

Численное моделирование словообразования

Кузнецов А.В., Заболеева-Зотова А.В.

Волг.ГТУ

art@tele-kom.ru, zabzot@vstu.ru

Аннотация

В данной работе рассмотрен вопрос о выработке формального аппарата, предназначенного для описания лингвистических знаний о естественном языке.

Для решения поставленной цели был применён метод комбинации морфологического и функционального описания системы.

В результате исследований было получено формальное описание лингвистических знаний, включающее в себя следующие понятия:

- 1) лингвистический эффект;
- 2) лингвистический объект;
- 3) лингвистическая система.

В работе предлагается использовать метод кодировки слов на основе места каждого символа в упорядоченном множестве $\{A, e\}$, где A - алфавит естественного языка, e - пустой символ (пробел). Что даёт возможность исследовать естественный язык при помощи формирования математических моделей, основанных на функциональных взаимосвязях над некоторыми мерами множеств символов естественного языка, которые описывают соответствующие лингвистические явления.

Полученные результаты могут применяться при построении программных систем, работающих с естественным языком.

1. Введение

Как известно, одной из задач, стоящих перед наукой Искусственный Интеллект является задача выработки некоторой формализации, при помощи которой можно было бы описывать достаточно большой круг явлений, происходящих в окружающем нас мире. Такое формальное описание позволит создать автоматизированные системы искусственного интеллекта близкие к естественному.

Одним из подходов, который позволяет подойти к рассмотрению данной проблемы является использование структурированной информации в виде Эффекта, т.е. *причинно-следственной взаимосвязи между изменением состояний взаимодействующих объектов*.

Очевидно, что следуя данному определению, мы следуем концепции детерминизма, с точки зрения которой, рассматривать реальные объекты необходимо с позиций их причинно-следственной взаимосвязи, из чего следует, что применение системного подхода является естественным шагом.

Известно, что лингвистическая информация может быть описана как система (например, слово может быть описана как система морфов, предложение, как система членов предложения, таких как: подлежащее, сказуемое, причастие и т.д., т.е., собственно, слов. Система предложений, в свою очередь, образует текст), а значит в ней, как и во всякой другой системе можно выделить элементы и связи между ними. Рассматривая такие системы в развитии мы наблюдаем смену их состояний, следовательно можно говорить о причинах и следствиях подобных процессов, т.е. можно говорить об эффектах (в описанном выше смысле).

Согласуем терминологию. Будем рассматривать понятие системы как *множество элементов, находящихся в отношениях и связях друг с другом, которые образуют определённую целостность*. В силу бесконечного разнообразия объективной реальности невозможно создать такую модель системы, которая учитывала бы все её аспекты, а следовательно, под системой некоторого феномена в дальнейшем мы понимаем лишь только отражение (в указанном выше смысле) этого феномена на некоторую абстрактную модель с точки зрения цели исследования.

Под элементом системы будем понимать *простейшую неделимую часть системы, т.е. предел членения системы с т.з. решения конкретной задачи и поставленных целей*. Интересно отметить, что: “Одни и те же элементы, взаимодействуя различными сторонами, могут образовывать различные системы” ([1], с. 39. Заметим, что под сторонами элемента в данном случае, по всей видимости, следует понимать различные его части, которые выделяются для описания системы с точки зрения поставленной цели).

Далее, можно дать понятие связи как *взаимодействие между двумя или несколькими элементами системы, отражаемое состоянием системы*. Причём, *состояние системы обычно описывает т.н. срез системы, момент в её развитии*.

Отметим, что как показано в ([11], с.31), тройка {причина, следствие, условие}, которая реализуется во взаимодействии, и есть тот инвариантный аспект системы, который мы называем связью.

Таким образом, определить все связи для данного элемента системы значит определить все возможные взаимодействия этого элемента с другими элементами.

Описывать систему с точки зрения взаимодействующих элементов принято через понятие структуры. Другими словами, под структурой системы будем понимать *устойчивую упорядоченность в пространстве и во времени ее элементов и связей*. Следует заметить, что: “Так как элементы участвуют во взаимодействии не всеми своими сторонами целиком, то структура в определённом смысле от них независима, то есть возможна замена элементов системы на качественно иные, но обладающие одним или несколькими

свойствами, сходными с взаимодействующими свойствами заменяемых элементов” ([1], с. 39).

Таким образом, описать систему можно путём определения следующей тройки: {множество элементов, множество связей, структура}, причём под множеством связей понимаются как связи внутри системы, так и связи системы с окружающей средой. В дальнейшем будем следовать данному подходу.

2. Понятие Лингвистического Эффекта

Рассмотрим понятие лингвистического эффекта (ЛЭ), с точки зрения изложенных выше позиций, как причинно-следственную связь между совпадением семантической и синтаксических форм объекта. Тогда это *взаимодействие*, а значит, ЛЭ может быть описан тройкой вида {причина, следствие, условие}. Договоримся называть эту модель представления ЛЭ, с некоторой степенью условности конечно, “*причинной моделью*”.

Таким образом формально определим ЛЭ как:

$J=(X, Z, f, P, I)$, где

X - множество входных воздействий;

Z - множество состояний ЛЭ;

f - функция переходов ЛЭ, т.е. отображение вида: $f: X \times Z \rightarrow Z$;

P - “предикат выполнимости” ЛЭ, т.е. формула исчисления предикатов принимающая значение Т на интерпретации I только тогда, когда соблюдаются условия применимости функции f для описания данного явления;

I - интерпретация для предикатной формулы P .

Лингвистический эффект, описанный в виде J представляет собой некоторый процесс, который происходит с объектом, например процесс словообразования. Если говорить упрощённо, то предложенную формулу можно прочесть так: под действием входных воздействий и с учётом текущего состояния, объект воздействия переходит в новое состояние (Z) по закону (f), причём мы имеем право применять данный закон (f) для описания этого явления только в том случае, если выполняются некоторые ограничения связанные, прежде всего, с видом этого закона (причём эти ограничения описаны в предикате P , принимающем соответствующие значения на интерпретации I , в которой фиксируется реальная ситуация на момент поступления входного сигнала).

Однако, в предложенной формализации не учитывается условие (в описанном выше смысле), как одна из компонент ЛЭ, что не удивительно, так как это зависит не от ЛЭ, как взаимодействия, а от объекта в котором происходит данное взаимодействие, т.е. от самого денотата. Таким образом, появилась необходимость описать *лингвистический объект (ЛО)*, как элементарную, неделимую, с т.з. поставленной задачи, часть исследуемой системы, в которой происходят некоторые лингвистические эффекты, следующим образом:

$W=(X_w, Z_w, J_w, f_w, g_w)$, где

X_w - “обобщённое” множество входных сигналов ЛО, т.е. $X_k \in X_w, k=1, 2, \dots, |J_w|$;

Z_w - “обобщённое” множество состояний ЛО, т.е. $Z_s \in Z_w, s=1, 2, \dots, |J_w|$;

J_w - множество лингвистических эффектов, реализуемых в данном лингвистическом объекте;

f_w - “функция перебора” для ЛО, т.е. $f_w: X_w \rightarrow J_w$;

g_w - “функция перехода” ЛО, т.е. $g_w: Z_w \rightarrow Z_w$;

Отметим, что множество Z_w может состоять не только из элементов, которые являются частью соответствующих множеств лингвистических эффектов, реализуемых в данном ЛО, т.е. собственно говоря слов на ЕЯ, но и из таких элементов, например, цепочек нетерминальных символов, которые туда не входят, а следовательно, отражают состояние лингвистического объекта (денотата) в целом. Аналогичная картина и с множеством X_w .

Опишем словесно, в упрощённой форме, формализацию ЛО в виде W . Пусть в некоторый момент времени на ЛО воздействует другой ЛО или группа ЛО-ов (например, мы составляем предложение). Тогда между ними произойдёт взаимодействие (взаимодействия), которое (которые) можно описать следующим образом: причиной взаимодействия являются входные воздействия, описываемые как элементы множества X_w , в это время ЛО находится в определённом состоянии (описываемом элементами множества Z_w) и если выполняются условия, описанные в f_w , то порождаются соответствующие ЛЭ-ты из множества J_w , которые, в свою очередь, изменяют состояние объекта, например трансформируют слово путём добавления аффиксов. Пусть в множестве состояний содержатся такие элементы, которые отражают состояние ЛО в целом (например, цепочки нетерминальных символов), тогда одновременно с описанным выше процессом взаимодействия, происходит изменение значений таких “переменных” по некоторому закону g_w , что может повлиять на порождение новых ЛЭ-ов.

Далее введём понятие *лингвистической системы* (ЛС), как совокупности лингвистических объектов как-либо связанных между собой.

В общем случае под ЛС будем понимать:

$S=(W_s, R_s, P_s)$, где

W_s - множество лингвистических объектов из которых состоит данная ЛС;

R_s - множество типов соединений ЛО;

P_s - предикат инцидентности, определённый для всех пар $(W_i, r_j), (W_i \in W_s, i=1, 2, \dots, n, n=|W_s|)$ $(r_j \in R_s, j=1, 2, \dots, m, m=|R_s|)$.

Следует отметить, что предложенное формальное описание ЛС в виде S не является, по мнению авторов, единственно возможным, а представляет собой один из вариантов такого описания на основе введённых формальных понятий ЛЭ и ЛО.

3. Пример описания лингвистического эффекта

Если рассматривать функцию переходов как функцию над некоторыми мерами множеств X и Z , то получим числовую функцию над числовыми аргументами, что позволит существенно оптимизировать соответствующие алгоритмы анализа и синтеза текста.

Очевидно, что понятие меры в данном случае не ограничивается простым количеством элементов соответствующих множеств, т.е. семантика функции переходов должна включать в себя формализацию процесса словообразования.

С этой целью используем метод кодировки слов на основе места каждого символа в упорядоченном множестве $\{A, e\}$, где A - алфавит естественного языка, e - пустой символ (пробел).

Поясним сказанное. Возьмём, например слово “*грач*”, используя тот факт, что основание рассматриваемой системы исчисления равно 34 и, сопоставляя каждому символу его порядковый номер в алфавите получим: “*грач*” = $34^3 \cdot 4 + 34^2 \cdot 18 + 34^1 \cdot 1 + 34^0 \cdot 25 = 178083$. Т.о. получаем, что “*грач*” в символах алфавита равен числу 178083 в символах 10-ричной системы исчисления.

Исходя из этого, можно ввести операции “сложения”, “вычитания”, “умножения” и “деления” над словами алфавита.

Например, $a+a=b$, $a+b=v$, ..., $a+я=ae$ (e - пустой символ или “пробел”).

Используя введённые понятия можно достаточно просто определить лингвистические эффекты, которые задаются, например, как отображения между словами алфавита путём оперирования численными мерами этих слов в 10-ричной системе исчисления.

Рассмотрим такой лингвистический эффект, как множественное число существительных в именительном падеже. “Следами действия” этого эффекта, с морфологической точки зрения, является то, что основы слов изменяются путём добавления, в зависимости от падежа, соответствующих окончаний (Прим.: е.ч., и.п. “*грач*”® “*грачи*” м.ч., и.п.). Чтобы проследить это изменение будем относить полученные словоформы в единственном числе, к соответствующим основам во множественном числе. Тогда получим некоторый коэффициент, разделив на который, можно из другой основы получить её словоформу во множественном числе.

Прим.: $грач=178083$, $грачи=6054832$, $K=грач/грачи=178083/6054832 \approx 0.029411716$; $окна=окн \cdot K$, $окна=18919/0.029411716=643247$.

Исследования проведённые авторами, показали, что хотя значение коэффициента до 3-го знака (0.029) одинаковы для всех основ, после него начинают просматриваться различия. Выделяются несколько следующих групп основ:

- 1) {окн, сел, грач}, $K=0.029411716$;
- 2) {окн, сел, мост, том}, $K=0.029411720$;
- 3) {книг, камн}, $K=0.029411747$;
- 4) {кол, кот, кон, кос}, $K=0.029410026$;
- 5) {гроб, гром}, $K=0.029411624$;

6) {рос, роз}, $K=0.029410591$ и т.д.

В заключении приведём пример формального описания ЛЭ. Опишем *множественное число существительных в именительном падеже для первой группы основ* как ЛЭ.

Определим соответствующие множества:

$X=\{...,g,...\}$, где g - численное значение любой комбинации любых символов алфавита русского языка, $Z=\{...,g',...\}$.

Тогда определим функцию переходов в виде: $f: g'=g*0.029411716$.

Опишем предикат выполнимости следующим образом: $P(x, y)=A(x, y)$, где

$A(x, y)=$ " x принадлежит группе y " - требование, говорящее о том, что данный вид закона применим, вообще говоря, только для определённой группы основ (В данном случае y - задается как группа основ типа: {окн, сел, грач, ...}).

4. Заключение

Рассмотрим основные результаты, полученные в данной работе.

Денотаты слов могут быть рассмотрены как элементы некоторой системы, например предложения. (В терминах введенных понятий - как лингвистические объекты лингвистической системы.) Являясь членами системы, объекты взаимодействуют между собой и, находясь в определенных состояниях, образуют некоторую целостность. (Конкретные словоформы суть состояния ЛО. Набор определенным образом упорядоченных словоформ образует предложение. Взаимодействие происходит через лингвистические эффекты.)

Оказывается, что каждая словоформа может быть рассмотрена, как некоторое число в символах соответствующей системы исчисления, основание которой равно числу букв в алфавите плюс пустой символ (пробел). Следовательно, лингвистические эффекты суть некоторые функции, устанавливающие соответствие между различными словоформами (числами).

Выяснив конкретный вид таких функций можно использовать их для построения систем, работающих с ЕЯ, как при синтезе, так и при анализе текстов. Например, если мы установим вид функции лингвистического эффекта множественного числа существительных, то оказывается, что её можно применять как для синтеза словоформы множественного числа из основы, так и для распознавания основы исходя из словоформы, используя её как обратную полученной.

Предложение есть совокупность конкретных словоформ. Словоформу, в свою очередь, можно представить как основу с применённой к ней функцией (ЛО). Тогда получаем, что предложение есть уравнение вида:

$y = f_1(x_1) \text{ } \ddot{f}_2(x_2) \text{ } \ddot{f}_3(x_3) \text{ } \ddot{f}_4(x_4) \text{ } \ddot{...} \text{ } \ddot{f}_m(x_n)$, где

у - предложение (, т.е. число, т.к. пробелы (пустые символы) тоже являются символами алфавита),

x_i - основа слова,

f_j - лингвистический эффект,

Ä - знак операции.

Таким образом, применение результатов, полученных в ходе данного исследования, позволяет подойти к проблемам обработки ЕЯ на ЭВМ с новых позиций, что даст возможность, в дальнейшем, построить эффективные алгоритмы работы с ЕЯ.

5. Литература

Аверьянов А. Н. Система: философская категория и реальность. - М: "Мысль", 1976. - 188 с.

Гиг Дж. ван. Прикладная общая теория систем: Пер. с англ. - М.: Мир, 1981. - Т.1. - 336 с.: Т. 2. - 736 с.

Гришин В. А., Камаев В. А. Математическое моделирование изделий и технологий. Учебное пособие. - Волгоград: Изд ВолгПИ, 1986. - 192 с.

Дружинин В. В., Конторов Д. С. Системотехника. - М.: Радио и связь, 1985. - 200 с., ил.

Заболеева-Зотова А. В. Структурный синтез сложных технических систем: формализация и алгоритмы - Волгоград: Волг.ГТУ, 1996. - 61 с.: - Деп. в ВИНТИ, №1487-В-95.

Казаринов М. Ю. Детерминизм в сложных системах управления и самоорганизации. - Л.: Издательство Ленинградского университета, 1990, 168 с.

Клир Дж. Системология. Автоматизация решения системных задач: Пер. с англ. - М.: Радио и связь, 1990. - 550 с., ил.

Корн Г., Корн Т. Справочник по математике (для научных работников и инженеров): Пер. с англ. - М.: Наука, 1974. - 832 с., ил.

Кузин Л. Т. Основы кибернетики: В 2-х т. Основы кибернетических моделей. - М.: Энергия, 1979. - Т. 2. - 548 с., ил.

Пуанкаре А. О науке: Пер. с франц. - М.: Наука, 1983. - 560 с.

Сейфуллаев Р. С. Концепция причинности и её функции в физике. - Новосибирск: "Наука", 1973. - 134 с.

Заболеева-Зотова А. В. Введение в системологию: Учебное пособие. - Волгоград: Волг.ГТУ, 1999. - 100 с.

Заболеева-Зотова А. В., Кузнецов А. В. Применение системного подхода к моделированию физических эффектов. III Всероссийский симпозиум "Математическое моделирование и компьютерные технологии". Тез. Докл. - Кисловодск: КИЭП, 1999. - т. 2, С. 5-7.

Моделирование языковой деятельности в интеллектуальных системах/ Под. ред. А.Е. Кибрика и А.С. Нариньяни; С предисловием А.П. Ершова -М.: Наука. Гл.ред.физ.-мат.лит., 1987.- 279с.

Гладкий А.В. Синтаксические структуры естественного языка в автоматизированных системах общения. - М.: Наука. Гл.ред.физ.-мат.лит., 1985.- 144с.

Лопатин В.В., Милославский И.Г., Шелякин М.А. Современный русский язык. Теоретический курс. - М.: "Русский язык", 1989. - 262 с.

Рождественский Ю. В. Типология слова. - М.: "Высшая школа", 1969. - 286 с.

Фомина М.И. Современный русский язык. Лексикология., 3-е изд., испр. и доп. - М.: "Высшая школа", 1990. - 415 с.

Пиотровский Р.Г. Текст, машина, человек. - Л.: "Наука", 1975. - 327 с.