Подход к реализации автоматизированной системы построения тезауруса

© Тарасов С.Д.

Балтийский Государственный Технический Университет им. Д.Ф.Устинова «ВОЕНМЕХ»
tarasov_sd@mail.ru

Аннотация

В статье рассматриваются практические аспекты создания автоматизированной системы построения тезауруса. Определяются основные проблемы организации современных тезаурусов и методы их построения. Рассматривается проблема автоматической генерации тезаурусов на основе лингвистических источников разного типа: результатов анализа корпусов текстов, дефиниций толковых словарей, данных ассоциативных словарей и т.д. Описывается предлагаемая архитектура автоматизированной системы генерации тезаурусов для любой предметной области на основе лингвистических источников и при помощи участия независимых асессоров.

Введение

В 60-е и 70-е годы основным подходом к представлению семантики языка был компонентный подход, в рамках которого значение каждого слова естественного языка должно было быть представлено в виде комбинации семантических универсалий. К середине 80-х годов стало ясно, что общепризнанный набор таких универсалий так и не удалось составить. Альтернативой компонентному подходу в семантике стала реляционная семантика. При этом подходе значения слов языка описываются заданием связей со значениями других слов, а вся понятийная система языка представляется как семантическая сеть [3].

Наиболее известным лингвистическим ресурсом, представляющим понятийную систему английского языка в виде семантической сети, является электронный словарь WordNet [3]. Основным методологическим источником для развития тезауруса русского языка стал Общественно-политический информационно-поисковый тезаурус для автоматического индексирования [3]. Общественно-политический тезаурус включает в себя терминологию экономической, политической, военной, финансовой, законодательной, социальной, культурной и других сфер деятельности [3]. Общественно-политический тезаурус является на сегодняшний день наиболее крупным проектом подобного рода. Существует также довольно большое число более мелких тезаурусов, соответствующих определенным более узким предметным областям (ПО). Именно такие тезаурусы, как правило, используются в информационно-поисковых системах (ИПС) для повышения полноты поиска [5]. По мнению, изложенному в [1], тезаурус является эффективным инструментом формирования запросов к универсальным машинам поиска (МП) Интернет и может существенно повысить качество информационного поиска в специализированных тематических областях. Для этого необходимо выполнения следующих условий [1]:

1. Тезаурус отражает терминологию достаточно узкой предметной области,
2. В тезаурусе используется набор сильно дифференцированных семантических отношений,
3. Тезаурус независим от МП.

Принцип использования тезауруса для улучшения качества поиска и построения более точных запросов довольно подробно описан в [1] и [2]. Однако сами вопросы получения тезауруса, по мнению автора, освещены в литературе довольно слабо. Ручное построение тезауруса, с одной стороны, довольно тривиальная задача, с другой,- крайне трудоемкая, так как требует привлечения специалистов(экспертов предметной области и лингвистов) для выделения семантических отношений между понятиями, что особенно сказывается при большом количестве терминов тезауруса. За время развития компьютерной лингвистики было предпринято довольно много попыток автоматической генерации тезаурусов путем преобразования описаний толковых словарей и энциклопедий в семантические сети. Однако в силу того, что такое преобразование само по себе является чрезвычайно сложным и трудоемким мыслительным процессом, эти попытки не увенчались успехом. Основные проблемы, которые возникают в процессе такого преобразования [3]:

1. Значения слов весьма расплывчаты, полные синонимы, удовлетворяющие критерию замены в синтаксическом контексте,

относительно редки. Два синонима одного и того же слова в одном и том же значении часто не синонимы между собой.
2. Различные словари при описании многозначных слов выделяют разное количество значений.
3. В процессе изучения толкований слов в толковых и энциклопедических словарях оказывается, что в значительном количестве таких толкований не содержится ссылка на родовое (вышестоящее) понятие. Кроме того, значительная доля понятий может иметь более одного вышестоящего понятия.

Основные принципы автоматизированной системы построения тезауруса

Таким образом, полностью автоматическое построение тезауруса является на текущий момент невозможным. Однако было предпринято довольно много успешных попыток частично автоматизировать данный процесс. К недостаткам существующих систем, известных автору, можно отнести:

1. Невозможность одновременной многопользовательской работы с системой.
2. Наличие только бинарных связей между понятиями не позволяющих каким-либо образом дифференцировать связи.
3. Набор связей между понятиями тезауруса ограничен традиционным списком РОД-ВИД, ЧАСТЬ-ЦЕЛОЕ, АССОЦИАЦИЯ, СИНОНИМ, АНТОНИМ.
4. Связь между понятиями определяется единожды одним экспертом.
5. Система не имеет статистики по работе экспертов, а также не хранит историю изменений.

Предлагаемая автором автоматизированная система позволяет упростить процесс построения тезауруса, а также обладает следующими отличительными чертами:

1. В построении тезауруса могут принимать участие, как специалисты (эксперты предметной области и лингвисты), так и обычные «асессоры». В таком случае возможно дифференцирование весов «оценок» различных групп участников.
2. В формировании каждой семантической связи между понятиями участвует сколь угодно много «мнений» экспертов и асессоров, что позволяет существенно повысить точность определения этих связей и практически полностью исключить флуктуации.
3. Набор возможных семантических связей между понятиями никак не ограничен. На первом этапе планируется использовать традиционные связи (РОД-ВИД, ЧАСТЬ-ЦЕЛОЕ, АССОЦИАЦИЯ, СИНОНИМ, АНТОНИМ)
4. Каждая семантическая связь имеет свой вес $w \in [0,1]$.
5. Система хранит всю историю изменений и позволяет в ручном или автоматическом режиме получать все статистические характеристики (мат. ожидание, ср. кв. отклонение и т.п.), а также исключать из результирующей оценки заведомо неправильные, по мнению эксперта, «мнения».

Тезаурус, для построения которого предназначена система, представляет собой иерархическую сеть понятий, соответствующих тем или иным значениям отдельных слов или текстовых выражений. Для описания связей между понятиями используется традиционная для тезаурусов система семантических отношений:

1. РОД-ВИД
2. ЧАСТЬ-ЦЕЛОЕ
3. АССОЦИАЦИЯ
4. СИНОНИМ
5. АНТОНИМ

Все семантические связи имеют весовой коэффициент $w \in [0,1]$. Набор различных видов семантических отношений никак не ограничивается системой. Планируется добавление в систему такого вида отношения, как ПАРОНИМ.

Исходные термины предметной области могут быть загружены через специальный интерфейс. Источником таких терминов может быть определенный словарь или заранее подготовленный список. В будущем возможно составление списка терминов посредством автоматизированного анализа исходных текстов с выделением терминов предметной области и исключением общеупотребительных слов. В случае применения автоматизированной системы для получения ассоциативного словаря общеупотребительной лексики в качестве исходного набора слов может быть использован любой словарь. Для заполнения тезауруса используются два основных подхода.

Метод оценки веса семантической связи

Пользователю системы (специалисту или асессору) предлагается случайная (псевдослучайная) выборка N терминов тезауруса $A^N \subset V$ и также случайная (псевдослучайная) выборка N других терминов $B^N \subset V$, так что $A_i \neq B_i, i = 1..N$.

Участнику необходимо проставить коэффициенты всех возможных семантических связей между понятиями A_i и B_i, например, при N=2 и

A = {“автомобиль”, “самолет”}
B = {“двигатель”, “крыло”}

Асессор должен произвести оценку семантических связей между понятиями “автомобиль” – “двигатель” и “самолет” - “крыло”. Строго говоря, принцип выборки двух понятий из общей массы терминов для оценки семантической связи между ними не столь важен для системы. Важно обеспечить более-менее равномерное распределение оценок между понятиями. В любой момент коэффициент конкретной семантической связи между понятиями может быть вычислен путем усреднения с учетом весов оценок всех пользователей. Также следует отметить, что система не накладывает никаких ограничений на выбор метода оценки результирующего коэффициента. Метод усреднения с учетом весов является на текущий момент самым простым и наглядным.

Предположим, пользователь поставит следующие оценки для семантической связи “автомобиль” – “двигатель”:

РОД-ВИД: 0.00
ЦЕЛОЕ-ЧАСТЬ: 1.00
АССОЦИАЦИЯ: 0.80
СИНОНИМ: 0.00
АНТОНИМ: 0.00

Метод ассоциаций

Пользователю системы (асессору) предлагается случайная (псевдослучайная) выборка N терминов тезауруса $A^N \subset V$.

Для каждого из терминов он должен выбрать некое множество $B_i^{M_i}$ ассоциаций. Такой подход применяется в лингвистике для построения вербальных ассоциативных словарей. В нашем случае такой метод позволяет быстро выделить основные семантические связи между понятиями.

Коэффициент семантической связи может быть вычислен по следующей методике:

$K(A_i, B_i) = f(m_i)$, где $f(m_i)$ - некоторая функция от m_i - позиции термина B_i в списке ассоциаций. Предполагается, что список будет естественным образом упорядочен по убыванию значимости семантической связи.

Особое внимание здесь необходимо уделить вопросу морфологического разбора, так как выдавать список всех возможных терминов не представляется технически возможным (при больших объемах последнего), а в общем случае список, формируемый асессором, может содержать понятия отнюдь не в нормальной форме. В таком случае исходный набор терминов тезауруса должен содержать список отдельных лексических единиц в нормальной форме и связанный с ним список различных форм лексем. Морфологические словари в электронном виде не являются дефицитом, так же как и готовые решения модулей морфологического разбора.

За одну сессию асессор или эксперт может произвести сколь угодно много оценок связей между понятиями. Каждый пользователь может открыть сколь угодно много сессий. Администратор системы при необходимости может аннулировать результаты какой-либо из сессий или все оценки какого-либо пользователя, например, чтобы исключить заведомо неправильные результаты или флуктуации.

Заключение

Построенная по такому принципу система обладает очень большой гибкостью. Единственным недостатком системы является необходимость большого числа независимых оценок для достижения большой точности. Если изначально база тезауруса насчитывает N понятий, то для получения коэффициентов семантических связей из M независимых оценок требуется не менее $M \cdot \frac{N \cdot (N-1)}{2}$ оценок. Однако при отсутствии возможности обеспечения такого количества независимых оценок, принцип может быть сведен к традиционному, когда каждая связь определяется единожды экспертом при помощи лингвистических источников. То же можно сказать и о коэффициенте семантической связи. Вещественный коэффициент всегда может быть приведен к бинарному (есть связь/нет связи) различными способами.

Полученная в результате функционирования системы иерархическая сеть понятий может быть использована для различных нужд, в том числе и в задачах информационного поиска. Семантическая сеть обеспечивает пользователям навигацию в терминологическом пространстве. Кроме того, семантическая сеть позволяет выполнять "интеллектуальные" преобразования терминологии, возникающие при формировании глоссариев, а также при формировании запросов к МП. Каждую вершину семантической сети можно рассматривать как потенциальный запрос, передающий информационно-поисковой системе термин-словосочетание. Такие потенциальные запросы активизируются пользователем по мере надобности.

Кроме того, система хранит всю историю изменения весов, что позволяет выполнить сколь угодно тонкую «подстройку» весов связей, а также вычислять статистические характеристики не только связей, но и пользователей. Вычисление таких характеристик и оценок крайне интересно, однако выходит за рамки данной статьи.

Литература

[1] Альшанский Г.А., Браславский П.И., Титов П.В. Формирование информационных запросов к машинам поиска интернета на основе тезауруса: семантико-ориентированный подход // Труды VIII Международной конференции по электронным публикациям "EL-Pub2003". 8 – 10 октября 2003 года, Новосибирск, Академгородок.

[2] Браславский П.И. Тезаурус для расширения запросов к машинам поиска Интернета: структура и функции // Компьютерная лингвистика и интеллектуальные технологии. Тр. Междунар. конференции Диалог’2003 (Протвино, 11-16 июня 2003 г.). М.: Наука, 2003. С. 95-100.

[3] Лукашевич Н.В., Добров Б.В., Описание понятийной системы русского языка в виде тезаурусноорганизованной семантической сети // Труды международной конференции "Знания-Диалог-Решение" - Спб. - 2001 - Т.2 - С.438-444.

[4] Лукашевич Н. В., От общеполитического тезауруса к тезаурусу русского языка в контексте автоматической обработки больших массивов текстов // Труды международного семинара Диалог-99, т.2, 1999, с.184 -190.

[5] Солтон Д. Динамические библиотечно-информационные системы. Пер. с англ. М. Мир 1979

The automated system of construction the thesaurus

Tarasov S.D.

In article practical aspects of creating of the automated system of construction of the thesaurus are considered. Major problems of the organization of modern thesauruses and methods of their construction are defined. The problem of automatic generation of thesauruses on the basis of linguistic sources of different type is examined: results of the analysis of cases of texts, definitions of the explanatory dictionaries, the given associative dictionaries, etc. The offered architecture of the automated system of generation of thesauruses for any subject domain on the basis of linguistic sources and by means of participation independent assessors is described.