УДК 004.02

Методы семантического аннотирования

В.В. Лопаткин[1]

[1]Московский государственный университет им. М.В. Ломоносова

В современной литературе существует разное понимание термина «аннотация ресурса». Обычно под аннотацией ресурса, понимаются некоторые метаданные (информация о данных) этого ресурса. Выделяют следующие три вида аннотаций: неформальные, формальные и онтологические. Неформальная аннотация описывается на естественном языке и поэтому обычно не обрабатывается с помощью программ. Формальные аннотации составляются с использованием специальных языков (таких как XML и RDF), что позволяет выполнять их программную обработку. Если аннотация составляется на основе некоторой семантической модели (онтологии), описывающей основные понятия и отношения той предметной области, к которой относится описываемый ресурс, то она называется онтологической. Формирование аннотаций ресурсов может выполняться с применением различных средств, таких как теги, наборы ключевых слов, наборы понятий и наборы триплетов.

При использовании тегов аннотирование выполняется путем их добавления в тексты документов. Это делается для того, чтобы обрабатывающая программа могла определить формальный смысл выделенных с их помощью частей. Например, данный способ может использоваться для аннотирования веб-страниц, в которые, кроме HTML тегов, включаются и микроформаты RDFa. Стандарт RDFa позволяет использовать URI идентификаторы для набора атрибутов, а также набор общепринятых словарей, таких как hCalendar, hCard или hAtom.

Другим способом аннотирования ресурсов является использование набора ключевых слов. Такое средство чаще применяется в электронных библиотеках для аннотирования научных статьей. Ключевые слова для конкретной статьи выбираются их авторами, а затем хранятся в базе данных для поддержки процесса поиска. Неоднозначность семантики ключевых слов приводит к использованию наборов понятий. Такие понятия могут выбираться из онтологий или тезаурусов, описываемых с помощью языка RDFS или OWL. При этом каждое понятие представляется в виде URI идентификатора.

Третьим способом аннотирования ресурса является его описание с помощью наборов утверждений, имеющих следующий формат: субъект–предикат–объект, которые также называются триплетами. Данный способ является самым новым, и его использование еще в достаточной степени не исследовано. В данном докладе как раз рассматривается подход к аннотированию ресурсов с использованием данного способа.

Обзор методов семантического аннотирования документов

Аннотирование может выполняться разными способами: вручную, полуавтоматически или автоматически. При ручном способе аннотирования составлением метаописания документа занимается специалист. Однако программная система может оказывать ему помощь, например, показывая онтологии и имеющиеся экземпляры, выполняя проверку их правильности. При полуавтоматическом методе аннотирования система автоматически составляет начальный вариант аннотации, а затем специалист может проверять и выполнить ее корректировку. Ручной и полуавтоматический способы трудно использовать при большом количестве документов, для документов больших размеров, а также при недостатке специалистов, которые могут своевременного выполнять такую работу. В этих случаях возникает потребность в автоматическом методе аннотирования.

Полуавтоматическое и автоматическое аннотирование.

В отличие от ручного метода, полуавтоматический метод создает частичную аннотацию документа с помощью анализа естественного языка документа. Извлечение информации является основной технологией для связывания неструктурированного текста с формальными описаниями, содержащимися в онтологиях. При извлечении информации часто используются такие компоненты обработки естественного языка, как разметчик частей речи (part-of-speech tagger), морфологический анализатор, сканеры именованных сущностей, полный (или поверхностный) синтаксический анализ и семантическая интерпретация. Существует два подхода к извлечению информации:

1. На основе использования наборов правил, разработанных специалистами по лингвистике, которые создают словари и описывают правила извлечения требуемой информации.

2. На основе использования методов машинногообучения, позволяющих выполнять автоматическое обучение для решения различных задачизвлечения информации.

Преимуществом первого подхода является то, что не требуется составлять обучающие выборки данных для создания правил и описания знаний предметной области. Благодаря созданным словарям и правилам системы аннотирования могут работать быстрее систем, разработанных с помощьюмашинного обучения. Для этого подхода обычнотребуется создание коллекции проаннотированныхчеловеком обучающих данных для достижения высокой точности. С одной стороны, на обучение такой системы требуется меньше затрат, чем на создание словарей и наборов правил извлечения информации из текста, но, с другой стороны, могут возникать проблемы, связанные с низким качеством составления аннотаций.

Метод полуавтоматического семантического аннотирования

В общем виде для составления триплета аннотирования документа необходимо вручную выбрать субъект, определять его предикат (отношение), на основе его описания в онтологии, а затем выбрать связанный с ним объект. Созданный триплет сохраняется в базе знаний.

Выбор субъектов и объектов триплетов выполняется в ходе решения задач поиска кандидатов и преодоления многозначности.С учетом изложенного выше, задача семантического аннотирования может быть структурированаследующим образом:

1. Выбор нужных понятий. Аннотирование документа выполняется в соответствии с некоторой онтологией предметной области, и при этом специалист имеет возможность ограничить некоторые категории понятий в онтологии предметной области (желаемые понятия) для преодоления многозначности.

2. Поиска кандидатов. Поиск в базе знаний основных кандидатов (понятий, предикатов или экземпляров). Для этого необходимо в документе находить термины (слова или словосочетания), которые совпадают с метками экземпляров или понятий онтологии или близки им в соответствии с некоторой оценкой семантической близости.

3. Преодоление многозначности. Данный шаг заключается в том, что из аннотации должны быть исключены все нерелевантные кандидаты.Эти кандидаты сходны с терминами документа по текстовым меткам, но в действительности не описывают содержание данного документа.

Решение задачи поиска кандидатов.

Обозначим набор понятий и экземпляров онтологии, хранящихся в базе знаний как $M_{CE} - \{ce_1, ce_2, \dots, ce_n\}$. Каждый элемент (ce) может иметь текстовые метки для их обозначения на разных естественных языках. Набор текстовых меток соответствующих элементов M_{CE} некоторой онтологий O обозначим как M=$\{m_1,m_2,\dots,m_n\}$, каждая метка может быть представлена в виде набора токенов в результате токенизации (tokenization). Под токенизацией понимается процесс разделения текста, содержащего метки экземпляров, понятий или документов, на последовательность токенов, при этом в качестве разделителя может использоваться знак пробела.

Таким образом, аннотируемый документ (илинабор документов D) может быть представлен в виде множества токенов. Для учета грамматики естественных языков возникает необходимость выполнять нормализацию токенов. Существуют два типа систем, выполняющих нормализацию токенов: лемматизатор (lemmatizer) и стеммер (stemmer).

Как правило, лемматизатор токена возвращает его исходную форму, а стеммер – его корень с помощью правил отсечения. При нормализации могут быть удалены стоп-слова (шумовые слова), которые не несут никакой смысловой нагрузки.

Решение задачи поиска кандидатов может быть разделено на следующие шаги:

• преобразование D и M в наборы нормализированных токенов, обычно отсортированных в алфавитном порядке;

• поиск в D набора токенов для каждой метки $m_i \in M$.

В результате выполнения этих шагов будет получен набор кандидатов M_K $(M_K \subset M_{CE})$

Решение задачи преодоления многозначности.

Существуют два подхода к автоматическому решению данной задачи: с помощью измерения семантической близости и с помощью измерения популярности. Идея подхода измерения семантической близости для решения многозначности заключается в том, что для набора найденных кандидатов с использованием онтологий вычисляются их семантические близости с понятиями, подходящими для аннотирования документа. В результате этого система может выбирать тех кандидатов, близость которых с желаемыми понятиями больше некоторого порогового значения.

Допустим, что имеется конечный набор кандидатов MK из онтологии O. Также имеется конечный набор желаемых понятий для аннотирования документа M_C={c_i ∈C⊂O}. Тогда релевантными будут считаться кандидаты из MK, удовлетворяющие следующему условию:

$$Sem(ce_i, M_C) = \max\left(Sem_{c_i \in M_C}(ce_i, c_j)\right) > \varepsilon \qquad \forall \varepsilon > 0,$$

где Sem – семантическая близость и ε – установленное пороговое значение.

Литература

1. *Oren E., Hinnerk Möller K., Scerri S., Handschuh S., Sintek M.* What are Semantic Annotations? URL: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.97.7985&rep=rep1&type=pdf

2. *Domingue J., Fensel D., Hendler J.A.* Handbook of Semantic web Technologies. – Heidelberg; Dordrecht; London; N.Y.: Springer, 2011. – 1077 p.

3. *Черний А.В., Тузовский А.Ф.* Развитие информационной системы организации с использованием семантических технологий // Знания–Онтологии–Теория: Матер. Всеросс. конф. с междунар. участием. – Новосибирск, 20–22 октября 2009. – Новосибирск: ЗАО «РИЦ Прайс-Курьер», 2009. – Т. 2. – С. 52–59