МОДЕЛИ КОГНИТИВНЫХ ПРОЦЕССОВ
(Вычислительные системы)
1998 год Выпуск 164

УДК 519.766.2

К ВОПРОСУ О КАНОНИЧЕСКОМ ПРЕДСТАВЛЕНИИ ТЕКСТА ЕСТЕСТВЕННОГО ЯЗЫКА[1]

Д.Е.Пальчунов

Проблема автоматического извлечения знаний из текста имеет много граней. Одна из них — проблема автоматизации семантического разбора предложения, проблема сведения смысла предложения к смыслу входящих в него слов. Безусловно, эту проблему значительно легче решать тогда, когда структура предложения достаточно проста.

Цель настоящей работы — выяснить, к каким наиболее простым предложениям можно свести произвольное предложение естественного языка.

Здесь мы будем ограничиваться только повествовательными, описательными предложениями, не содержащими сложных логических конструкций. Предложения, которые мы будем рассматривать, не содержат, во-первых, кванторов "все", "для всех", "для каждого", "существует ... такой, что...", "найдётся ... такой, что..." и т.д., во-вторых, условных конструкций "если ..., то ...", в-третьих, не содержат конструкций вида "..., или, ..." и "либо ..., либо ..." и, в-четвёртых, сложных отрицаний "не верно, что ...<некоторое утверждение>". Рассмотрение таких достаточно несложных предложений является первым шагом к решению

[1]Работа частично поддержана: грант СО РАН № 3; грант РФФИ № 96-06-80570; грант РФФИ № 96-01-01525.

общей проблемы. На наш взгляд то, что в данной работе проделывается для описательных предложений, можно перенести и на случай произвольных предложений.

Идею канонического представления предложений естественного языка можно перенести из исчисления предикатов первого порядка. Наиболее подходит, на наш взгляд, представление формулы языка исчисления предикатов первого порядка в виде приведённой нормальной формы.

Напомним, что приведённой нормальной формой называется такая формула $\psi(x_1, \ldots, x_n)$, которая, во-первых, имеет вид $\psi(x_1, \ldots, x_n) = Q_1 y_1 \ldots Q_n y_n \, \phi(x_1, \ldots, x_n, y_1, \ldots, y_n)$, где $Q_i \in \{\exists, \forall\}$, $i \leq n$, а формула $\phi(x_1, \ldots, x_n, y_1, \ldots, y_n)$ является бескванторной (при выполнении только этого условия говорят, что формула $\psi(x_1, \ldots, x_n)$ является предварённой нормальной формой), и, во-вторых, каждая минимальная подформула формулы $\phi(x_1, \ldots, x_n, y_1, \ldots, y_n)$ содержит не более одного сигнатурного символа (т.е., формула $\phi(x_1, \ldots, x_n, y_1, \ldots, y_n)$ составлена с помощью логических связок $\&$, $\vee$, $\rightarrow$ и $\neg$ из бескванторных формул, содержащих не более одного сигнатурного символа). Для произвольной формулы φ существует формула ψ, называемая приведённой нормальной формой формулы φ, такая что $\psi \equiv \varphi$, и ψ является приведённой нормальной формой.

Заметим, что, поскольку рассматриваемые нами предложения не содержат выражений "все", "для всех", "существует ... такой, что...", и т.д., мы должны проводить аналогию с бескванторными формулами исчисления предикатов. Далее, так как наши предложения не содержат условных конструкций "если ..., то ...", конструкций вида "..., или, ..." и "либо ..., либо ..." и сложных отрицаний "не верно, что ...", соответствующие формулы исчисления предикатов составлены из атомарных формул при помощи конъюнкции $\&$ и отрицания $\neg$, причём отрицание стоит только перед атомарными подформулами.

Как правило, в исчислении предикатов последовательность формул $\varphi_1, \ldots, \varphi_n$ означает (семантически, по смыслу) то же самое, что и их конъюнкция $\varphi_1 \& \ldots \& \varphi_n$. Поэтому в качестве простейших формул достаточно рассматривать (в нашем случае) только атомарные формулы (в каждую из которых входит не

более одного сигнатурного символа). Вместо конъюнкции у нас будет следование одних формул за другими.

Обратим внимание, что, хотя отношение следования не является коммутативным (φ следует за ψ — это не то же самое, что ψ следует за φ), а логическая связка конъюнкция является коммутативной, мы ничего не нарушаем при такой замене. Дело в том, что в естественном языке союз "и", как и отношение следования, не является коммутативным, причём по той же самой причине. Для примера достаточно рассмотреть два предложения: "Вася пошёл домой и заснул" и "Вася заснул и пошёл домой". В естественном языке союз "и" содержит оттенок временного следования.

Наконец, отрицания предикатов мы можем рассматривать как новые предикаты. В некоторых случаях такое рассмотрение является даже необходимым. Действительно, смысл фразы "Фрося не любит Васю" отличается от "Не верно, что Фрося любит Васю" (здесь не работает закон исключённого третьего — Фрося к Васе может быть равнодушна). Здесь мы видим, что в естественном языке отрицание предиката может иметь несколько иной смысл, чем простое логическое отрицание, оно является усиленным. Поэтому вместо отрицания предиката $\neg P$ более разумно рассматривать новый предикат P'. Таким образом, в качестве самых простых формул мы можем рассматривать только позитивные атомарные формулы.

Итак, наша цель — свести текст естественного языка к последовательности предложений, логическим аналогом которой будет последовательность позитивных атомарных формул. В случае исчисления предикатов каждая такая атомарная формула содержит не более одного сигнатурного символа.

В этом месте мы несколько отойдём от аналогии с логикой предикатов. Дело в том, что в естественном языке практически нет переменных (исключением являются местоимения кто-то, где-то куда-то и т.д.). Поэтому, введение новых переменных не упростит текст, а, наоборот, сделает его более громоздким. С другой стороны, слова естественного языка в некотором смысле удобнее мыслить не как предикаты, а как константы, обозначающие определённые объекты. В качестве единственного предиката

можно рассматривать двуместное отношение, означающее, что два слова (т.е., два понятия, две константы) связаны друг с другом. Например, фразы "Поезд едет", "Цветок красный", "Идёт в школу" можно мыслить соответственно как предикаты P(поезд, едет). P(цветок, красный) и P(идёт, в школу).

Таким образом, в качестве самых простых предложений, к последовательности которых сводится произвольный текст естественного языка, мы можем рассматривать предложения, состоящие из двух слов (не считая предлогов и т.п.). Смыслом данного предложения мы будем считать связь этих двух слов (двух понятий). Соответствующими формулами исчисления предикатов будут формулы вида $P(c, d)$, где P — предикат связи, а c и d — две константы, обозначающие два произвольных слова, понятия естественного языка.

Обратим внимание, что мы не можем просто так оставлять константы, соответствующие определённым словам естественного языка. Дело в том, что одним и тем же словам могут соответствовать разные понятия (объекты). Например, после преобразования фразы "Вася пошёл в кино, а Петя пошёл в лес." в последовательность <"Вася пошёл.", "Пошёл в кино.", "Петя пошёл.", "Пошёл в лес."> два вхождения слова "пошёл" имеют разное значение. Это особенно легко понять, если мы рассмотрим преобразование фразы "Вася медленно пошёл в кино, а Петя быстро пошёл в лес." в последовательность <"Вася пошёл.", "Пошёл в кино.", "Пошёл медленно.", "Петя пошёл.", "Пошёл в лес.", "Пошёл быстро."> . Здесь слово пошёл имеет два противоположных свойства — "быстро" и "медленно". Очевидно, что мы здесь имеем дело не с одним, а с двумя разными действиями, имеющими соответственно и разные свойства.

Для того чтобы избежать коллизии, мы должны на эти слова ставить разные метки, например, нумеровать их. На первый взгляд это ведёт к загромождению текста. Но на самом же деле никакой потери здесь не происходит, поскольку и без преобразования, упрощения текста мы имеем дело с последовательностью предложений, в которых одинаковые слова имеют разные значения. Поэтому в любом случае, даже не упрощая предложения,

мы не сможем обойтись без маркировки разных вхождений одного и того же слова, имеющих разные значения.

Рассмотрим теперь преобразование текста естественного языка на примере упрощения одного предложения. В случае последовательности предложений процедура будет аналогичной.

Возьмём предложение:

"Глупый мальчик Вася пошёл в тёмный дремучий лес собирать разные грибы, но он забыл взять с собой лукошко и не знал, какие грибы хорошие, а какие — плохие, т.е. поганки."

Преобразование предложения будем проделывать в три этапа.

1. Разбивка сложного предложения на простые и добавление пропущенных слов.

"Глупый мальчик Вася пошёл в тёмный дремучий лес собирать разные грибы. Он забыл взять с собой лукошко. *Он* не знал, какие грибы хорошие. *Он не знал,* какие *грибы* плохие, т.е. поганки."

2. Рассмотрение предложения в контексте предыдущих предложений:

а) Отождествление слов (понятий), которые означают одно и то же:

"Глупый мальчик Вася пошёл в тёмный дремучий лес собирать разные грибы. Он[1] забыл взять с собой лукошко. Он[1] не знал, какие грибы хорошие. Он[1] не знал, какие грибы плохие, т.е. поганки."

б) Разделение одинаковых слов (понятий), имеющих разное значение (нумерация синонимов):

"Глупый мальчик Вася пошёл в тёмный дремучий лес собирать разные грибы[1]. Он[1] забыл взять с собой лукошко. Он[1] не знал[1], какие[1] грибы[2] хорошие. Он[1] не знал[2], какие[2] грибы[3] плохие, т.е. поганки."

в) Означивание ссылок, т.е., замена переменных (местоимений и др.) константами (определёнными словами, понятиями):

"Глупый мальчик Вася[1] пошёл в тёмный дремучий лес собирать разные грибы[1]. Вася[1] забыл взять с собой лукошко. Вася[1] не знал[1], какие[1] грибы[2] хорошие. Вася[1] не знал[2], какие[2] грибы[3] плохие, т.е. поганки."

3. Разбивка предложений на минимальные, означающие, по существу, связь двух понятий:

"Вася[1] мальчик[1]. Мальчик[1] глупый. Вася[1] пошёл[1]. Пошёл[1] в лес[1]. Лес[1] тёмный. Лес[1] дремучий. Пошёл[1] собирать[1]. Собирать[1] грибы[1]. Грибы[1] разные. Вася[1] забыл[1]. Забыл[1] взять[1]. Взять[1] с собой. Взять[1] лукошко. Вася[1] не знал[1]. Не знал[1], какие[1]. Какие[1] грибы[2]. Какие[1] хорошие. Вася[1] не знал[2]. Не знал[2], какие[2]. Какие[2] грибы[3]. Какие[2] плохие[1]. Плохие[1], т.е. поганки."

Мы получили последовательность предложений, каждое из которых достаточно примитивно (практически, предельно примитивно). Тем не менее, если читать полученный текст по порядку, его смысл воспринимается (хотя, конечно, не так легко, как смысл исходного предложения).

В целом, полученный текст сохраняет смысл исходного предложения, хотя, конечно, некоторые оттенки смысла теряются. Например, такая потеря произошла на первом этапе, когда мы, разбивая исходное предложение на простые, удалили союз "но".

С другой стороны, на наш взгляд, изменение оттенков смысла при переходе к некоторой канонической форме текста неизбежно по существу. Действительно, разные формы одного и того же по смыслу предложения тем и отличаются, что ставят ударение, оттеняют разные детали. Например, "Вася ловит рыбу" и "Рыбу ловит Вася". При переходе к канонической форме текста может естественным образом произойти утрата таких оттенков.

Обратим также внимание, что на третьем этапе, при сведении простых предложений к минимальным, мы имели определённый произвол при определении порядка, в котором будут следовать полученные минимальные предложения. Более того, произвол у нас был и при выборе пар. Например, вместо последовательности:

"Вася[1] мальчик[1]. Мальчик[1] глупый. Вася[1] пошёл[1]. Пошёл[1] в лес[1]. Лес[1] тёмный. Лес[1] дремучий. Пошёл[1] собирать[1]. Собирать[1] грибы[1]. Грибы[1] разные. Вася[1] забыл[1]. Забыл[1] взять[1]. Взять[1] с собой. Взять[1] лукошко. Вася[1] не знал[1]. Не знал[1], какие[1]. Какие[1] грибы[2]. Какие[1] хорошие. Вася[1] не знал[2]. Не знал[2], какие[2]. Какие[2] грибы[3]. Какие[2] плохие[1]. Плохие[1], т.е. поганки."

Мы могли бы записать последовательность:

"Вася[1] мальчик[1]. Мальчик[1] глупый. Мальчик[1] пошёл[1]. Пошёл[1] в лес[1]. Лес[1] тёмный. Лес[1] дремучий. Пошёл[1] собирать[1]. Собирать[1] грибы[1]. Грибы[1] разные. Мальчик[1] забыл[1]. Забыл[1] взять[1]. Взять[1] с собой. Взять[1] лукошко. Мальчик[1] не знал[1]. Не знал[1], какие[1]. Какие[1] грибы[2]. Какие[1] хорошие. Мальчик[1] не знал[2]. Не знал[2], какие[2]. Какие[2] грибы[3]. Какие[2] плохие[1]. Плохие[1], т.е. поганки."

Или:

"Мальчик[1] Вася[1]. Мальчик[1] глупый. Мальчик[1] пошёл[1]. ... "

При помощи изменения порядка минимальных предложений, а также используя имеющийся у нас некоторый произвол в выборе минимальных пар, мы можем даже в канонической форме менять смысловые оттенки текста.

Обратим внимание ещё на один момент использования канонической формы текста. Выписанная последовательность минимальных предложений задаёт, кодирует определённый порядок обращения локуса внимания на отдельные части предложения. Последовательность минимальных предложений определяет последовательность восприятия смысловых единиц, из которых состоит текст. Последовательное восприятие минимальных смысловых единиц текста порождает определённый контекст, в котором происходит понимание предложения. Различные последовательности минимальных предложений могут задавать различные контексты, различные способы восприятия одного и того же предложения.

Возможно, именно в этом феномене кроется причина того, почему разные люди воспринимают одно и то же предложение по-разному (и даже один человек может понять предложение различным образом в разные моменты времени).

Таким образом, выписанную в явном виде каноническую последовательность минимальных предложений можно понимать как формализацию определённого способа восприятия данного сложного предложения.

Формально построенная нами каноническая последовательность минимальных предложений выглядит так:

$$P(c_1, c_2), P(c_3, c_4), P(c_5, c_6), \ldots,$$

где P — предикат связи, а c_i — некоторые слова, понятия, возможно повторяющиеся. Эту последовательность (опустив порядок следования) можно изобразить в виде неупорядоченного графа, в вершинах которого стоят константы-понятия c_i, а каждый предикат $P(c_i, c_j)$ превращается в ребро, связывающее вершины c_i и c_j.

Такой граф является видоизменением сети понятий, подробно изучаемой в работах по методу ГАБЕК [1–9]. Разница состоит в том, что в нашем случае в узлах графа находятся только понятия, там нет предложений. В принципе, предложениями можно метить рёбра графа, если в этом есть необходимость. Такое помеченное ребро показывает не только то, что два понятия являются связанными, но также указывает, какое предложение осуществляет эту связь.

Таким образом, построенный нами граф показывает не синтаксическую близость предложений [10, 11], а семантическую близость понятий.

В заключение кратко отметим, как от рассмотрения описательных предложений можно перейти к произвольным. В случае описательных предложений, не содержащих сложных логических конструкций, по существу, единственной связкой между минимальными предложениями было отношение следования — одно минимальное предложение следует за другим. Если же мы будем иметь дело с произвольными предложениями, которые могут содержать сложные отрицания "не верно, что ...", конструкции вида "..., или, ...", условные конструкций "если ..., то ...", кванторы "для всех", "существует ... такой, что..." и др., то, очевидно, одного отношения следования уже будет не достаточно.

В качестве минимальных предложений естественно рассматривать те же предложения, что и раньше, но к ним необходимо будет применять логические связки.

Поэтому, каноническим представлением текста будет не последовательность минимальных предложений, а последовательность булевых комбинаций минимальных предложений.

Литература

1. ZELGER J. Ein ganzheitliches Verfahren zur Bewältigung sprachlich erfaßbarer Komplexität (GABEK). — Innsbruck, 1990. — (Preprint /Institut für Philosophie der Universität; 10).

2. ZELGER J. GABEK, a new method for qualitative evaluation of interviews and model construction with PC-support// Enchanging human capacity to solve ecological and socio-economic problems /Stuhler E.,Suilleabhain M.O., (eds.). — Munchen, Mering: Rainer Hampp Verlag, 1993. — P. 128-172.

3. ZELGER J. GABEK I: Constituting conceptual networks. II: Evaluation of conceptual networks. III: From conceptual networks to linguistic Gestalten. — Innsbruck, 1990. — (Preprints /Institut für Philosophie der Universität Innsbruck; 15, 16, 17).

4. ZELGER J. Von sprachlichen Gestalten zur Hypergestalt //Philosophie und Verfarhren kreativer Selbstorganisation, Innsbruck, 1993. — (Preprint /Institut für Philosophie der Universität Innsbruck; 26).

5. ZELGER J. Sprachlichen Gestaltbildung durch das PC-unterstützte Verfahren GABEK. — Innsbruck: Universität, Institut für Philosophie, Manuskript 1993.

6. ZELGER J. Qualitative Auswertung sprachlicher Äußerungen. Wissensvernetzung, Wissensverarbeitung und Wissensumsetzung durch GABEK. //Begriffliche Wissensverarbeitung: Grundfragen und Aufgaben //Wille R., Zickwolff M. (Hrsg.): Mannheim (B.I. Wissenschaftsverlag), 1994. P. 239–266.

7. ПАЛЬЧУНОВ Д.Е. Анализ текстов естественного языка с помощью метода ГАБЕК //Модели когнитивных процессов. — Новосибирск, 1977. — Вып. 158: Вычислительные системы. — С. 149–166.

8. ПАЛЬЧУНОВ Д.Е. Алгебраическое описание смысла высказываний естественного языка //Модели когнитивных процессов. — Новосибирск, 1997. — Вып. 158: Вычислительные системы. — С. 127–148.

9. PAL'CHUNOV D.E. Algebraishe Beschreibung der Bedeutung von Auberungen der naturlichen Sprache. Hrsg.: Josef Zelger & Martin Maier, Qualitative Forschung: Auf dem Weg vom Entdecken zum Verwerten mit dem Verfahren GABEK, 1998.

10. **ПАЛЬЧУНОВ Д.Е.** Синтаксическая близость предложений языка первого порядка //Измерение и модели когнитивных процессов. — Новосибирск, 1998 — Вып. 162: Вычислительные системы. — С. 58–80.

11. **PAL'CHUNOV D.E.** On a logical analysis of GABEK. Problem-& Konfliktbearbeitung durch GABEK. 2 Internationales GABEK-Symposiums, Sterzing, 1998.

Поступила в редакцию
30 декабря 1998 года