Знания-Онтологии-Теории (ЗОНТ-09)

Обнаружение и извлечение причинно-следственных закономерностей из текста на естественном языке

П.П. Маслов

Новосибирский государственный технический университет, пр. К. Маркса, 20, г. Новосибирск, 630092, Россия

mpp84@rambler.ru

Аннотация. *В статье приведено описание метода извлечения причинно-следственных (ПС) закономерностей из текста на естественном языке. Введена модель представления ПС- и иных связей над объектами текста. Описаны этапы декларативной обработки текста. Для фрагментационного анализатора описаны общая схема функционирования, типы выделяемых фрагментов, а также алгоритмы для их выделения. Для синтаксического анализатора указана общая схема функционирования, виды синтаксических групп, формируемых посредством правил синтаксического связывания. Разработан язык описания синтаксических правил, позволяющий без перестроения программы адаптировать анализатор под определенную предметную область либо для выделения отдельных языковых конструкций.*

Ключевые слова: обработка естественного языка, извлечение фактов, причинно-следственные закономерности

1 Введение

Современные информационные системы характеризуются постоянно растущими объемами неструктурированных данных, представленных различными документами на естественном языке (ЕЯ). Эта тенденция способствует возникновению средств обработки ЕЯ-текстов, направленных на формирование определенного сжатого представления содержания анализируемых документов. Такое со-держание представляет собой разнообразную фактографическую информацию, структура которой определяется различными факторами (предметная область, особенности обрабатываемых языковых механизмов и т.д.).

Актуальным в datamining является направление разработок, целью которых является извлечение продукционных (if-then) правил. Однако в контексте обработки ЕЯ-текстов (textminig) наблюдается явная нехватка подобных средств, что обуславливается несовершенством методов обработки текста, а также достаточно узкой специализацией прикладных разработок.

Предлагаемы в работе метод направлен на решение задачи извлечения фактографической информации, описывающей if-then правила, содержащиеся в тексте. В контексте работы if-then правила будут обозначаться как причинно-следственные закономерности (ПС-закономерности). Фактографической информацией, отражающей причинно-следственные закономерности, обозначаются сведения, описывающие некоторые синтактико-семантические объекты текста, являющиеся причинами соответствующих им других объектов (следствий). И наоборот: объектов-

следствий, соответствующих объектам-причинам. Важным с точки зрения автора является наличие в тексте нетривиальных ПС-закономерностей, описываемых различными отношениями, носящими характер транзитивности и эквивалентности [1].

2 Представление причинно-следственных закономерностей

Для деловой прозы характерны жесткие средства выражения, однозначность передаваемой информации, экономичность языковых средств и четкость функции каждого сообщения. 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 сведениями, содержащимися непосредственно в анализируемом ЕЯ-тексте.

Под фактами, описывающими причинно-следственные закономерности, понимаются объекты текста $s_i \in S$ (множество вершин именных, генетивных идругих синтаксическихь групп, а также единичных лексем), семантически связанные отношениями $R_C \subseteq S \times S$, $R_A \subseteq S \times S \times S$ и группой отношений $RE \subseteq S \times S$. Далее 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:

$S^N = \left\{ s_1^N, ..., s_k^N \right\}$ - множество вершин именных групп (единичных лексем)

$S^V = \left\{ s_1^V, ..., s_l^V \right\}$ - множество сказуемых, для которых выполняется $\forall s_i^V \in S^V, i = \overline{1,l} : \exists s_j^N$ согласованный синтаксически с s_i^V.

$S^D = \left\{ s_1^D, ..., s_m^D \right\}$ - множество определений, для которых выполняется $\forall s_i^A \in S^A, i = \overline{1,m} : \exists s_j^N$ согласованный синтаксически с s_i^D.

Введем конечное множество отношений между объектами s_i

$R = \left\{ r_1, ..., r_0 \right\} = R_A \vee R_C \vee RE = R_A \vee R_C \vee R_{ED} \vee R_{AD} \vee R_{DC}$:

1. $R_A = r_i \left(s_{i1}^V, s_{i2}^N, s_{i3}^V \right) \subseteq S^V \times S^N \times S^N$ - множество связей, описывающих сказуемые s_{i1}^V, синтаксически согласованные с подлежащими s_{i2}^V и дополнениями s_{i3}^V (s_{i2}^V или s_{i3}^V по отдельности могут быть пустыми).

2. $R_C = r_j \left(s_{i1}^N, s_{i2}^V \right) \subseteq S^N \times S^N$ - множество причинно-следственных связей, для которых $\forall s_{i2}^N \in S^N, \exists s_{i1}^N \in S^N : s_i^N$ является семантической причиной (предпосылкой, условием и т.д.) для s_{i2}^N.

3. $R_{ED} = r_k \left(s_{i1}^N, s_{i2}^N \right) \subseteq S^N \times S^N$ - 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. $R_{AD} = r_l \left(s_{i1}^N, s_{i2}^N \right) \subseteq S^N \times S^N$ - 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. $R_{DC} = r_m\left(s_{i1}^N, s_{i2}^N\right) \subseteq S^N \times S^N$ - 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 $s_i^N \in S^N$, 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.

Введем подмножества объектов, $S_1^{SN},...,S_h^{SN}$ для которых существуют отношения RE. На каждом подмножестве необходимо определить число $n_i, i = \overline{1,\left|S_p^{SN}\right|}$, $p = \overline{1,h}$ вхождений в текст данной лексемы. В каждом множестве S_k^{SN} выделяются подмножества S_k^{SSN}, состоящие из имен собственных (имена, географические названия и т.п.). Множество S_k^{SN} упорядочивается следующим образом: $\left\{s_1,...,s_g,s_{g+1},...,s_f\right\}$, $s_1,...,s_g \in S_k^{SSN}$, $s_g,...,s_f \in S_k^{SN} \setminus S_k^{SSN}$, $w_i = \frac{n_i}{\left|S_k^{SN}\right|}$, $w(s_i) \geq w(s_{i+1})$.

Упорядоченные указанным способом наборы лексем являются аргументами причинно-следственных фактов выводимых из текста.

3 Обработка текста и метод формирования причинно-следственных закономерностей

Обработка текстов в предлагаемом методе можно представить следующей последовательностью: графематический, морфологический, фрагментационный, синтаксический и постсинтакцический анализы текста и последующий вывод закономерностей. Программно метод реализован в виде двух компонент: компоненты обработки естественного языка (русского) и компоненты логического вывода (реализованной в среде prolog).

В качестве базы компоненты обработки естественного языка взята более ранняя разработка - информационно-поисковая система с лингвистической обработкой текста ISS3[2]. Далее приведено описание этапов анализа текста.

Графематический анализ. Из хранилища документов выбирается очередной файл, содержащий русскоязычный ЕЯ-текст. На этом этапе для каждой графемы определяется ее месторасположение в тексте (порядковые номера абзаца, предложения и графемы) а так же ее графематический тип (по аналогии с [3]).На этапе графематического анализа для каждого обработанного файла создается выходной файл, содержащий графематическое представление ЕЯ-текста.

Морфологический анализ. На этапе морфологического анализа, для каждой лексемы строится множество лемм с атрибутами. Каждая лемма представляет собой нормальную форму слова, а атрибуты – набор дескрипторов (часть речи, число, падеж и.д.). Для реализации морфологического анализа используется библиотека mcr.dll [4,5] и морфологический словарь А.А. Зализняка.

Фрагментационный анализ. На этапе фрагментационного анализа из предложений текста выделяются синтаксически целостные конструкции. Разработанный фрагментационный анализатор распознает следующие виды фрагментов:

- фрагменты в скобках или кавычках;
- причастные или деепричастные обороты;

- вводные слова или словосочетания;
- придаточные предложения со словами: который, какой, чей; что, чтоб, чтобы;
- фрагменты с однородными членами.

На вход фрагментационному анализатору подается фрагмент, состоящий целиком из предложения текста, к которому последовательно применяются правила фрагментации. При выполнении условий выделения фрагмента того или иного типа, происходит перестроение родительского фрагмента таким образом, что выделенные элементы удаляются из родительского фрагмента и образуют новый фрагмент. В родительском фрагменте добавляется новый, указывающий на субфрагмент. Исходя из того, что после выделения фрагментов, из родительского фрагмента возможно выделение других фрагментов с учетом уже выделенных, перестроенный родительский фрагмент заново подается на вход фрагментационному анализатору. Если в результате применения правил к фрагменту, он целиком удовлетворяет определенным правилам выделения, то перестроение не выполняется, а к ранее обработанному фрагменту добавляется новый тип.

При просмотре фрагмента на возможность применения правил фрагментации, возможно, что в качестве очередного элемента будет элемент, ссылающийся на субфрагмент. В этом случае процедура поиска запускается так же и для субфрагмента. При этом такая ситуация порождает новый цикл проверок для каждого субфрагмента. Если для таких субфрагментов ни одно правило не применимо, то они определенным образом помечаются как «бесперспективные» для дальнейшей фрагментации. Таким образом, значительно сокращается время анализа при повторной проверке фрагмента.

Синтаксический анализ. На вход синтаксическому анализатору подается дерево фрагментов, соответствующее одному предложению входного текста. Перебор субфрагментов в дереве начинается с выделения самого «низшего» субфрагмента в иерархии. После анализа очередного субфрагмента, ему, помимо исходно типа фрагмента, полученного на этапе фрагментационного анализа, присваивается синтаксический тип, соответствующий синтаксическому типу корневого элемента из корневой синтаксической группы. Таким образом, при дальнейшем анализе «вышестоящих» фрагментов, элемент, ссылающийся на дочерний элемент, может быть проинтерпретирован по синтаксическому типу его вершины. Если фрагмент имеет неполное связывание (см. далее), то синтаксический анализ для элемента, ссылающегося на такой фрагмент, выполняется только по типу фрагмента, полученного на этапе фрагментационного анализа.

Лексема фрагмента может быть связана синтаксическим правилом, если выполняются следующие условия:

- Графематический тип фрагмента соответствует лексеме, состоящей из букв русского алфавита (всего таких типа три, см. гл. 34234).
- Элемент является ссылкой на другой фрагмент (согласование по фрагментационному или синтаксическому типам вершины фрагмента).
- Элемент не связан каким-либо синтаксическим правилом.
- Элемент входит в какую-либо синтаксическую группу (покрыт правилом) и является корневым элементов в этой группе.

Результат синтаксического анализа представляет собой журнал последовательного применения правил к элементам субфрагмента (в общем случае фрагмента). При этом такой журнал позволяет без дополнительного описания интерпретировать «висячие» группы (группы, которые не связаны с общим деревом синтаксических групп во фрагменте). Таким образом, на этапе последующего постсинтаксического анализа для каждого элемента можно установить связь с другими элементами.

Правила синтаксического связывания. Связывание элементов фрагментов выполняется исходя из правил синтаксического связывания, последовательно применяемых на этапе синтаксического анализа. Для простоты описания и возможности изменения без перестроение программы синтаксические правила имеют внешнее текстовое описание в виде файла.

Структура синтаксических правил разработана таким образом, чтобы максимально охватить возможные типы элементов фрагментов и их свойства. При этом в качестве типов элементов могут выступать графематический, фрагментационный и синтаксический, а в качестве свойств элементов

могут рассматриваться как морфологические свойства конкретной лексемы, так и фиксированные и динамические (формируемые для группы на этапе синтаксического анализа) свойства. Ограничения, накладываемые на выделяемые цепочки лексем из фрагмента могут быть комбинированы с различным их учетом. Каждое правило также описывает синтаксический тип группы связанных элементов, корневой элемент, морфологические свойства и т.д.

В общем случае структура синтаксического правила имеет следующий вид:

n_1 n_2 *restrict-set_1 ... restrict-set_n1* *GBO* v_1 *R* v_2 *GID* v_3 *GD1* v_4 *GD2* v_5

Где указанные директивы имеют следующий смысл:

- n1 – количество наборов ограничений restrict-set_k
- n1 – элементы фрагмента покрываемых правилом должны быть расположены контактно друг к другу.
- restrict-set – набор ограничений к элементам фрагмента (см. далее)
- GBO v1 – режим последующего связывания корня группы, значение v1=1 соответствует режиму связывания только по идентификатору группы, v1=0 – связывание как по идентификатору группы так и по морфологическим признакам корневого элемента. Введение такого параметра необходимо, чтобы, например, для группы «числовой комплекс» с корневым элементом числительным, этот элемент не рассматривался при дальнейшем анализе как числительное и не образовывались неправильные связи. При этом синтаксические правила упорядочены таким образом, что в поднаборах сходных правил в первую очередь объявлены правила, рассчитаны на выделение наиболее полной структуры. Например, правило «дата» может охватывать число+месяц+год, либо в менее полной форме число+месяц.
- R v2 – корневой элемент группы. Значение v2 соответствует порядковому номеру элемента в группе, либо последнему элементу, в случае если v2=100.
- GID v3 – синтаксический идентификатор сформированной группы.
- GD1 v4 – фиксированные характеристики первого типа.
- GD1 v5 – фиксированные характеристики второго типа. Морфологические характеристики поделены на два типа по аналогии с результатом, выдаваемым морфологическим анализатором mcr.dll.

Параметр restrict-set представляет собой описание набора ограничений на элемент (элементы, см. далее), который заключен либо в (…), либо в ([…]), либо в ({…}), где:

- (…) – набор ограничений к одному элементу.
- ([…]) – набор ограничений к одному элементу, выполнение которого не обязательно. То есть, если текущий элемент не удовлетворяет такому набору ограничений, то анализатор проверяет следубщий набор ограничений для текущего элемента. Например, в группе «числовой комплекс» соответствующей цепочке «одна тысяча миллионов» первая графема может отсутствовать – «тысяча миллионов». В этом случае, «упрощенная» группа будет выделена, и для нее не потребуется дополнительного описания в виде правила.
- ({…}) – набор ограничений применимый к каждому элементу цепочки однотипных элементов, но хотя бы к одному элементу. Такие ограничения применимы к цепочке элементов, например, «сто двадцать три тысячи семьсот», где все элементы могут быть рассмотрены как цепочка числительных или существительных (см. далее). Но при этом цепочка не может быть пустой, как в предыдущем типе ограничений.

Структура набора ограничений restrict-set , на примере выделенного в (…), имеет следующий вид:

(v6 restrict1 ... restrictl SS v7 SSG v8), где:

- v6 указывает на количество ограничений restrict в наборе.

- restrict – ограничение на элемент (см. далее).
- SS v7 – список типов морфологических характеристик для согласования. Значение v7 составляет набор чисел от 1 до 5, разделенных символом «_». Значение чисел соответствует следующим типам морфологических характеристик: 1 – род, 2 – число, 3 – падеж, 4 – время, 5 – лицо. Например, значение параметра 1_2_3 – требует от проверяемого элемента фрагмента согласование по роду, числу и падежу с другими элементами, обозначенными директивой SS. Если у двух и более элементов, удовлетворяющих ограничениям с параметром SSG, значения соответствующих типов морфологических характеристик совпадает, то элементы считаются согласованными. Подробнее смотрите ниже на примере одного из правил синтаксического связывания.
- SSG v8 – аналогично параметру SS, только значение v8 определяет значения соответствующих типов морфологических характеристик, приписываемых в качестве динамических морфологических характеристики группы. Значения характеристик берутся от той лексемы, которая удовлетворяет набору ограничений, в который входит параметр SSG.

Параметр restrict представляет собой группы конкретных ограничений из списка, накладываемых на проверяемый элемент. Группы могут быть заключены либо в […], либо в [[…]], либо в [[[…]]], где:

- […] – группа ограничений, обязательных для удовлетворения.
- [[…]] – группа ограничений, в которой достаточно удовлетворения хотя бы одному ограничению.
- [[[…]]] – группа ограничений, которым обязательно не должен удовлетворять проверяемый элемент.

В качестве конкретных ограничений группы могут быть указаны ограничения следующего типа:

- TM – по графематическому типу (см. таблицу графематических кодов в гл. 2.32423)
- TMF – по типу фрагмента для элементов ссылающихся на другой фрагмент (см. типы фрагментов в гл. 2.33453452423)
- TMG – по синтаксическому типу корневого элемента ранее выделенной синтаксической группы.
- TM1 – по морфологическим характеристикам первого типа (см. морфологический анализатор.)
- TM2 – по морфологическим характеристикам второго типа (см. морфологический анализатор.)
- LIST – список конкретных значений в общем случае графемы, разделенных знаком «_».

Постсинтаксический анализ. На вход анализатору поступают проанализированные на этапе синтаксического анализа фрагменты с привязанными к ним журналами, описывающими последовательность применения синтаксических правил. Последовательность применения правил журнала обеспечивает наилучшее связывание элементов фрагмента. Таким образом, для любого элемента любого фрагмента, если элемент связан каким-либо правилом, можно однозначно установить взаимоотношения с другими элементами фрагмента.

В методе извлечения причинно-следственных закономерностей ключевыми являются связи Rc и RE. При этом остальные связи «квалифицируют», т.е. присваивают некоторые синтаксические признаки объектов, входящих в связей Rc, RE.

Задачей постсинтаксического анализа является формирование связей Rc, RE по журналу применения синтаксических правил к фрагменту. При этом структурными элементами выделенных связей должны быть конкретные элементы фрагментов, а не группы. Вместе с этим, если в связях Rc, RE структурные элементы (по журналу) связаны другими синтаксическими связями, то такие «квалифицирующие» связи так же формируются.

Функционирование постсинтаксического анализатора начинается с поступления на вход фрагментов, обработанных на этапе синтаксического анализа, и журналов применения правил синтаксического связывания. В этих журналах выполняется поиск применения синтаксических связей, соответствующих связям Rc, RE, при обнаружении которых создается запись, имеющая вид:

R{S1,…,Sn}, где

- R – одна из указанных выше связей.
- S1,…,Sn – объекты связей.

Извлечение причинно-следственных закономерностей. Сформированный на этапе постсинтаксической обработки фрагментов набор ключевых и квалифицирующих связей поступает на вход процедуре вывода причинно-следственных закономерностей. Указанная процедура, реализованная в среде prolog (PDC Visual Prolog 5.2), выполняет логический вывод над указанным набором связей .

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.

В результате анализа в первую очередь выдаются непосредственные причины следствий, затем косвенные. При этом вывод может выполняться как для конкретных причин или следствий (например, когда необходимо определить последствия или предпосылки каких-либо событий, отраженных в тексте), а так же полностью для всех закономерностей.

4 Заключение

В работе предложен метод извлечения и описания причинно-следственных закономерностей из текстов жанра деловой прозы на русском языке. Актуальность работы в том, что извлекаемые сведения, выражающие причинно-следственные связи, могут, например, являться источником экспертных знаний для различных алгоритмов предсказания [6].

Анализ систем АОТ показал наличие разнообразных средств, имеющих специализированное назначение, и как правило закрытый исходный код. Фактическое наличие информационно-поисковой системы ИПС ISS3 позволило не начинать разработку компоненты обработки текста «с нуля», а, выполнив определенную модернизацию, использовать уже имеющуюся систему.

Разработанный язык описания правил синтаксического связывания позволяет улучшить качество разбора, а так же выполнить адаптацию анализатора, на выделение конкретных языковых механизмов. Такая возможность обеспечивается за счет представления правил во внешнем файле.

Описанная модель представления ПС-закономерностей, а применение prolog для логического вывода, позволяют использовать средства искусственного интеллекта для формирования ПС-закономерностей.

В рамках дальнейшего развития метода планируется выполнить следующее:

- Расширить наборы правил синтаксического и постсинтаксического связывания.
- Разработать интерфейс к системе, позволяющий наглядно, наподобие систем поддержки принятия решений, указывать позиции ПС-закономерностей в тексте.
- Ввести процедуру обучения анализатора. Под обучение подразумевается определение параметров, преимущественно фрагментационного анализатора (например, линейная длинна оборотов), характерных для анализируемых тексов.

Литература

[1] Маслов П.П. Модель описания причинно-следственных фактов. // Материалы международной конференции «Горизонты прикладной лингвистики. MegaLing - 2008». Украина, Крым. 2008

[2] Gennadiy Lbov, Nikolai Dolozov, Pavel Maslov. Analysis and coordination of expert statements in the problems of intellectual information search //International journal "Information theories & applications", Bulgaria, Varna, FOI ITHEA. – 2007. - Vol.14, Iss. 1. - P. 95-99.

[3] Проект русско-английского машинного перевода "Диалинг" // http://www.aot.ru.

[4] Выдрин Д., Поляков В. (2002) "Реализация электронного словаря на основе н-грамм". Труды III Международной научно-практической конференции "Искусственный интеллект–2002" - Кацевелли, том 2, с. 79-84 Изд. "Институт проблем искусственного интеллекта" http://iai.donetsk.ua

[5] Выдрин Д., Громов С., Поляков В. (2004) "Метод сравнения библиографических описаний, представленных в различных форматах". Обработка текста и когнетивные технологии №9. VII Междуународная конференция - Варна. М: Издательство "Учеба", с. 166-172

[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, СС. 59-62.