Оценка эффективности автоматического и гибридного подхода к извлечению знаний из текстов на ЕЯ на примере модели логического анализа фраз

Крайванова В.А.[1], Крючкова Е.Н.[1]

[1] *Алтайский государственный технический университет им. И.И. Ползунова, пр. Ленина, д. 46, Барнаул, 656038, Россия*

krayvanova@yandex.ru

Аннотация. *В работе рассматриваются механизм извлечения из текста на ЕЯ знаний для модели логического вывода, основанной на правилах. Оценивается эффективность предложенного механизма при автономной работе с неадаптированными текстами и в гибридной человеко-машинной системе.*

Ключевые слова: человеко-машинный интерфейс, естественные языки, извлечение знаний, гибридные системы.

1 Введение

Логический анализ является одним из наиболее эффективных подходов к интеллектуальной обработке текстовой информации на естественных языках(ЕЯ) [1]. Поскольку многие прикладные области анализа ЕЯ не требуют глубокого понимания дискурса, мы будем говорить о буквальном понимании ЕЯ.

Предлагаемая модель логического анализа фраз на ЕЯ основана на правилах контекстной замены и реализует комбинацию подходов на основе шаблонов и семантических словарей[2]. Рассматривается лишь та семантика, которая может быть непосредственно извлечена из формальной составляющей языка - синтаксиса.

Эффективность любой интеллектуальной системы определяется, прежде всего, полнотой и неизбыточностью базы знаний. Построение и поддержание в актуальном состоянии такой базы требует больших усилий и временных затрат. Ниже предложен механизм извлечения знаний из текста на ЕЯ для модели логического анализа и анализ эффективности этого механизма.

2 Математическая модель

Для моделирования естественного языка необходимо задать следующие :

- лексикон как множество известных в модели слов;
- формальное представление фразы ЕЯ на основе ее синтаксической структуры;
- представление команд системы управления;

- механизм анализа фраз.

В качестве элемента лексикона возьмем синтаксически неделимую единицу Ц слово. Из множества W слов предметной области выделим множество N понятий естественного языка. Нечеткое **отношение обобщения** $Gen : W \times N \rightarrow [0..1]$ определяет степень уверенности модели в том, что некоторое понятие $p \in N$ обобщает слово $w \in W$. Нечеткое **отношение синонимичности** $Syn : W \times W \rightarrow [0..1]$ определяет степень уверенности системы в том, что некоторое слово $w_A \in W$ является синонимом к слову $w_B \in W$.

Формально, фразу на естественном языке определим как функциональную форму, в которой в качестве функциональных символов выступают слова лексикона W, а в качестве типов аргументов - подчинительные синтаксические отношения. Пусть Rel - конечное множество типов аргументов модели. Функциональную форму ϕ некоторой ЕЯ-фразы можно также представлять в виде дерева, вершины которого помечены словами $w \in W$, а ребра - типами аргументов $r \in Rel$. Пусть u - узел помеченного дерева фразы. Тогда $Word(u)$ - пометка узла u, $Relation(u)$ - пометка ребра, соединяющего узел u с родительским, $Args(u)$ - множество дочерних узлов узла u. $Negation(u)$ - признак отрицания узла u.

Получить такое помеченное дерево фразы из предложения на ЕЯ, можно, например, с помощью библиотек Dialing [3].

Команда системы управления представляется парой: $e = \langle \nu, Parameters \rangle$, где ν - функциональная форма, задающая определение команды на ЕЯ. Некоторые функциональные символы в ν могут быть помечены не словами из W, а параметрами из множества $Parameters$. $Parameters = \{p | p = \langle name_i, N_i \rangle, N_i \subseteq N\}$ - множество параметров команды, где $name_i$ - имя параметра, N - множество понятий.

В узлы дерева ν, помеченные параметром $name_i$, может быть подставлено некоторое дерево ϕ, такое, что $\exists n : Gen(Word(Root(\phi)), n) > 0, n \in N_i$.

Задача логического анализа заключается в том, чтобы свести некоторую начальную фразу ϕ к множеству фраз, определяющих команды и понятных объекту управления. Механизм преобразования фраз описывается конечным множеством правил контекстной замены $Rules$. В общем случае правило задается следующим образом:

$r_L = \langle \nu \rightarrow \sigma, Parameters \rangle$,

где ν и σ - гипотеза и следствие правила, представленные двумя суперпозициями, некоторые функциональные символы которых, возможно, помечены не словами ЕЯ, а параметрами из множества $Parameters$. $Parameters$ - множество параметров правила, аналогичное множеству параметров команды.

Команда в этом случае является частным случаем правила - заключительным правилом, для которого следствие σ - пустое дерево.

Цепочка вывода из функциональной формы ϕ - это конечная последовательность функциональных форм $< \phi_0, \phi_1, ...\phi_n >$, таких, что $\phi_0 = \phi$, форма и ϕ_{i+1} непосредственно выводима из ϕ_i для $0 < i < n - 1$ с помощью некоторого правила $r_i \in Rules$, форма ϕ_n получена из формы ϕ_{n-1} с помощью заключительного правила $r_n \in Rules$.

На основе введенных определений задачу логического анализа функциональной формы ϕ можно сформулировать следующим образом: найти множество E_ϕ всех заключительных функциональных форм, выводимых из данной функциональной формы ϕ с помощью цепочек вывода длины не более $Lenght$. Очевидно, что при такой постановке задачи множество E_ϕ конечно.

Если рассматривать данную модель с точки зрения практической применимости, возникает проблема, типичная для интеллектуальных систем: высокая трудоемкость построения базы знаний. Далее предложен механизм пополнения логических знаний модели и оценка его эффективности.

3 Задача извлечения знаний

Правило контекстной замены может быть описано фразой на естественном языке. Рассмотрим в качестве объекта управления саму модель. Множество параметров для заключительных правил в этом случае состоит из двух элементов: $HYPOTHESIS и $CONSEQUENCE. В эти параметры могут быть подставлены любые функциональные формы. Результатом логического анализа в этом случае являются пары пропозициональных (т.е. не содержащих параметров) функциональных форм $\langle \phi_{hyp}, \phi_{con} \rangle$.

Рассмотрим механизм выделения параметров. Пусть по завершении вывода в параметры подставлены следующие функциональные формы:

$Values = \{\langle \$HYPOTHESIS, \phi_{hyp} \rangle, \langle \$CONSEQUENCE, \phi_{con} \rangle\}$

Тогда получаемое правило будет иметь вид:

$r = \langle \phi'_{hyp} \to \phi'_{con}, Parameters \rangle$,

где ϕ'_{hyp} и ϕ'_{con} получены из ϕ_{hyp} и ϕ_{con} заменой некоторых слов на параметры. Множество параметров строится следующим образом:

$$\begin{aligned} Parameters = \quad & \{p_i | p_i = \langle name_i, N_i \rangle : \\ & \exists u \in \phi_{hyp} : \exists w \in W : Gen(w, Word(u)) > 0, \\ & N_i = \{Word(u)\}, \\ & N_i \neq N_j \text{ при } i \neq j\} \end{aligned}$$

Узел u из ϕ_{hyp} или из ϕ_{con} помечается параметром $p_i = \langle name_i, N_i \rangle$, если $Word(u) \in N_i$.

Предложенный механизм позволяет пополнять знания модели как на начальном этапе построения базы знаний, так и в процессе эксплуатации. Поскольку извлечение знаний для модели представляет требует более глубокого анализа текста, чем просто буквальное понимание, предложенный механизм имеет следующие ограничения:

- правило должно описываться одним предложением;
- при формировании параметров не учитываются синонимы и местоимения.

Далее представлена оценка эффективности предложенного механизма для извлечения знаний из необработанного текста (самостоятельная автономная работа) и для адаптированного текста (гибридная человеко-машинная система).

4 Вычислительный эксперимент

Эффективность модели извлечения знаний проанализирована в серии вычислительных экспериментов.

Исходные данные для экспериментов - тексты советов по различным тематикам, взятые из сети Интернет. Рассмотрены три предметные области:

1. советы по ведению домашнего хозяйства;
2. советы по уходу за комнатными растениями;
3. советы начинающим программистам.

Тексты для каждой предметной области составлены на основе нескольких сайтов. В первой предметной области правила сформулированы наиболее четко. Тексты третьей предметной области фактически не содержат четко сформулированных правил.

В связи с особенностями представления информации в Интернет исходные тексты прошли предварительную техническую подготовку, которая позволяет избежать ошибок на этапе синтаксического анализа. Это удаление нетекстовых объектов (рисунков, таблиц) и исправление орфографических и синтаксических ошибок. Авторские формулировки полностью сохранены. Под необработанными текстами ниже подразумевается текст, прошедший предварительную техническую подготовку.

Роль человека в гибридной человеко-машинной системе извлечения знаний для предложенной модели логического анализа заключается в адаптации необработанного текста. Это устранение местоимений, упрощение формулировок, выделение неявных определений в отдельные предложения. При адаптации текста фразы, не содержащие правил, сохранялись в тексте, чтобы смоделировать поведение модели в процессе эксплуатации.

Для оценки результатов извлечения знаний введем обозначения.

R_{expert} - количество правил, обнаруженных экспертом. Учитываются только правила, полностью содержащиеся в одном предложении, например, «Чтобы сервировочные ножи блестели, их нужно почистить» - это правило, а такая конструкция: «Для той же цели подойдет разбавленный настой чая» - нет.

R_{auto} - количество правил, извлеченных автоматической системой.

$R_{relevant}$ - количество релевантных правил, извлеченных автоматической системой.

Качество извлеченных знаний оценивалось по двум параметрам:

- $Precision = R_{relevant}/R_{auto}$ - точность;
- $Recall = R_{relevant}/R_{expert}$ - полнота;

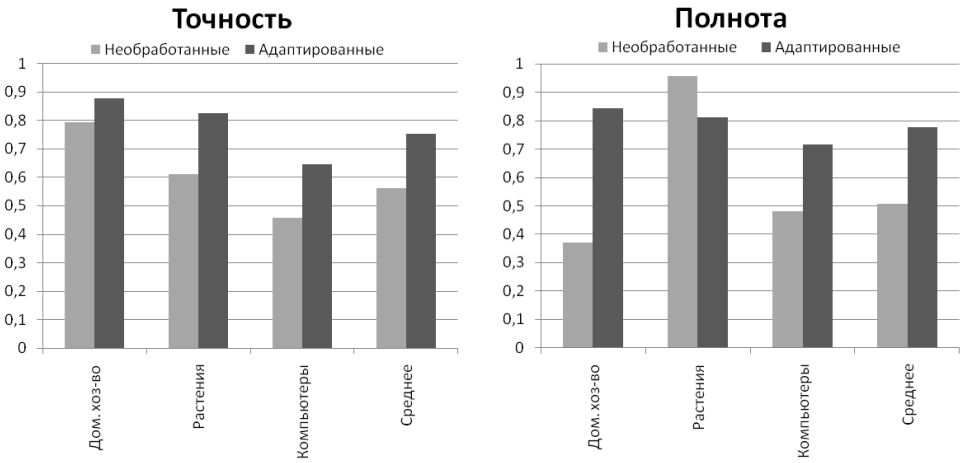

Рис. 1. Точность и полнота

Как видно из рисунка 1, поведение модели на необработанных текстах существенно зависит от текста. Полнота результатов анализа адаптированных текстов мало зависит от типа текста и составляет в среднем $0,77$, что значительно выше аналогичного показателя для необработанных текстов. Точность анализа адаптированных текстов зависит от сложности текста, но при этом заметно выше точности для необработанных текстов.

Важным количественным показателем в эксперименте является $R_{relevant}$ - количество полученных релевантных правил. Как видно на рисунке 2, использование гибридного подхода (т.е. предварительной адаптации текста) позволяет извлечь примерно в $2,6$ раза больше правил.

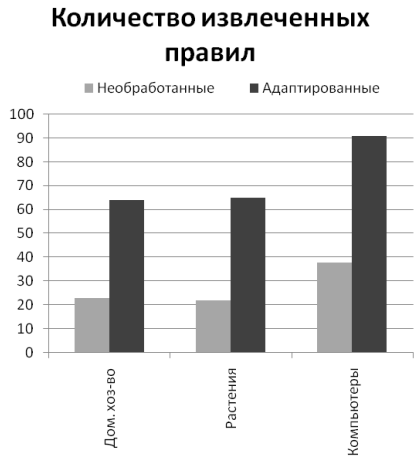

Рис. 2. Количество извлеченных правил

Выводы

Результаты вычислительного эксперимента показывают, что качество извлечения знаний из необработанного текста существенно зависит от структуры этого текста. Применение гибридной системы позволяет извлечь значительно большее количество правил.

Список литературы

[1] *Люгер Д. Ф. Искусственный интеллект: стратегии и методы решения сложных проблем 4-е издание. - М.: Вильямс, 2003.*

[2] *Найханова Л.В., Евдокимова И.С. Методы и алгоритмы трансляции естественно-языковых запросов к базе данных в SQL-запросы. - Улан-Удэ: Издательство ВСГТУ, 2004.*

[3] *Сайт рабочей группы "Автоматическая обработка текстов"[Электронный ресурс] / Режим доступа: http://aot.ru/*