Использование онтологии SUMO для пополнения онтологии InTez

Фадеева М.В., Чижик А.В.
СПбГУ, факультет искусств, кафедра информационных систем в искусстве и гуманитарных науках
fadeevamf@gmail.com, afrancuzova@mail.ru

В статье описана специфика онтологии SUMO, обоснован выбор между онтологиями DOLCE, Cyc и SUMO для пополнения InTez. Описана специфика онтологии InTez. Обозначена проблема при работе с онлайн-версией SUMO. Описан алгоритм программы для извлечения подклассов заданного класса SUMO из документа, содержащего описание онтологии на языке KIF.

Введение

На сегодняшний день в различных областях человеческой деятельности накоплено огромное количество информации, имеющей практическую ценность для развития науки и производства. Информация объединяется в базы знаний по соответствующим областям. Однако базы знаний имеют разную организацию, информация в них представлена в разных форматах, что затрудняет ее совместное использование несколькими системами. Решение этой проблемы связано с задачей представления знаний, в настоящее время находящейся в центре внимания многих исследователей. Сегодня в инженерии знаний существует ряд средств представления знаний, одним из которых является онтология.

Онтология — это формализованное представление основных понятий определенной области и связей между ними. В настоящее время наиболее активно используются *онтологии предметной области*, описывающие конкретную область знания. Кроме того, идет работа над построением и развитием онтологий, содержащих *общие знания* — онтологий *верхнего уровня* (top-level ontology).

Можно выделить три наиболее развитые онтологии верхнего уровня:

- SUMO;
- Cyc;
- DOLCE.

Постановка задачи

В настоящее время на кафедре информационных систем в искусстве и гуманитарных науках СПбГУ ведется работа над созданием онтологической системы InTez. С 2010 года некоторые компоненты системы находятся в открытом доступе в сети Интернет. При реализации проекта учитывается опыт уже существующих больших онтологий верхнего уровня. Было принято решение использовать онтологию SUMO для пополнения онтологии InTez.

Онтология SUMO была выбрана для пополнения InTez как наибольшая из существующих на сегодняшний день открытая онтология верхнего уровня. Следует отметить, что SUMO значительно полнее, чем DOLCE, и в отличие от Cyc она открыта для доступа, поэтому работа с ней наиболее проста и доступна.

Обзор SUMO

SUMO (Suggested Upper Merged Ontology) принадлежит IEEE (Institute of Electrical and Electronics Engineers). Онтология используется для исследований и создания приложений в таких областях, как извлечение информации, лингвистика, логический вывод.

В основе SUMO лежит принцип подключения необходимых отраслевых онтологий. SUMO — онтология верхнего уровня, MILO — онтология среднего уровня, а к ним подключаются онтологии предметных областей (финансы, страны и регионы, медиа, транспорт и так далее). При подключенных онтологиях предметных областей SUMO содержит около 20000 терминов и 70000 аксиом. На сегодняшний день онтология SUMO полностью синхронизирована с системой WordNet (каждому понятию SUMO поставлен в соответствие синонимический ряд WordNet). SUMO реализована на диалекте языка KIF, KIF-SUO.

Обзор InTez

InTez — это универсальная словарная лексически интерпретированная (поддерживающая связь с лексической системой естественного языка) онтология, ориентированная на достаточно детальное описание концептов и связей между ними. Онтология основана на модели знаний, формулируемой средствами *логического языка InfoL*.

Отличительной особенностью онтологии является способ представления базовой иерархии классов в формате *«Дерево признаков»*. Для каждого простого класса в качестве непосредственного хозяина в иерархии указывается классификационный признак, определяющий, по какому основанию выделен данный подкласс из вышестоящего класса. Классификационный признак подчинен в иерархии классов вышестоящему классу, разбиение которого образуют нижестоящие — подчиненные одному и тому же классификационному признаку. Вершинам классов могут быть подчинены не только классификационные признаки, но и признаки других типов.

Таким образом, в дереве признаков отображаются базовые (элементарные) классы объектов и все признаки, определенные для объектов.

Хранение концептов, их характеристик, значений, связей между концептами, примеров словосочетаний и так далее осуществляется в таблицах базы Microsoft Access в локальной версии, и в MS SQL Server – в сетевой версии.

Работа по пополнению InTez

При работе над наполнением InTez терминами из SUMO планируется заполнять три таблицы InTez:

- *TEZ0* — список всех концептов и их унарные характеристики, состоит из полей:
    - *Dscr — идентификатор концепта (число);*
    - *SC — семантическая категория;*
    - *ST — семантический тип;*
    - *LV, B1, B2, B3, B4 — поля для представления словарных признаков, имеющих переменную семантику;*
    - *KEW0 — стандартный термин;*
    - COMMENT — *комментарий;*
    - DN — *предметная область;*

- DateChange — *дата изменения (заполняется автоматически);*
- Who Changed — *автор изменения (заполняется автоматически);*

- *TzAd — таблица связей между концептами. Она имеет следующие поля:*
  - *DSCR1* — идентификатор первого концепта;
  - *TYP* — тип связи;
  - UVS1 — указатель вида связи между концептами;
  - UVS2 — указатель вида связи между концептами;
  - *ADSR* — идентификатор второго концепта;
- *WordDscr* — список всех слов, определяющих возможное лексическое представление концепта. Таблица определяет связь между концептуальной системой онтологии и лексической системой естественного языка. «Фиктивные» (искусственно построенные термины, необходимые для системной организации концептов) завершаются звездочкой. Принято решение, что при переносе концептов из SUMO в InTez термины будут отмечаться звездочкой, так как не все термины SUMO совпадают с нормальной формой слова, а к некоторым из них невозможно подобрать подходящее словесное обозначение. Например, многие классы SUMO имеют название (одно слово), составленное из двух слов, где второе слово начинается с заглавной буквы (*PublicSchool, SocialInteraction* и т.п.).Таблица состоит из следующих полей:

  - *Word* — словоформа;
  - *Word_Id* — идентификатор пары {словоформа, концепт};
  - *Dscr* — идентификатор интерпретируемого концепта;
  - *Comment* – комментарий;
  - *WLemma* – лемма;
  - *PS_Id* — идентификатор части речи слова;
  - *CollocId* — идентификатор словосочетания (если концепт интерпретирует одно слово, = 0);
  - *CollocFth* — ссылка на идентификатор слова — синтаксического хозяина в словосочетании (для одиночного слова и синтаксического хозяина = 0);
  - *LangId* — идентификатор языка;
  - *WdStyleId* — идентификатор стилевой пометы;
  - *GrFeatId*— идентификатор дополнительной грамматической характеристики слова;
  - *StdTerm* — вхождение в стандартный термин.

При импорте пар концептов «родитель-потомок» из текстового файла SUMO в таблицы MS Access будут созданы следующие записи:

1. концепт-родитель и концепт-потомок будут добавлены в таблицу *TEZ0 с* созданием соответствующих идентификаторов;

2. концепт-родитель и концепт-потомок будут добавлены в таблицу *WordDscr*. в таблице *TEZ0* будет создан новый концепт – основание деления (признак, по которому выделяется заданный подкласс-потомок);

3. после создания соответствующих идентификаторов концептов будет заполняться таблица связей, *TZAD*, где концепт-родитель и концепт-потомок будут связаны через концепт – основание деления.

В редакторе InTez для представления таксономии понятий используется графический интерфейс TreeView. Планируется использовать функции этого

интерфейса для наглядного сопоставления онтологий SUMO и InTez и импорта концептов SUMO в InTez.

Официальный сайт SUMO дает возможность скачивания онтологии и текстового представления онтологии на языке KIF. Также существует онлайн-версия системы - браузер Sigma Knowledge Engineering Environment. Однако было выяснено, что онлайн-версия системы не всегда корректно отображает подклассы заданного класса. Для более эффективной работы с системой была создана программа на языке JAVA, извлекающая подклассы заданного класса SUMO из документа, содержащего описание онтологии на языке KIF. В основе работы программы лежит использование регулярных выражений, то есть пакет import java.util.regex. Пользовательский интерфейс реализован с помощью пакетов javax.awt.*, javax.swing.* и info.clearthought.layout.TableLayout. Для загрузки исходных текстовых файлов в программу использован пакет import java.io.*.

На данном этапе разработки программа работает по следующему алгоритму:

1. запуск программы;
2. пользователь выбирает файл для загрузки нажатием кнопки «open» — нажатие кнопки запускает класс, позволяющий выбрать файл из заданной папки (опция введена для возможности выбирать, с файлом какой онтологии в данный момент нужно работать — SUMO или MILO);
3. выбранный файл загружается в окно JTextPane;
4. пользователь просматривает файл и вводит в поле JTextArea название того концепта, для которого хочет получить дочерние узлы;
5. нажатие на кнопку «find» запускает основную программу — поиск данного концепта в файле и вывод всех его подклассов и экземпляров в окно JTextArea;
6. полученные подклассы можно сохранить в файл subclass.txt нажатием на кнопку «save».

Заключение

В дальнейшем планируется продолжить разработку программы и использовать ее для автоматического наполнения баз данных InTez терминами SUMO. На данный момент планируется пополнение InTez терминами из области «мир социального».

Дальнейшее развитие онтологии InTez предполагает создание средств для удаленного пополнения и редактирования, разработку средств для энциклопедического импорта (пополнения онтологии из машиночитаемых словарей), подключение библиотеки функций, обеспечивающей доступ к онтологии для различных интеллектуальных информационных технологий.

Список литературы:

1) Официальный сайт Suggested Upper Merged Ontology: http://ontologyportal.org/
2) Официальный сайт проекта InTez: http://inttez.ru/
3) Пивоварова Л.М., Рубашкин В.Ш. Компоненты онтологических систем и их реализация в современных проектах // Труды конференции «Интернет и современное общество-2007», СПб, 2007.
4) Шилдт, Г. Полный Справочник по Java, 7-ое издание. М.: "Вильямс". 2007. - 1040с. – ISBN 978-5-8459-1168-1