Вестник РУДН Серия *Математика. Информатика. Физика.* № 3 (1). 2010. С. 61–68

Информатика

УДК 004.891

Рекурсивное решение проблемы обусловленности

Л. В. Савинич, В. Л. Стефанюк

Институт проблем передачи информации им. А.А. Харкевича РАН
Большой Каретный переулок, 19, стр.1, ГСП-4, 127994, Москва, Россия

Данная статья продолжает исследование способов формализации естественно-языковых текстов с целью представления их в виде совокупности строгих логических правил, пригодных для экспертных систем. На примере текстов регламентирующего характера, с опорой на их типологические особенности, анализируется синтаксическая структура предложений в соотношении с их смысловым содержанием.

Ключевые слова: обусловленность, типология текста, тема, рема.

1. Введение

Большая часть знаний, накопленных людьми, сегодня хранится в форме текстов на естественном языке. Особенно это характерно для гуманитарных дисциплин или для общедоступных популярных изложений научно-технических дисциплин. В последнем случае изложения часто ориентированы на богатую интуицию человека, который в состоянии все более глубоко проникать в смысл текста по мере того, как в этом возникает необходимость, иногда исправляя или уточняя сказанное. Ярким примером может служить область законодательства, которое приходится время от времени корректировать.

Даже такая чёткая и формализованная область, как программирование, нуждается в постоянном уточнении [1].

Современные применения компьютеров в прикладных задачах связаны с необходимостью введения отдельного специалиста — инженера по знаниям, который должен уметь извлекать необходимые продукционные правила из текстов и других источников, опираясь на свои собственные знания и интуицию. В наших работах сделана попытка автоматизировать процесс порождения продукционных правил естественно-языкового текста, которые позволят, не прибегая к услугам инженера по знаниям, своевременно вносить уточняющие изменения в продукционные правила в соответствии с новыми публикациями в данной предметной области.

2. Предисловие

Данная статья является продолжением наших исследований, имеющих своей целью: во-первых, обозначить возможные способы выявления конструкций со значением обусловленности в текстах естественного языка; во-вторых, придать найденным конструкциям формат, используемый в экспертных системах. И, наконец, в-третьих, ожидается, что трансформация любого (или только заданного типа) текста в совокупность конструкций со значением обусловленности позволит осуществить поиск требуемых конструкций при помощи компьютера. Иными словами, задача состоит в том, чтобы чисто формальными методами преобразовывать тексты естественного языка в совокупность строгих логических правил (или продукций), типичных для интеллектуальных систем [2–4].

В основе логических правил для интеллектуальной системы лежит понятие *импликации* (если ..., то). Но так как в лингвистике импликация имеет другое

Статья поступила в редакцию 9 марта 2010 г.

Работа частично финансировалась РФФИ по проекту 09-07-00233-а, а также программой Президиума РАН по проекту №211.

значение, а именно нечто подразумеваемое, мы используем иной термин — *отношение обусловленности.* Обусловленность, или каузальность, т.е. причинность в широком смысле слова, объединяет в себе существенно более богатый спектр значений, таких как основание, обоснование, доказательство и др. Весь этот круг отношений предполагает такую связь ситуаций, при которой одна служит основанием для реализации другой [5]. Таким образом, в нашу задачу требуемого трансформирования текстов в правила входит также задача формального разграничения логически связанных между собой ситуаций.

Отражая тем или иным образом логическую структуру высказывания, синтаксические отношения образуют в языке формальные признаки слов, обозначающих предмет высказывания, и противопоставленные им признаки слов, типично обозначающих предикацию. С целью представления для интеллектуальных систем формальных признаков слов, с одной стороны, и выявления понятийного аппарата текста, с другой стороны, нами был предпринят анализ лексического состава текста, в результате чего выделены характерные для выбранного типа текста классы лексических единиц и форматы, маркирующие их (см. ниже).

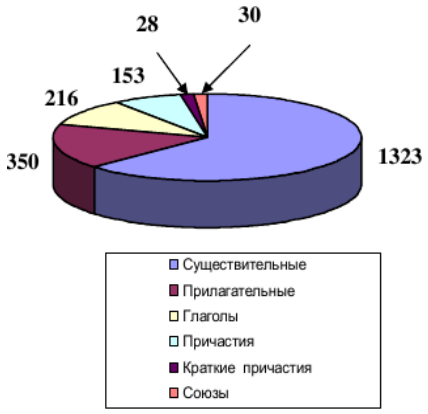

Рис. 1. **Категориальный состав анализируемой совокупности текстов**

Анализ подкреплялся статистическими данными, собранными на конкретном образце делового текста достаточного объёма, поскольку эти количественные показатели могут свидетельствовать и о типологических особенностях текста. Результаты подобного лексического и статистического анализов представлены на рис. 1.

Цифрами на этом рисунке обозначено количество употреблений лексических единиц различных классов в анализируемом тексте.

На первом этапе работы нами были выявлены элементы, маркирующие отношение обусловленности в естественном языке, включая общеизвестные условные союзы *если, в случае, в случае когда*, а также их субституты [2, 6].

На основе данных формальных показателей был составлен алгоритм, автоматически преобразующий текст в набор продукционных правил для использования в интеллектуальной системе [7].

Однако выявленные формальные компоненты предложения, маркирующие отношение обусловленности, не исчерпывают всех случаев смыслового выражения обусловленности, как, например, в следующих предложениях:

Граждане и юридические лица по своему усмотрению осуществляют принадлежащие им гражданские права.

Граждане (физические лица) и юридические лица приобретают и осуществляют свои гражданские права своей волей и в своём интересе.

Поэтому для обнаружения иных маркеров отношения обусловленности был предпринят анализ синтаксической структуры предложений в его соотношении со смысловым содержанием.

Мы исходили из того, что тип текста в значительной степени предопределяет его синтаксическую структуру, поэтому обратились к описанию структуры деловых текстов для выявления их типологических особенностей.

3. Основные черты делового текста регулирующего характера

Выбранный для анализа текст представляет собой свод положений, предписываемых к исполнению и регулирующих взаимоотношения участников в различных официально-деловых ситуациях. Такая *регулирующая* функция выражается в тексте рядом характерных грамматических особенностей.

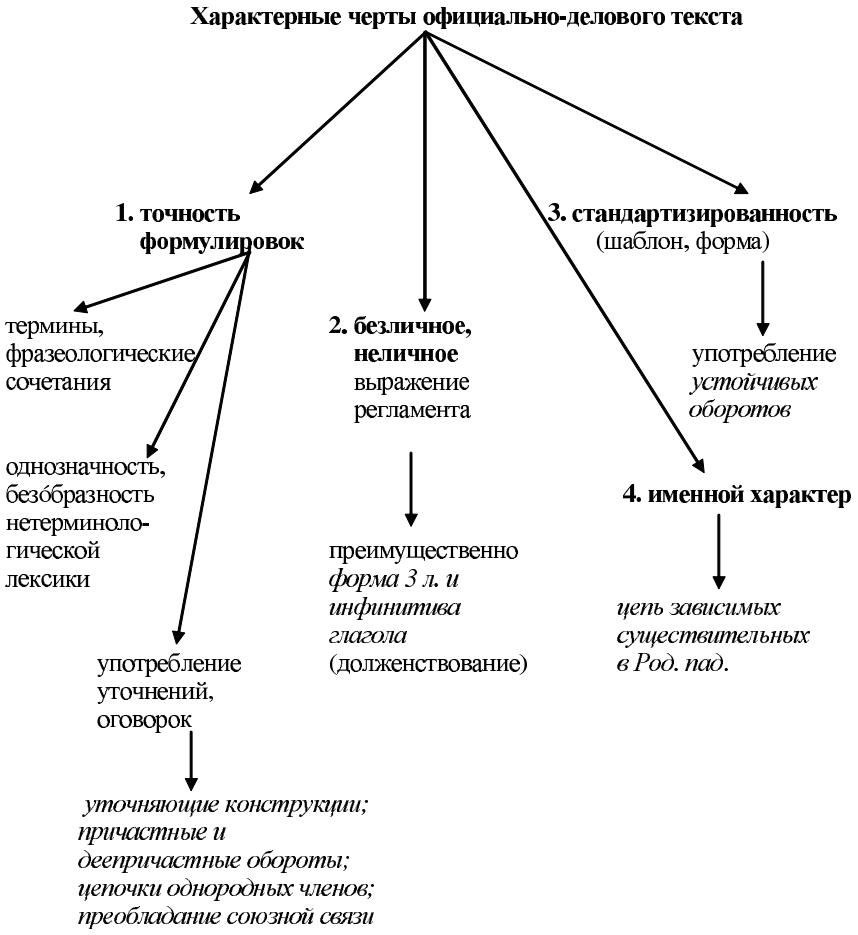

Рис. 2. **Характерные черты делового текста и способы их выражения**

Прокомментируем данную схему. Основной стилевой чертой данного типа текстов является точность формулировок, не допускающая инотолкования [8] в описании ситуации и её участников. Эта черта реализуется путём использования специальных терминов и фразеологических сочетаний, характерных для данной сферы деятельности, в однозначности и безóбразности нетерминологической лексики:

физические лица, юридические лица, предпринимательская деятельность, гражданские права, третейский суд, наркотические средства и др.

Это исключает использование синонимов, которые могут выражать иные смысловые оттенки. Поэтому особенностью деловой речи является неоднократное повторение одних и тех же слов, в основном терминов.

Помимо этого, выражению точности способствует употребление различных уточнений и оговорок, что проявляется в использовании многочисленных уточняющих синтаксических конструкций, причастных и деепричастных оборотов, цепочек однородных членов. А для выражения точности и логичности высказывания для данного типа текстов характерно преобладание союзной связи. Причём, в большинстве своём используются составные отымённые союзы и предлоги:

в случае; в случае, когда; в соответствии с; вследствие; в результате.

Другой стилевой чертой текстов регулирующего, регламентирующего характера является *безличное, неличное* выражение предписываемого регламента. Данная черта проявляется в отсутствии форм глагола 1-го и 2-го лица, с другой стороны — в преимущественном употреблении форм 3-го лица и инфинитива. Форма 3-го лица настоящего времени имплицитно выражает долженствование:

Товары, услуги и финансовые средства свободно перемещаются (т.е. должны свободно перемещаться) *на всей территории Российской Федерации.*

Допускается (т.е. должна допускаться) *самозащита гражданских прав.*

Эта форма употребления глагола называется *настоящим временем предписания.*

Как правило, характер долженствования в деловой речи проявляется в частом использовании кратких прилагательных модального значения долженствования (*должен, обязан, обязателен*) или инфинитивов глагола (*Приказываю: 1) Командировать...; 2) Повысить производительность...; 3) Установить...* и т.п.). Например:

Способы самозащиты должны быть соразмерны нарушению и не выходить за пределы действий, необходимых для его пресечения.

Однако в анализируемых нами текстах регулирующего, регламентирующего характера черта долженствования значительно смягчена и выражается также другими сочетаниями: глаголом *может* + инфинитив глагола; модальным словом *вправе* + инфинитив, имеющими предписывающий характер:

При осуществлении процедуры признания банкротом индивидуального предпринимателя его кредиторы по обязательствам, не связанным с осуществлением им предпринимательской деятельности, также вправе предъявить свои требования.

За несовершеннолетних, не достигших четырнадцати лет (малолетних), сделки ... могут совершать от их имени только их родители, усыновители или опекуны.

Далее, учитывая ситуацию делового общения, следует выделить типичную для деловой речи *стандартизированность* (шаблон, форма), употребление устойчивых для деловой сферы общения оборотов:

на основании и во исполнение настоящего закона; защита гражданских прав; компенсация морального вреда; возмещение причинённых убытков; признание брака недействительным; осуществление предпринимательской деятельности; удовлетворять требования (граждан, кредиторов); сохраняют силу; нести ответственность; без уважительной причины и т.д.

Типичным признаком деловой речи является её именной характер (т.е. широкое использование имён существительных и имён прилагательных). Употребление в ней имён существительных значительно превышает использование единиц других лексических категорий (см. рис. 1): прилагательных — более чем в 3,5 раза; глаголов — в 6 раз; причастий — в 8,5 раз и т.д. За отглагольными существительными часто следует цепь зависимых существительных в родительном падеже, создавая устойчивые обороты деловой речи. То есть в деловой речи для существительных характерно употребление не только в свойственной им номинативной функции (называния), но и очень часто в атрибутивной (функции определения), образуя ряды номинативных конструкций: *права авторов произведений науки; порядок осуществления права собственности; ограничение дееспособности гражданина* и др.

Таким образом была представлена характеристика делового типа текста с его синтаксическими и грамматическими особенностями.

4. Синтаксическая структура делового типа текста регулирующего характера

Возвращаясь к приведённой выше схеме состава лексических единиц в тексте (рис. 1), ещё раз отметим, что выявленные формальные маркеры отношений обусловленности (союзы *если; в случае; в случае, когда*) (см. [2, 6]) не исчерпывали всех случаев выражения импликации для интеллектуальной системы. Поэтому нами был использован иной подход для определения обусловленности, выраженной имплицитно (не явно).

С учётом прагматической значимости порядка слов в предложении были определены статистические данные для выделения исходной позиции предиката. С этой целью, с одной стороны, был (1) предпринят анализ линейной синтаксической структуры предложений. С другой стороны, (2) проведён анализ коммуникативной структуры предложения.

Для выполнения первой поставленной задачи (1), при помощи произвольно взятых условных обозначений, были последовательно составлены схемы всех предложений значительного по объёму отрывка текста. В данном случае были важны не общепринятые обозначения субъекта (S), глагола (V), объекта (O) и т.д., а именно произвольные знаки, например геометрические фигуры, для обозначения лексических единиц, имеющих категориальную маркировку для интеллектуальных систем. В схему были также включены так называемые операторы — знаки пунктуации и союзы, — могущие выступать в роли маркеров. Например, следующее предложение может быть представлено в виде схемы, изображённой на рис. 3:

Товары, услуги и финансовые средства свободно перемещаются на всей территории Российской Федерации.

Рис. 3. **Линейная синтаксическая структура предложения**

В данной схеме прямоугольник условно обозначает существительное; пунктирная линия — определение, выраженное прилагательным; овал — обстоятельство; треугольник — сказуемое (глагольное).

В схеме, показанной на рис. 4, условные геометрические фигуры представляют также именные группы:

Способы самозащиты должны быть соразмерны нарушению и не выходить за пределы действий, необходимых для его пресечения.

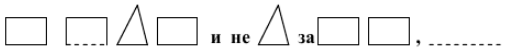

Рис. 4. **Линейная синтаксическая структура предложения**

В данной схеме первый прямоугольник обозначает подлежащее, выраженное существительным (*Способы*); второй прямоугольник — определение, выраженное существительным в атрибутивной функции (*самозащиты*); первый треугольник — составное именное сказуемое (*должны быть соразмерны*), а пунктирная линия — распространённое обособленное определение (*необходимых для его пресечения*). Деление определения на составляющие компоненты в данном примере не существенно.

Таким образом текст был представлен в виде условных графических схем, иллюстрирующих линейную структуру предложений и отражающих порядок следования входящих в предложения компонентов с их маркировкой. В большинстве

предложений порядок следования был однотипным: существительное, являющееся подлежащим и обозначающее предмет высказывания, предшествовало глагольной группе, выражающей предикацию. Проведённый нами анализ предложений значительного по объёму текста позволяет нам утверждать, что в данном типе текста регулирующего характера преобладает однотипная, жёсткая, упорядоченная синтаксическая структура.

5. Анализ коммуникативной структуры предложений

Исходя из прагматически релевантного актуального членения, было проанализировано деление предложений на *тему* и *рему* высказывания. *Тема* — «компонент актуального членения предложения, исходный пункт сообщения, — то, относительно чего нечто утверждается в данном предложении» [9, с. 507]. Как показал анализ, тема преимущественно занимает начальную позицию в предложении. *Рема* — «компонент актуального членения предложения, то, что утверждается или спрашивается об исходном пункте сообщения — *теме* — и создаёт предикативность, законченное выражение мысли» [9, с. 410]. Как свидетельствует анализ, рема занимает конечную позицию.

|*Граждане* (тема) | || | *могут иметь имущество на праве собственности* (рема)|; ...

|*Акты гражданского законодательства* (тема)| || |*не имеют обратной силы и применяются к отношениям, возникшим после введения их в действие.* (рема)|

Таким образом, анализ коммуникативной структуры позволяет нам сделать очевидный вывод: тема высказывания в подобных текстах регулирующего характера преимущественно занимает начальную позицию в предложении и заканчивается перед первым глагольным или именным сказуемым, которое маркируется свойственными глаголам и именам окончаниями. Причём, разнообразие окончаний ограничивается только формами инфинитива и глаголами 3-го лица настоящего времени, что, как указывалось в Разделе 3 выше, свойственно текстам делового типа и поэтому существенно упрощает поиск таких компонентов в предложении.

Интересны в этом отношении позиции упомянутых в Разделе 3 союзов. Для союза *если*, например, характерно использование в рематической части предложения со значением аргументированности и мотивировки:

Страховая сумма выплачивается, || *если в течение года ... наступит постоянная утрата трудоспособности; Ограничения перемещения товаров и услуг могут вводиться в соответствии с федеральным законом,* || *если это необходимо для обеспечения безопасности, защиты жизни и здоровья людей...* Союзы *в случае*, *в случае когда* используются преимущественно в составе темы предложения: *В случае несоблюдения требований, предусмотренных пунктом 1 настоящей статьи,* || *суд, арбитражный суд или третейский суд может отказать лицу в защите принадлежащего ему права.*

6. Локализация рематической части предложений в интеллектуальной системе

Как свидетельствует проведённый анализ синтаксической структуры предложений типа текстов регулирующего характера [см. Раздел 4], существительное, обозначающее предмет высказывания, предшествует глагольной (редко именной) группе, выражающей предикацию.

Способ формального выделения интеллектуальной системой глагола по морфологическим показателям — так называемый *вербоцентрический подход* — не является новым. Однако в текстах подобного типа глагол выступает для нас не

только формальным разграничителем связанных между собой ситуаций, но также несёт основную смысловую нагрузку в регулировании их отношений.

Поэтому на следующем этапе работы был предпринят рекурсивный анализ предложения для выявления компонентов, находящихся в препозиции к глаголу и влияющих на его семантику.

При рекурсивном анализе разрабатываемая нами интеллектуальная система учитывала, во-первых, отрицательную частицу *не*. Во-вторых, учитывались модальные слова со значением долженствования, уже упомянутые в Разделе 3: *обязан, должен, вправе, не вправе.* В-третьих, — обстоятельства образа действия, также находящиеся в препозиции к глаголу: *самостоятельно, свободно, соответственно, по своему усмотрению*. Например:

Однако такой гражданин | самостоятельно несёт имущественную ответственность по совершённым им сделкам и за причинённый им вред.

(Вертикальная черта показывает, что место членения предложения на тему и рему высказывания сдвинута от глагола влево — перед обстоятельством.)

7. Инверсия рематической части

Как показал синтаксический анализ, в данном типе текста регулирующего характера преобладает однотипная, жёсткая, упорядоченная структура предложений. Тем не менее, при однотипной синтаксической структуре исключительно редко встречаются предложения с инверсией глагола. Например:

Не допускается |использование гражданских прав в целях ограничения конкуренции, а также злоупотребление доминирующим положением на рынке.

Вынесение глагола в начальную позицию прагматически обосновано и усиливает категоричность предписания.

8. Синтаксический модуль: бессоюзное сложное предложение

Такие предложения характерны для официально-деловой речи. Они используются при перечислении различных реалий в деловой сфере общения, условий взаимодействия сторон и т.д. Обязательным оператором (знаком пунктуации) перед перечисляемыми элементами является двоеточие. Основная синтаксическая функция двоеточия заключается в разграничении связанных ситуаций, из которых одна служит основанием для реализации другой. Таким образом, двоеточие представляет собой формальный маркер для интеллектуальной системы при выделении отношений обусловленности. Помимо этого, при перечислении после двоеточия часто используется цифровая нумерация перечисляемых компонентов и пунктуационный знак — точка с запятой — после каждого компонента, которые также могут служить дополнительными маркерами. Например:

Государственной регистрации подлежат следующие акты гражданского состояния:

1) *рождение;*
2) *заключение брака;*
3) *расторжение брака;*
4) *усыновление (удочерение);*
5) *установление отцовства;*
6) *перемена имени;*
7) *смерть гражданина.*

В данном примере одна ситуация реализуется на основании обусловливающих её перечисленных ситуаций. Это пример обратной последовательности в отношениях обусловленности. Однако, если двоеточию предшествуют слова с модальным значением долженствования — *должен, обязан, вправе, имеет право*, — то отношение обусловленности выстраивается в прямом порядке. Например:

Участники хозяйственного товарищества или общества вправе:
участвовать в управлении делами товарищества или общества. . . ;
получать информацию о деятельности товарищества или общества. . . ;
принимать участие в распределении прибыли;

получать в случае ликвидации товарищества или общества часть имущества, оставшегося после расчётов с кредиторами, или его стоимость.

9. Заключение

Впервые такая задача автоматизации извлечения экспертных знаний из текста возникла в наших работах по созданию динамических экспертных систем в 1990-х годах, когда знания специалистов по сейсмологии стали недоступны, в связи с некоторыми особенностями проходившей в стране «перестройки». При этом, однако, имелась достаточно богатая литература по вопросу сейсмопрогноза. В то время казалось естественным использование инженеров по знаниям, которые, исходя из текстового материала, строили продукционные правила, предназначенные для динамической экспертной системы.

Примерно тогда и возник вопрос об автоматическом извлечении знаний в форме правил. Однако опыт показал, что эта задача оказалась намного сложнее, чем можно было ожидать, что и потребовало интенсивных исследований.

Отметим также, что трудность извлечения правил из текста носит вполне объективный характер, даже в такой глубоко продуманной предметной области, как юриспруденция. Поэтому разрабатываемый в наших публикациях подход может оказаться полезным и для авторов создаваемых текстов, требуя от них уточнения и даже переосмысления излагаемого ими материала.

Литература

1. `http://update.microsoft.com/windowsupdate/`.
2. *Савинич Л. В., Стефанюк В. Л.* Выражение обусловленности в естественном языке // Информационные технологии и вычислительные системы. — 2008. — № 1. — С. 30–37.
3. *Савинич Л. В., Стефанюк В. Л.* К извлечению знаний об отношениях обусловленности // Труды конгресса AIS-IT'09. — Т. 1. — М.: Физматлит, 2009. — С. 391–398.
4. *Стефанюк В. Л., Жожикашвили А. В.* Сотрудничающий компьютер: проблемы, теории, приложения. — М.: Наука, 2007. — 274 с.
5. Русская грамматика. — М.: Изд. АН СССР. Институт русского языка. Наука, 1980. — Т. II.
6. *Савинич Л. В., Стефанюк В. Л.* Представление конструкций со значением обусловленности // Труды 2-й международной конференции «Системный анализ и информационные технологии (САИТ-2007)». — Т. 1. — М.: ЛКИ/URSS, 2007. — С. 171–173.
7. Отчет по теме РФФИ, грант №07-07-00391-а (рукопись). — 2008.
8. *Кожина М. Н.* Стилистика русского языка. — М.: Просвещение, 1983.
9. Лингвистический энциклопедический словарь / под ред. В. Н. Ярцевой. — М.: Советская энциклопедия, 1990.

UDC 004.891

Recursive Solution for the Conditionality Problem

L. V. Savinitch, V. L. Stefanuk

Institute for Information Transmission Problems of Russian Academy of Science
Bolshoy Karetny per. 19, Moscow, 127994, Russia

The paper continues the research of the methods of formalization of natural language texts with the goal to represent the texts as collections of strict logical rules suitable for use in expert systems. Using regulation character texts with their typological peculiarities the syntactic structure of a phrase is analyzed with respect to its semantics.

Key words and phrases: conditionality, text typology, topic, focus.