ПРОЕКТ UWN: ПЛАТФОРМА ДЛЯ УСКОРЕННОЙ РАЗРАБОТКИ ЛИНГВИСТИЧЕСКИХ ПРИЛОЖЕНИЙ

Никоненко А. А.
Киевский национальный университет им. Т. Шевченко
Факультет кибернетики
г. Киев-680, 03680, Украина
e-mail: andrey.nikonenko@gmail.com

Аннотация — В рамках проекта UWN создана технологическая платформа, обеспечивающая совместную работу онтологических баз и лингвистической логики. Данный подход позволил создать ряд лингвистических модулей, содержащих базовый функционал по работе с онтологиями и набор программных пакетов, содержащих серверную логику для прикладных приложений. Повторное использование элементов ранее разработанных программ позволяет значительно ускорить и облегчить создание новых лингвистических приложений. Интеграция серверной логики с данными позволяет добиться высокой эффективности систем описанного вида, а клиент-серверная архитектура допускает создание множества разнотипных (толстых/тонких) клиентов для каждого приложения.

I. Введение

На сегодняшний день в задачах обработки естественного языка существует четыре основных подхода: статистический, морфологический, синтаксический и семантический (детальнее в [1]). Статистические методы применяются в основном в системах, обрабатывающих большие массивы информации (поисковые машины, статические переводчики и т. д.); морфологические методы обычно выступают в роли вспомогательных и используются совместно с другими методами; синтаксические методы базируются на создании синтаксических анализаторов и позволяют выделять в предложениях значимые элементы и связанные с ними слова; семантические методы нацелены на определение смысловых элементов текста. Семантические методы базируются на лексических ресурсах, содержащих знание о структуре, взаимосвязи и значении элементов языка.

Созданию такого лексического ресурса (онтологии) для трех языков (английский, украинский, русский) и посвящен проект UWN [2]. UWN предоставляет свободный доступ к своей онтологии и является онлайн-проектом, пополняемым и улучшаемым волонтерами. Технология совместного доступа в проекте UWN (детально описана в [3]) разработана на базе идеологии проектов Wikipedia [4] и ConceptNet [5].

II. Архитектура системы

Все приложения в проекте строятся на основе клиент-серверной архитектуры. В качестве сервера выступает база данных, которая отвечает за хранение данных и выполнение серверных функций, описанных на встроенном языке. Расположение данных и логики, осуществляющей их обработку в одном месте позволяет существенно сократить время выполнения ресурсоемких операций. Расположение основных элементов системы изображено на рисунке 1.

Все приложения используют единообразную модель доступа к системе:

1) На первом шаге для установления соединения используется технологический пользователь ***ua_guest.***

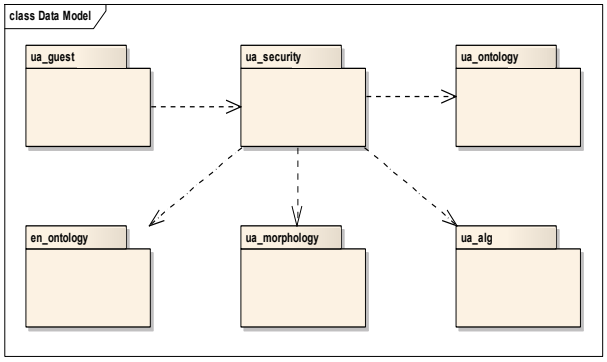

Рис. 1. Схема взаимодействия основных элементов UWN.

Fig. 1. Scheme of interaction of the UWN basic elements

2) На втором шаге приложение вызывает процедуру регистрации в системе, которая проводит авторизацию приложения с последующей выдачей ему прав, определенным профилем безопасности для данного класса систем. Данные процессы происходят в схеме ***ua_security***.

3) На третьем шаге приложение получает доступ к серверной логике в соответствии с предоставленными ему правами. На этом этапе возможен вызов функций из схем онтологий (***ua_ontology, en_ontology***) или схем с логикой (***ua_alg***).

4) На четвертом шаге начинается непосредственная работа приложения с системой, здесь допустим как вызов отдельных лингвистических функций (например, определения семантической близости двух слов), так и использование готовых пакетов логики, предназначенных для реализации серверной части приложений (например, так в UWN организованы системные утилиты, детально изложено в [6]).

UWN не предоставляет прямого доступа к данным, вместо этого, внешним системам предоставляется доступ к API специально разработанной системы, которая управляет данными. Создание такой системы позволяет обеспечить корректное функционирование большого количества многопользовательских разнотипных систем без возникновения коллизий. Основное ядро механизма управления доступами к данным расположено в схеме ***ua_security***, вспомогательные элементы разнесены по управляемым схемам. Все данные в системе находятся под контролем средств обеспечения целостности и мониторинга изменений.

III. Проектирование лингвистических приложений

Создание приложений в UWN происходит следующим путем:

1) Сначала определяется тип создаваемого приложения: отдельная система, подсистема или приложение. Отдельная система — это приложение,

2011 21st Int. Crimean Conference "Microwave & Telecommunication Technology" (CriMiCo'2011). 12-16 September, Sevastopol, Crimea, Ukraine
 ISBN: 978-966-335-351-7. IEEE Catalog Number: CFP11788

полностью решающее определенную лингвистическую задачу и развернутое на базе UWN. Примерами отдельных систем могут служить приложения, проводящие реферирование текстов, определяющие тематику, производящие улучшение переводов и т. д. Как правило, отдельные системы размещаются в специально предназначенной для них схеме либо, при высокой сложности и наличии дополнительных источников данных, в отдельной схеме. Подсистема — это система, решающая ограниченную задачу или часть более общей задачи. Подсистема не может быть предоставлена в работу пользователю, т. к. не несет достаточный для решения прикладных задач функционал. Однако подсистемы могут успешно использоваться в качестве частей отдельных систем и приложений. Примерами подсистем являются: подсистема проверки орфографии, подсистема поиска вариантов исправления слов с орфографическими ошибками, подсистема выдачи информации о синсетах и т. д. Под приложением в данном случае подразумевается особый вид отдельной системы, который не решает определенной лингвистической задачи, но предназначен для обслуживания взаимодействия пользователей с онтологией UWN. Примерами подсистем являются онтокорректоры и онторедакторы. Такие приложения ориентированы на разделение уровней доступа пользователей и содержат большой набор разнообразных функций, на основе комбинирования которых может быть создано большое число различных прикладных систем. Как правило, приложения располагаются в схемах с данными, поскольку ориентированы на интенсивную работу с ними.

2) На втором этапе определяются задачи разрабатываемой системы и необходимые для их выполнения программные модули и средства.

3) Далее определяется наличие уже реализованных элементов будущей системы в UWN в рамках других систем/подсистем/приложений. Производится конфигурирование доступов.

4) Решение о размещении отсутствующих на данный момент элементов системы принимается на базе анализа их сходства и простоты реализации в качестве частей уже существующих подсистем. Необходимые новые элементы, схожие с которыми отсутствуют, группируются в новые подсистемы.

5) Последним этапом формируется ядро системы, которое отвечает за синхронизацию и внутреннюю логику работы системы. В случае реализации простого функционала допустимо создание систем без ядра (например, лингвистических приложений, осуществляющих простую трехшаговую обработку текста: загрузка данных →обработка → результат).

IV. Заключение

Проект UWN представляет собой удобную площадку для разработки новых лингвистических приложений. Созданные в проекте механизмы позволяют разграничивать доступы к данным и логике, поддерживают разные уровни доступа пользователей и прикладных систем. Также в UWN доступно повторное использование готовых элементов кода и целых подсистем для решения различных прикладных лингвистических задач. Использование UWN в качестве научно-исследовательской площадки учеными открывает разработчикам доступ к уже реализованному и постоянно растущему множеству современных лингвистических методов.

V. Список литературы

[1] *Никоненко А. О.* Семантичний аналіз природномовних текстів: підходи та засоби // Міжнародна науково-технічна конференція «ОБЧИСЛЮВАЛЬНИЙ ІНТЕЛЕКТ - 2011 (результати, проблеми, перспективи) ОІ — 2011»: тези конференції. (Черкаси, 10—13 травня 2011 р.). 2011.

[2] *Сайт* проекта UWN. URL: http://lingvoworks.org.ua (дата обращения: 10.05.2011).

[3] *Никоненко А. О.* Проект UWN: Методи спільного редагування онлайн онтологій // Міжнародна науково-практична конференція «Інформаційні технології та комп'ютерна інженерія»: тези конференції. (Харків, 26—27 травня 2011 р.). 2011.

[4] *Wilkinson D. M., Huberman B. A.* Cooperation and quality in wikipedia // Proceedings of the 2007 international symposium on Wikis. WikiSym '07. (New York, USA, 2007). NY: ACM, 2007. P. 157—164.

[5] *Liu H., Singh P.* Conceptnet — a practical commonsense reasoning tool-kit // BT Technology Journal 2004. No. 22(4). P. 211—226.

[6] *Никоненко А. О.* Проект UWN: Досвід створення універсальної онлайн онтології української мови // Міжнародна наукова конференція ISDMCI'2011 «Интеллектуальные системы принятия решений и проблемы вычислительного интеллекта»: тези конференції. (Євпаторія, 16—20 травня 2011 р.). 2011.

UWN PROJECT: RAPID LINGUISTIC APPLICATIONS DEVELOPMENT PLATFORM

Nykonenko A. A.
National Taras Shevchenko University of Kyiv,
Faculty of Cybernetics
Kiev-680, 03680, Ukraine
e-mail: andrey.nikonenko@gmail.com

Abstract — A technological platform that enables collaboration of ontological bases and linguistic logic has been created in UWN project. A number of linguistics modules with ontology functions and a set of software packages with server-side logic based on the platform were created.

I. Introduction

There are four main approaches in the natural language processing tasks: statistical, morphological, syntactic and semantic (see details in [1]). Semantic methods are based on lexical resources that contain knowledge about the structure, relationship and meaning of language elements. The UWN project [2] concerns the creation of such lexical resource (ontology) for three languages (English, Ukrainian and Russian).

II, III. Main part

All applications in the project are based on client-server architecture. The database works as a server, which is responsible for data storage and execution of server-side functions. The location of the system main elements is shown in Figure 1.

UWN does not provide direct access to the data; instead of it, the external systems have access to the specifically designed data management system API. Such system availability can ensure the correct functioning of a large number of multiuser heterogeneous systems without causing collisions.

IV. Conclusion

UWN project represents a convenient platform for the new linguistic applications development. The mechanisms created in the project allow differentiating the access to the data and logic, supporting different access levels for users and applications. Also UWN allow reusing code elements and whole linguistic subsystems. Using UWN as scientific research platforms offers developers access to the set of modern linguistic methods.

2011 21st Int. Crimean Conference "Microwave & Telecommunication Technology" (CriMiCo'2011). 12-16 September, Sevastopol, Crimea, Ukraine