Применение метода ассоциативных правил для открытого извлечения семантических отношений ♣

Р.М. Гареев, В.Д. Соловьев

Казанский (Приволжский) федеральный университет
gareev-rm@yandex.ru, maki.solovyev@mail.ru

Аннотация

В статье предлагается модель открытого извлечения экземпляров отношений из естественно-языковых текстов с использованием базы знаний онтологии, где открытое извлечение подразумевает поиск всевозможных отношений в тексте. В основе модели лежит поиск устойчивых сочетаний языковых единиц и их лексических, грамматических и семантических (онтологических) свойств. Для реализации поиска устойчивых сочетаний предлагается применить метод ассоциативных правил.

Предлагаемая модель обладает следующими преимуществами: не требует избыточности информации в корпусе, не требует глубокого лингвистического анализа, не требует множества исходных примеров отношений, не является ориентированной на фиксированные структуры языковых выражений.

1. Введение

Проблема автоматического извлечения структурированной информации из естественно-языковых ресурсов становится все более актуальной в связи с непрерывным ростом количества доступных текстовых источников информации. Класс задач, решаемых традиционными методами информационного поиска, ограничен, поскольку эти методы оперируют данными на уровне документов, ранжируя их по релевантности в соответствии с ключевыми словами запроса пользователя. В классических подходах информационного поиска не проводится выделения описываемых в тексте сущностей, их взаимосвязей и свойств. Представление смыслового содержимого текста документа с помощью специализированных структур и методов извлечения знаний позволяет выражать информацию, необходимую пользователю, в явном виде, в отличие от систем традиционного информационного поиска, возвращающих лишь список источников необходимой информации (т.е. документов), каждый из которых, вне зависимости от объема, пользователю еще предстоит обработать вручную для удовлетворения своей информационной потребности. Семантическая интерпретация текста позволяет решать такие задачи, как фактографический поиск, поиск ответа на вопрос, реферирование, соотнесения различных текстов по смыслу (англ. text entailment, т.е. задача определения, можно ли смысл одного текста вывести из смысла другого).

Одним из способов структурирования извлекаемой информации являются онтологии. Для задач анализа естественного текста онтологии могут служить концептуальной моделью, к которой привязываются результаты применения разнотипных методов извлечения информации – распознавания именованных сущностей, извлечения фактов и пр. В процессе извлечения осуществляется отображение текстовых единиц в элементы онтологии (сущности и отношения).

В данной работе предлагается подход к извлечению семантических отношений между выделенными в тексте онтологическими сущностями. Предлагаемый подход основан на концепции «открытого извлечения отношений» (англ. Open Information Extraction [3]), где модель извлечения не является направленной на конкретные отношения, а ориентирована на выявление всевозможных отношений в корпусе. Подобные подходы не требуют от пользователей каких-либо обучающих примеров.

В сравнении с существующими работами в области «открытого» извлечения предлагаемый подход отличается следующими преимуществами:

- не требует избыточности информации в корпусе,
- не требует глубокого лингвистического анализа,
- не требует множества исходных примеров отношений,
- не является ориентированной на фиксированные структуры языковых выражений (например, глагольно-ориентированные).

Труды XIV Всероссийской объединенной конференции «Интернет и современное общество» (IMS-2011), Санкт-Петербург, Россия, 2011.

В качестве исходных данных был выбран корпус статей Википедии и онтология DBPedia, лексикон которой служит основой для разметки текста статей онтологическими сущностями.

2. Современное состояние области

На данный момент в области извлечения информации существует несколько постановок задачи извлечения отношений, зависящих от объекта извлечения. Некоторые работы [7, 9] ставят своей целью выявление факта упоминания в тексте некоторого типа отношения между двумя классами онтологии и номинацию (т.е. именование) этого отношения. При этом конкретные экземпляры отношений (т.е. тройки “сущность-отношение-сущность”) в явном виде не извлекаются. Большое количество работ [8, 10] ориентировано на извлечение экземпляров предварительно заданных одного или нескольких отношений. В таком случае, предполагается, что пользователи модели извлечения имеют представление о целевых отношениях и могут привести примеры сущностей, связанных этими отношениями. Однако, необходимость обучать модель под каждое отдельное отношение и подбирать множество примеров ведѐт к проблемам масштабирования. Рост трудоемкости пропорционален росту числа отношений, к тому же возникает необходимость решать проблему связности извлеченных независимо друг от друга экземпляров отношений. Подобные выводы привели к появлению концепции “открытого извлечения информации”, где модель извлечения не является направленной на конкретные отношения, а ориентирована на выявление всевозможных отношений в корпусе. Подобные подходы не требуют от пользователей каких-либо обучающих примеров. К существующим разработкам в области открытого извлечения можно отнести систему TextRunner [3], систему WOE [11], LUCHS [5], работу [4].

Отметим определенные особенности перечисленных работ в сравнении с нашим подходом.

Подходы TextRunner [3] и [4] выделяют языковой шаблон, как выражающий отношение, основываясь на избыточности появления этого шаблона в текстах корпуса, т.е. одно и то же имя (существительное или именная фраза) должно быть связано этим шаблоном несколько раз. Заметим, что во всех приведенных работах речь идет об именах без привязки к онтологическим сущностям. В нашем подходе предлагается использование избыточности на уровне “шаблон + классы сущностей”, что ослабляет требования к объему корпуса и необходимости дублирования информации в нем.

Подход WOE [11] в качестве шаблонов рассматривает цепочки слов, связывающих пару имѐн в синтаксическом представлении предложения. Стоит отметить ограниченность подобного подхода ввиду сложности и неточности синтаксического анализа, в особенности, на сложных предложениях (что отмечают и сами авторы) и для текстов на русском языке.

Подходы WOE [11] и LUCHS [5] используют значения атрибутов инфобоксов Википедии для начальной классификации предложений. В нашем подходе предлагается полностью автономный подход к выявлению шаблонов, обрабатывающий только текст статей, что вполне вероятно позволит применять подход на других корпусах (при наличии в них семантической разметки).

3. Постановка задачи

Пусть $D = \{d_1, d_2, \ldots, d_N\}$ – множество документов корпуса. В нашем случае его элементами являются статьи Википедии. Основываясь на определении онтологии в работе [6], зададим онтологию как семерку $O = (C, \leq_C, R, \sigma, E, ref_E, rel)$, где

- C и R – два непересекающихся множества, элементы которых называются классами и отношениями соответственно,
- частичный порядок $\leq_C$ на множестве C, называемый иерархией классов или таксономией,
- отображение $\sigma : R \to C \times C$, называемое сигнатурой. Сигнатура обозначает домен (область определения) и диапазон (область значений) отношения,
- E – множество сущностей (также называемых экземплярами классов),
- $ref_E \subseteq E \times C$ – отношение, назначающее сущность соответствующему классу,
- $rel = \{rel_r\}, r \in R$ – семейство множеств, где rel_r – множество пар вида (e_1, e_2), где $(e_1, C_1) \in ref_E, (e_2, C_2) \in ref_E$, где классы C_1 и C_2 связаны равенством $\sigma(r) = (C_1, C_2)$. Т.е. каждое множество rel_r представляет собой множество пар сущностей онтологии, связанных отношением r.

Далее, разметка текста документов сущностями онтологии представляет собой индекс I, который является множеством элементов вида (d_i, e_j, i_1, i_2), обозначающим, что сущность e_j встречается в тексте документа d_i в позиции от i_1 до i_2.

Требуется пополнить экстенсионал rel на основе анализа корпуса текстов D следующим образом. Для каждого нового экземпляра $l_i = (e_{i1}, e_{i2}), l_i \in rel_r$ отношения r должна существовать четверка вида $(d_{i3}, l_i, y_{i4}, y_{i5})$,

обозначающая, что в документе d_{i3} имеется фрагмент текста в позиции от y_{i4} до y_{i5}, содержащий данный экземпляр отношения l_i. Множество подобных четверок $\{(d_{i3}, l_i, y_{i4}, y_{i5}) \mid l_i \in rel_r\}$ для каждого отношения r будем именовать индексом Y_r.

4. Предлагаемый подход

В качестве исходного корпуса в данной работе используется русскоязычная Википедия. Стоит отметить, что методы основанные на избыточности информации не являются подходящими для извлечения отношений из статей Википедии, поскольку каждая статья является уникальной, т.е. в большинстве случаев единственным текстом, описывающим конкретную сущность и еѐ отношения, что, очевидно, уменьшает количество дублирующейся информации. Разметка текста сущностями онтологии (индекс I) организована путем сопоставления внутренних ссылок (ссылок на другие статьи, т.е. другие сущности) в статьях с лексиконом базы знаний онтологии DBPedia.

В рассмотренных подходах «открытого» извлечения отношений отсутствовала селекция выявленных сущностей в тексте в явном виде, т.е. рассматривалась совокупность всех выявленных языковых выражений (описывающих некоторый набор отношений) между любыми сущностями. В нашем подходе предлагается выявлять всевозможные отношения с учетом онтологического описания контекста их появления, посредством явного указания классов сигнатуры отношения.

В работе [1] были проведены эксперименты, показывающие, что слова, являющиеся частыми для контекстов конкретной пары классов, и редкими - для других контекстов, являются значимыми с точки зрения обозначения семантического отношения, понятного для человека. Например, для контекста (Educational Institution, Person) были получены следующие слова: “поступил(а)”, “окончил(а)”, “учился”, “изучал”, “профессором”, “факультет”, “диссертацию” и прочие. Следующим этапом является выявление языковых шаблонов с детализацией, достаточной для утверждения наличия в тексте семантического отношения между онтологическими сущностями, заполняющими поля шаблона.

Предполагается, что извлекаемые отношения выражаются в тексте конечным числом шаблонов. Под шаблоном здесь понимается структура ограниченного фрагмента текста с выделенными в еѐ элементах значениями определенных свойств и взаимосвязями (лексическими, грамматическими). Так как в выбранном представлении онтологии каждое отношение связывает 2 класса, то можно предположить, что языковое представление отношения в тексте связывает имена сущностей этих классов. Для выявления шаблонов, выражающих отношения между двумя определенными классами онтологии, предлагается анализировать контексты совместного появления сущностей этих классов. Целью данного анализа является поиск наиболее характерных (т.е. частотных) шаблонов. Частотные шаблоны являются таковыми:

- либо по причине общей распространенности в языке (точнее, языке обрабатываемого корпуса). Эту гипотезу для каждого шаблона можно проверить путем анализа других фрагментов текста в корпусе. В случае подтверждения можно рассматривать частные случаи данного шаблона путем усиления ограничений на его элементы или добавления новых ограничений.
- либо они являются значимыми конкретно для выбранных классов онтологии и являются индикаторами способа выражения в тексте онтологического отношения (возможно, нескольких отношений).

Таким образом, предлагается группировать отношения, объединѐнные одними и теми же классами сигнатуры. Обозначим группу отношений, состоящих из отношений с сигнатурой (C_1, C_2) либо (C_2, C_1) через $\{C_1, C_2\}$. Далее будем рассматривать процесс извлечения для конкретной группы $\{C_1, C_2\}$.

Определим задачу формирования шаблонов на основе анализа текстовых фрагментов. Ограничим множество возможных фрагментов текста предложениями. Множество всех предложений корпуса обозначим T. Множество всех предложений корпуса, содержащих хотя бы одну сущность класса C_1 и хотя бы одну сущность класса C_2, будем обозначать через T_{C_1,C_2}.

Определим шаблон p как функцию принимающую значение истины или ложности на текстовых фрагментах, т.е. $p: T \to \{0,1\}$. Поддержкой (support) шаблона p на множестве фрагментов $Q \subseteq T$ будем называть долю текстовых фрагментов этого множества, на которых p принимает значение истины:

$$\sup_Q(p) = \frac{1}{|Q|}\sum_{s \in Q} p(s)$$

Задача формирования шаблонов заключается в поиске шаблонов с максимальной поддержкой, выражающих одно конкретное онтологическое отношение. На данный момент мы не вводили механизмов, различающих выражение одного отношения в тексте от выражения другого. Более того, формализация этого механизма и его реализация вычислительной процедурой не представляется тривиальной. С учетом данного замечания, предлагается упорядочивать множество

шаблонов P по значению поддержки на множестве T_{C_1,C_2}. В дальнейшем, просматривая полученный список по порядку, можно вручную проводить соответствия шаблонов онтологическим отношениям. Если отношению r_i будет установлен в соответствие шаблон p_j, то мы можем пополнить онтологию по следующему алгоритму:

ДЛЯ КАЖДОГО $s \in T_{C_1,C_2}$

ЕСЛИ $p_j(s) = 1$ ТО

$l = (e_a, e_b)$

Добавить l в rel_{r_i}

Добавить $(doc(s), l, bpos(s), epos(s))$ в Y_{r_i}

Где:

- $doc(s)$ – документ, содержащий фрагмент s
- $epos(s)$ – функция, возвращающая позицию конца фрагмента s в документе
- $bpos(s)$ – функция, возвращающая позицию начала фрагмента s в документе

5. Метод ассоциативных правил

В нашей работе предлагается подход к спецификации и формированию шаблонов, основанный на методе ассоциативных правил. Данный метод [2] является инструментом поиска закономерностей в массиве данных в виде наборов бинарных признаков. Если представить текстовые фрагменты корпуса изучаемыми объектами, а различную лексическую, морфологическую, синтаксическую и семантическую информацию о фрагментах текста представить множеством значений бинарных признаков, то в результате применения метода ассоциативных правил ожидается получить ассоциативные правила, интерпретация которых будет обозначать какие слова, синтаксические конструкции и концепты обычно используются в текстах статей для выражения отношения между сущностями.

Основываясь на работе [2], адаптируем исходную терминологию метода ассоциативных правил для задачи анализа множества текстовых фрагментов Q. Для этого каждый текстовый фрагмент представим транзакцией (transaction). Пусть $F = \{f_1, \ldots, f_n\}, f_j : Q \to \{0,1\}, j = 1 \ldots n$ – множество бинарных признаков над текстовыми фрагментами, также называемых элементами (items). Если $f_j(s) = 1$, то говорят, что транзакция s содержит признак f_j (f_j встречается в s). Каждая транзакция $s \in Q$ представляет собой бинарный вектор, где $s[j] = f_j(s)$. Пусть $\varphi \subseteq F$ – набор признаков. Если для каждого $f_j \in \varphi$ выполняется $s[j] = 1$, то говорят, что транзакция s содержит набор φ (признаки набора φ встречаются совместно в s).

Частота встречаемости или поддержка (support) набора φ (обозначим её $\sup(\varphi)$) определяется долей транзакций содержащих набор во множестве всех транзакций Q.

Пара непересекающихся наборов $\varphi, \psi \subseteq F$ называется ассоциативным правилом (association rule) $\varphi \to \psi$, если выполнены два условия:

$$\sup(\varphi \cup \psi) \geq MinSupp$$

$$conf(\psi \mid \varphi) \equiv \frac{\sup(\varphi \cup \psi)}{\sup(\varphi)} \geq MinConf$$

где $MinSupp$ – минимальная поддержка, $conf(\psi \mid \varphi)$ – достоверность правила (confidence), $MinConf$ – минимальная достоверность. $MinSupp$ и $MinConf$ являются параметрами алгоритма поиска ассоциативных правил. Приведем пример ассоциативного правила (ожидаемого для контекста (Person, Educational Institution)):

(1) фрагмент содержит глагол “поступать”
→ (2) фрагмент содержит предлог “в” перед именем второй сущности (класса Educational Institution)
и (3) имя второй сущности во фрагменте находится в родительном падеже

Содержательно это правило можно интерпретировать следующим образом: в проанализированном множестве транзакций (т.е. множестве текстовых фрагментов):

- признаки (1), (2) и (3) встречаются совместно достаточно часто (с частотой выше MinSupp)
- если (1) встретилось, то в значительной доле случаев (не менее MinConf) встретится (2) и (3).

Аппарат ассоциативных правил предоставляет способ задания шаблонов в виде наборов элементарных признаков. Одним из преимуществ подобного представления является простота интерпретируемости, поскольку отбор итоговых шаблонов и их привязка к отношениям будет осуществляться человеком вручную.

Таким образом, если применить алгоритм поиска ассоциативных правил на множестве T_{C_1,C_2}, то полученные правила можно рассматривать как искомые шаблоны, при этом значение поддержки шаблона выражается поддержкой соответствующего ассоциативного правила.

В данной работе предлагаются к использованию следующие виды бинарных признаков для формирования ассоциативных правил.

1. Признаки, обозначающие присутствие словоформы в тексте:

$wf_w(s) = 1$,если $w \in s$, иначе 0,

где w – конкретная словоформа. При этом не рассматриваются предлоги и другие стоп-слова.

2. Признаки наличия определенных морфологических атрибутов у имени сущности:

$morphEnt_{gr}(s) = 1$, если имя сущности в s имеет грамматическое значение gr , иначе 0.

3. Признаки, обозначающие наличие конкретных словоформ непосредственно перед именем сущности или непосредственно после имени сущности (в отличие от признаков вида (1) здесь возможны любые словоформы, включая предлоги):

$wBeforeEnt_{wf}(s) = 1$, если $\exists i : t_i = wf, t_{i+1}$ – первый токен упоминания сущности, иначе 0,

$wAfterEnt_{wf}(s) = 1$, если $\exists i : t_i = wf, t_{i-1}$ – последний токен упоминания сущности, иначе 0.

Категория признаков наличия словоформы (1) ограничивает языковые конструкции лишь условием наличия словоформы в тексте, однако, место еѐ в тексте игнорируется. Рассмотрим некоторые примеры контекстов появления в тексте сущностей класса «Университет» (в квадратных скобках указаны границы именование сущности):

1. Родился в Чикаго, учился на факультете электроники [Корнелльского университета].
2. Специалисты [Университета в Торонто] исследовали манеру письма Кристи в эти годы и выдвинули предположение, что ... .
3. В молодые годы основал и возглавил в [Гарварде] социалистический кружок, от членства в котором позднее отказался.

В первом примере глагол «родился» не связан с Корнелльским университетом. Второй пример, несмотря на наличие глагола «исследовать» и именования университета, не говорит о том, что Агата Кристи занималась исследованиями. Третий пример показывает, что не всегда наличие во фрагменте текста глагола «основал» и именования университета говорит о факте основания университета.

Приведенные примеры показывают необходимость добавления признаков категории (2) и (3), связывающих слова с именованием сущности онтологии в тексте. Эти признаки имеют большую значимость, поскольку указывают на роль сущности в тексте. К примеру:

1. Первый официальный герб Казани был утверждѐн 18 октября 1781 года. (Казань как владелец сущности упомянутой непосредственно перед еѐ именем в родительном падеже.)
2. В Казани за год перевезено почти 300 миллионов пассажиров. (Казань в роли локации, предлог «в» перед названием, родительный падеж.)
3. Казань расположена на левом берегу р. Волги. (Казань как субъект утверждения, именительный падеж.)

6. Заключение

В данной работе предложен новый подход «открытого» извлечения семантических отношений. Был проведен анализ ряда существующих работ в этой области, в результате которого были выявлены конкретные их недостатки, связанные как с применяемым инструментарием, так и требованиями к исходным данным. Предлагаемый подход обладает определенными преимуществами по этим критериям.

Так предлагаемый подход не требует избыточности информации на уровне конкретных экземпляров отношений. Введение семантической разметки текста сущностями онтологии позволяет анализировать частотность языковых конструкций во фрагментах текста, содержащих сущности одних и тех же классов. При этом избыточность появления в тексте конкретной сущности не является сколь либо значимой – важным является количество фрагментов текста, содержащих пару сущностей, принадлежащих конкретным классам сигнатуры отношений.

Предлагаемый подход не формирует исходное множество примеров отношений на основе атрибутов инфобоксов статей Википедии, что, вероятно, позволит применять его на других подобных корпусах, лишенных инфобоксов.

Процесс формирования шаблонов в предлагаемом подходе не завязан на конкретные языковые структуры, например, ориентированные на наличие глагола. Используемый в данном подходе метод ассоциативных правил направлен на поиск сочетаемостей любых словоформ. На конкретных примерах из статей Википедии обоснована необходимость предлагаемого набора признаков. При этом вычисление значений этих признаков не требует глубокого лингвистического анализа.

Дальнейшая работа включает в себя реализацию предлагаемого подхода, проведение экспериментов, оценку качества получаемых шаблонов и извлекаемых отношений. Кроме того, с целью повышения автономности процесса извлечения, предполагается рассмотреть адаптацию существующих методов обнаружения парафраз (paraphrase discovery) для автоматической группировки найденных шаблонов по выражаемому семантическому отношению.

Литература

[1] Гареев, Р.М. Извлечение онтологических отношений из семантически размеченного корпуса текстов Википедии / Гареев Р.М., Иванов В.В. // Системный анализ и семиотическое моделирование: материалы первой всероссийской научной конференции с международным участием (SASM-2011). Казань: Академия Наук РТ, 2011. С. 33-40.

[2] Agrawal, R. Mining association rules between sets of items in large databases / Agrawal R., Imielinski T., Swami A. // Proceedings of the 1993 ACM SIGMOD international conference on Management of data. ACM, 1993. P. 207-216.
[3] Banko, M. Open information extraction from the web / Banko M., Cafarella M., Soderland S., Broadhead M., Etzioni. O. // Proceedings of the 20th International Joint Conference on Artificial Intelligence. IJCAI 2007, Proceedings of the 20th International Joint Conference on Artificial Intelligence, Hyderabad, India, January 6-12, 2007. P. 2670-2676.
[4] Davidov, D. Unsupervised Discovery of Generic Relationships Using Pattern Clusters and its Evaluation by Automatically Generated SAT Analogy Questions / Davidov D., Rappoport A. // Proceedings of ACL-08: HLT. Association for Computational Linguistics, 2008. P. 692-700.
[5] Hoffmann, R. Learning 5000 relational extractors / Hoffmann R., Zhang C., Weld D. S. // Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, 2010. P. 286-295.
[6] Maedche, A. Bootstrapping an ontology-based information extraction system / Maedche A., Neumann G., Staab S. // Intelligent exploration of the web. Heidelberg: Physica-Verlag GmbH, 2003. P. 345-359.
[7] MaedcheЮ A. Discovering conceptual relations from text / Maedche A., Staab S. // ECAI 2000, Proceedings of the 14th European Conference on Artificial Intelligence / Horn, W., editor. Berlin, Germany, August 20-25, 2000. IOS Press, 2000. P. 321-325.
[8] Sarawagi, S. Information extraction // Foundations and Trends in Databases. 2008. Vol. 1. № 3.
[9] Schut,z A. RelExt: A Tool for Relation Extraction from Text in Ontology Extension / Schutz A., Buielaar P. // The Semantic Web–ISWC 2005. Vol. 3729. Springer, 2005. P. 593-606.
[10] Weikum, G. From Information to Knowledge: Harvesting Entities and Relationships from Web Sources / Weikum G., Theobald M. // Proceedings of the twenty-ninth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems of data. ACM, 2010. P. 65-76.
[11] Wu, F. Open information extraction using Wikipedia / Wu F., Weld D. S. // Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, 2010. P. 118-127.

Open extraction of semantic relationships using association rules mining

R. Gareev, V. Solovyev

Currently most of existing methods of relationship extraction are relation-specific and use supervised techniques to learn extractors. But there is also another extraction paradigm, called Open Information Extraction, which is basic for our proposed approach. We introduce new extraction model based on patterns learning. The model's input is represented by encyclopedic corpora with automatically generated semantic markup using knowledge base of ontology. We use association rules mining to learn patterns of the most frequently mentioned relations between different classes of entities. We expect following advantages of proposed extraction model:

– it doesn't require information redundancy in corpora,
– it doesn't require deep linguistic analysis,
– it doesn't require seed set of relation instances and doesn't utilize infobox values to make seeds,
– it isn't focused on particular phrasal structures (e.g. verb-centric).

We are going to use Russian Wikipedia as input corpora and DBPedia as source ontology.

♣ Работа выполнена при поддержке РФФИ (проекты №№ 09-07-97007-р_поволжье_а и 10-07-00445-а)