ISSN 2079-3316 ПРОГРАММНЫЕ СИСТЕМЫ: ТЕОРИЯ И ПРИЛОЖЕНИЯ №1(28), 2016, с. 135–152

УДК 004.89:004.912

И. В. Трофимов

Особенности задачи мелкогранулярного извлечения фактографической информации из текста

Аннотация. В работе сформулирована постановка задачи мелкогранулярного извлечения фактографической информации из текста. Рассматриваются проблемы, дополнительно возникающие при такой формулировке задачи.

Ключевые слова и фразы: анализ естественного языка, извлечение информации, MUC, фрейм, имплицитная информация в тексте.

Введение

Область извлечения информации из текста охватывает довольно разнообразные задачи, как по постановке, так и по методам их решения. Чтобы это проиллюстрировать, достаточно рассмотреть несколько примеров.

(1) Задача структурирования библиографических ссылок [1] относительно проста. Для решения таких задач не требуются сколь-нибудь значительные предметные знания или сложные методы лингвистического анализа текста. Не ставятся подзадачи нормализации извлечённой информации и разрешения кореферентности.
(2) В задачах извлечения информации из таблиц [2] на первый план выдвигается проблема распознавания и анализа структуры таблицы — построение её физической, функциональной, структурной и семантической моделей.

Работа выполнена в рамках проекта РФФИ 14-07-00368 А «Исследование проблем и методов моделирования и использования общих знаний для разрешения кореферентности».

(3) Задачи анализа относительно крупных фрагментов связного текста [3] типично требуют многоаспектного лингвистического анализа и общих (иногда и специализированных) знаний о мире. В качестве примеров можно привести извлечение информации из новостных сообщений, досье, статей, отчётов, протоколов, жалоб и др.

Такая разноплановость задач извлечения информации привела к тому, что в современной литературе определение термина извлечение информации стало носить весьма обобщённый характер [4]. Поэтому исследователи вынуждены уточнять различные аспекты задачи извлечения, которую они решают.

В качестве отправной точки мы возьмем определение, сформулированное в рамках конференций MUC [3]. В нём постулируется следующее:

- информация извлекается из текста;
- информация извлекается в форме текстовых строк и обработанных (processed) текстовых строк;
- извлечённые сведения помещаются в слоты, наименования которых говорят о том, какой вид информации может их наполнять.

Данное определение приводит к вопросу, какая именно информация остается неструктурированной (в форме текстовых строк)? Приведённая [3] BNF-конструкция, специфицирующая устройство целевого шаблона для сущностей, фактически состоит всего лишь из четырех содержательных слотов[1]: название, тип, категория и дескриптор. Для мест (location) определено 3 слота: название, тип и страна. Это говорит о том, что в задачах MUC-7 значительная часть описательной (атрибутивной) информации не подлежит структурированию (в лучшем случае, она помещается в слот *дескриптор* в виде текста — именной группы).

В отличие от задач MUC мы будем рассматривать такую постановку задачи извлечения информации, в рамках которой бóльшая часть описательной информации, относящейся к целевым сущностям, подлежит категоризации и структуризации. Для этого шаблон целевой сущности должен обладать «богатой» слотовой структурой, способной охватить всё разнообразие описательной информации, которая может характеризовать сущность данного типа. Данную целевую установку будем называть мелкогранулярностью извлечения.

[1] Если исключить служебные OBJ_STATUS и COMMENT.

Тип фрейма:	**лицо-человек**
ФИО:	Иван Иванов
Должность:	заместитель генерального директора ОАО «Якорь»

Рис. 1. Пример крупногранулярного извлечения

Далее мы сформулируем характеристики задачи мелкогранулярного извлечения информации и проанализируем, какие проблемы влечёт за собой такая постановка задачи. Ограничимся рассмотрением лишь фактографического извлечения информации (структурирование сведений о конкретных событиях и сущностях).

1. Мелкогранулярное извлечение фактографической информации

Обычно при фактографическом извлечении информации нас интересуют какие-то определённые объекты, ситуации и события реального мира, описываемые (или просто упоминаемые) в текстах. Таким образом, структурирование представляется как процесс построения неких *информационных моделей* этих объектов, ситуаций и событий. Язык описания таких моделей должен быть с одной стороны достаточно формальным, чтобы допускать дальнейшую машинную обработку извлечённой информации, а с другой — удобным для восприятия человека. Поэтому в качестве контейнера для извлекаемой информации имеет смысл использовать фреймовую нотацию. Под фреймом мы будем понимать множество пар атрибут-значение (именуемых также слотами). У фрейма будем выделять специальный обязательный атрибут — «тип фрейма».

Как было отмечено ранее, создаваемые в процессе анализа информационные модели могут быть различной степени гранулярности. Так, для одной и той же фразы

(*1*) *заместитель генерального директора ОАО «Якорь» Иван Иванов*

можно построить различные конструкции фреймового типа. Например, можно построить единственный фрейм следующего вида (рис. 1).

В этом случае значительная доля информации остаётся неструктурированной (текстовой) и, следовательно, её машинная обработка будет затруднена (хотя для человеческого восприятия такой вид фрейма может оказаться приемлемым и даже предпочтительным). Альтернативный путь — построить несколько взаимосвязанных фреймов (рис. 2).

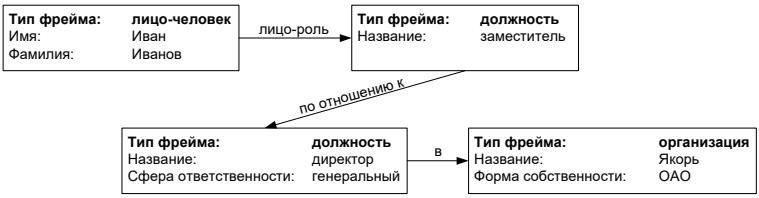

Рис. 2. Пример мелкогранулярного извлечения

Нас будет интересовать задача извлечения информации, в которой необходимо добиться как можно более мелкой гранулярности. Гранулярность характеризует два измерения:

- распределение информации по фреймам;
- распределение информации по атрибутам.

Обычно (но не всегда) текстовому фрагменту[2] соответствует фрейм, если этот фрагмент отсылает (имеет референцию) к отдельной сущности мира дискурса (или группе, множеству сущностей, упоминаемых в высказывании как единое целое). Предпосылками для представления сущности отдельным фреймом являются:

- фактическое или потенциальное наличие у сущности собственной атрибутивной информации;
- сущность может являться участником ситуации или события.

Приведем пример, демонстрирующий потребность моделирования фреймами сущностей, которые обычно осмысливаются нами как отношения (а не объекты). Для ситуации «Иван является сыном Петра» родственное отношение «сын» следует моделировать отдельным фреймом (рис. 3), так как оно потенциально может обладать дополнительной атрибутивной информацией (например, старший/приёмный/внебрачный сын).

В то же время компоненты описательной информации с семантикой интенсивности, степени выраженности атрибута и т.п. мы не относим к предпосылкам для порождения отдельного фрейма. Например, фрейм для автомобиля может иметь слот «скоростная характеристика» с индикатором интенсивности (в информационной модели данный индикатор представляется как составная часть слота). В этом случае

[2]Минимальными текстовыми фрагментами будем считать токены (слова, числа, различные символьные последовательности со специальной структурой и семантикой, такие как номера телефонов, форматированные записи даты, email и др.).

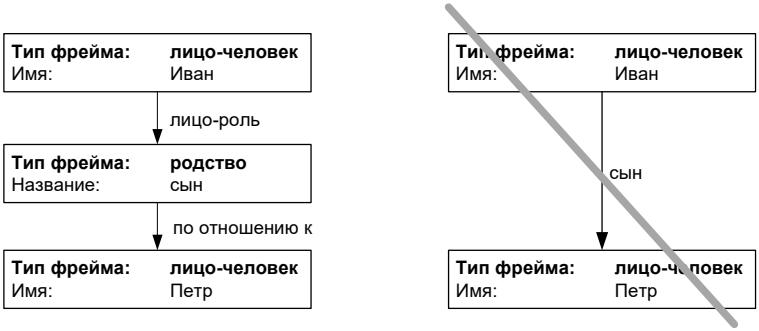

Рис. 3. Система фреймов для отношения родства

текстовые выражения *быстрый, очень быстрый, невероятно быстрый* помещаются в данный слот следующим образом: сам слот будет иметь значение «быстрый», а индикатор интенсивности принимает различные значения (∅, очень, невероятно).

При мелкогранулярном подходе к представлению извлечённой информации порой возникает необходимость строить фреймы, в которых полностью отсутствуют собственные идентифицирующие признаки объекта. Такое может происходить, когда объект упоминается в тексте как участник ситуации или события и идентифицируется лишь через отношения с другими объектами. Будем называть такие объекты и их модели внешнеидентифицированными. Рассмотрим пару фраз (*2*)–(*3*):

(*2*) *одна из дочерей Петрова;*

(*3*) *дочь Петрова.*

Информационной моделью обеих фраз будет следующая конструкция (рис. 4).

Про саму дочь в данной ситуации ничего не известно, кроме факта её существования, определяемого через отношение родства с отцом. Характерный для лиц идентифицирующий признак — имя или фамилия — отсутствует. Тем не менее в рамках прикладных задач извлечения информации потребность работать и с такими неполными сведениями обычно существует, так как модель внешнеидентифицированного объекта может выполнять связующую роль в структуре модели ситуации или события. См., например, высказывание (*4*) и соответствующий фрагмент модели (рис. 5).

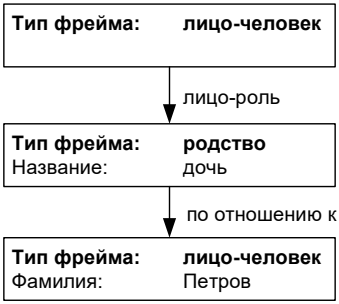

Рис. 4. Пример системы фреймов с внешнеидентифицируемым объектом

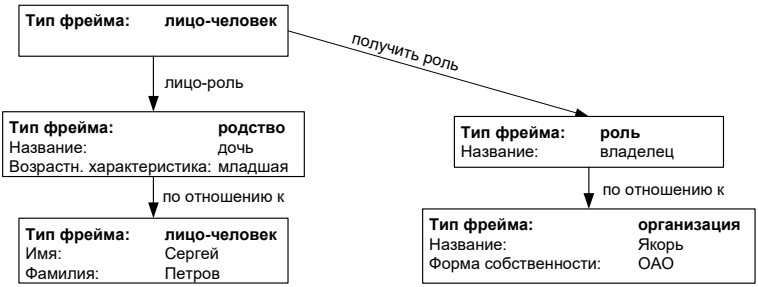

Рис. 5. Система фреймов для высказывания (*4*)

(*4*) *владельцем ОАО «Якорь» стала младшая дочь Сергея Петрова*

Причиной внешней идентифицированности могут выступать различные факторы. Возвращаясь к примерам фраз (*2*) и (*3*), можно допустить следующие причины:

- автор сам не обладает всей полнотой информации;
- автор не считает эти сведения существенным элементом изложения и поэтому опускает их для краткости;
- автор предпочитает не раскрывать эту информацию.

Отметим также, что существуют категории объектов, для которых внешний способ идентификации закреплен в менталитете и языке. Такие объекты идентифицируются через обязательное отношение с другим(и) объектом(ами). Например, должности (штатные единицы) идентифицируются через организации (подразделения, геополитические единицы), в которых они учреждены. Аналогично, правительства,

Тип фрейма:	**организация**
Концепт:	перевозчик
Род деятельности:	услуги
Вид услуг:	транспортировка
Среда транспортировки:	водный
Вид акватории:	морской

Рис. 6. Фрейм для фразы «морской перевозчик»

парламенты и т.п., хотя и могут иметь собственные названия, типично идентифицируются через геополитическую единицу. Для не обладающего энциклопедическими знаниями, «среднестатистического» носителя русского языка понимание фразы *парламент Боливии* не составит труда, чего нельзя сказать о фразе *Национальный конгресс Боливии*.

Теперь перейдем к рассмотрению проблем, которые возникают при мелкогранулярном подходе к извлечению.

2. Сложности мелкогранулярного извлечения

2.1. Проблема структуры фрейма

Постановка прикладной (конкретной) задачи извлечения информации сводится к разработке системы фреймов. Фреймы определяют, какая именно информация является целевой и в какой форме она подлежит структуризации.

Традиционно слотовая структура фрейма определяется типом фрейма и представляется как статическая. При мелкогранулярном подходе статический фрейм не кажется хорошим решением. Рассмотрим это на примере фрейма для организации.

Мелкогранулярный подход требует от нас категоризации самой разнородной описательной информации. Пусть для фраз *морской перевозчик* и *нефтеперерабатывающий комбинат* мы хотим получать структурированную информацию в следующем виде — рис. 6 и рис. 7, соответственно.

Наборы слотов, характеризующие упомянутые во фразах организации, отличаются и оказываются в какой-то мере зависимыми от значений определённых слотов. Так слот «вид акватории»[3] актуален только для описания перевозок водными видами транспорта (а также для портов).

[3]Морской, речной.

Тип фрейма:	**организация**
Концепт:	комбинат
Род деятельности:	переработка
Сырьё:	нефть

Рис. 7. Фрейм для фразы «нефтеперерабатывающий завод»

При статическом подходе фрейм для организации должен иметь настолько обширную слотовую структуру, чтобы быть способным разместить в себе описательную информацию для довольно непохожих друг на друга организаций: партий, аптек, профсоюзов, спортивных клубов, межгосударственных организаций и т.д. Это приводит к тому, что таким фреймом становится сложно управлять: определять новые слоты, ограничения на их значения, зависимости означивания, не допускать дублирования слотов (имена разные, а предполагаемое содержание одинаковое) и т.д.

С другой стороны, выделять отдельные типы фреймов для организаций различного рода также затруднительно. В тексте упоминание организации конкретного типа может осуществлять посредством более общего понятия и при этом иметь описательную информацию, характерную для конкретного типа. Например, фрейм для понятия «концерн», очевидно, не характеризуется слотом «сырьё», но для фразы *нефтеперерабатывающий концерн* наличие такого слота необходимо. Таким образом, возникает проблема поиска подходящего фрейма.

В результате мы приходим к тому, что нам нужен единственный, но динамический фрейм для организации. Состав его слотовой структуры находится в зависимости от уже означенных слотов. Это обеспечивает лучшую управляемость на этапе проектирования фрейма. Однако сам процесс извлечения описательной информации из-за необходимости индуцировать добавление слотов во фрейм, а не только присваивать им значения, становится более сложным. Здесь показателен пример с *морским перевозчиком*, когда за одной лексической единицей (*морской*) скрываются значения сразу для двух слотов, причём добавление во фрейм одного из них (вид акватории) зависит от значения другого (среда транспортировки).

2.2. Проблема носителя описательной информации

В случаях, когда одно слово в тексте влечёт построение нескольких фреймов на уровне информационной модели, возникает вопрос,

как с ними соотносится описательная информация, синтаксически связанная с исходным словом? Рассмотрим примеры с адъективами:

(*5*) *трёхлетний сын, голубоглазый сын;*

(*6*) *внебрачный сын, младший сын.*

Если фреймовая модель для этих примеров будет построена по подобию той, что изображена на рис. 3, то различие примеров (*5*) и (*6*) будет состоять в следующем. Слоты для *трёхлетний* и *голубоглазый* относятся к фрейму «лицо-человек» (свойства лица, как человека), в то время как, *внебрачный* и *младший* — к «родству» (свойства лица, как сына).

Таким образом, при мелкогранулярном подходе требуется такая интерпретация признаковой информации, которая позволит не только определить, в какой слот данная признаковая информация должна попасть, но также фрейм, в котором этот слот находится. Неоднозначность в выборе фрейма типично обусловлена неоднозначностью интерпретации самого слова-признака. Например, в (*7*)

(*7*) *малый сын*

слово малый может означать как младший, так и малолетний (как правило, совмещает оба значения).

2.3. Проблема границ экспликации

Возможность построения внешнеидентифицируемых сущностей приводит к вопросу моделирования имплицитных сведений, выводимых из текста. Возвращаясь к теме родства[4], зададимся вопросом, как на уровне информационной модели следует представить следующую фразу (*8*)

(*8*) *внучатый племянник Петрова по бабушкиной линии*[5]

Моделирование на уровне толкований понятий (один из возможных вариантов изображён на рис. 8) в большинстве случаев избыточно, хотя все упомянутые в тексте в явном виде сведения, в том числе бабушка, нашли своё отражение на схеме.

---

[4]Будем полагать, что целевыми фреймами являются «лицо-человек» и «родство».

[5]Отметим, что бабушка в данной фразе может быть упомянута и при помощи существительного, потенциально имеющего референта — «внучатый племянник Петрова по линии бабушки».

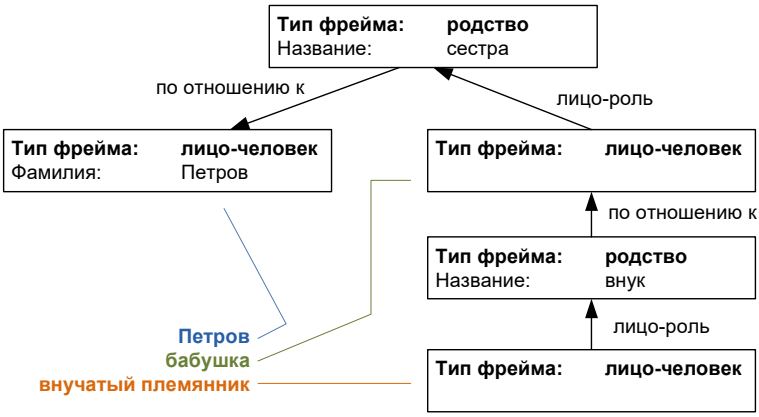

Рис. 8. Информационная модель, построенная через обращение к толкованию фразы (*8*)

В (*8*) бабушка упомянута скорее в атрибутивной, а не референтной функции. Её упоминание служит лишь уточнению (с целью идентификации) родственной цепочки, но не имеет целью сделать бабушку темой обсуждения. Таким образом, фраза (*8*) отсылает адресата к единственному родственному отношению между Петровым и его внучатым племянником, а не к целому фрагменту действительности, включающему, например, одного из родителей внучатого племянника или общего родителя для Петрова и бабушки. Это фрагмент действительности становится доступен адресату благодаря его общим знаниями, а не сведениям, содержащимся в тексте. Поэтому представление в следующем виде (рис. 9) с бабушкой-идентификатором, кажется более предпочтительным.

Проблема экспликации вышеприведённого типа может возникать при выполнении следующих условий:

- упомянутое в тексте слово служит для обозначения «объекта-системы», состоящего из элементов и связей между ними;
- синтаксически связанная с данным словом описательная информация либо характеризует отдельный «внутренний» элемент системы (иногда подмножество элементов), либо служит для идентификации такого элемента (подмножества) в системе.

Рассмотрим ещё один случай подобного рода, порождённый языковой экономией. Во фразе:

(*9*) *украинский спикер* (= ‘спикер парламента Украины’)

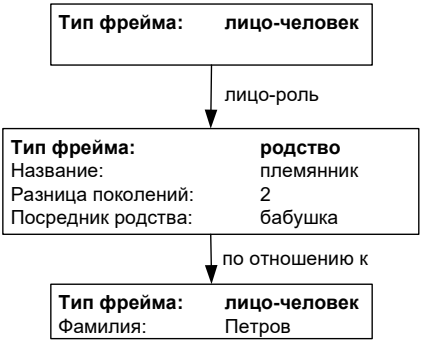

Рис. 9. Компактная информационная модель для фразы (*8*)

целый объект — парламент, — связывающий признак *украинский* со *спикером*, не представлен в тексте в явном виде. Адресат при необходимости может его «восстановить», пользуясь своими знаниями о мире. В каком-то смысле спикера здесь можно рассматривать как объект-систему, подразумевающую существование парламента (как часть своего толкования), хотя, строго говоря, скорее спикер является элементом в системе парламента.

Отметим, что проблема экспликации для объектов-систем может и не возникать по причине невозможности моделирования структуры системы. Упоминавшееся выше понятие «внучатый племянник» определяет строгую структуру родственных отношений, которая в принципе может быть эксплицирована. В отличие от него, такое понятие, как, например, «торговая сеть» в бытовом представлении не имеет строгой структуры и поэтому экспликация[6] элементов толкования затруднительна. Рассмотрим фразу:

(*10*) *мелкооптовая розничная сеть*

Мы понимаем, что это сеть предприятий торговли (магазинов, аптек и т.п.) и характеристики «мелкооптовость» и «розничность» характеризуют именно это подмножество элементов в системе, а не иные сущности в составе сети как коммерческой организации (например, подразделения, обеспечивающие логистику или маркетинг). Но по причине нечёткости внутреннего устройства сети (системы сущностей и отношений), экспликация компонентов сети невозможна, а

[6]Можно разве что говорить об её ограниченной экспликации (каком-то отношении между сетью и её типичным основным компонентом — предприятиями торговли).

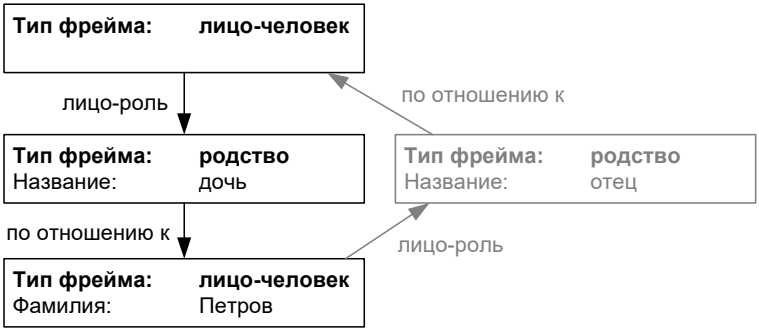

Рис. 10. Двунаправленность родственного отношения

характеристики «мелкооптовости» и «розничности» вынужденно приписываются к сети в целом.

Кроме случая с объектом-системой (племянник, спикер, сеть) рассмотрим еще одну типичную ситуацию, в которой приходится принимать решение об экспликации, — конверсивные конструкции. Если вернуться к рис. 4, то с точки зрения проблемы экспликации возникает вопрос, при каких условиях требуется в явном виде строить фрейм типа «родство», отражающий родственное отношение с точки зрения отца (рис. 10)?

Проблема границ экспликации указывает на то, что необходимы какие-то принципы разумной компактности в духе бритвы Оккама. Попробуем сформулировать такие принципы для сущностей в атрибутивном контексте и конверсивных конструкций.

Для сущностей в атрибутивном контексте (*по бабушкиной линии*) можно предложить следующее. Сущность целевого типа, упомянутая в атрибутивном контексте, не подлежит представлению на уровне информационных моделей в форме фрейма за исключением следующих случаев:

(1) экспликации данных сведений требует прикладная задача;
(2) в другом фрагменте текста та же сущность упомянута в неатрибутивном контексте (т.е. в любом случае будет представлена фреймом на уровне информационной модели);
(3) имеющаяся в текстовом фрагменте описательная информация не может быть размещена в слотах фрейма для сущности, к которой данная информация относится. В то же время:

- описательная информация отсылает к единичной сущности (не к классу, не к типичному представителю класса и не к множеству сущностей);
- существует целевое отношение, пригодное для связи фрейма сущности, характеризуемой этой описательной информацией, с фреймом, при помощи которого предполагается данную описательную информацию смоделировать.

Сделаем замечание касательно референции как индикатора, указывающего на упоминание в тексте сущности, которая может подлежать извлечению. На уровне информационной модели представление в форме фрейма имеют не все даже референтно употреблённые целевые сущности. Рассмотрим два вида текстовых фрагментов.

(1) Целевая сущность упомянута как один из аргументов какого-либо из предикатов высказывания. В этом случае моделирование целесообразно осуществлять в форме фрейма, т.к. соотношение объекта с ситуацией обычно реляционно. Примеры:
   a. *В марте Россия будет председательствовать в Совбезе ООН.*
   b. *Влияние России на Ближнем востоке нарастает.*
   c. *Российское председательство* (= 'председательство России') *в Совбезе ООН поможет переломить сложившуюся напряжённую обстановку.*

(2) Целевая сущность упомянута в атрибутивном контексте (для идентификации или характеризации другой сущности). В этом случае вопрос моделирования посредством фрейма не столь однозначен. Примеры:
   a. *Министерство иностранных дел России* (ср. *российский МИД*);
   b. *президент России* (ср. *российский президент*);
   c. *дядя по линии матери*;
   d. *санитарно-эпидемиологические нормы в странах Африки* (ср. *африканских странах*).

Для упоминаний второго типа возможно моделирование как в форме отдельного фрейма, так и в форме значения слота. Референция здесь уже не может служить индикатором средства моделирования.

Для конверсивов принцип разумной компактности может звучать так: если извлечённая ситуация (событие) полностью выражена с точки зрения хотя бы одного из её участников, то нет необходимости в явном виде моделировать взгляд на ситуацию со стороны других её участников. Полностью в данном случае говорит о том, что вся

содержащаяся в тексте описательная информация уже нашла своё отражение в информационной модели.

Таким образом, если мы имеем текстовый фрагмент:

(*11*) *приёмная дочь Петрова,*

то вопрос экспликации фрейма для отца (рис. 10) будет зависеть от изначально заданной системы целевых фреймов. Если предусмотрен слот для «приёмности» во фрейме дочери, то экспликация не требуется. Если то же самое свойство предполагается выражать только слотом во фрейме для отца («биологический отец»: да/нет), то структурирование потребует построения фрейма для отца.

2.4. Проблема интерпретации ассоциативного

Немалая доля описательной информации служит для формирования аморфного ассоциативного поля у адресата сообщения, что вызывает трудности при структурировании. Типичным примером здесь может послужить описательная информация, обозначающая сферу ответственности:

(*12*) *министр нефти / немецкая сеть парфюмерии;*

(*13*) *нефтяной магнат / алюминиевый концерн.*

Такого рода сведения сложно представить в результирующей информационной модели отличным от строки образом. Моделирование в виде фрейма оправданно разве что в целях унификации: *министр нефти* и *нефтяной магнат* оба связаны с абстрактной нефтяной сферой.

2.5. Проблема референции в описательной информации

Мелкогранулярный подход требует разграничения референциальных статусов [**5**] именных групп в составе описательной информации. Типично наличие референтных и родовых употреблений. Например, если текст повествует о каком-нибудь холдинге, его интродукция в тексте уже осуществлена и затем встречается фраза:

(*14*) *директор холдинга,*

то имеет место референтное употребление. Референтно употребляются почти[7] все собственные имена, входящие в описательную информацию (*президент России*). В то же время аналогичные синтаксические конструкции могут отсылать не к конкретному объекту, упоминаемому в повествовании, а к типичному представителю объектов данного класса (родовая референция):

[7] За исключением автонимных употреблений — бизнесмен по имени Лю.

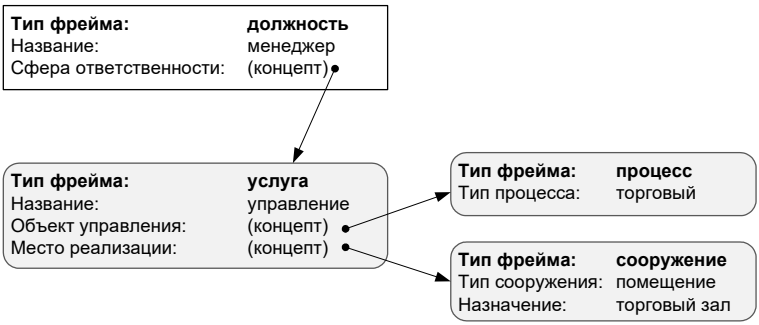

Рис. 11. Информационная модель, включающая модели концептов

(*15*) *менеджер торгового зала;*

(*16*) *менеджер по работе с корпоративными клиентами;*

(*17*) *сеть круглосуточных аптек.*

В этом случае «вложенной» именной группе не соответствует конкретного объекта, поэтому они не должны представляться фреймами в результирующей информационной модели, построенной по тексту.

Структурирование информации с родовым референциальным статусом возможно посредством специального типа фреймов, задача которых описать не информационный объект, а концепт вместе с уточняющей его информацией (корпоративный, круглосуточный). Модель концепта в этом случае может быть использована в качестве значения слота у целевого информационного объекта (менеджера, сети) или другой модели концепта. Однако разработка системы таких специальных фреймов является сложной задачей. Возможный способ представления для фразы (*15*) представлен на рис. 11.

Заключение

Уменьшение гранулярности при извлечении информации из текста способствует расширению возможностей в реализации поисковых и аналитических операций над содержащимися в тексте сведениями. Однако при мелкогранулярном подходе возникает ряд проблем, не характерных для задач извлечения в постановке MUC. Наиболее сложными, на наш взгляд, являются проблемы границ экспликации, представления ассоциативной информации и распознавания типов

референции в описательной информации. Кроме того, мелкогранулярный подход требует тщательного проектирования целевой системы фреймов.

Мелкогранулярность также обостряет проблему унификации извлечённой информации. Действительно, если информация в значительной степени структурирована, то это структурирование для схожих сущностей и атрибутов должно выполняться схожим образом. Если в текстах мы встречаем фразы *президент РФ* и *российский президент*, то на уровне информационной модели этим упоминаниям должны в идеале соответствовать одинаковые конструкции, так как содержание приведённых фраз эквивалентно. Однако требование унификации может противоречить сформулированному в статье возможному решению для проблемы экспликации.

Список литературы

[1] A. C. Álvarez, A. A. Lopes. “Information Extraction from Tagged Bibliographical References”, II Workshop on Web and Text Intelligence (WTI) (São Carlos, SP, Brazil, 2009), pp. 1–9, URL: http://www.icmc.usp.br/~alneu/papers2007_2009/WTI09Caceres_Lopes.pdf ↑ 135

[2] M. Hurst. *The interpretation of tables in text*, PhD. Thesis, School of Cognitive Science, Informatics, The University of Edinburgh, United Kingdom, 2000, URL: https://www.era.lib.ed.ac.uk/bitstream/handle/1842/7309/515564.pdf ↑ 135

[3] N. Chinchor, E. Marsh. “MUC-7 Information Extraction Task Definition”, Proceedings of the Seventh Message Understanding Conference (Fairfax, 1998), URL: http://www.itl.nist.gov/iaui/894.02/related_projects/muc/proceedings/ie_task.html ↑ 136

[4] S. Sarawagi. “Information Extraction”, *Foundations and Trends in Databases*, **1**:3 (2008), pp. 261–377, URL: http://pages.cs.wisc.edu/~anhai/courses/784-fall15/ieSurvey.pdf ↑ 136

[5] Е. В. Падучева. *Высказывание и его соотнесенность с действительностью*, Наука, М., 1985, 272 с. ↑ 148

Рекомендовал к публикации *к.т.н. Е. П. Куршев*

Об авторе:

Игорь Владимирович Трофимов

Старший научный сотрудник Исследовательского центра искусственного интеллекта ИПС им. А. К. Айламазяна РАН, специалист по технологиям автоматической обработки естественного языка, извлечения информации, автоматического планирования

e-mail: itrofimov@gmail.com

Пример ссылки на эту публикацию:

И. В. Трофимов. «Особенности задачи мелкогранулярного извлечения фактографической информации из текста», *Программные системы: теория и приложения*, 2016, **7**:1(28), с. 135–152.

URL: http://psta.psiras.ru/read/psta2016_1_135-152.pdf

Igor' Trofimov. *Peculiarities of fine-grained factual information extraction from text.*

Abstract. The paper states the problem of fine-grained factual information extraction form text. The author reveals some additional issues that arise from such formulation of the task. (*In Russian*).

Key words and phrases: natural language processing, information extraction, MUC, frame, implicit information in text.

References

[1] A. C. Álvarez, A. A. Lopes. "Information Extraction from Tagged Bibliographical References", II Workshop on Web and Text Intelligence (WTI) (São Carlos, SP, Brazil, 2009), pp. 1–9, URL: http://www.icmc.usp.br/~alneu/papers2007_2009/WTI09Caceres_Lopes.pdf

[2] M. Hurst. *The interpretation of tables in text*, PhD. Thesis, School of Cognitive Science, Informatics, The University of Edinburgh, United Kingdom, 2000, URL: https://www.era.lib.ed.ac.uk/bitstream/handle/1842/7309/515564.pdf

[3] N. Chinchor, E. Marsh. "MUC-7 Information Extraction Task Definition", Proceedings of the Seventh Message Understanding Conference (Fairfax, 1998), URL: http://www.itl.nist.gov/iaui/894.02/related_projects/muc/proceedings/ie_task.html

[4] S. Sarawagi. "Information Extraction", *Foundations and Trends in Databases*, **1**:3 (2008), pp. 261–377, URL: http://pages.cs.wisc.edu/~anhai/courses/784-fall15/ieSurvey.pdf

[5] E. V. Paducheva. *Utterance and its relation to reality*, Nauka, M., 1985 (in Russian), 272 p.

Sample citation of this publication:

Igor' Trofimov. "Peculiarities of fine-grained factual information extraction from text", *Program systems: theory and applications*, 2016, **7**:1(28), pp. 135–152. (*In Russian*). URL: http://psta.psiras.ru/read/psta2016_1_135-152.pdf