ПРИМЕНЕНИЕ ПРАВДОПОДОБНЫХ РАССУЖДЕНИЙ ДСМ – МЕТОДА ДЛЯ ПОПОЛНЕНИЯ СЕМАНТИЧЕСКОГО СЛОВАРЯ

JSM - METHOD'S PLAUSIBLE REASONING APPLICATION TO SEMANTIC DICTIONARY COMPLETION

О.С Кожунова *(kozhunovka@mail.ru)*

Институт проблем информатики Российской академии наук

Целью работы являлась разработка и программная реализация макета интеллектуальной системы пополнения семантического словаря. Пополнение словаря происходит при помощи обучения на примерах – основной процедуре ДСМ–метода. Для обработки текстов использован COM-объект нормализации слов в предложениях.

Введение

Семантический (концептуальный) словарь, предназначен для использования в информационных технологиях, связанных с пониманием текста (обнаружение и исправление ошибок в текстах, распознавание письменного и устного текста, аннотирование и реферирование, текстовый информационный поиск, извлечение фактов из текста на естественном языке и др.). Существующий подход к построению такого словаря основан на совместном использовании формального логического языка [1] и естественного языка. В данной работе основной моделью семантического словаря для разработки макета интеллектуальной системы пополнения семантического словаря является структура описаний понятий В.Ш.Рубашкина, описанная в [1]. Помимо словаря [1] в работе был использован «Русский семантический словарь» [2], явившийся своеобразным практическим руководством при макетировании интеллектуальной системы пополнения семантического словаря.

Основными единицами описания в концептуальном словаре являются не слова, а понятия. При этом источником слов принято считать корпус текстов, написанных на данном языке. Каждое слово вычленяется из текста достаточно просто: это часть текста, ограниченная с двух сторон пробелами (или пробелом и знаком препинания).

В данной работе описаны:

- разработка модели пополнения семантического словаря по объему понятий и/или лексического расширения словарной статьи (понятие и примеры этого понятия формируют словарную статью);
- проблемно-ориентированная подсистема, предназначенная для ввода, обработки и отображения текстов, используемых для формирования ДСМ - гипотез и их последующего применения к необработанным текстам;
- макетирование интеллектуальной системы пополнения семантического словаря;
- реализация интерфейса для работы с системой.

Реализованная система носит демонстрационный характер и не является готовым лингвистическим программным продуктом. В данной разработке предполагалось продемонстрировать возможности ДСМ – метода (его упрощенного аналога) применительно к задаче пополнения семантического словаря. Структура словаря значительно упрощена в аспектах иерархии понятий и их объема.

Особенности подхода

Разработанная система пополнения семантических словарей основана на идеях машинного обучения. Система дополняет работу эксперта при процедуре пополнения объема понятий и лексического расширения словарной статьи, что осуществляется при помощи обучения на примерах - основной процедуре ДСМ - метода. Интеллектуальные системы типа ДСМ основаны на инструментальных средствах, которые могут быть применимы в различных областях науки, где знания слабо формализованы, данные хорошо структурированы, а в БД содержатся как положительные, так и отрицательные примеры некоторых объектов. Однако, наиболее эффективным, как показывает обзор литературы [3-8], оказалось применение ДСМ – метода в области machine learning.

Начальные данные, с которыми работает система, находятся в: базе понятий, корпусе текстов, базе фактов и базе знаний. База понятий состоит из понятий (контрпонятий) и примеров понятий (контрпонятий) (например, понятие - «природная катастрофа», пример этого понятия – «смерч»; контрпонятие для понятия выбирается пользователем из предложенного списка). Корпус текстов (хранилище) содержит предложения естественного

языка, которые нормализуются при помощи встроенного COM-объекта с функцией нормализации и затем используются для поиска новых примеров понятий (контрпонятий) (пример предложения из корпуса текстов - «смерч обрушился на южные селенья», его нормализованный вариант – «смерч обрушиться на южный селение»).

В базу фактов поступают уже нормализованные предложения, из которых исключены примеры понятий (контрпонятий) и заменены спецсимволом. Такие предложения называются +-примерами и –примерами для понятий и контрпонятий соответственно (к примеру, «$ обрушиться на южный селение» - является +-примером, поскольку сформирован из предложения, содержащего пример понятия). +-примеры и –примеры подвергаются процедуре индуктивного обобщения для формирования из них +-гипотез и –гипотез соответственно. В данной работе процедура индуктивного обобщения представлена алгоритмом Норриса, который подробно рассматривается в соответствующем разделе.

База знаний содержит сформированные +-гипотезы и –гипотезы, которые предварительно проверяются процедурой абдуктивного объяснения. В случае успешного завершения этой процедуры гипотезы сохраняются в базе знаний (например, «$ обрушиться на южный селение» - +-гипотеза) и подвергаются процедуре аналогии, т.е. накладываются на нормализованный корпус текстов на предмет совпадения. При совпадении гипотезы (шаблона) и предложения позиция в гипотезе, соответствующая спецсимволу, накладывается на новый пример понятия (контрпонятия). В результате происходит пополнение исходно выбранного пользователем понятия (контрпонятия).

Кроме того, разработанная система проводит верификацию (принятие) и фальсификацию (опровержение) модели, что позволяет выдавать результат – статьи словаря, пополненные или не пополненные новыми примерами понятий. В результате проведения процедур верификации и фальсификация модели возможно применение модели для новых текстов, что существенно облегчает процедуру пополнения семантического словаря.

Участие пользователя в пополнении словаря заключается в следующем: выборе понятий и примеров понятий, активации соответствующих кнопок и адекватной реакции на сообщения программы.

ИНТЕЛЛЕКТУАЛЬНЫЕ СИСТЕМЫ, ОСНОВАННЫЕ НА ДСМ - МЕТОДЕ АВТОМАТИЧЕСКОГО ПОРОЖДЕНИЯ ГИПОТЕЗ(АПГ)

В соответствии с [8] система считается интеллектуальной, если она способна:

Выделить существенное в массиве данных, упорядочить их;

Вывести из исходных данных новые знания посредством рассуждения (дедуктивного, а также правдоподобного индуктивного или абдуктивного);

Оценить результат собственной работы;

Объяснить полученный результат;

Использовать синтез различных средств получения знаний;

Быть адаптированной под конкретную задачу и предметную область;

Аргументировано делать выводы;

Порождать гипотезы;

Обучаться и использовать память.

Согласно [8] в состав интеллектуальной системы входит Решатель (рассуждатель, вычислитель, синтезатор), Информационная среда (база данных и база фактов) и Интерфейс. Рассуждатель – это синтез трех процедур: индукции (порождение гипотез о причинах), аналогии (раскрытие неопределенности) и абдукции (объясняет полученные результаты).

Интеллектуальные системы типа ДСМ основаны на инструментальных средствах ДСМ-метода. Такие системы применимы в областях науки, где знания слабо формализованы, данные хорошо структурированы, а в БД содержатся как положительные, так и отрицательные примеры некоторых свойств объектов.

ДСМ – метод автоматического порождения гипотез был предложен В.К. Финном в конце 70-х годов. Название метода составляют инициалы известного английского философа, логика, историка и социолога Джона Стюарта Милля, чьи “методы здравомыслящего естествоиспытателя” частично формализованы в ДСМ - методе.

ДСМ-метод формализует схему правдоподобного и достоверного вывода, называемую ДСМ - рассуждением. Если говорить о ДСМ - методе как о технологии интеллектуального анализа данных, то смысл ДСМ - рассуждения состоит в том, чтобы на анализе исследуемых данных извлекать знания двух видов:

- знания о структурных причинах исследуемых свойств;
- знания о том, какими свойствами могут обладать исследуемые объекты.

Модель интеллектуальной системы пополнения семантического словаря

При формировании модели использованы следующие основные термины:

Якорь – пример понятия в предложении (понятие «природные катастрофы», пример понятия – «оползень»).

Контекст – здесь граммемы, окружающие якорь в рамках предложения. (Например, в предложении «В результате землетрясения пострадало много людей» контекст – «В результате», «пострадало», «много», «людей»; якорь – «землетрясение»).

Шаблон – контекст без якоря, используемый для формирования гипотезы в ДСМ-методе, а также сами гипотезы – в некотором смысле шаблоны, которые в рамках данной работы накладываются на текст с целью поиска соответствия. (Например, «В результате землетрясения пострадало много людей», шаблон - «В результате пострадало много людей»).

Разработанная система создана при соблюдении следующих условий:

пополняемое понятие задает якоря (a priori заданные термины, соответствующие этому понятию); (понятие «природные катастрофы», пример понятия (якорь) – «оползень»);

- обучающие примеры представлены предложениями (текстами), содержащими якоря;
- обучающая выборка состоит из обучающих примеров;
- получаемая модель (использования якорей в контекстах) аналогична множеству гипотез 1-го рода ДСМ-метода АПГ (автоматического порождения гипотез). Для ее порождения используется процедура пересечения, где встречается якорь;
- индуктивное обобщение примеров строится с помощью алгоритма Норриса;
- в результате индуктивного обобщения порождаются гипотезы;
- гипотезы представлены текстовыми шаблонами;
- в процессе индуктивного обобщения осуществляется фальсификация порожденных гипотез относительно корпуса текстов;
- на основании гипотез (шаблонов) будут находиться новые якоря, встречающиеся внутри порожденных шаблонов. Эти якоря и будут пополнять объем исследуемого понятия. Это, в некотором смысле, есть аналог правил 2-го рода ДСМ-метода;
- процедура абдуктивного объяснения исходных обучающих примеров является средством верификации построенной модели и ее принятия в случае успешной верификации.

В данной работе операция сходства определяется операцией нахождения максимального общего подсписка в списках правого и левого контекстов якорей для + -примеров и - -примеров (например, при осуществлении операции сходства над «На Африка налететь сильный $, вызвать многочисленный жертва среди население» и «$ вызвать многочисленный жертва рогатый скот», получим ««$ вызвать многочисленный жертва»).

Пересечение текстов определяется следующими правилами:

- пересекаются слова текстов, соответствующих (+)-примерам, затем слова нормализуются с использованием встроенного нормализатора COM-объекта;
- вхождение якоря в текст задает разделение предложения на левый и правый контексты;
- **последовательно справа налево сравниваются слова левого контекста одного текста со словами левого контекста другого текста на предмет совпадения. То же для правых контекстов (они пересекаются слева направо). В результате, в текст гипотезы попадут совпадающие наборы слов правой и левой частей (+)-примера относительно якоря.**

Один и тот же текст не может относиться к примерам разных групп (другими словами, не может быть + и – примером одновременно: +-пример «На Африка налететь сильный $», $ - якорь (циклон в данном случае); – пример для этого +-примера (как и сам +-пример) выбирается пользователем «В результат $ разрушить целый селение»). Поскольку + и – примеры являются значимой частью процедур ДСМ-метода, множества + и – примеров не допускают совпадений, с другой стороны, примеры понятий, входящие в каждое из множеств могут обладать синонимией, но это всего лишь частные случаи в общем потоке, обрабатываемом указанным методом.

Для компьютерной обработки текстов на естественном языке с целью их нормализации используется COM-объект [9].

Технология COM – объектно-ориентированная программная спецификация, предложенная Microsoft. Программный объект, созданный согласно спецификации COM называется COM-объектом.

Алгоритм Норриса для рассматриваемого случая

Одна из центральных процедур правдоподобных рассуждений разработанного макета, процедура индуктивного обобщения, проводится в соответствии с модифицированной версией алгоритма Норриса [10].

Реализация этого алгоритма в данной системе использует в качестве выходного параметра строку в файле Hyp_pos_i.txt. Название файла выбрано в соответствие с его содержимым: Hypotheses positive (Hyp_pos_i) – положительные гипотезы, i – номер гипотезы, сохраненной в файле с соответствующим названием; для каждой группы файлов с гипотезами генерируется папка, также отражающая их название. Значения i пробегают число полученных гипотез, имеющих формат (для случая положительных гипотез)

$<i, Y, y_i, + >$,

где i – номер текущей гипотезы,

Y – номера (+)-примеров, образующих (+)-пересечение,

y_i –собственно (+)-пересечение (пересечение +-примеров),

+ - метка (+)-гипотезы в нашем случае выражена в названии директории Hyp_Pos и в названиях файлов с гипотезами, лежащих в данной директории.

Входные параметры – файлы с (+)-примерами, где лежат нормализованные тексты (в каждом файле – отдельный (+)-пример, для каждой группы примеров (+ и -) генерируется своя папка). Составляющие текстов не словоформы, а леммы. Например, после нормализации словоформы «жертвы» получим «жертва».

Алгоритм Норриса разбивает входную строку (текст (+)-примера) на левый и правый контекст, каждый из образовавшихся контекстов разбивает на отдельные слова, выбрасывая при этом пробелы. Делается это для того, чтобы можно было с максимальной точностью, т.е. пословно, сравнивать контексты (подпроцедура сходства процедуры Норриса): набор слов левого контекста 1-го текста сравнивается с набором слов левого контекста 2-го текста (для случая n = 2) и аналогично с правыми контекстами. Порядок слов в предложении заведомо фиксированный, т.е. при наличии в подвергаемых процедуре сходства предложениях двух одинаковых наборов слов мы получим их пустое пересечение. При этом набор совместных примеров понятий будет пополнять формирование основного понятия, то есть, понятие «лес» как экосистема будет пополняться соответствующими примерами из, например, «в лесу водятся грибы», «в лесу водятся зайцы», «в лесу лежат поваленные деревья» и т.д. Но релевантность грибов, зайцев и поваленных деревьев по отношению к понятию «лес», естественно, не подразумевает их синонимичности.

После проведения процедуры сходства в рамках алгоритма Норриса получаем пересечение (+)-примеров, каждую часть контекстов которых записываем в строку (текст) (+)-гипотезы, добавляя между словами пробелы. Происходит это также при фиксированном порядке, заданном изначально в текстах (+)-примеров. Количество файлов с положительными гипотезами варьируется в зависимости от количества пересечений исходных (+)-примеров (одиночные пересечения, т.е. пример, пересекающийся сам с собой, двойные примеры, т.е. пересекаются два разных примера и т.д.). Алгоритм Норриса, таким образом, порождает максимальный набор неповторяющихся пересечений примеров по принципу «все со всеми» (неповторяемость достигается за счет двух условий алгоритма: относительной каноничности и каноничности).

Пример работы системы пополнения семантического словаря и алгоритма Норриса как одного из центральных алгоритмов системы (для +-примеров):

Например, в начале работы системы пользователь хочет пополнить объем понятия «природные катастрофы». Для этого он выбирает некоторое число примеров этого понятия (якорей): тайфун, оползень и смерч. Система считывает выбранные якоря и ищет в корпусе текстов (хранилище) предложения, содержащие их. Если такие предложения не найдены, то система сообщает пользователю об ошибке. В случае если эти предложения обнаружены, система передает их для обработки морфологическому нормализатору: «На Африку налетел сильный тайфун, вызвавший многочисленные жертвы среди населения», «В результате оползня разрушено целое селение», «Смерч вызвал многочисленные жертвы среди рогатого скота». Нормализованные предложения: «На Африка налететь сильный $, вызвать многочисленный жертва среди население», «В результат $ разрушить целое селение», «$ вызвать многочисленный жертва среди рогатый скот».

Нормализованные предложения (+-примеры) записываются в базу фактов. Затем они обрабатываются процедурой индуктивного обобщения – алгоритмом Норриса:

База фактов ((+)/(-)примеры) /здесь +-примеры, содержимое папки _plus/:

plus_1.txt: На Африка налететь сильный $, вызвать многочисленный жертва среди население

plus_2.txt: В результат $ разрушить целое селение

plus_3.txt: $ вызвать многочисленный жертва среди рогатый скот

База знаний ((+)/(-) гипотезы) /здесь +-гипотезы, содержимое папки Hyp_pos/:

База знаний содержит +-гипотезы, сгенерированные в результате пересечения +-примеров в соответствие с алгоритмом Норриса (для данной работы) и встроенной в него процедурой сходства.

1.txt: На Африка налететь сильный $, вызвать многочисленный жертва среди население – 1 пример, взятый целиком (не является подмножеством других множеств и не входит ни в один пример целиком) – гипотеза 1

2.txt: В результат $ разрушить целое селение – 2 пример целиком; – гипотеза 2

3.txt: $ вызвать многочисленный жертва среди рогатый скот – 3 пример целиком; – гипотеза 3

4.txt: 0 (то есть файл пуст)– пустое пересечение 1 и 2 примеров; – гипотеза 4

5.txt: $ вызвать многочисленный жертва – пересечение 1 и 3 примеров; – гипотеза 5

После работы алгоритма Норриса сгенерированы гипотезы, которые передаются для проверки процедуре абдукции (у каждой гипотезы должен быть пример, из которого она образована, иначе абдукция неверна и дальнейшая работа с этими гипотезами невозможна). В случае успешного завершения абдукции гипотезы из базы знаний передаются процедуре аналогии, которая и осуществляет их наложение на текст, т.е. на предложения, которые могут содержать потенциальные якоря для пополняемого понятия. Например, «Потоп вызвал многочисленные жертвы». Это предложение совпадает с гипотезой № 5. Значит, «потоп», совпадающий с $ в предложении гипотезы, и есть искомый пример понятия «природные катастрофы» и будет успешно добавлен в базу понятий.

Результаты

В результате разработки создан программный макет интеллектуальной системы, в которой были реализованы процедуры правдоподобных рассуждений для пополнения семантического словаря.

В процессе создания и применения макета:

1. разработана модель пополнения семантического словаря по объему понятий и/или лексического расширения словарной статьи;
2. создана проблемно-ориентированная подсистема, предназначенная для ввода, обработки и отображения текстов, используемых для формирования ДСМ-гипотез и их последующего применения к необработанным текстам.

Кроме того, в процессе реализации системы решены следующие задачи:

- для ввода, обработки и отображения текстов, используемых для формирования гипотез и их последующего применения к необработанным текстам, реализована операция сходства, порождающая шаблоны;
- осуществлена реализация правдоподобных рассуждений (аналог рассуждений ДСМ - метода);
- разработана проблемно - ориентированная подсистема нормализации текстов с применением технологии COM;
- реализован алгоритм Норриса для проведения процедуры индуктивного обобщения, подтвердивший свою эффективность в рамках поставленной задачи.

Список литературы

Рубашкин В.Ш., Лахути Д.Г. *Семантический (концептуальный) словарь для информационных технологий. Ч. 1 // НТИ. Сер. 2.—1998.— № 1.— С.19-24.*

Караулов Ю.Н., Молчанов В.И., Афанасьев В.А., Михалев Н.В. *Русский семантический словарь. Опыт автоматического построения тезауруса: от понятия к слову // под ред. Бархударова С.Г. М.: Наука, 1982.*

Финн В.К. *О роли машинного обучения в интеллектуальных системах // НТИ. Сер. 2.— 1999 .— № 12 .— С.1-3.*

Финн В.К. *О некоторых проблемах логики и методологии интеллектуальных систем // НТИ. Сер. 2.— 1999 .— № 1-2 .— С. 2.*

Deitterich T., Michalski R. *A comparative review of selected methods for learning from examples //Machine learning: an artificial intelligence approach / Eds. R.S. Michalski et al.— Berlin: Springer, 1984.*

Michalski R. T*heory and Methodology of Inductive Learning // Artificial Intelligence — 1983.—Vol. 20, № 2.*

Cardie C., Mooney R.J. *Guest Editors' Introduction: Machine Learning and Natural Language //Machine Learning — 1999 .— Vol. 34 .— P. 5-9.*

Финн В.К. *О базах знаний интеллектуальных систем типа ДСМ // II Всесоюзная конференция «Искусственный интеллект-90», Минск, 1990 – с.180-182.*

Сокирко А.Г., Панкратов Д.В.*Проект ДИАЛИНГ, COM-объект Goldrml. (www.aot.ru).*

Финн В.К., Виноградов Д.В., Кожунова О.С. *Интеллектуальная система пополнения семантических словарей // Программные продукты и системы, № 2, 2006.*

Харари Ф. *Теория графов. М.: Мир, 1973.*