УДК 681.3

СПОСОБ ФОРМАЛИЗАЦИИ СВЯЗЕЙ В КОНСТРУКЦИЯХ ТЕКСТА ПРИ СОЗДАНИИ ЕСТЕСТВЕННО-ЯЗЫКОВЫХ ИНТЕРФЕЙСОВ

И. С. Лебедев,
канд. техн. наук, преподаватель
Санкт-Петербургское высшее военное училище радиоэлектроники

Предложен способ, основанный на использовании структур, характеризующих лексические единицы текста, позволяющий вычислять ответы в тексте на вопросы, заданные в естественном виде. Рассмотрены вопросы разделения текста на семантически связанные единицы.

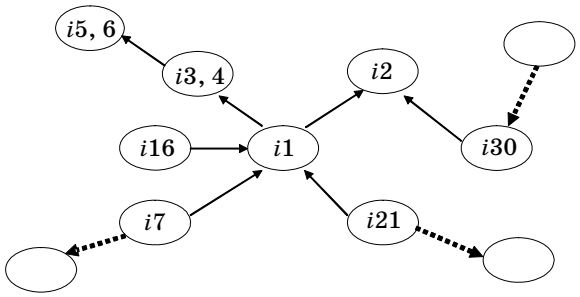
A method based on the usage of structures characterizing lexical units of a text is proposed which allows to calculate answers in the text to questions put in a natural form. Questions of text division into semantically connected units are considered.

Автоматический анализ текстовой информации приобретает огромную актуальность в связи с развитием глобальных вычислительных сетей и формированием больших объемов распределенных данных.

Современные системы автоматической обработки текстов, доступные широкому кругу пользователей, например информационно-поисковые машины в глобальных вычислительных сетях, в основном сталкиваются с проблемой классификации документов по запросу. На сегодняшний день существуют довольно приемлемые решения, дающие хорошие результаты при анализе всего содержания документа в целом. Однако существует огромный класс программного обеспечения, где необходимо создание естественно-языкового (ЕЯ) интерфейса, позволяющего искать ответ по тексту на вопрос, заданный в удобном для пользователя виде. К таким системам относятся в первую очередь контролирующие, обучающие, вопросно-ответные, где необходимо приближать диалог к естественному для человека виду. Во многих случаях подобный интерфейс полезен информационно-справочным, поисковым системам помощи для анализа вопроса, заданного пользователем, и предоставления точного адекватного ответа в реальном режиме времени. Поэтому разбиение текста на смысловые составляющие, определение семантических связей между словами является актуальной задачей.

Большинство ее решений связано с использованием языков разметки, что требует либо предварительной обработки текста экспертом, либо наличия жесткой структуры. Альтернативные подходы заключаются в том, что текст представляется в виде информационного потока, и по нему строится граф отношений, содержащий объекты текста и связи между ними. Объекты текста, которые для простоты могут быть представлены словами, обозначаются соответствующими информационными элементами. В работе представлены результаты, на основе которых был создан реально работающий анализатор, позволяющий построить граф предложения. Его демонстрационная версия находится в сети Интернет по адресу *www.semlp.com*. В основу анализатора была положена «семантическая модель естественного языка» профессора Санкт-Петербургского государственного университета В. А. Тузова. При реализации были созданы морфологический, синтаксический словари и словарь, содержащий семантическую информацию о 100 000 исходных форм слов. Простейшие системы, использующие подобные подходы, могут не содержать никаких словарей или тезаурусов, что позволяет достичь скорости обработки за счет качества, но для более точного определения связей необходимо проводить морфологический, синтаксический и семантический анализ. На рис. 1 приведен пример графа текста со словами $i_1, ..., i_n$.

В результате анализа графа, построенного по тексту, видно, что максимальное количество связей образуют несколько информационных элементов, которые являются основными для определения принадлежности тематики этого текста. Построение систем, дающих возможность обрабатывать запросы к тексту на естественном языке, требует как можно более точного определения связей между словами и семантического значения этих слов. Роль и значение слова в предложении определяет часть речи, поэтому необходима формали-



■ ***Рис. 1.*** *Граф текста*

зация существительных, прилагательных, глаголов, наречий, предлогов, частиц.

Наиболее сильно предложение характеризуют существительные. Они и представляют элементарные аргументы предложения и могут быть представлены в виде структуры, содержащей несколько полей, например:

$$S(k_1, ..., k_n), \qquad (1)$$

где S – объект на основе существительного, а аргументы k присоединяются с помощью связей *какой? сколько? чей? чего? кого? кем? чем?*

Применительно к тексту, на котором проводится поиск, объекты можно условно классифицировать по нескольким типам.

1. Существительное, стоящее в тексте:
знания, тестирование.

2. Существительное, уточненное прилагательным:
мероприятия контрольные.

3. Существительные, уточняющие себя другими существительными в родительном или творительном падеже:
результаты тестирования.

4. Существительные с прилагательными, уточняющие себя другими существительными в родительном или творительном падеже:
тестирование путем проведения мероприятий контрольных.

Такое деление является относительным, но если в запросе пользователя, заданного в любой форме, выделяются подобные группы на основе какого-то объекта, то релевантный документ в своем тексте должен содержать слова уточняющей группы при нем.

На основе тех форм запросов, которые выдают пользователи, используя выражение (1) и подключив словарь синонимов, возможно задавать перефразировки. Например, для запроса «результаты тестирования», используя электронный словарь синонимов [1], находим описания:

результат	*следствие, последствие, след, итог, плод, сумма;*
тестирование	*проверка, испытание.*

Подставив их в выражение (1), получаем следующие перефразировки:

результаты тестирования, результаты проверки, результаты испытания, следствия тестирования, следствия проверки, следствия испытания и т. д.

К подобным перефразировкам нужно относиться с осторожностью, так как в результате может исказиться смысл высказывания или возникнуть избыточность информации. Чтобы свести к минимуму возникновение подобных ситуаций, словари синонимов должны подключаться только в соответствии с той тематикой, стилем и жанром, которые являются основными для текста. Современный пользователь устроен таким образом, что желает увидеть в ответе те же словоформы, что и в запросе, поэтому приоритет необходимо отдавать исходным словам.

Базой конструкции предложения, к которой прикрепляются все основные члены, является глагол. Если глагола в предложении нет, то его можно заменить «пустым глаголом». Каждый глагол аналогично существительному может быть также рассмотрен в виде предиката

$$N(G(S_1(k_1, ..., k_n) ... S_m(k_1, ..., k_n))), \qquad (2)$$

где N — наречие, отвечающее на вопрос *как? когда? куда? где? откуда? как долго?*; G — глагольная функция; S – объект на основе существительного.

Аналогичным образом описываются и другие части речи. Более подробно это отражено в работе [2].

Основной материал для анализа текста предоставляют существительные или объекты на основе их [3, 4]. Если в тексте нет существительного или его синонима, которое встретилось в вопросе к тексту, то маловероятно, что в тексте будет содержаться конкретный ответ на вопрос.

Каждая стрелка в графе текста (см. рис. 1) определена совокупностью вопросов, которую можно задать от одного объекта к другому. Условно их можно разбить на две группы. Первая группа основывается на падежных вопросах (*кто? что? кого? чего? кому? чему? кем? чем? о ком? о чем?*). Она практически однозначно определяется предложно-падежной формой, и ее формализация зависит только от морфологической информации. Зная, в каком падеже стоит, например, существительное или прилагательное, всегда можно подобрать вопрос падежа и сформулировать вопрос к словоформе или словосочетанию.

Пришел (из чего?) из деревни.

Вторая группа – это смысловые вопросы, которые гораздо сложнее анализировать.

Пришел (откуда?) из деревни.

Пришел (почему?) из вежливости.

Для формализации смысловых вопросов, которые можно задать к тексту, необходимо вычленить элементарные структуры, внутри которых необходимо описать связи. В данном примере в качестве элементарной единицы рассматривается граф, изображенный на рис. 2.

Вершины этого графа составляют глагол G, прилагательное $Pril$, предлог $Predl$, существительное S, наречие Nar.

Создание формулы предложения состоит в том, чтобы определить каждый аргумент предложения и каждому слову приписать его семантико-грамматический тип. В случае построения такого предиката каждому аргументу можно задать вопрос от глагола, который будет определяться морфологической информацией, что позволит использовать падежный вопрос, а, с другой стороны — смысловой вопрос, стоящий в прямой зависимости от семантики слова.

Многие вопросы, заданные в естественном виде, будут содержать вопросительное слово либо падежей, либо смыслового вопроса. Например, на вопрос к тексту, состоящий из одного вопросительного аргумента *какой?*, могут отвечать не только прилагательные в именительном падеже единственного числа мужского рода, но и существительные в родительном с предлогами *от* и *из*, дательном с предлогами *по*, творительном с предлогом *с*.

Доклад (какой?) от 5-го числа.

Для описания смысловых вопросов необходимо приписать каждому существительному индекс некоторого класса. При описании этих классов с целью вычисления смысловых вопросов за основу было принято описание семантики предлогов русского языка [3]. Число классов может колебаться в зависимости от объема словаря, точности требуемого описания, но оно всегда недалеко от тридцати. Ниже приведены некоторые из них:

дата, направление, свойство, содержание, элемент, действие, материал, множество, мера, число, объект, отношение, чувство, время, емкость, расстояние, закон, часть, инф. источник.

Каждый смысловой вопрос к существительному, независимо от какой части речи он задается, можно выразить по формуле

(ПРЕДЛОГ + ПАДЕЖНЫЙ ВОПРОС) ⊗ Семантика слова = СМЫСЛОВОЙ ВОПРОС,

где семантика слова определяется классом, к которому принадлежит обозначаемое им понятие.

Существительное с предлогом рассматривается как единое целое. На графе рис. 2 стрелками показаны основные связи, которые необходимо формализовать для вычисления ответа на ЕЯ-вопрос.

Однако несмотря на довольно строгое применение предлога для вычисления смыслового вопроса в предложении, человек, задающий вопрос к неизвестному предложению или даже тексту, может изначально ставить его в неправильной форме, поэтому при разработке ЕЯ-интерфейсов необходимо расширять варианты поиска. Исходя из этого ниже приведен подход к формализации для некоторых смысловых вопросов. Необходимо отметить, что приводимая формализация находится в стадии внесения изменений и не является окончательной. Всего в русском языке существует около 25 вопросительных слов, в приведенном ниже примере (для вопроса *почему?*) показывается смысловой вопрос, предлог с падежным вопросом, формула согласно рис. 2, показывающая часть речи, от которой задается вопрос, и особенности существительных (падеж и класс), к которым вопрос ставится.

1. Вопрос *почему?*

1.1 почему? (от чего? от кого? с чего? из чего? из-за чего?)

1.1.1 с S образуется связь «элемент от S»:

серый (почему? от чего? с чего? из чего? из-за чего?) от (из-за) пыли

$$Pril + \{Predl = от, с, из, из\text{-}за \quad S^{Р.п} = класс\ «объект»\}$$

1.1.2 образуется связь «чувство»

ушел (почему? от чего? с чего? из чего? из-за чего?) из вежливости

$$G + \{Predl = от, с, из, из\text{-}за \quad S^{Р.п} = класс\ «чувство»\}$$

1.2 почему? (по чему? по кому? как?)

1.2.1 образуется связь «по закону»:

трактовал (почему? по чему? как?) по закону

$$G + \{Predl = по \quad S^{Д.п} = класс\ «закон»\}$$

1.3 почему? (на что? на кого?)

1.3.1 образуется связь «действие»:

закрыли (почему? на что? зачем?) на ремонт

$$G + \{Predl = на \quad S^{В.п} = класс\ «действие»\}$$

1.4 почему? (за чем?)

1.4.1 образуется связь «объект»:

шел (почему? за чем?) за неимением денег

$$G + \{Predl = за \quad S^{Т.п} = класс\ «объект»\}$$

Таким образом, рассмотрим два выражения, где подчиненные существительные стоят в родительном падеже:

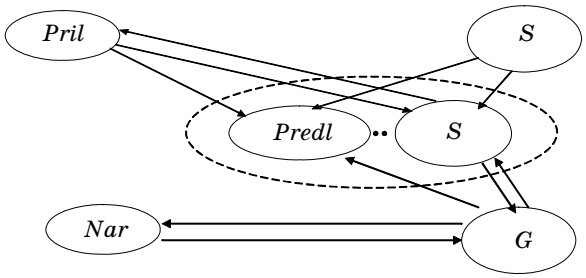
пришел из вежливости
пришел из деревни

Анализатор выдает следующие конструкции:

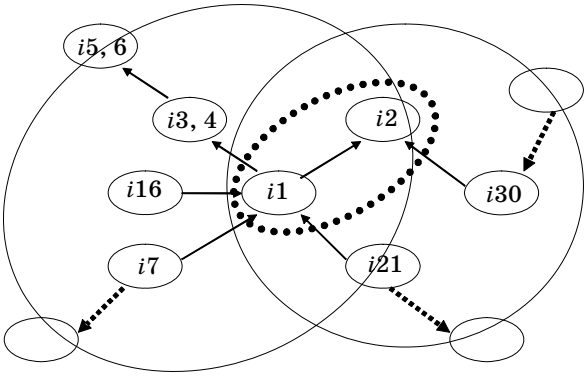
@Глагол Пришел (@Почему из (@Род вежливости))

@Глагол Пришел (@Откуда из (@Род деревни))

Видно, что к первому словосочетанию можно поставить вопрос *Пришел почему? от чего?* и получить в качестве ответа *из вежливости, от безысходности, от горя* — существительные класса «чувства» в родительном падеже. Второе словосочетание отвечает на вопрос *откуда?* Его ответом будут существительные класса «объект» в родительном падеже: *из деревни, из дома, с поля.*



■ ***Рис. 2.*** *Связи между единицами графа текста*



■ ***Рис. 3.*** *Объекты и их окрестности*

После того как построен граф текста, дуги, соединяющие объекты, могут принимать значения либо смысловых вопросов, либо вопросов падежей, в которых находятся объекты.

Цель создания ЕЯ-интерфейса состоит в том, чтобы на запрос пользователя вычислить информацию, адекватную запросу на семантическом уровне. Причем в зависимости от модели диалога, структур, алгоритмов обрабатываемая и выдаваемая информация может быть словом, синтаксической конструкцией, предложением и частью связанного текста. Рассматривая синтаксические конструкции вопроса, получаем некоторый портрет взаимосвязанных предложений, который во многих случаях является ответом. Зачастую здесь содержится основная информация об объекте текста, которую можно узнать, привлекая только формально морфологические признаки. На рис. 3 приведен пример, содержащий объекты и их окрестности.

Однако последнее время все чаще встречается мнение, что достичь качественного прорыва с применением одних только математических и лингвистических методов анализа текста не удается, и все больше исследователей приходит к мнению о том, что необходимо подключать прагматику.

Например, предложения:

Он прочитал газету.
@Глагол прочитал (*@Им Он, @Вин газету*).
Он просмотрел газету.
@Глагол просмотрел (*@Им Он, @Вин газету*).
Он подержал газету.
@Глагол подержал (*@Им Он, @Вин газету*).

можно анализировать как ответы на вопросы:

Он прочитал газету?
@Глагол прочитал (*@Им Он, @Вин газету*)?
Он ознакомился с газетой?
@Глагол ознакомился(*@Им Он, @сТв с газетой*)) ?
Он взял газету?
@Глагол взял (*@Им Он, @Вин газету*)?

В любом из перечисленных предложений необходимо сравнивать предикаты G_1(*Он, газета*), G_2(*Он, газета*). Если упростить анализ глагольных функций, то система будет выдавать все предложения из примера на любой из вопросов, в противном случае возрастает риск пропустить адекватную информацию. Для борьбы с подобными явлениями необходимо описывать модель реальной действительности, модель ситуаций, где должны содержаться правила сравнения.

В заключение хочется отметить следующее. В основе конструкции семантического языка находятся объекты, образующие между собой связи. Идентификация объектов и вычисление значения их связей основываются на модели представления естественного языка, на способе представления текстовой информации и являются зависящими друг от друга. Не вычислив связи, нельзя определить, является ли множество слов семантической конструкцией, и наоборот, не определив объект, сложно говорить о связях, которые он может образовывать с другими объектами. Формализация связей, способность их вычисления — основная проблема, от решения которой зависит построение адекватных правил работы с текстом.

Связь предложений в тексте в случае ее формализации дает возможность определить границы текста, где можно анализировать несколько предложений в качестве ответа на вопрос. Для анализа текста в вопросно-ответных системах необходимо получить как можно более полный и точный граф предложений.

При создании ЕЯ-интерфейсов огромная роль принадлежит формализации вопросов, задаваемых на естественном языке. Здесь необходимо учитывать и то, что пользователь может ставить семантически правильные вопросы в неправильной форме с точки зрения семантики синтаксиса.

Вычисление смыслового вопроса к предложно-падежной форме сводится к поиску конкретного атрибута присоединяющего слова.

Литература

1. Информационный сервер г. Набережные Челны. Электронный словарь синонимов. <http://www.chelni.ru/slovari/sinonim>
2. **Кондратьев А. В., Кривцов А. Н., Лебедев И. С.** Анализаторы текстов формальной модели русского языка для компьютера // Процессы управления и устойчивость: Тр. XXIX науч. конф. студентов и аспирантов факультета ПМ-ПУ / НИИ химии СПбГУ. СПб., 1998. С. 142–154.
3. **Тузов В. А.** Компьютерная семантика русского языка. СПб.: Изд-во СПбГУ, 2004. 400 с.
4. **Комаров И. И., Кривцов А. Н., Лебедев И. С.** Принципы построения семантической модели текста и ее применение в системах лингвистического обеспечения // Процессы управления и устойчивость: Тр. XXXIII науч. конф. студентов и аспирантов факультета ПМ-ПУ / НИИ химии СПбГУ. СПб., 2002. С. 373–382.