РАСПОЗНАВАНИЕ КОЛИЧЕСТВЕННОЙ ИНФОРМАЦИИ В ЕЯ-ТЕКСТАХ[1]

QUANTITATIVE DATA RECOGNITION AT NLP

В. Ш. Рубашкин *(vrub@mail.nw.ru)*

Б. Ю. Чуприн *(boris@vr4591.spb.edu)*

СПбГУ, С.-Петербург

Рассмотрены методы распознавания количественной информации в системах типа Information Extraction. Обсуждаются следующие вопросы: что такое "количественная информация"; способы ее представления; задачи, которые должен решать алгоритм анализа; требования к средствам словарной поддержки; методы программной реализации и опыт тестирования.

Значимость количественной информации в научно-технических и вообще в деловых текстах не требует развернутых пояснений. Но - хотя одна из главных целей систем извлечения фактографической информации из делового текста как раз и состоит в нахождении, интерпретации и стандартизации числовых и, более широко, количественных данных, - в большинстве публикаций, даже тех, которые представляют наиболее продвинутые проекты, этот аспект анализа не представлен вовсе (В качестве примера можно сослаться на такие авторитетные и обстоятельные работы как [1] и [2]; ср. также [3]). Исключения немногочисленны и также не содержат развернутых описаний объектов и техники анализа (ср. [4]). Такое положение можно объяснить, видимо, только отсутствием четко оформленного социального заказа, поскольку сами по себе модели представления количественной информации неоднократно и подробно обсуждались, в том числе и в отечественной литературе (ср. [5] и [6]).

В докладе рассматриваются как общие вопросы анализа количественной информации, так и, конкретно, методы, использованные авторами при разработке системы такого типа, ориентированной на анализ текстов свободного стиля (публикации общественно политической тематики в СМИ).

1. Количественная информация и количественные группы

Рассмотрим стандартный пример контекста, в котором, несомненно, имеется количественная информация: *Жесткие диски емкостью до 100 ГБ.*

Здесь можно выделить следующий набор значимых для результата анализа элементов [1]:

1) имя объекта: *жесткие диски;*
2) наименование признака: *емкость;*
3) количественное (в данном примере числовое) значение: *100;*
4) единица измерения: *ГБ*
5) модификатор значения: *до*.

Некоторые из элементов могут отсутствовать: *Жесткие диски до 100 ГБ.*

Собственно **количественной группой** мы будем называть фрагмент, содержащий элементы 2 – 5 (не включая сюда имя характеризуемого объекта.). Пару, составленную из имени (классификационного описания) объекта и количественной группы, естественно назвать **количественным фактом.**

Разные варианты количественных групп могут быть охвачены следующей простой классификацией, выстроенной по различению типа используемой оценки упомянутого в группе признака (для каждого типа указана возможная логическая интерпретация):

А. Числовые группы

1) Точечные:
мощностью 100 вт → *МОЩНОСТЬ_вт (x, v) & v = 100*
2) Типа "пятно":
*мощностью **около** 100 вт* → *МОЩНОСТЬ_вт (x, v) & v ≈ 100*
3) Интервальные:
- зона, ограниченная снизу: *мощностью свыше 100 вт*;
- зона, ограниченная сверху: *мощностью до 100 вт*;
- собственно диапазон
мощностью от 100 до 1000 вт → *МОЩНОСТЬ_вт (x, v) & v >= 100 & v <= 1000*

[1] Доклад подготовлен при частичной поддержке РФФИ (проект № 03-06-80109)

4) Парциальные: *В Латвии сейчас до трети населения неграждане*.
5) Представляющие числовую оценку динамики изменения:
- «на сколько» - абсолютная оценка:
мощность увеличена на 100 вт → *МОЩНОСТЬ_вт (x, v) & Увеличение_на (v, 100)*
- «на сколько_%» - относительная оценка:
мощность упала на 20 % → *МОЩНОСТЬ_вт (x, v) & Уменьшение_на_% (v, 20)*
- «во сколько»:
мощность выросла в 1,5 раза → *МОЩНОСТЬ_вт (x, v) & Увеличение_в (v, 1,5)*

Б. Нечисловые группы.

1) Нормативно-оценочные:
большой мощности → *МОЩНОСТЬ_вт (x, v) & Большая_величина (v)*
2) Представляющие вербальную оценку динамику изменения:
мощность (резко) растет → *МОЩНОСТЬ_вт (x, v) & Увеличение (v)*
Приведенное выше и есть перечень тех основных различений, которые должен уметь проводить алгоритм анализа, имея дело с количественными группами.

2. Основные задачи, решаемые анализатором

1) Преобразование вербальных и вербально-цифровых значений в числовой формат (с предварительным восстановлением сокращенных обозначений элементов числа):

тысяча сто → *1100; 10 млн.* → *10 000 000* и т. д.

2) Интерпретация именованного числа как значения признака; пересчет значения к стандартной единице измерения:

10 квт → *10 000 вт (мощность)*

3) Разграничение ***величины, количества и парциальной оценки***:

20 кг ***vs*** *20 человек* ***vs*** *треть населения*

В первом случае должен быть построен признак *ВЕС_г (x, v) & v = 100,*

во втором – выражение *ЧЕЛОВЕК (x) & ЧИСЛЕННОСТЬ_СОВОКУПНОСТИ (x, v) & v=100,*

в третьем - выражение *НАСЕЛЕНИЕ (x) & ДОЛЯ_СОВОКУПНОСТИ (x, v) & v = 0,33*

4) Присваивание признаку значения; уточнение наименования признака:

толщиной 100 мкм (первоначально восстановленный по единице измерения обобщенный признак *линейный размер* уточняется как *толщина*).

1) Устранение смысловой избыточности. Учет лексически опосредованных связей между элементами количественных групп:

Финансовая помощь ЕС Литве в 2004-2006 гг. планируется на уровне 2,5 млрд. евро (Здесь число является оценкой признака *финансовая помощь,* но не признака *уровень*)

6) Прикрепление количественной группы к имени объекта: (например, связь между ИГ *жесткие диски* и количественной группой *емкостью до 100 ГБ*, интерпретируемая на уровне ЯПЗ как конъюнктивная связь.)
В некоторых ситуациях это может представлять нетривиальную проблему, что можно проиллюстрировать сопоставлением следующих двух фраз:

(а) Возьмите деревянный брусок с отверстием ***диаметром 30 мм****.*

(б) Возьмите деревянный брусок с отверстием ***весом 300 г****.*

Разрешение такого рода коллизий требует не только определенной различительной силы алгоритмов, но, прежде всего, специальной поддержки со стороны концептуального словаря.

3. Алгоритмический аспект

В разработанной нами системе анализа все названные алгоритмические задачи (и не только эти) решаются внутри единого переборного механизма, реализующего все процедуры семантической интерпретации. На вход интерпретатора подается синтаксически размеченный текст. Наличие разметки наиболее критично для тех фрагментов входного текста, которые, как в случае количественных групп, должны получить точную интерпретацию. Однозначность разметки не предполагается. Разрешение синтаксической и оставленной

синтаксисом лексической омонимии производится единообразно путем перебора и оценки результатов интерпретации всех имеющихся вариантов. Выбор процедуры интерпретации управляется семантическими характеристиками синтаксического хозяина и синтаксического слуги.

4. Словарная поддержка процедур анализа

Функциональность и структуру концептуального словаря (онтологии), востребованную процедурами анализа можно кратко охарактеризовать следующим перечнем: связи *признак – единицы измерения*; *стандартные – нестандартные единицы измерения* (с указанием алгоритма пересчета); *признак – релевантный класс объектов*; наличие функциональных термов, характеризующих всю номенклатуру возможных значений (при этом нужно позаботиться о достаточно полном отражение лексических способов их выражения в толковом словаре системы). Так, для одного только "функционального" смысла *(очень) большая величина* нужно иметь весьма обширный список текстовых эквивалентов: *большой, весомый, высокий, крупный, немалый, астрономический, безграничный, бесчисленный, большущий, великий, гигантский, гипертрофированный, головокружительный, грандиозный, громадный, здоровенный,…*и т.д., и т.п.

5. Вопросы реализации

В фактографической системе логическое представление, как правило, упрощается до сетевой нотации. (Семантическую сеть можно интерпретировать как логический язык с существенными ограничениями допустимых правил построения высказываний.)

В сетевой нотации различаются объектные и предикатные узлы; они соединяются ролевыми дугами, либо дугами, представляющими типовые отношения предметной области (*часть-целое, локализация, предмет-функция* и т. п.). Каждый узел характеризуется множеством терминов, разделяемых на два подкласса - термины-свойства и термины, представляемые парой < признак, значение >.

Количественные группы обычно характеризуют именно объектные узлы (хотя возможны и признаки, характеризующие предикатные узлы – процессы (*скорость, длительность* и т.п.)

В используемой нами нотации рабочим языком представления является табличный язык стандартной РСУБД. Семантическая сеть представляется тремя основными таблицами. Все термы, опознанные в тексте и релевантные целевой системе знаний, отображаются в ***таблице термов***; отнесение терма к тому или иному узлу маркируется в поле *Номер узла*, используемом как эквивалент референциальных индексов в логической записи. Вторая таблица хранит ***значения признаков***. В ней представлены все значения (в частности, количественные). Понятно, что помимо собственно значения, таблица должна, во-первых, специфицировать и его тип - в номенклатуре и различиях, приведенных в разделе 1. И, во-вторых, связывать это значение с определенным наименованием признака, записанном в *таблице термов*. Еще одна таблица должна представлять (именованные) связи между узлами сети.

Словарь и собственно анализатор ориентированы на анализ как количественной, так и чисто вербальной информации и реализованы средствами *CBuilder 6.0.* Текущее состояние – тестирование разработчиком на текстах свободного стиля при одновременном пополнении словарей. Результативность анализа, разумеется, полностью зависит от наличия в словарях соответствующей лексики и от точного соответствия ее описания принятым в модели анализа спецификациям. При этих условиях анализ гарантированно успешен, так что многократное тестирование однотипных конфигураций, в сущности, лишено смысла. Более осмысленный подход состоит в поиске в потоке текстов лексико-грамматических конфигураций не охваченных алгоритмами анализа.

Список литературы

1. *Ralph Grishman*. Information extraction: Techniques and challenges // Maria Teresa Pazienza, editor. Information Extraction. Springer-Verlag, Lecture Notes in Artificial Intelligence, Rome, 1997
2. *Nirenburg S., Raskin V.* Ontological Semantics. – Cambridge, MA: MIT Press, 2004
3. *А. Е. Ермаков*. Поиск фактов в тексте // Мир ПК, № 2, 2005
4. http://www.osp.ru/pcworld/2005/02/068.htm
5. *Appelt, Douglas E., Jerry R. Hobbs, John Bear, David Israel, Megumi Kameyama, and Mabry Tyson,* The SRI MUC-5 JV-FASTUS Information Extraction System // Proceedings Fifth Message Understanding Conference (MUC-5), Baltimore, Maryland, August 1993
6. *Рубашкин В. Ш*. Признак и значение // Научно-техническая информация. Сер. 2. - М., 1976. № 3
7. *Семенова С.Ю.* Алгоритм извлечения информации о параметрах из текстов рефератов и первичных документов // Научно-техническая информация. Сер. 2. - М., 1991. № 6.