# ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ И СИСТЕМЫ

УДК 001.103:002

И. Е. Воронина

## КОЛИЧЕСТВЕННЫЕ ОЦЕНКИ ПРИ МОДЕЛИРОВАНИИ ЯЗЫКОВОЙ СИСТЕМЫ

Рассматривается задача количественного оценивания сочетаемости языковых единиц при проведении исследований в области формализации естественного языка.

***Ключевые слова:*** *компьютерная лингвистика, обработка естественного языка, сочетаемость языковых единиц, компьютерное моделирование языковых объектов.*

Для языкознания характерно соперничество системно-классификационного (номотетического) и индивидуально-идиографического методов [1]. Первый используется естественными и техническими науками и ориентирован на выявление в исследуемом материале основных закономерностей, которые могут быть представлены в виде упрощенных (обедненных), но легко формализуемых схем. Второй подход используется для полного описания отдельно взятого объекта или его свойства, имеющего особое значение для понимания сущности всего явления.

На начальном этапе развития так называемой инженерной лингвистики с логико-лингвистическими исчислениями появилось большое количество системных лингвистических исследований. Но анализ неудач на пути создания систем автоматической обработки текста показал, что естественный язык (ЕЯ) является открытой коммуникативной системой. Разумным компромиссом вышеупомянутых подходов могло бы послужить создание набора исследовательских инструментов, которые бы, с одной стороны, были ориентированы на поиск закономерностей, выявление и формализацию правил ЕЯ, а затем и на их программное подтверждение, с другой — не отвергали исследовательских методик языковедов-традиционалистов. Для выработки, выявления тенденций, способных непосредственно повлиять на принятие решения, количественных характеристик исследовательского процесса необходимы однозначные критерии количества.

Глобальная цель всех проводимых лингвистических исследований — „постичь“ структуру языка. Уровни структуры языка — это синтаксические предложения, слова, морфемы, фонемы. Изучать язык можно путем анализа и синтеза, ибо выявленные правила синтеза могут способствовать проведению анализа, и наоборот. Все языковые уровни характеризуются наличием базовых элементов. Так, например, на комбинаторику фигур (букв и слогов) накладываются ограничения в сочетаемости простых знаков — морфем. В свою очередь, комбинаторика морфем ограничивается сочетаемостью знаков более высокого порядка — слов. Затем, по мере развертывания текста, на комбинаторику слов накладываются ограничения в сочетаемости словосочетаний и предложений, а на них — экстралингвистические композиционно-сюжетные ограничения [2]. Тем самым выявляется направление укрупнения лингвистических

объектов. Речь идет о выявлении и программном подтверждении правил сочетаемости языковых единиц.

Выявлению правил может способствовать наличие программно реализованного набора инструментов, позволяющего максимально автоматизировать данный процесс. Актуальность проводимых исследований определяется тем, что правила сочетаемости языковых единиц не только играют важную роль при синтезе текста, но и могут лечь в основу анализаторов разного уровня (морфемно-морфологического, синтаксического, семантического). Ни одно из известных решений для создания анализаторов (наиболее результативные относятся к области морфологии и синтаксиса) не получило общего признания. Вот почему исследования в этом направлении могут представлять профессиональный интерес. Традиционно любая формализация подразумевает наличие совокупности правил, позволяющих строить описание объекта на декларативном или функциональном уровне. По сути дела, эти правила позволяют ответить на вопрос „как можно“ (построить, описать, сделать и т.д.). Возможен подход к формализации, основанный на системе правил „как нельзя“ [3, 4]. Правила вида „как нельзя“ разбиваются на группы. Каждая группа правил определяет фильтр. Каждый фильтр — это подсистема запретов на сочетаемость структурных единиц, весь предлагаемый инструментарий ориентирован на применение опыта и интуиции исследователя, подкрепляемых использованием математических оценок для принятия решения в случае недостаточно определенной сочетаемости структурных единиц.

Рассмотрим задачу построения лексических цепочек на заданном языковом уровне. Предлагается формулировать правила в виде запретов на сочетаемость базовых единиц каждого языкового уровня. Формулировать правила могут только эксперты, т.е. выбор, обоснование и оценка решений не могут быть выполнены на основании точных расчетов вследствие их качественной новизны и сложности. 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 либо заведомо не реализуется, либо в момент принятия решения его не удается доказать. Основные трудности обусловлены неполнотой имеющейся информации или ее недостаточно высоким качеством.

В недостаточно определенных ситуациях исследователь может самостоятельно оценить возможности сочетания тех или иных структурных единиц. Эта оценка может носить лингвистический характер. Явное сходство с анкетами, которые заполняют респонденты в ходе социальных исследований, позволяет взять за основу методы детерминационного анализа [5], а также воспользоваться идеями, изложенными в работе [6].

В нашем случае первичные эмпирические данные будут представлять собой совокупность отображений вида $E \to X_i$, $i \in 1, \ldots, n$, где $E$ — множество объектов, $X_i$ — множество значений переменной $x_i$, $i$ — индекс, нумерующий переменные, участвующие в эмпирическом исследовании или эксперименте. Используется всего одна переменная $X = \{x_1, x_2, x_3, x_4, x_5\} : x_1$ — да, $x_2$ — нет, $x_3$ — не знаю, $x_4$ — скорее да, чем нет, $x_5$ — скорее нет, чем да.

Полученную функцию можно представить в виде табл. 1. Множество строк — это множество исследуемых объектов: слов (как сочетаний морфем), словосочетаний, предложений и любых других объектов, для которых уместны подобные оценки.

*Таблица 1*

**Представление функции**

| Объекты | Оценка |
|---|---|
| $e^{(1)}$ | $x^{(1)}$ |
| $e^{(2)}$ | $x^{(2)}$ |
| … | … |
| $e^{(k-1)}$ | $x^{(k-1)}$ |
| $e^{(k)}$ | $x^{(k)}$ |

Множества $E$, $X_i$, $i \in 1, \ldots, n$, дискретны и конечны, что является прямым следствием их номинальности. При использовании предложенного метода учитывается процесс коммуникации (диалога), поэтому проводимые измерения являются номинальными, или качественными.

Рассматриваются объекты уровня $n$, обеспечивающие универсальность контекста. Для каждого из этих объектов определяется значение переменной $x$, таким образом, задается отображение $E \to X_j$, $j \in 1, ..., n$. При этом значение переменной $x$ задается путем опроса. Исследователь принимает решение самостоятельно, используя собственный опыт и интуицию.

Вторым шагом будет являться выделение тех составляющих объекта, сочетание которых представляет интерес. Сочетаемость, собственно говоря, и определяется правилом ЕСЛИ $a$, ТО $b$ ($a \to b$). Здесь $a$ — это утверждение вида $\mathrm{Comp}_1$ & $\mathrm{Comp}_2$, где $\mathrm{Comp}_1$ и $\mathrm{Comp}_2$ — те составляющие объекта, о сочетаемости или несочетаемости которых надо принять решение; $b$ — утверждение о том, что такое сочетание имеет место.

Следует заметить, что в нашем случае правило на самом деле имеет вид: ЕСЛИ $a$, ТО *возможно* $b$ ($a \to b$).

Интерпретация правила: *сочетаемость* (<фиксированная единица (объект) уровня $n - 1$ > И < единица (объект) уровня $n - 1$ >) *имеет место с определенной долей уверенности.*

Для большей наглядности, не изменяя семантику правила, будем записывать его следующим образом: $\mathrm{Comp}_1 \to \mathrm{Comp}_2$, или, при необходимости, $\mathrm{Comp}_1 \xrightarrow{x_i} \mathrm{Comp}_2$, когда подразумевается степень уверенности $x_i$.

В зону определенности попадают отображения со значениями переменной $x_1$ и $x_2$, все остальные — в зону неопределенности.

Лингвистические оценки (значения переменной $X$) наделяются весами (коэффициентами уверенности). При этом шкала весов должна быть настраиваемой. Настройки должен осуществлять сам исследователь. Фиксируются лишь диапазоны для каждой из переменных зоны неопределенности: 0, ..., 1. Вполне понятно, что коэффициент „1“ соответствует значению „ДА“, а „0“ — „НЕТ“. Далее, используя интенсивность каждого правила и соответствующий весовой коэффициент, можно получить усредненную картину по всем исследуемым объектам. Исследователь может интуитивно установить пороговое значение, которое должен превысить полученный результат, для того чтобы считаться положительным и чтобы исследователю начать поиски объяснения сочетаемости, используя собственные знания и опыт. Если поиски увенчаются успехом, то будет получено очередное правило, которое в дальнейшем станет составляющей фильтра.

Учитывая вышеизложенное, можно считать, что имеется качественная шкала рассматриваемого показателя $X$, подобно [6]. Эта шкала может стать количественной при задании весовых коэффициентов. Поскольку у нас под объектом понимается сочетаемость двух структурных составляющих, можно сопоставить одной из альтернатив выбор (предпочтение) конкретного объекта. Полученные результаты могут быть сведены в таблицу, где на пересечении строки и столбца можно поставить либо 1, либо 0, что будет означать наличие или отсутствие оценки $x_i$ (табл. 2).

*Таблица 2*

**Оценки сочетаемости**

| Переменная | $\mathrm{Comp}_1$ | $\mathrm{Comp}_2$ | … | $\mathrm{Comp}_j$ | … | $\mathrm{Comp}_m$ |
|---|---|---|---|---|---|---|
| $x_1$ | | | | | | |
| $x_3$ | | | | | | |
| $x_3$ | | | | | | |
| $x_4$ | | | | | | |
| $x_5$ | | | | | | |
| Итого | | | | | | |

**Оценки сочетаемости.** Итоговое значение для каждого столбца будет представлять собой суммарный вес каждого правила, а выделенная итоговая строка представляет результаты распределения по шкале $X$.

Если обозначить через $q(x_i)$ значение весового коэффициента для переменной $x_i$, а количество объектов, для которых было определено значение переменной $x_i$ как $N(x_i)$, то каждый элемент строки „итого“ будет содержать величину

$$S_{\mathrm{Comp}_j} = \sum_{i=1}^{k} q(\chi_i) N^{(j)}(\chi_i), \tag{1}$$

где $k$ — количество переменных, участвующих в эмпирическом обследовании (в данном случае $k = 5$).

Строго говоря, существует лишь один случай, когда имеется полная и однозначная определенность: это те ситуации, когда значение $x$ есть „ДА“, т.е. речь идет о $x_1$. При этом можно вынести точный вердикт о сочетаемости структурных единиц. Однако такой случай очень редок, поскольку работа происходит в условиях изучения объекта, эволюционирования модели, когда полная формализация правил образования объекта (модели) еще не прошла.

Заметим, что предполагается

$$\sum_{x_i \in X} q(x_i) = 1. \tag{2}$$

Введем некоторые обозначения. $\mathrm{Sel}(\mathrm{Comp}_j)$ — это множество, состоящее только из тех переменных $x_i$, которые были задействованы в эмпирическом обследовании для компонента $\mathrm{Comp}_j$:

$$\mathrm{Sel}\left(\mathrm{Comp}_j\right) \subset X = \bigcup_{i=1}^{k} x_i. \tag{3}$$

Напомним, что $N(e)$ — это общее количество правил универсального контекста, т.е. практически это количество исследованных объектов уровня $n$. При проведении исследования на сочетаемость представляется разумным фиксировать один из компонентов (тот, в отношении которого надо принять решение о его сочетаемости или несочетаемости с какими-либо другими компонентами). Обозначим его как $\mathrm{Comp}_{\mathrm{fixed}}$. Можно сказать, что изучение поведения $\mathrm{Comp}_{\mathrm{fixed}}$ является *целью* исследования. В таком случае общее количество правил вида $\mathrm{Comp}_{\mathrm{fixed}} \to \mathrm{Comp}_j$, $j \in 1, \ldots, m$ ($m$ — количество компонентов, которые проверяются на сочетаемость с $\mathrm{Comp}_{\mathrm{fixed}}$), совпадет с $N(e)$. Сюда войдут правила и с отрицательным заключением (в случае положительного заключения ответ однозначен и дальнейшие действия теряют смысл). Заключение „НЕТ“ не приводит к прекращению процесса исследования и отбрасыванию $\mathrm{Comp}_j$ как возможного претендента на сочетаемость, поскольку данное заключение выносится в отношении объекта более высокого уровня (универсального контекста). Количество правил, когда $\mathrm{Comp}_{\mathrm{fixed}}$ с той или иной долей уверенности демонстрировал тип поведения $\mathrm{Comp}_j$, есть $\sum_{\mathrm{Sel}\left(\mathrm{Comp}_j\right)} N^{(j)}(x_i)$. Тогда суммарная интенсивность правил для отдельного компонента $\mathrm{Comp}_j$ есть

$$I_{\Sigma}\left(\mathrm{Comp}_{\mathrm{fixed}} \to \mathrm{Comp}_j\right) = \sum_{\mathrm{Sel}\left(\mathrm{Comp}_j\right)} N^{(j)}(x_i) \Big/ N(e), \tag{4}$$

суммарная интенсивность состоит из интенсивностей правил

$$\mathrm{Comp}_{\mathrm{fixed}} \xrightarrow{x_i} \mathrm{Comp}_j,$$

причем каждая переменная $x_i$ имеет свой вес $q(x_i)$.

Назовем *взвешенной интенсивностью* правила с переменной $x_i$ произведение интенсивности и весового коэффициента данного правила:

$$I_{W_i} = N(x_i)/N(e) - q(x_i) = I\left(\text{Comp}_{\text{fixed}} \xrightarrow{x_i} \text{Comp}_j\right) - q(x_i),\ \forall\, i \in 1, ..., k,\ \forall\, j \in 1, \ldots, m,$$

тогда суммарная взвешенная интенсивность будет следующей:

$$I_{\sum W} = \sum_i I_{W_i} \Big/ N(e) = \frac{\sum_{\text{Sel}(\text{Comp}_j)} N^{(j)}(x_i) q(x_i)}{N(e)} = S/N(e). \qquad (5)$$

Исследователь сам может установить пороговое значение, сравнение с которым позволит отсеять часть претендентов на сочетаемость, оставив материал для размышления и изучения. Принятие решения будет заключаться в формулировке правила сочетаемости (фильтра).

Но и отвергнутый материал может быть исследован. Визуализация распределения по значимости каждого правила, т.е. взвешенных интенсивностей $I_{W_i}$, позволит получить картину, которая может косвенно быть полезна при принятии решения.

Для оценивания на основе вычислительного эксперимента был создан программный инструментарий, позволяющий:

— создавать персональный отчет для каждого исследователя, в котором накапливаются экспертные оценки (веса) рассматриваемых сочетаний слов;

— настраивать весовые коэффициенты, приписываемые элементам качественной шкалы;

— просматривать, добавлять, удалять оценки для соответствующих словосочетаний;

— на основе полученных результатов для списка заданных словосочетаний рассматривать его как набор альтернатив и оценивать указанный выше критерий для каждого из словосочетаний, что устанавливает транзитивные отношения между значениями критерия для каждого словосочетания из списка.

В качестве примера вычислительного эксперимента можно привести обработку словосочетаний со словом „свобода“. Источником информации послужил „Морфемно-морфонологический словарь языка А. С. Пушкина“ [7], цель: выбрать наиболее подходящие словосочетания исходя из экспертных оценок.

Указав часть речи, авторы получили список подходящих слов для изучения сочетаемости с исходным словом (в данном случае — список прилагательных). После проведения экспертизы, задания весовых коэффициентов, порогового значения получены следующие результаты для неформализованной сочетаемости (табл. 3). Принятие решения основано исключительно на субъективных экспертных оценках при полном отсутствии правил формирования словосочетаний.

*Таблица 3*

**Результаты обработки с пороговым значением 2,5**

| Критерий | $\text{Comp}_{\text{fixed}}$ | $\text{Comp}_j$ |
|---|---|---|
| 4 | СВОБОДА | ЖЕЛАННЫЙ |
| 3,85 | СВОБОДА | БЕСКОРЫСТНЫЙ |
| 3,7 | СВОБОДА | БЕЗЗАКОННЫЙ |
| 3,7 | СВОБОДА | НЕВЫМЫШЛЕННЫЙ |
| 3,65 | СВОБОДА | НЕСОМНЕННЫЙ |
| 3,6 | СВОБОДА | КРАЕУГОЛЬНЫЙ |
| 3,5 | СВОБОДА | ЗАМЕШАННЫЙ |
| 3,5 | СВОБОДА | МЕРЗКИЙ |
| 3,5 | СВОБОДА | ОПЫТНЫЙ |
| 2,65 | СВОБОДА | ЗАВЕТНЫЙ |

*Продолжение таблицы 3*

| Критерий | $Comp_{fixed}$ | $Comp_j$ |
|---|---|---|
| 2,5 | СВОБОДА | ЗАСЛУЖЕННЫЙ |
| 2,5 | СВОБОДА | ДОЛГОВЕЧНЫЙ |
| 2,5 | СВОБОДА | НЕМИНУЕМЫЙ |
| 2,5 | СВОБОДА | ПРОНЗИТЕЛЬНЫЙ |
| 2,5 | СВОБОДА | НЕОБЪЯТНЫЙ |

Представленный пример иллюстрирует подход к отбору исследовательского материала для того, чтобы попытаться сформулировать правила сочетаемости. И если в случае сочетаемости слов важность экспертных оценок не так очевидна, то при изучении сочетаемости морфем в словообразовательных процессах [8] значимость экспертизы весьма понятна.

Следует заметить, что не требуется затрат на сбор и обобщение знаний специалистов: программный инструментарий ориентирован на отдельного эксперта и помогает найти необходимое решение, предоставляющим возможность оценивания достоверности по апостериорным данным. Эту оценку можно использовать в качестве априорных данных для дальнейших экспертиз при создании языковых фильтров.

## СПИСОК ЛИТЕРАТУРЫ

1. *Пиотровский Р. Г.* Лингвистический автомат (в исследовании и непрерывном обучении). СПб: Изд-во РГПУ, 1999. 256 с.

2. *Пиотровский Р. Г.* Инженерная лингвистика и теория языка. Л.: Наука, 1979. 112 с.

3. *Воронина И. Е.* Компьютерное моделирование лингвистических объектов. Воронеж: Изд.-полиграф. центр Воронежского гос. ун-та, 2007. 177 с.

4. *Воронина И. Е.* Актуальность моделирования лингвистической среды // Мат. Междунар. науч. конф. „Проблемы компьютерной лингвистики – 2009“. Воронеж: Изд.-полигр. центр Воронежского гос. ун-та, 2009.

5. *Чесноков С. В.* Детерминационный анализ социально-экономических данных. М.: Наука, 1982. 168 с.

6. *Жаке-Лагрез Э.* Применение размытых отношений при оценке предпочтительности распределенных величин // Статистические модели и многокритериальные задачи принятия решений. М.: Статистика, 1979. С. 168—183.

7. *Кретов А. А., Матыцина Л. Н.* Морфемно-морфонологический словарь языка А. С. Пушкина: ок. 23 000 слов. Воронеж: Центрально-Черноземное книжное изд-во, 1999. 208 с.

8. *Воронина И. Е.* Использование программных средств моделирования словообразовательных процессов в научно-исследовательской и педагогической практике // Проблемы компьютерной лингвистики: сб. науч. тр. Воронеж: Изд.-полиграф. центр Воронежского гос. ун-та, 2008. Вып. 3. С. 42—62.

***Сведения об авторе***

***Ирина Евгеньевна Воронина*** — канд. техн. наук; Воронежский государственный университет, кафедра программного обеспечения и администрирования информационных систем; E-mail: irina.voronina@gmail.com

Рекомендована кафедрой программного обеспечения и администрирования информационных систем

Поступила в редакцию 18.02.11 г.