УДК 004.82

ИЗВЛЕЧЕНИЕ ПРИЧИННО-СЛЕДСТВЕННЫХ ЗАКОНОМЕРНОСТЕЙ ИЗ ЕСТЕСТВЕННО-ЯЗЫКОВЫХ ТЕКСТОВ

© Маслов П.П.

Новосибирский государственный технический университет
факультет прикладной математики и информатики
пр-т К. Маркса, 20, г. Новосибирск, 630092, Россия

e-mail: *mpp84@rambler.ru*

Abstract. The paper considers analysis methods of texts in natural language. It proposes 1) the model of cause-and-effect relations in Russian text; 2) application of derived relations as expert statements in the algorythm of extremal situations forecasting.

Введение

Для современного состояния информационных систем и информационных ресурсов характерны всевозрастающие объемы неструктурированной информации, которая представлена различного рода текстовыми документами на естественном языке (ЕЯ).

Эффективность системы обработки естественного языка (ОЕЯ) определяется ее целями и методами, применяемыми для формализации и обработки ЕЯ-информации. Системы ОЕЯ, ориентированные на извлечение определенного смысла из текстов, в той или иной степени используют синтактико-семантические объекты текста. Среди работ в этой области необходимо отметить такие как «Модель извлечения фактов из естественно-языковых текстов и метод ее обучения» [1] ; систему комплексного смыслового анализа «TextAnalyst» [2]; «методы и программные средства для анализа документов на основе модели предметной оболасти» [3]; «Диалинг» [4]. Благодаря свободнораспространяемому семантическому анализатору группы разработчиков aot.ru [5] этот программный компонент был взят за основу предлагаемого подхода.

В работе предлагается способ извлечения и описания причинно-следственных фактов из текстов жанра деловой прозы на русском языке. Актуальность работы в том, что извлекаемые факты, выражающие причинно-следственные связи, могут, например, являться источником экспертных знаний для алгоритмов предсказания экстремальных ситуаций [6].

1. Модель представления причинно-следственных фактов

Деловой прозе свойственны жесткие средства выражения, однозначность передаваемой информации, экономичность языковых средств, четкость функции каждого сообщения и другие полезные свойства. Такой жанр, как правило, содержит информацию об объектах (событиях, явлениях, лицах и т.д.), которая практически не требует дополнительных сведений для их описания и может быть представлена набором фактов содержащихся непосредственно в анализируемом ЕЯ-тексте. Под фактами, описывающими причинно-следственные закономерности, понимаются объекты текста $s_i \in S$ (множество вершин именных групп (единичных лексем), сказуемых и

определений, согласованных синтаксически с подлежащими), семантически связанные отношениями $R_C \subseteq S \times S$, $R_A \subseteq S \times S \times S$ и группой отношений $RE \subseteq S \times S$. Ниже приведено более подробное описание объектов и связей между ними.

В качестве объектов будем рассматривать конечное множество $S = S^N \vee S^V \vee S^D$, где:

1. $S^N = \{s_1^N, ..., s_k^N\}$ – множество вершин именных групп (единичных лексем).
2. $S^V = \{s_1^V, ..., s_l^V\}$ – множество сказуемых, для которых выполняется $\forall s_i^V \in S^V, i = \overline{1,l} : \exists s_j^N \in S^N$ согласованный синтаксически с s_i^V.
3. $S^D = \{s_1^D, ..., s_m^D\}$ – множество определений, для которых выполняется $\forall s_i^D \in S^D, i = \overline{1,m} : \exists s_j^N \in S^N$ согласованный синтаксически с s_i^D.

Введем конечное множество отношений между объектами

$$R = \{r_1, ..., r_o\} = R_A \vee R_C \vee RE = R_A \vee R_C \vee R_{ED} \vee R_{AD} \vee R_{DC} :$$

1. $R_A = r_i(s_{i1}^V, s_{i2}^N, s_{i3}^N) \subseteq S^V \times S^N \times S^N$ – множество связей, описывающих сказуемые s_{i1}^V, синтаксически согласованные с подлежащими s_{i2}^N и дополнениями s_{i3}^N (s_{i2}^N или s_{i3}^N по отдельности могут быть пустыми)
2. $R_C = r_j(s_{i1}^N, s_{i2}^N) \subseteq S^N \times S^N$ – множество причинно-следственных связей, для которых $\forall s_{i2}^N \in S^N, \exists s_{i1}^N \in S^N : s_{i1}^N$ является семантической причиной (предпосылкой, условием и т.д.) для s_{i2}^N.
3. $R_{ED} = r_k(s_{i1}^N, s_{i2}^N) \subseteq S^N \times S^N$ – множество связей, устанавливаемых между эквивалентными (посредством знаков препинания «-», «:» , таких слов-объектов как "быть", "являться"и т.д.) по тексту объектами $s_{i1}^N, s_{i2}^N \in S^N$.
4. $R_{AD} = r_l(s_{i1}^N, s_{i2}^N) \subseteq S^N \times S^N$ – множество анафорических связей, таких, что s_{i1}^N, s_{i2}^N ссылаются на один и тот же по тексту объект (в частности $s_{i1}^N \subset S^{PN} \subseteq S^N$, где S^{PN} – множество именных групп (единичных лексем) с местоимением в качестве главного элемента).
5. $R_{DC} = r_m(s_{i1}^N, s_{i2}^N) \subseteq S^N \times S^N$ – множество связей, таких, что s_{i1}^N, s_{i2}^N эквивалентны по тексту, при этом s_{i1}^N принадлежит главному , а s_{i2}^N придаточному предложениям , связанным посредством таких слов-объектов, как «быть», «являться» и т.д. в сочетании с союзами и союзными словами или без таковых.

Если для объектов $s_i \in S^N$, связанных отношением R_C, существуют другие отношения R_C, RE, R_A, то в этом случае возможно выявление дополнительных причинно следственных закономерностей, элементы которых на семантическом уровне связаны иерархически (R_C, R_A), либо эквивалентны RE.

Указание свойств связей осуществляется посредством атрибутов $A = a\{r, v\} \subseteq R \times V$, где V – множество допустимых значений атрибутов. Атрибуты делятся на $A_A \subseteq A$ для описания свойств симметричности, транзитивности, рефлексивности и т.д. и, $A_V \subseteq A$ для указания значений стандартных типов (string, float, integer), например, для указания вероятностных характеристик причинно-следственных связей.

Введем подмножества $S_1^{SN}, ..., S_h^{SN}$ для которых существуют симметричные, рефлексивные и транзитивные отношения RE. На каждом подмножестве необходимо определить число n_i, $i = \overline{1, |S_p^{SN}|}$, $p = \overline{1,h}$ вхождений в текст данной лексемы.

В каждом множестве S_h^{SN} выделяются подмножества S_h^{SSN}, состоящие из имен собственных (имена, географические названия и т.п.). Множество S_h^{SN} упорядочивается следующим образом: $\{s_1, ..., s_g, s_{g+1}, ..., s_f\}$, $s_1, ..., s_g \in S_h^{SSN}$, $s_{g+1}, ..., s_f \in S_h^{SN} \backslash S_h^{SSN}$, $w = \frac{n_i}{|S_h^{SN}|}$, $w(s_i) \geq w(s_{i+1})$.

Упорядоченные указанным способом наборы лексем являются аргументами причинно-следственных фактов выводимых из текста, где каждый такой набор эквивалентных объектов помечается уникальным идентификатором. Формирование результата осуществляется за счет поиска всех возможных подстановок в аргументы причинно-следственных связей R_C, с учетом упорядочения объектов. При этом в первую очередь выводятся факты, аргументы которых имеют максимальный вес, затем соответственно по уменьшению весов. Для причинно-следственных связей $r_1(s_{i,1}^N, S_{j,1}^N), ..., r_k(s_{i,n}^N, S_{j,n}^N)$, где для аргумента-следствия (результата) l-й связи и аргумента-причины (предпосылки, условия и т.д.) $l+1$-й существую эквивалентные отношения RE, выполняется аналогичный вывод, упорядоченный в соответствии с количеством связей R_C, обладающих указанными условиями. Это позволяет упорядочить факты в зависимости от того является ли объект непосредственной причиной в факте или косвенной (аналогично для результата), а также от весов объектов.

Заключение

В настоящее время предлагаемый подход находится в стадии практической реализации, и выполнен в виде системы логического вывода причинно-следственных закономерностей в среде разработки PDC Visual Prolog 5.2. На данном этапе реализованы отношения R_C, R_A, R_{ED}, частично выполнены учет весов объектов и атрибутов первого типа. Помимо указанного выше способа упорядочения объектов в наборах в дальнейшем предполагается использовать метод извлечения доминантных словосочетаний [7].

Список литературы

1. *Андреев А.М., Березкин Д.В., Симаков К.В.* Модель извлечения фактов из естественно-языковых текстов и метод ее обучения // 6-ая Всероссийская научная конференция RCDL'2004.
2. TextAnalyst tm, Microsystems, Ltd. http://www.analyst.ru/
3. *Сидорова Е.А.* методы и программные средства для анализа документов на основе модели предметной обоасти: Дисс. ... канд. ф.-м. наук: 05.13.11. – Новосибирск, 2006. – 125 с.
4. *Группа разработчиков aot.ru* Проект русско-английского машинного перевода «Диалинг» // http://www.aot.ru.
5. *Группа разработчиков aot.ru* Пример вызова поверхностной семантики на Delphi // http://www.aot.ru/docs/delphi_seman_test.html.
6. *Лбов Г.С., Бериков В.Б.* Прогнозирование экстремальных ситуаций на основе анализа многомерных разнотипных временных рядов и экспертных высказываний // Материалы всероссийской конференция с международным участием "Знания-Онтологии-Теория"(ЗОНТ-07), том 1, С. 59-62.
7. *Чанышев О.Г.* Автоматическое извлечение доминантных словосочетаний // Материалы всероссийской конференция с международным участием "Знания-Онтологии-Теория"(ЗОНТ-07), том 1, С. 236-245.

Статья поступила в редакцию 30.04.2008