УДК 004.82

СОЧЕТАЕМОСТЬ ЛИНГВИСТИЧЕСКИХ ОБЪЕКТОВ В ПРОБЛЕМЕ ОБРАБОТКИ ЕСТЕСТВЕННОГО ЯЗЫКА

И. Е. Воронина, Т. М. Леденева

Воронежский государственный университет

Поступила в редакцию 20.06.2013 г.

Аннотация: Рассматривается подход к исследованию сочетаемости языковых единиц структурного уровня естественного языка с привлечением экспертов. Цель – формирования правил в виде запрета на сочетаемость.
Ключевые слова: компьютерная лингвистика, обработка естественного языка, сочетаемость языковых единиц, лингвистические объекты.
Annotation. The article is devoted to the approach to the study of the compatibility of language units of the structural level of natural language with the assistance of experts. Proposed to build rules in the form of restrictions on the compatibility for each skill level.
Keywords: Computer Linguistics, Natural Language Processing, the compatibility of language units, language objects.

ВВЕДЕНИЕ

В условиях информационного общества взаимодействие разных пользователей, государственной службы и социальной сферы с виртуальными ресурсами существенным образом зависят от адекватности лингвистической среды, общение в которой происходит на естественном языке (лингвистическая среда – вид коммуникативного социокультурного пространства, в котором реализуется общение).

На принятие решений в современном обществе сильнейшее влияние могут оказывать информационные воздействия, реализуемые средствами массовой информации, особенности современного законодательства, уровень образования, доступность мировых информационных ресурсов. В то же время невозможно отрицать стремительный рост объемов самой информации, причем налицо преобладание неструктурированных данных, и высокая динамика распространения неструктурированной информации. Понимая под лингвистическим обеспечением информационных процессов совокупность языковых средств общения и технологий их реализации, можно с уверенностью утверждать, что будущее за развитием естественно-языковых технологий со всеми вытекающими проблемами формализации естественного языка.

Таким образом, создание удобного и эргономичного пользовательского интерфейса, реализация эффективного поиска в телекоммуникационных сетях, совершенствование далеко не идеальных систем машинного перевода, обработка неструктурированной информации, развитие образовательных возможностей за счет не только пополнения электронного контента, а путем создания автоматизированных обучающих систем, опирающихся на анализ и принятие решения, – все это все это требует фундаментальных исследований в области естественного языка.

Формализация естественного языка является нетривиальной задачей и обладает всеми особенностями слабоструктурированных проблем. Прикладные научные исследования в области формализации естественного языка характеризуются тем, что обычные способы сбора и обработки информации не обеспечивают необходимой быстроты, полноты и качества ее переработки. Отсутствие диагностического инструментария, позволяющего количественно оценить степень приближения получаемых результатов к реальности, также не способствует повышению эффективности и качества исследований. Рассматривая в качестве системы-объекта естественный язык необходимо не только проанализировать подходы, проблемы и достижения на пути построения теоретической системы (Н. Хомский, Т. Виноград,

А.С. Нариньяни, Бодуэн де Куртене, А.С. Гердт, А.Г. Белоногов, В.В. Налимов, Д.А. Поспелов, Н.Н. Перцова, Р.Г. Пиотровский, Р.С. Гиляревский, Ю.И. Шемакин, А.И. Кузнецова, Г.П. Мельников, А.А. Кретов), но и представить собственное развитие методологии исследовательского процесса, разработав математическое, алгоритмическое и программное обеспечение его поддержки.

ПРОБЛЕМА СОЧЕТАЕМОСТИ ЛИНГВИСТИЧЕСКИХ ОБЪЕКТОВ

Введем понятие лингвистического объекта как элемента определенного уровня языковой структуры (рис. 1). Лингвистический объект – единица определенного структурного уровня языка, исследуемая как относительное самостоятельное, но всё равно во всех его взаимосвязях.

Формализация ЕЯ является нетривиальной задачей и обладает всеми свойствами слабоструктурированных проблем. Глобальная цель всех проводимых лингвистических исследований – разобраться в структуре языка. Уровни структуры языка – это синтаксические предложения, слова, морфемы, фонемы. Все языковые уровни характеризуются наличием базовых элементов.

На каждом языковом уровне возникает задача порождения правильных языковых цепочек. В терминологии формальных грамматик Хомского «правильная» означает соответствие правилам грамматики, а цепочки образуются путем конкатенации базовых элементов. В нашем случае в качестве базовых элементов выступают языковые единицы (объекты) заданного уровня, а цепочки образуются путем сочетания этих единиц, результатом чего является порождение объекта следующего уровня. Таким образом, рассматривается задача порождения правильных языковых цепочек на заданном языковом уровне. Для формирования правил в виде запретов на сочетаемость базовых единиц каждого уровня привлекаются эксперты. Принятие решений обычно предполагает, что информация, используемая для их обоснования, достоверна и надежна. Но для задач, которые по своему характеру являются качественно новыми, это предположение зачастую не выполняется. Основные трудности обусловлены неполнотой имеющейся информации или ее недостаточно высоким качеством. Для проблем, в отношении которых информационный потенциал недостаточен для уверенности в истинности выдвигаемых гипотез, должны использоваться модели, ориентированные на обработку качественной информации. В сложных ситуациях каждый эксперт должен определить возможность сочетания тех или иных структурных единиц, используя качественные оценки, основанные на используемом понятии лингвистической переменной.

В рамках исследования водится лингвистическая переменная Comp = (СОЧЕТАЕМОСТЬ, T, [0, 1], G, F), где $T = \{T_i\}_{\overline{i=1,n}}$ – упорядоченное терм-множество значений лингвистической переменной, которое, по сути, образует лингвистическую шкалу. В рамках проведенного исследования предполагается, что шкала имеет следующий вид: $T = \{t_1 = нет; t_2 = скорее\ нет; t_3 = не\ знаю; t_4 = скорее\ да; t_5 = да\}$. Каждому терму t_i ставится в соответствие весовой коэффициент q_i, *так что* $\sum_{t_i \in T} q_i = 1, \forall i(q_i \in [0,1])$. При проведении исследования данные коэффициенты на-

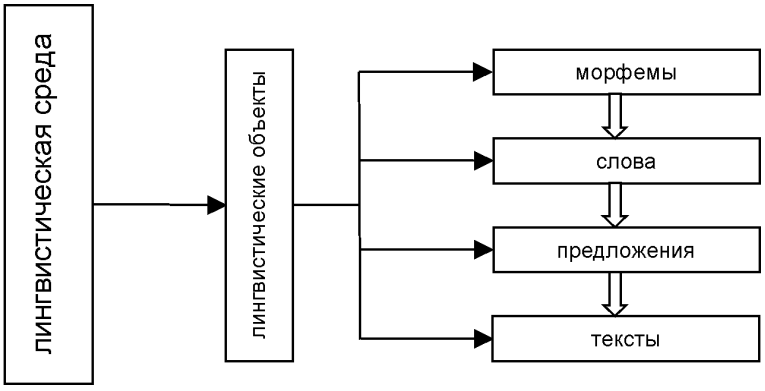

Рис. 1. Лингвистические объекты

страиваются на основе эмпирического опыта эксперта с помощью специальной программы.

Для лингвистических объектов оценка сочетаемости осуществляется в процессе коммуникации. Если из n объектов выделяется подгруппа таких, которые обладают некоторым свойством a, и К объектов данной группы демонстрирует тип поведения b, то будем считать, что правило $a \to b$ имеет *интенсивность* $I(a \to b) = K$.

При исследовании проблемы моделирования и принятия решений в задачах анализа и синтеза лингвистических объектов, в правилах сочетаемости в качестве объекта выступает единица, которая стоит на более высоком уровне иерархии, нежели единицы, рассматриваемые на предмет совместимости: если объект – это текст, то составляющими текста будут предложения, на уровне предложений составляющими являются слова, на уровне слова – морфемы. В результате исследования формируется множество правил, каждому из которых ставится в соответствие степень, которой можно придать различную интерпретацию (уверенность). Правила сочетаемости задаются в виде: $R_l = E_j \xrightarrow{t_i} E_k$, где E_j, E_k – ЕЯ, t_i – значение лингвистической переменной «СОЧЕТАЕМОСТЬ», или $E_j -> E_k$. Интерпретация правила: **СОЧЕТАЕМОСТЬ(** <фиксированная единица (объект) E_j уровня n – 1 > **И** < единица (объект) E_k уровня n – 1 >) ИМЕЕТ МЕСТО СО СТЕПЕНЬЮ t_i.

МЕТОД ИНТУИТИВНОЙ ОПТИМИЗАЦИИ

Для принятия решения в случае неочевидной сочетаемости ЯЕ предлагается метод интуитивной оптимизации, при этом под оптимизацией будем понимать сокращение числа шагов, необходимых для принятия решения.

Рассматриваются объекты уровня n. Выделим языковые единицы уровня n - *1,* сочетаемость которой предстоит исследовать (E_{fixed}). Эксперты оценивают возможность сочетания E_{fixed} с ЯЕ ее уровня. Метод заключается в усреднении лингвистических оценок сочетаемости языковых единиц на основе интенсивности каждого правила и соответствующего весового коэффициента и включает следующие шаги.

Шаг 1. В процессе экспертизы каждый эксперт $spec_k (k = \overline{1, \varepsilon})$ заполняет матрицу оценок сочетаемости E_{fixed} с каждой из языковых единиц $E_j (j = \overline{1, m})$ (строки соответствуют термам $\{t_1, t_2, t_3, t_4, t_5\}$, а столбцы – E_j). При наличии сочетаемости E_{fixed} с E_j с оценкой t_i элемент экспертной матрицы полагается равным 1, иначе 0.

Шаг 2. Вычислить обобщенную матрицу $spec = \sum_{k=1}^{\varepsilon} spec_k$, в результате чего получим, что элемент $spec_{ij} = N(t_i)$, где $N(t_i)$ – количество экспертов, которые оценили сочетаемость E_{fixed} с E_j оценкой t_i.

Шаг 3. Сформировать множество активных оценок:

$Sel(E_j) = \{t_i: N(t_i) \neq 0\} = \{t^*_{i_1}, \ldots, t^*_{i_{r_j}}\}$, r_j – количество активных оценок для E_j.

Шаг 4. Для каждой единицы E_j вычислить взвешенную интенсивность сочетаемости с E_{fixed} по формуле $S_{E_j} = \frac{1}{m} \sum_{\{i: t_i \in Sel(E_j)\}} spec_{ij} \cdot q_i$.

Шаг 5. Задать пороговые значения на сочтаемость/несочетаемость $bound^+$ и $bound^-$ и сформировать множества положительных и отрицательных правил.

Результаты работы предложенного алгоритма позволяют сформировать исходный материал для исследования возможностей формирования правил фильтрации в виде запрета на сочетаемость.

Следует заметить, что и отрицательный материал может быть подвергнут исследованию. Визуализация распределения по значимости каждого правила, то есть взвешенных интенсивностей S_{E_j}, позволит получить профиль сочетаемости E_{fixed} с E_j, которая может косвенно быть полезна при принятии решения о выборе предпочтительного сочетания и формализации правила выбора.

ЗАКЛЮЧЕНИЕ

Предложенный алгоритм был программно реализован и апробирован на тестовых примерах. Например, было обработано 4810 словосочетаний со словом «*свобода*» (источник информации: «Морфемно-морфонологический словарь языка А.С. Пушкина» – около 23000 слов) и выбраны предпочтительные сочетания с заданными порогами [1].

СПИСОК ЛИТЕРАТУРЫ

1. *Воронина И.Е.* Моделирование и алгоритмизация исследования лингвистической реальности / И.Е. Воронина - LAP LAMBERT Academic Publishing GmbH & Co. KG,Saarbrücken, Germany, 2011 – 263 с.

Воронина Ирина Евгеньевна – к. т. н., доцент кафедры программного обеспечения и администрирования информационных систем факультета ПММ, Воронежский государственный университет. E-mail: irina.voronina@gmail.com

Леденева Т. М. – зав. каф. вычислительной математики факультета Прикладной математики, механики и информатики Воронежского государственного университета, д.т.н., профессор. Тел. (4732) 208-282. E-mail: dean@amm.vsu.ru

Voronina Irina Ye. – Associated Professor of Software & Information System Administering Chair, Department of Applied Mathematics, Computer Science & Mechanics, Voronezh State University. E-mail: irina.voronina@gmail.com

Ledeneva Tatyna Michaylovna – Doctor of Technic Sciences, Professor, The dept. of the Mathematical Methods of Oreration Research, Voronezh State University. Tel. (4732) 208-282. E-mail: dean@amm.vsu.ru