных с улучшением качества искусственной русской речи, является разрешение омографии [1]. Омографы [2] – это слова, которые совпадают в написании, но различаются в произношении, к примеру: кирка (церковное) – кирка́ (инструмент), порты (мн.ч. порт) – порты́ (штаны) и т.д.

В силу того, что компьютер рассматривает слово как набор символов и не может четко определить его семантику, автоматизация обработки омографов вызывает существенную сложность, однако имеется несколько методов разрешения неоднозначности в омографах.

В иностранных языках явление омографии не связано с изменением ударного слога в словах, а рассматривает слова одинакового написания с одинаковым ударным слогом, но различным произношением. Кроме того, явление омографии рассматривают наравне с омонимией, когда произношение слов одинаковое, но смысл разный [3], например: cáп (мочь) – cáп (бидон), báт (летучая мышь) – báт (ракета), létтер (письмо) – létтер (точность). В русском языке понятия омонимии и омографии, как правило, имеют четкое разделение [4].

Большинство методик разрешения омографии (как и омонимии) в английском языке базируется на формализованной структуре предложения, где каждая часть речи занимает определенное положение [5]. Описанные методики не рассматривают семантические связи за пределами одного предложения.

Высокая формализация структуры английского языка позволяет активно применять методы машинного обучения для решения задач омографии и омонимии [6].

Один из первых методов разрешения омографии, применяемых в системах автоматизированного синтеза русской речи, заключался в использовании словарей, содержащих устойчивые выражения с

омографом [7], например: кукурузная мука́, со́рок разбойников, строительные леса. Однако качество работы такого метода полностью зависело от размера словаря. В силу сложности составления словаря, который мог охватить все богатство русского языка и рассмотреть все возможные случаи омографии, данный метод только частично решал поставленную задачу.

В связи с появлением синтаксических анализаторов текста, Б.М. Лобановым была проведена серия исследований по разрешению омографии на основе синтаксического и морфологического анализа предложения. Предложенный метод заключался в поиске статических закономерностей между результатами синтаксического и морфологического анализа текстовых отрывков синтаксическим анализатором ЭТАП-3 и проявлением «Ё»-омографов в различных текстах [1]. Метод не решал поставленную задачу в полной мере, так как требовал большого количества дополнительной информации, получаемой на основе морфологического, синтаксического, семантического и прагматического анализа текстов.

В 2013 г. научными сотрудниками Санкт-Петербургского национального исследовательского университета информационных технологий был разработан новый метод снятия омографии [8]. В его основе лежит анализ контекста предложения. В предложении с омографом проводится поиск ключевых слов или выражений. Сначала обрабатываются слова, стоящие в непосредственной близости с омографом, затем анализируется состав предложения целиком. Далее проводится исследование грамматического окружения на уровне классов словоформ. Были введены специальные грамматические правила для формализации данного принципа. Они увеличивают условный «вес» словоформы в зависимости от ее окружения. На основе данного метода был разработан алгоритм и проведена его оценка. В 95% слу-

чаев система успешно разрешила омографию. Основными источниками ошибок стали сложные случаи, для решения которых требовался анализ дистанционных синтаксических связей.

Несмотря на то, что последний метод довольно успешно разрешает омографическую неопределенность, анализ данных в нем (как и в остальных методах) проводится только в пределах одного предложения. Однако существуют случаи, когда предложение содержит недостаточно информации для исследования, т.е. в нем отсутствуют ключевые слова и выражения, например:

Что такое замок?

В приведенном примере анализ ближайшего окружения омографа в рамках одного предложения не даст достаточно данных для того, чтобы разрешить неоднозначность, так как невозможно определить семантику вопроса. Однако можно установить значение омографа, исследуя текстовый отрывок, представляющий собой сверхфразовое единство. Сверхфразовое единство (СЕ) – отрезок текста в форме последовательности двух и более предложений, которые группируются в смысловые блоки на основе структурных связей. В данной работе используется минимальное СЕ, состоящее из двух предложений – вопроса и ответа [9, 10]. Например:

Что такое замок? Крепость.

При сопоставлении слова смежного предложения «Крепость» с омографом «замок» разрешается неоднозначность, что дает возможность установить ударный слог в омографе (первый).

Для разрешения неоднозначности в омографах на основе данных смежного предложения необходимо определить семантическую связь между омографом и словом смежного предложения внутри СЕ [11].

Методика автоматического разрешения омографии на основе семантической связи слов смежных предложений

Пару предложений можно представить в виде двух множеств:

$$P_1 = \{C_{П11}, C_{П12} \dots C_{П1n}\},$$

$$P_2 = \{C_{П21}, C_{П22} \dots C_{П2n}\},$$

где P_1 – множество слов первого предложения; P_2 – множество слов второго предложения. P_1 и P_2 обрабатываются семантическим анализатором текста, который представляет выходные данные в виде списка слов и их семантических характеристик на основе данных семантических словарей – семантического класса S слова C , представленного в виде множества. Каждый элемент множества S определяет значение слова на определенном уровне иерархии:

$$S_c = \{V_{1c}, V_{2c} \dots V_{nc}\}.$$

Например, множество значений слова «самовар» представлено в следующем виде:

$$S_c = \{\text{физический объект, вещь, утварь, инвентарь, кухонный}\}.$$

Некоторые слова в русском языке – омографы и омонимы [1] – могут иметь несколько семанти-

ческих классов, например, для омонима «коса» представлен следующий набор семантических классов:

$$S_{c1} = \{\text{жизнь, части живого, голова, волосы}\};$$

$$S_{c2} = \{\text{физический объект, вещь, утварь, инвентарь, сельскохозяйственный}\};$$

$$S_{c3} = \{\text{природа, природные зоны, ландшафт, берег}\}.$$

А для омографа «замок» представлены следующие семантические классы:

$$S_{c1} = \{\text{физический объект, вещь, утварь, инструменты}\};$$

$$S_{c2} = \{\text{физический объект, поселения, постройка, жилье, дом}\}.$$

Для определения семантической связи между омографом и ключевым словом смежного предложения (ключевым относительно исследуемого омографа) проводится поиск ключевого слова во втором предложении. В данной работе рассматриваются три модели предложений, которые легко поддаются формализации и машинной автоматизированной обработке:

1. Предложение содержит только одно слово – ключевое (КС).

$$P_2 = \{КС\}.$$

Например (предлоги не учитываются):

Где гвоздики? У молотка.

$$P_1 = \{\text{где, гвоздики}\}, P_2 = \{\text{у, молотка}\},$$

$$КС = \{\text{молотка}\}.$$

2. Предложение содержит только два слова: вспомогательное (ВС) и ключевое.

$$P_2 = \{ВС, КС\}.$$

Вспомогательное слово представляет собой согласие или отрицание. Например:

Он купил муку? Нет, пшено.

$$(КС = \{\text{пшено}\}).$$

3. Предложение, в котором присутствует уточняющее слово (УС):

$$P_2 = \{C_{П21}, C_{П22}, УС, КС, \dots C_{П2n}\}.$$

Уточняющее слово выражается наречиями «точно», «конечно», «только» и т.д. Оно будет стоять перед ключевым словом, так как собеседник делает акцент именно на этом слове. Пример:

Где купить гвоздики? Их можно купить только в цветочном магазине.

$$(КС = \{\text{цветочном}\}).$$

Для определения семантической близости между омографом и ключевым словом смежного предложения необходимо сопоставить семантические классы омографа с семантическим классом ключевого слова смежного предложения. Для числового представления семантической близости омографа и ключевого слова смежного предложения был введен коэффициент g_{oi} , который определяется по следующей формуле:

$$g_{oi} = \frac{|S_{oi} \cap S_{kc}|}{\max(|S_{oi}|, |S_{kc}|)}, \quad (1)$$

где g_{oi} – весовой коэффициент семантической близости омографа, который имеет порядковый номер i , и

ключевого слова смежного предложения; S_{oi} – семантический класс омографа с порядковым номером i ; S_{kc} – семантический класс ключевого слова смежного предложения. Пересечение множеств S_{oi} и S_{kc} позволяет выделить те значения омографа и ключевого слова смежного предложения, которые совпали. Выбор множества с наибольшей мощностью позволяет исключить ситуацию, в которой один из исследуемых семантических классов содержит недостаточно значений слова. Например, множеству слов СЕ «Где гвоздики? У молотка», соответствуют следующие семантические классы (предлоги, частицы и союзы опускаются, так как они не имеют семантических классов):

Для слова C_{11} = «гвоздики»:

S_{o1} = {физический объект, вещь, утварь, детали};

S_{o2} = {жизнь, пища, блюда, соусы};

S_{o3} = {природа, растения, цветы}.

Для ключевого слова смежного предложения «молотка»:

S_{kc} = {физический объект, вещь, утварь, инструменты, дляремонта и строительства}.

При сопоставлении элементов семантических классов омографа и ключевого слова смежного предложения рассчитываются коэффициенты семантической близости:

$$g_{o1}=0,6, \quad g_{o2}=0, \quad g_{o3}=0.$$

Наиболее семантически близкое к ключевому слову смежного предложения значение омографа будет иметь наибольшее значение коэффициента семантической близости. В рассматриваемом примере максимальный весовой коэффициент имеет порядковый номер 1.

$$\max(g_{oi}) = g_{o1} = 0,6.$$

Если между омографом и ключевым словом смежного предложения существует семантическая связь, то значение омографа определяется семантическим классом с максимальным коэффициентом семантической близости $\max(g_{oi})$. В рассматриваемом примере тип семантической связи определяется как «физический объект, вещь, утварь». Эта связь была установлена между словами «гвоздики» и «молотка», значение омографа определяется как

S_{o1} = {физический объект, вещь, утварь, детали}.

Определив семантический класс омографа, можно установить ударный слог в этом слове. В рассмотренном примере ударение в омографе падает на первый слог (гвоздики). Для исключения семантически слабосвязанных слов, между которыми не наблюдается семантической связи (т.е. ключевое слово смежного предложения никак не связано с рассматриваемым омографом), был введен коэффициент значимости p , который оценивает наличие семантической связи между словами. Например, слова «леса» и «Артем» имеют следующие семантические классы:

Для омографа «леса»:

S_{o1} = {физический объект, материалы, строительные};

S_{o2} = {природа, природные зоны, ландшафт, лес}.

Для слова «Артем»:

S_c = {физический объект, живой человек, личность, ФИО, имя}.

Несмотря на то, что у данных слов совпадают первые элементы семантических классов, S_{o1} и S_c , они не связаны по смыслу. Коэффициент значимости рассчитывается эмпирическим путем. Если $\max(g_{oi}) \geq p$, то верный в данном контексте семантический класс омографа определяется как семантический класс, порядковый номер которого совпадает с порядковым номером максимального коэффициента семантической близости g_{oi} . Если $\max(g_{oi}) < p$, то семантическая связь между омографом и ключевым словом смежного предложения отсутствует.

Система автоматического разрешения омографии на основе семантической связи слов смежных предложений

Для реализации описанной методики была разработана система автоматического разрешения омографии на основе семантической связи слов смежных предложений. Функциональная схема системы приведена на рис. 1.

На вход системы подается пара предложений. С помощью блока разбиения текста на предложения они разбиваются на два предложения. Каждое предложение преобразуется в список слов. Список слов предложения, содержащего омограф (P_1), передается блоку выделения омографа, который на основе словаря омографов [12] выделяет во множестве слов P_1 омограф. Выделенный омограф передается блоку

Рис. 1. Функциональная схема системы автоматического разрешения омографии на основе семантической связи слов смежных предложений

семантического анализа омографа для определения списка его семантических классов. Список слов второго предложения (Π_2) передается блоку выделения ключевого слова предложения, который на основе описанных в методике моделей предложений выделяет ключевое слово предложения. Ключевое слово предложения передается блоку семантического анализа ключевого слова предложения для определения его семантического класса.

В качестве инструмента для проведения автоматического семантического анализа слов использовался семантико-синтаксический анализатор Semsin [13–15], основанный на семантическом словаре Тузова [16]. Набор семантических классов омографа (S_{oi}) и семантический класс ключевого слова предложения (S_{kc}) передаются блоку расчета весовых коэффициентов семантической близости слов, кото-

рый рассчитывает весовые коэффициенты g_{oi} , используя формулу (1). Весовые коэффициенты g_{oi} передаются блоку определения семантической связи между омографом и ключевым словом предложения и выделения ударного слога в омографе. Данный блок определяет наличие семантической связи между омографом и ключевым словом предложения и номер ударного слога в омографе ($n_{уд}$). Номер ударного слога в омографе определяется семантическим классом омографа. Для автоматизации определения $n_{уд}$ был разработан специальный словарь, примеры элементов которого представлены в табл. 1.

В первом столбце табл. 1 находится омограф, во втором столбце – один из множества семантических классов омографа – одно из его семантических значений. Третий столбец содержит номер ударного слога в омографе.

Таблица 1

Пример определения ударного слога в омографе на основе его семантического класса

№	Омограф	Семантический класс омографа (S_{oi})	Номер ударного слога в омографе ($n_{уд}$)
1	Атлас	Знания, литература, книга	1 (атлас)
2	Атлас	Физический объект, материалы, волокно, ткани	2 (атлас)
3	Замок	Физический объект, поселения, постройка, жилье, дом	1 (замок)
4	Замок	Физический объект, вещь, утварь, инструменты	2 (замок)

Номер ударного слога в омографе передается блоку SSML-разметки текста, который отмечает омограф специальным SSML-маркером. SSML-разметка обработанного СЕ может быть передана синтезатору русской речи, который поддерживает стандарты SSML [17] для конвертации в искусственную речь.

Наличие семантической связи между омографом и ключевым словом предложения характеризуется параметром Con_o . Если между омографом и ключевым словом смежного предложения присутствует семантическая связь, то значение Con_o равно единице, иначе – нулю.

На базе реализованной системы был проведен эксперимент с использованием выборки, состоящей из 110 элементов – пар предложений. Выборка была разработана с использованием национального корпуса русского языка [18]. Первое предложение содержало исследуемый омограф, второе – ключевое слово, связанное по смыслу с данным омографом. Примеры пар предложений и результатов их обработки приведены в табл. 2.

Второй столбец табл. 2 содержит пары предложений, использовавшихся для проведения эксперимента, третий и четвертый столбцы содержат результаты обработки предложений системой.

Таблица 2

Примеры результатов обработки пар предложений системой

№	Пара предложений	Результаты автоматической обработки пар предложений системой	
		Con_o (наличие семантической связи)	$n_{уд}$ (номер ударного слога в омографе)
1	Какой орган? Сердце	1 (есть)	1 (орган)
2	Какой орган? Симфонический	1 (есть)	2 (орган)
3	Где атлас? У географа	1 (есть)	1 (атлас)
4	Где атлас? Помидор	0 (нет)	– (отсутствует)

Оценка результатов работы системы автоматического разрешения омографии на основе семантической связи слов смежных предложений

Для возможности оценки качества работы системы был разработан дополнительный программный модуль. Функциональная схема данного модуля представлена на рис. 2.

На вход модуля подается элемент выборки, который обрабатывается вручную экспертом и автоматически системой. И эксперт, и система определяли два параметра: наличие семантической связи между омографом и ключевым словом смежного предло-

жения (Con_o) и номер ударного слога в омографе ($n_{уд}$). Эксперт анализировал данные на основе личного опыта работы со связным текстом и разрешением омографии, система использовала описанную методику.

В качестве экспертов были привлечены сотрудники отделения русского языка школы базовой инженерной подготовки Томского политехнического университета.

Все множество исследуемых пар предложений A было разбито на пять подмножеств:

1. Множество, содержащее пары предложений, в которых эксперт вручную разрешил омогра-

фию ($n_{уд}$) и выделил семантическую связь (Con_o) между омографом и ключевым словом смежного предложения (A_{ST}).

2. Множество, содержащее пары предложений, в которых эксперту не удалось выделить семантические связи между омографом и ключевым словом смежного предложения (и соответственно определить номер ударного слога в омографе) в силу их отсутствия (A_{SF}).

Рис. 2. Функциональная схема модуля оценки качества работы системы автоматического разрешения омографии на основе семантической связи слов смежных предложений

Для количественной оценки результатов работы системы были предложены четыре меры.

Мера первая – доля совпадения данных о парах предложений, содержащих Con_o (без учета $n_{уд}$), полученных на основе оценки эксперта и системы

$$\varepsilon_{AT} = \frac{|A_{AT} \cap A_{ST}|}{|A_{ST}|} \times 100\%.$$

Мера вторая – доля совпадения данных о парах предложений, которые содержат и Con_o и $n_{уд}$, которые были получены на основе оценки эксперта и системы:

$$\varepsilon_{ATT} = \frac{|A_{ATT} \cap A_{ST}|}{|A_{ST}|} \times 100\%.$$

Мера третья – доля совпадения данных о парах предложений, в которых отсутствуют семантические связи между омографом и ключевым словом смежного предложения. Они были получены на основе оценки системы и эксперта:

$$\varepsilon_{AF} = \frac{|A_{AF} \cap A_{SF}|}{|A_{SF}|} \times 100\%.$$

Мера четвертая – общая доля совпадения оценок системы и эксперта.

$$\varepsilon = \frac{|(A_{ATT} \cap A_{ST}) \cup (A_{AF} \cap A_{SF})|}{|A|} \times 100\%.$$

Для используемой в эксперименте выборки были вычислены значения мер:

$$\varepsilon_{AT} = 82,3\%; \varepsilon_{ATT} = 76,2\%; \varepsilon_{AF} = 87,5\%; \varepsilon = 74,1\%.$$

Значения всех четырех мер показывают, что в большинстве случаев вычисленные системой в автоматическом режиме значения Con_o (наличие семантической связи) и $n_{уд}$ (номер ударного слога в омографе) совпали с соответствующими значениями, предложенными экспертом.

По результатам проведенного эксперимента были сделаны следующие выводы:

3. Множество, содержащее пары предложений, в которых система автоматически определила Con_o , но не смогла определить $n_{уд}$ (A_{AT}).

4. Множество, содержащее пары предложений, в которых система не смогла определить $n_{уд}$ и Con_o (A_{AF}).

5. Множество, содержащее пары предложений, в которых система смогла определить и Con_o , и $n_{уд}$ (A_{ATT}).

Семантико-синтаксический анализатор Semsin не всегда формирует полный список семантических классов для исследуемых омографов. По этой причине система не может определить $n_{уд}$ при определении семантической связи Con_o . Например, в СЕ:

Где простынь? На кровати.

Для омографа «простынь» Semsin подобрал следующий список семантических классов:

$S_{o1} = \{\text{жизнь, болезни, болезни человека}\};$

$S_{o2} = \{\text{физический объект, температура, тепло}\}.$

Однако данный омограф имеет еще один семантический класс:

$S_o = \{\text{физический объект, вещь, белье, постель}\}.$

Так как ключевое слово смежного предложения имеет семантический класс

$S_{kc} = \{\text{физический объект, вещь, мебель}\},$

то отсутствие семантического класса S_o в списке семантических классов омографа влияет на результат работы системы (так как S_o наиболее близок к S_{kc}).

При сопоставлении семантических классов омографа с семантическим классом ключевого слова смежного предложения система определила неверное значение Con_o . Максимальный весовой коэффициент семантической близости для омографа был меньше коэффициента значимости $\max(g_{oi}) > p$. Для решения данной проблемы необходимо получить дополнительную информацию о зависимости коэффициента значимости от исследуемых семантических классов омографов.

3. В определенных случаях сопоставление семантических классов слов не даст нужного результата, так как слова могут быть связаны по смыслу, но не удовлетворять условию $\max(g_{oi}) > p$. Для решения данной проблемы необходимо провести более глубокий анализ контекста, в том числе и за пределами исследуемого текстового отрывка. Например:

Где жаркое? У повара.

Для омографа «жаркое» анализатор Semsin подобрал следующий список семантических классов:

$S_{o1} = \{ \text{физический объект, природа,}$
 погода, температура воздуха };

$S_{o2} = \{ \text{жизнь, пища, блюда, нерастительные} \}.$

Ключевое слово смежного предложения «повар» имеет семантический класс

$S_{kc} = \{ \text{физический объект, живой человек,}$
 личность, профессия, повар }.

При сопоставлении семантических классов S_{o2} и S_{kc} система определит значение $Con_o = 0$, однако в данном контексте между словами «жаркое» и «повар» существует семантическая связь.

Заключение

В данной статье описывается методика автоматического разрешения неоднозначности в омографах на основе данных смежного предложения. На базе методики построена система автоматического разрешения омографии на основе семантической связи слов смежных предложений. На специальной выборке проведен ряд экспериментов и получена оценка качества работы системы. В результате проведенных экспериментов были выявлены недостатки системы, связанные с невозможностью анализа слов в рамках контекста и неполнотой информации при автоматическом семантическом анализе.

Система автоматического разрешения неопределенности в омографах может быть использована в качестве дополнительного модуля в лингвистическом процессоре системы синтеза русской речи для повышения качества искусственной речи.

Анализ данных за пределами одного предложения позволяет существенно повысить качество автоматической разметки текста для синтеза речи.

Литература

1. Лобанов Б.М. Проблема разрешения «Ё»-омографов при синтезе речи по тексту // Тр. междунар. конф. «Компьютерная лингвистика и интеллектуальные технологии». – М.: Наука, 2009. – С. 330–338.
2. Розенталь Д.Э. Словарь-справочник лингвистических терминов / Д.Э. Розенталь, М.А. Теленкова. – М.: Просвещение, 1985. – 399 с.
3. Common Homographs in English (With Examples) [Электронный ресурс]. – Режим доступа: <https://7esl.com/homographs/>, свободный (дата обращения: 15.09.2018).
4. О разной трактовке соотношения омонимов, омофонов, омографов [Электронный ресурс]. – Режим доступа: <http://www.russkayarech.ru/files/issues/2012/3/09-ruchimskaja.pdf>, свободный (дата обращения: 21.09.2018).
5. Homograph Disambiguation in Text-to-Speech Synthesis [Электронный ресурс]. – Режим доступа: http://www.cs.toronto.edu/~gpenn/csc2518/progress_chp12.pdf, свободный (дата обращения: 12.09.2018).
6. The automatically built up homograph dictionary – a component of a dynamic lexical system [Электронный ресурс]. – Режим доступа: <http://www.aclweb.org/anthology/C73-2032>, свободный (дата обращения: 17.09.2018).
7. Рыбин С.В. Синтез речи: учеб. пособие. – СПб.: Университет ИТМО, 2014. – 92 с.
8. Хомицевич О.Г. Использование лингвистического анализа для нормализации текста и снятия омонимии в системе синтеза русской речи / О.Г. Хомицевич, С.В. Ры-

бин, И.М. Аничкин // Изв. вузов. Приборостроение. Тематический выпуск «Речевые информационные системы». – 2013. – № 2. – С. 42–46.

9. Солганик Г.Я. Синтаксическая стилистика. Сложное синтаксическое. – М.: Высшая школа, 1991. – 182 с.

10. Левковская Н.А. В чём различие между сверхфразовым единством и абзацем? // Филологические науки. – 1980. – № 1. – С. 75–78.

11. Филимонов О.И. Скрепа-фраза в языке. – М.: АГРУС, 2017. – 170 с.

12. Гребенкова Ю.Н. Словарь омографов русского языка. – Ливны: Издатель Г.В. Мухаметов, 2012. – 278 с.

13. Боярский К.К. Выявление анафорических отношений при автоматическом анализе текста. / К.К. Боярский, Е.А. Каневский // Научно-технический вестник информационных технологий механики и оптики. – 2013. – № 5. – С. 108–112.

14. Боярский К.К. Предсинтаксический модуль в анализаторе Semsin / К.К. Боярский, Е.А. Каневский // Интернет и современное общество. – 2013. – № 1. – С. 280–286.

15. Боярский К.К. Семантико-синтаксический анализатор Semsin / К.К. Боярский, Е.А. Каневский // Научно-технический вестник информационных технологий механики и оптики. – 2015. – № 5. – С. 869–876.

16. Тузов В.А. Компьютерная семантика русского языка. – СПб.: Изд-во СПбГУ, 2004. – 400 с.

17. Speech Synthesis Markup Language (SSML) Version 1.1 [Электронный ресурс]. – Режим доступа: <http://www.w3.org/TR/speech-synthesis/>, свободный (дата обращения: 15.05.2018).

18. Гришина Е.А. Национальный корпус русского языка: 2006–2008. Новые результаты и перспективы // Нестор-История. – 2009. – № 6. – С. 150–174.

Чемерилов Владимир Викторович

Аспирант отделения информационных технологий Инженерной школы информационных технологий и робототехники (ИШИТР) Томского политехнического университета (ТПУ) Ленина пр-т, д. 30, г. Томск, Россия, 634050
 Тел.: +7-923-410-93-87
 Эл. почта: Vchemerilov@gmail.com

Фадеев Александр Сергеевич

Канд. техн. наук, доцент отделения информационных технологий ИШИТР ТПУ Ленина пр-т, д. 30, г. Томск, Россия, 634050
 Тел.: +7 (382-2) 70-63-30
 Эл. почта: fas@tpu.ru

Chemerilov V.V., Fadeev A.S.

System of automatic homography resolution based on the semantic connection words of adjacent sentences in a text passage

This paper describes a technique to automatically resolve homographs in interrogative sentences based on a semantic connection with the answer. The results of the semantic analysis of the Semsin parser are used as input data. Based on this technique, a software algorithm was developed that allows to mark automatically homographs with an SSML tag that changes the acoustic characteristics of synthesized speech. On

a special test sample consisting of phrase units, an experiment conducted to evaluate the quality of the algorithm, which gave positive results. In the course of work, shortcomings related to errors in the semantic analysis of the Semsin parser and the inability to recognize the context within a single sentence were revealed.

Keywords: Russian text-to-speech systems, semantic connection of sentences, homography resolution, homographs, interrogative sentence, super phrase units, Semsin parser.

doi: 10.21293/1818-0442-2018-21-3-42-48

References

1. Lobanov B.M. The problem of resolving «E»-homographs in the synthesis of speech in the text. *Proceedings of the International Conference «Computer Linguistics and Intellectual Technologies»*, 2009, pp. 330–338.
2. Rosenthal D.E., Telenkova M.A. *Slovar-spravochnik lingvisticheskikh terminov*. Enlightenment, 1985, 399 p.
3. Common Homographs in English (With Examples). URL: <https://7esl.com/homographs/>.
4. About different interpretation of homonyms, homophones, homographs. URL: <http://www.russkayarech.ru/files/issues/2012/3/09-ruchimskaja.pdf>
5. Homograph Disambiguation in Text-to-Speech Synthesis. URL: http://www.cs.toronto.edu/~gpenn/csc2518/progress_chp12.pdf
6. The automatically built up homograph dictionary – a component of a dynamic lexical system. URL: <http://www.aclweb.org/anthology/C73-2032>.
7. Rybin S.V. *Sintez rechi. Uchebnoe posobie*. St. Petersburg: ITMO University, 2014, 92 p.
8. Khomitsevich O.G., Rybin S.V., Anichkin I.M. Use of linguistic analysis for the normalization of the text and the removal of homonymy in the system of synthesis of Russian speech. *Thematic issue «Speech Information Systems»*, 2013, pp. 42–46.
9. Solganik G.Y. *Sintaksicheskaya stilistika. Slozhnoe sintaksicheskoe*, 1991, 182 p.
10. Levkovskaya N.A. What is the difference between superphrase unity and a paragraph? *Philological Sciences*, 1980, pp. 75–78.
11. Filimonov O.I. *Skrepa-fraza v yazyke*. AGRUS, 2017, 170 p.
12. Grebenkova Y.N., Kanevsky E.A. *Slovar omografov russkogo yazyka*, 2012, 278 p.
13. Boyarsky K.K. Identification of anaphoric relations in automatic text analysis. *Scientific and Technical Herald of Information Technologies of Mechanics and Optics*, 2013, pp. 108–112.
14. Boyarsky K.K., Kanevsky E.A. Presyntactic module in the Semsin. *Inter-no and modern society*, 2013, pp. 280–286.
15. Boyarsky K.K., Kanevsky E.A. Parser Semsin *Scientific and Technical Herald of Information Technologies of Mechanics and Optics*, 2015, pp. 869–876.
16. Tuzov V.A. *Kompyuternaya semantika russkogo yazyka*. St. Petersburg: Publishing house of St. Petersburg State University, 2004, 400 p.
17. Burnett D., Walker M., Hunt A. Speech Synthesis Markup Language (SSML) Version 1.1, W3C Recommendation, 2014.
18. Grishina E. A. National building of the Russian language: 2006–2008. New results and prospects. *Nestor-History*, 2009, pp. 150–174.

Vladimir V. Chemerilov

PhD student, School of Information Technology and Robotics, Tomsk Polytechnic University (TPU)
30, Lenin pr., Tomsk, Russia, 634050
Phone: +7-923-410-93-87
Email: Vchemerilov@gmail.com

Alexander S. Fadeev

PhD, Associate Professor, School of Information Technology and Robotics, Tomsk Polytechnic University (TPU)
30, Lenin pr., Tomsk, Russia, 634050
Phone: +7 (382-2) 70-63-30
Email: fas@tpu.ru