

УДК 519.688

АЛГОРИТМ АВТОМАТИЧЕСКОГО ВЫДЕЛЕНИЯ ИНТОНАЦИОННОГО ЦЕНТРА В ВОПРОСИТЕЛЬНОМ ПРЕДЛОЖЕНИИ БЕЗ ЯВНОГО ВОПРОСИТЕЛЬНОГО СЛОВА НА ОСНОВЕ СЕМАНТИЧЕСКИХ СВЯЗЕЙ ПРЕДЛОЖЕНИЙ ПРИ АВТОМАТИЧЕСКОМ АНАЛИЗЕ ТЕКСТА ДЛЯ СИНТЕЗА РУССКОЙ РЕЧИ

Чемерилов В.В., Фадеев А.С., Мишунин О.Б.

*ФГБОУ ВПО «Национальный исследовательский Томский политехнический университет», Томск,
e-mail: vchemerilov@gmail.com*

В статье описаны правила выделения интонационного центра в вопросительном предложении без явного вопросительного слова с использованием его семантической связи с ответом на основе данных семантического анализа семантико-синтаксического анализатора Semsin. На основе приведенных правил был создан программный алгоритм, позволяющий вставлять в вопросительное предложение перед ударным словом специальный маркер, меняющий голосовые характеристики синтезатора. Был проведен ряд экспериментов, подтвердивших работоспособность алгоритмов на тестовой выборке сверхфразовых единств и пар предложений, кроме того, были выявлены недостатки, связанные с ошибками семантического анализа современных анализаторов текста и невозможностью определения семантического класса слова на основе контекста за пределами одного предложения. Использование семантических данных оказалось недостаточно для качественного решения поставленных задач. Наряду с семантическим анализом необходимо использовать морфологический и синтаксический анализы текста.

Ключевые слова: автоматический синтез русской речи, прагматика, семантическая связь предложений, интонационный центр, вопросительное предложение, сверхфразовое единство, анализатор Semsin

THE AUTOMATIC SELECTION ALGORITHM OF INTONATION CENTER IN A QUESTION WITHOUT AN EXPRESSIVE QUESTION WORDS USING BASIS OF SENTENCES SEMANTIC CONNECTIONS WITH THE AUTOMATIC ANALYSIS OF THE TEXT FOR RUSSIAN SPEECH SYNTHESIS

Chemerilov V.V., Fadeev A.S., Mishunin O.B.

National Research Tomsk Polytechnic University, Tomsk, e-mail: vchemerilov@gmail.com

The article describes the rules for allocation of the intonation center in interrogative sentence without an expressive question word using its semantic connection with the answer based on semantic analysis of the Semsin parser. Based on these rules, a software algorithm was created that allows inserting into the question before the expressive word a special marker that changes the voice characteristics of the synthesizer. A number of experiments were performed that confirmed the operability of the algorithms on the test sample of hyper phase units and pairs of sentences, in addition, shortcomings related to errors in the semantic analysis of modern text analyzers and the inability to determine the semantic class of words on the basis of the context outside of one sentence were revealed.

Keywords: Russian synthesis speech, pragmatics, semantic connection of sentences, intonation center, question sentence, hyper phase, parser Semsin

В процессе диалога, для уточнения информации, человек обычно использует различные интонационные средства. Одним из примеров такого явления может быть интонационный центр в вопросительных предложениях – слово, которое придает весь смысл вопросу. Вопросительные предложения относительно интонационного центра можно поделить на два типа: с явным вопросительным словом и неявным. Интонационный центр вопросительного предложения с явным ударным словом сконцентрирован на одном конкретном слове – на другие слова вопроса интонация падать не может [1]. Пример:

О ком она думает? Кто он такой?

В вопросительных предложениях с неявным вопросительным словом положение

интонационного центра зависит от семантики вопроса. Пример:

Ты сегодня ела кашу? Ты СЕГОДНЯ ела кашу?

Ты сегодня ЕЛА кашу? Ты сегодня ела КАШУ?

В первом случае интонационный центр падает на местоимение. Собеседник спрашивает второго участника беседы о том, кто именно потреблял сегодня продукт (ты ел сегодня кашу или не ты). Во втором случае логическим ударением выделяется наречие, обозначающее дату. Спрашивающий хочет узнать, когда собеседник употреблял в своем рационе продукт. В третьем случае специфической интонацией выделяется глагол – собеседника спрашивают о том, произошел ли процесс или нет. В четвертом случае спраши-

вает о виде продукта, который собирался употребить ответчик. Таким образом, одно и то же предложение формирует четыре ситуации и интонационный центр может находиться на любом слове данного предложения.

Современные системы синтеза речи способны выделить интонационный центр в предложениях с явным вопросительным словом [2]. Однако выделить ударное слово в вопросе с неопределенным интонационным центром они не в состоянии. В пределах вопросительного предложения с неявным ударным словом разрешить неопределенность интонационного центра невозможно, так как невозможно установить ситуацию. Ее можно установить, исследуя сверхфразовое единство.

Сверхфразовое единство (СЕ) – отрезок речи в форме последовательности двух и более предложений, которые группируются на основе структурных связей в смысловые блоки. В данной работе используется минимальное СЕ, состоящее из двух предложений – вопроса и ответа [3, 4].

Для решения задачи определения интонационного центра вопроса необходимо выделить семантическую связь [5] между предложениями.

Алгоритм автоматического выделения интонационного центра в вопросительных предложениях без явного вопросительного слова

Для реализации алгоритма использовался семантико-синтаксический анализатор Semsin [6–8], который проводил:

1. Морфологический анализ каждого слова в предложении. Определял часть речи, род, число, падеж и некоторые индивидуальные признаки частей речи.
2. Синтаксический анализ. Строил синтаксическое дерево зависимости каждого анализируемого предложения.
3. Семантический анализ. На основе семантического словаря В.А. Тузова [9] система определяла семантический класс S

слова C , представленный в виде множества. Каждый элемент такого множества S представляет собой значение слова на определенном уровне иерархии.

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

Например, для слова «ноябрь» Semsin подобрал следующее множество значений:

$$S_c = \{\text{Время, Конкретное время, Месяц}\}.$$

На рисунке приведена функциональная схема системы выделения интонационного центра в вопросительном предложении без явного вопросительного слова:

На вход алгоритма подается СЕ, состоящее из вопроса и ответа. Вопрос и ответ можно представить в виде множеств, состоящих из слов.

$$P_v = \{C_{v1}, C_{v2}, \dots, C_{vn}\},$$

$$P_o = \{C_{o1}, C_{o2}, \dots, C_{on}\}.$$

СЕ обрабатывается семантико-синтаксическим анализатором Semsin, который представляет выходные данные в виде списка слов ответа и вопроса и их морфологическими, синтаксическими и семантическими характеристиками. На первом этапе алгоритма определяется ключевое слова ответа (слово, определяющее смысл ответа). Для этого было выбрано три различные модели ответа (задача заведомо была ограничена и было выбрано три модели ответа, которые легко поддаются формализации и машинной автоматизированной обработке):

1. Прямой ответ. Состоит из одного ключевого слова.

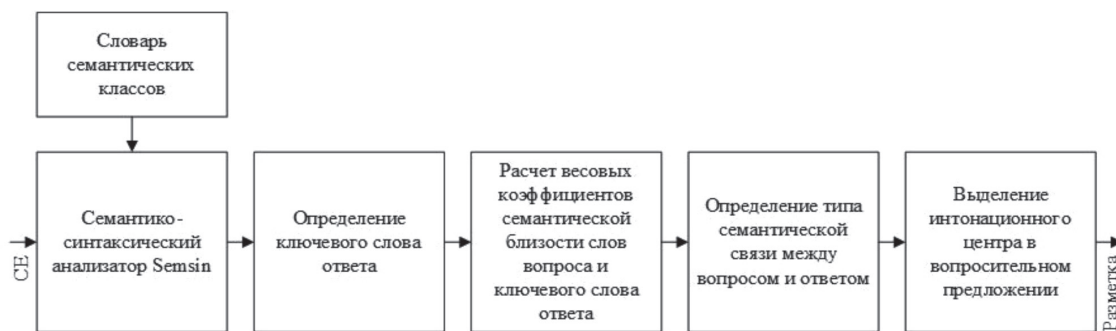
$$P_o = \{KC\}.$$

Например:

Олимпиаду выиграла Анна? Анна.

2. Прямой ответ состоит из двух слов: вспомогательного и ключевого.

$$P_o = \{BC, KC\}.$$



Функциональная схема системы выделения интонационного центра в вопросительном предложении без явного вопросительного слова

Вспомогательное слово выражает согласие или отрицание. Например:

Андрей вчера был у нас в гостях? Конечно Андрей.

Алексей работает программистом? Нет, юристом.

3. Косвенный ответ, в котором присутствует уточняющее слово.

$$P_o = \{C_{o1}, C_{o2}, UC, KC, \dots C_{on}\}.$$

Уточняющее слово может быть выражено наречиями «конечно», «точно», «только» и т.д. В этом случае ключевое слово будет стоять после слова уточнения, так как ответчик делает акцент именно на этом слове. Пример:

Ты сегодня пойдешь на физику? Нет, я сегодня пойду только на математику.

Для возможности определения семантической близости между словами был введен коэффициент, который рассчитывается по следующей формуле (выражает семантическую близость слов исследуемых предложений).

$$g_i = \frac{|S_{свi} \cap S_{кc}|}{\max(|S_{свi}|, |S_{кc}|)},$$

где g_i – весовой коэффициент семантической близости слова вопроса с порядковым номером i и ключевого слова ответа, $S_{свi}$ – семантический класс слова вопроса с порядковым номером i , $S_{кc}$ – семантический класс ключевого слова ответа.

На третьем этапе алгоритма из списка весов определяется элемент с максимальным значением семантической близости $\max(g_i)$, который сравнивается с коэффициентом значимости (p). Коэффициент значимости рассчитывается эмпирическим путем. В данной работе он равен 0,5. Если $\max(g_i) \geq p$, то тип семантической связи вопроса и ответа определяется как тип семантического класса на последнем уровне иерархии, где совпадают семантика слова вопроса и ключевого слова ответа. Если $\max(g_i) < p$, то семантической связи между вопросом и ответом нет, то есть эти предложения не составляют СЕ. Если между вопросом и ответом существует семантическая связь, то интонационный центр вопросительного предложения концентрируется на слове с максимальным коэффициентом семантической близости $\max(g_i)$.

Оценка качества работы алгоритма автоматического выделения интонационного центра в вопросительных предложениях без явного вопросительного слова

Статистический анализ качества работы системы автоматического определения

семантической связи и выделения интонационного центра в вопросительном предложении проводился на основе исследования выборки, содержащей 81 СЕ и 19 пар предложений, между которыми смысловая связи не наблюдалось. Данная выборка была разработана на основе национального корпуса русского языка [10]. СЕ состояла из вопроса (он повторялся 5 раз, но порядок слов в нем был изменен в каждом случае) и ответа (в каждом случае был разный ответ). На первом этапе работы выборка предложений была обработана экспертом вручную: данные эксперта включали в себя наличие/отсутствие смысловой связи между предложениями и местоположение интонационного центра в вопросе (вопросительного слова). На втором этапе выборка была обработана системой. Качество автоматической разметки определялось следующим образом. Все множество исследуемых предложений A было разбито на пять подмножеств:

1. Множество пар предложений, в которых эксперт вручную выделил семантические связи и интонационные центры (A_{AT}).

2. Множество пар предложений, в которых эксперт не смог выделить семантические связи в силу их отсутствия (A_{AE}).

3. Множество пар предложений, в которых система автоматически определила семантические связи и интонационные центры (A_{MTT}).

4. Множество пар предложений, в которых система смогла выделить семантические связи, но не смогла определить интонационные центры (A_{MTF}).

5. Множество пар предложений, в которых система не смогла выделить семантические связи (A_{MFF}).

По данным множествам вычислялись четыре меры. Первая мера – доля совпадения данных о парах предложений, содержащих семантические связи и интонационные центры, полученных на основе оценки эксперта и системы.

$$\epsilon_{TT} = \frac{|A_{AT} \cap A_{MTT}|}{|A_{AT}|} * 100\%.$$

Вторая мера – доля совпадения пар предложений, в которых данные эксперта и системы о семантических связях совпадают, но разнятся данные о местоположении интонационных центров.

$$\epsilon_{TF} = \frac{|A_{AT} \cap A_{MTF}|}{|A_{AT}|} * 100\%.$$

Третья мера – доля совпадения данных о парах предложений, в которых отсутству-

ют семантические связи, полученные на основе оценки эксперта и системы.

$$\epsilon_{FF} = \frac{|A_{MFF} \cap A_{AF}|}{|A_{MFF}|} * 100\%.$$

Четвертая мера – общая доля совпадения оценок системы и эксперта.

$$\epsilon = \frac{|(A_{AT} \cap A_{MTT}) \cup (A_{MFF} \cap A_{AF})|}{|A|} * 100\%.$$

Результаты обработки данных были сведены в таблицу.

Первый столбец содержит СЕ или пары предложений для оценки. Второй и третий столбец содержит данные оценки предложений экспертом, третий и четвертый – системой. В первой и четвертой строке оценки эксперта и системы полностью совпали, в третьей – не совпали. Во второй строке данные о наличии семантических связей совпали, но не совпали данные об интонационных центрах.

Результаты оценки качества алгоритма автоматического выделения интонационного центра в вопросительных предложениях без явного вопросительного слова

$$\epsilon_{TT} = 83,95\%; \epsilon_{TF} = 7,4\%; \epsilon_{FF} = 94,73\%; \\ \epsilon = 89,34\%; \epsilon^* = 90,56\%.$$

В большинстве случаев система правильно определила наличие семантической связи и местоположение интонационного центра.

Анализ работы системы выявил ряд ошибок и недостатков:

1. Semsin не всегда дает полную информацию о семантических классах слова, например, СЕ «Олег пил сегодня чай? Нет, Максим.» программа обработала следующим образом: ключевое слово ответа «Максим» имеет только один семантический класс – «Физический объект, Неодушевленное, Вещь, Утварь, Инструменты, Оружие стрелковое, Название», однако данное сло-

во имеет еще как минимум один семантический класс – «Физический объект, Живой человек, Личность, ФИО, Имя». Таким образом, сопоставляя данные семантического анализа ключевого слова ответа «Максим» и ключевого слова вопроса «Олег», система не смогла выделить семантическую связь несмотря на ее присутствие в данном СЕ.

2. Семантический словарь В.А. Тузова нельзя использовать для нахождения семантической близости глаголов, например, исследуя СЕ «Воздушный шарик улетел? Нет, лопнул.» система разделила семантические классы слов «улетел» и «лопнул» на втором уровне иерархии, и алгоритм не смог выделить семантическую близость данных слов для нахождения семантической связи.

3. Данных семантического анализа в некоторых случаях, оказалось недостаточно. Например, в СЕ «Кукловод сломал куклу Миши? Нет, Алексей.» слова «Миша» и «Алексей» имеют большую семантическую близость, чем слова «кукловод» и «Алексей», однако логическое ударение падает на слово «кукловод». Помимо семантического анализа, также надо использовать морфологический, в данном случае сопоставить падежи имен «Миша» и «Алексей». Для того чтобы логическое ударение падало на слово «Миша», надо, чтобы у имен совпадали падежи. Если интонационный центр падает не на одно слово, а на целую фразу, для выделения всех ударных слов, необходимо учесть данные синтаксического анализа.

4. В большинстве случаев ошибки возникли из-за невозможности определения семантического класса слова в контексте СЕ (если исключить случаи, когда Semsin неправильно определил семантический класс слова, то $\epsilon^* = 90,56\%$). Современные системы анализа текста, в том числе и Semsin, проводят семантический анализ данных в рамках одного предложения. Они не могут определить семантический класс слова, опираясь на данные соседнего предложения.

Пример результатов сравнения данных эксперта и системы

СЕ или пара предложений (вопрос – ответ)	Результаты ручной обработки СЕ или пары предложений экспертом		Результаты автоматической обработки СЕ или пары предложений системой	
	семантическая связь	интонационный центр	семантическая связь	интонационный центр
Ольга принесла подарки во вторник? Нет, в среду.	есть	вторник	есть	вторник
Кукловод сломал куклу Миши? Нет, Алексей.	есть	кукловод	есть	Миша
Воздушный шарик улетел? Нет, лопнул.	есть	улетел	нет	–
Андрей купил новый костюм? Нет, клубника.	нет	–	нет	–

Порядок слов в вопросительном предложении никак не повлиял на результаты исследований. Одной из важных особенностей русского языка является то, что в некоторых предложениях можно менять слова местами и смысл предложения при этом не изменится. В английском языке, например, эта операция недопустима. Так как ключевое слово ответа сопоставляется со всеми словами вопроса, их порядок при анализе не учитывается.

Заключение

В ходе исследования метода выделения интонационного центра в вопросительном предложении без явного вопросительного слова были сделаны следующие выводы:

1. Так как в основе данного метода лежит принцип сопоставления семантических классов на всех общих видах иерархии, то определить семантическую связь удалось не только из СЕ, которые содержали положительные ответы, но и из СЕ, содержавших отрицательные.

2. Для повышения качества оценки метода необходимо сравнивать не только семантические признаки слов, но и морфологические и синтаксические.

3. Семантический словарь В.А. Тузова не подходит для сопоставления семантики глаголов. Для данной операции лучше использовать другие специализированные словари, например семантический словарь глаголов Л.Г. Бабенко.

4. Для повышения качества оценки метода необходимо доработать методы семантического анализа Semsin.

5. Порядок слов в вопросительном предложении никак не повлиял на качество оценки исследования метода.

Список литературы

1. Одинцова И.В. Звуки. Ритмика. Интонация / И.В. Одинцова. – М.: Наука, 2008. – 368 с.
2. Рыбин С.В. Синтез речи. Учебное пособие / С.В. Рыбин. – СПб.: Университет ИТМО, 2014. – 92 с.
3. Левковская Н.А. В чём различие между сверхфразовым единством и абзацем? / Н.А. Левковская // Филологические науки. – 1980. – № 1. – С. 75–78.
4. Солганик Г.Я. Синтаксическая стилистика. Сложное синтаксическое целое / Г.Я. Солганик. – М.: Высшая школа, 1991. – 182 с.
5. Филимонов О.И. Скреп-фраза в языке / О.И. Филимонов. – М.: АГРУС, 2017. – 170 с.
6. 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.
7. Боярский К.К. Предсинтаксический модуль в анализаторе Semsin / К.К. Боярский, Е.А. Каневский // Интернет и современное общество. – 2013. – № 1. – С. 280–286.
8. 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.
9. Тузов В.А. Компьютерная семантика русского языка / В.А. Тузов. – СПб.: Изд-во СПбГУ, 2004. – 400 с.
10. Гришина Е.А. Национальный корпус русского языка: 2006–2008. Новые результаты и перспективы / Е. А. Гришина // Нестор-История. – 2009. – № 6. – С. 150–174.