МЕТОД И СИСТЕМА ИЗВЛЕЧЕНИЯ СЕМАНТИЧЕСКИХ ОТНОШЕНИЙ ИЗ СТАТЕЙ ВИКИПЕДИИ НА ОСНОВЕ КОМПОНЕНТНОГО АНАЛИЗА[57]

А. И. Панченко (Лован, Бельгия)
Catholic University of Louvain
panchenko.alexander@gmail.com
Ю. Н. Филиппович, С. А. Адейкин,
А. В. Романов, П. В. Романов (Москва)
Московский государственный технический университет
им. Н. Э. Баумана
y_philippovich@mail.ru

Мы представляем методы извлечения семантических отношений из статей Википедии на основе компонентного анализа [Филиппович, Прохоров, 2001], алгоритмов ближайших соседей KNN и MKNN (MutualKNN) и двух метрик семантической близости. Мы анализируем методы и приводим результаты их использования. Точность извлечения с помощью одного из методов достигает 83%. Кроме этого, мы представляем систему с открытым исходным кодом, которая реализует описанные методы.

Под семантическими отношениями в рамках нашей работы понимаются отношения синонимии, гиперонимии и ко-гиперонимии (слова имеющие общий гипероним). Мы представляем метод, который находит пары слов связанных этими отношениями, но не указывает тип найденной связи. Семантические отношения фиксируются в различного типа лингвистических ресурсах, к числу которых относятся, прежде всего, тезаурусы и идеографические словари, терминологические классификаторы и кодификаторы, словари синонимов, а также конкретные компьютерные онтологии.

Семантические отношения успешно применяются в задачах *автоматической обработки текста*, *автоматического перевода, поиска информации* при разрешении омонимии, расширении / сжатии поискового запроса, классификации текстовых документов, создании вопросно-ответных систем и т. п. Однако существующие лингвистические ресурсы часто недоступны или недостаточны для конкретного приложения или предметной области. Ручное создание и обновление ресурсов – дорогостоящий и трудоемкий процесс. Это делает актуальной задачу разработки методов автоматического извлечения семантических отношений. Цель

[57] Работа выполнена в рамках гранта РГНФ №12-04-12039в

метода извлечения семантических отношений – найти для множества входных слов *C* пары семантически связанных слов *R*.

В качестве материала для верификации разрабатываемых методов автоматического извлечения отношений из текста используются статьи Википедии (www.wikipedia.org). Данный ресурс привлекателен для анализа в силу того, что он достаточно полно покрывает основные предметные области и основные языки, а также постоянно пополняется пользователями. Сенлар [Senellart and Blondel, 2008] приводит обширный обзор методов извлечения семантических отношений. Струбе и Понзетто [Strube and Ponzetto, 2006], Габрилович и Маркович [Gabrilovich and Markovich, 2007], Зесч, Мюллер и Гуревич [Zech, Muller, and Gurevich, 2008] разработали методы нахождения семантически связанных слов и отношений на основе Википедии.

Исходными данными для извлечения отношений является множество статей Википедии, для каждого из входных слов. Мы используем данные, доступные на DBPedia.org, для того чтобы построить множество «определений» английских слов. Для каждого входного слова мы строим множество пар <слово; определение>, где «слово» – это заголовочное слово статьи Википедии, а «определение» – предварительно обработанный текст первого параграфа этой статьи. Обработка состояла в следующем: во-первых, из текста была удалена разметка и специальные символы; во-вторых, был произведен морфологический анализ с помощью анализатора TreeTagger, в результате чего каждое слово было представлено в виде тройки <токен#ЧАСТЬ-РЕЧИ#лемма>.

Для наших экспериментов мы подготовили два набора данных – малый, содержащий определения 775 слов (824 Кб), и большой, содержащий определения 327167 слов (237Мб) доступные по адресу http://cental.fltr.ucl.ac.be/team/~panchenko/def/.

Разработанные методы извлечения семантических отношений основаны на компонентном анализе, принцип которого заключается в том, что семантически близкие слова имеют подобные определения. Предложенные алгоритмы используют одну из двух метрик подобия определений – количество общих слов или косинус угла между векторами определений. В качестве входных данных алгоритмы извлечения семантических отношений принимают множество слов *C* между которыми необходимо вычислить отношения по их определениям *D*. Допустим, что на вход алгоритма поступило 5 слов, например, *C*={*alligator, animal, building, house, telephone*}, тогда задача алгоритма – распознать множество семантических отношений *R* ={<*alligator, animal*>, <*building, house*>} из всех 10 возможных пар слов. Первый алгоритм вычисляет семантические отношения с помощью метода ближайших соседей (KNN), второй с помощью

метода взаимных ближайших соседей (MKNN). Метапараметр алгоритмов – количество ближайших соседей *k*.

Работа алгоритмов состоит в следующем. Сначала вычисляется мера семантической близости всех возможных пар определений. На основе вычисленного значения заполняется массив наиболее близких слов для каждого определения. При этом мы поддерживаем число элементов этого массива равным k (количеству ближайших соседей) – это позволяет значительно сократить объем используемой памяти без потери информации о связности слов. После заполнения массива наиболее близких слов каждого определения, для получения результирующего набора отношений *R* необходимо в методе KNN заполнить выходное множество, а для метода MKNN – дополнительно проверить для каждого определения, входит ли оно в массив наиболее близких слов свой пары, и если входит, добавить в результирующее множество. Сложность разработанных алгоритмов при фиксированном количестве ближайших соседей *k* пропорциональна количеству поданных на вход слов |*C*|. Временная сложность равна O(|C^2|), пространственная сложность равна O(|*C*|).

Программное решение реализовано в виде консольного приложения на языке C++ и доступно для платформ Windows и Unix. Основные функции программы заключаются в: (1) загрузке файлов стоп-слов и слов, между которыми нужно найти отношения *C*; (2) загрузке с учетом стоп-слов и слов *C* файла дефиниций *D;* (3) вычислении семантической близости; (4) формировании списка наиболее близких слов *R*. Система Serelex имеет открытый исходный код, доступный на условиях лицензии GPLv3 по адресу https://github.com/jgc128/Serelex.

Мы исследовали работу алгоритмов KNN и MKNN с двумя описанными выше метриками близости и различными значениями количества ближайших соседей *k*. Полученные результаты свидетельствуют о практически линейном росте количества найденных отношений в зависимости от параметра *k* для обоих алгоритмов. При этом, количество найденных отношений мало зависит от используемой метрики подобия, а алгоритм KNN извлекает больше отношений чем MKNN. Мы также провели оценку точности работы алгоритмов KNN и MKNN для $k = 2$ на множестве из 775 определений. Для этого мы разметили вручную файлы с извлеченными отношениями и вычислили точность извлечения как количество верных отношений к общему количеству извлеченных отношений.

В таблице представлены полученные результаты экспериментальной оценки точности извлечения семантических отношений для k=2 из 775 слов.

Алгоритм	Мера подобия	Извлечено	Правильных	Точность
KNN	Косинус угла между определениями	1548	1167	0.754
	Количество общих слов	1546	1176	0.761
MutualKNN	Косинус угла между определениями	652	499	0.763
	Количество общих слов	724	603	**0.833**

Примеры извлеченных отношений между множеством из 775 слов с помощью алгоритма MKNN (*k=2*) и количества общих слов в качестве метрики подобия: *<acacia, pine>, <aircraft, rocket>, <alcohol, carbohydrate>, <alligator, coconut>, <altar, sacristy>, ... , <watercraft, boat>, <watermelon, berry>, <weapon, warship>, <wolf, coyote>,<wood, paper>*.

ТРАНСФОРМАЦИЯ ПРЕЦЕДЕНТНЫХ ТЕКСТОВ КАК ОТРАЖЕНИЕ ЯЗЫКОВОЙ КАРТИНЫ МИРА (НА МАТЕРИАЛЕ ГАЗЕТНЫХ ЗАГОЛОВКОВ)

Е. Б. Пономаренко (Москва)
Российский университет дружбы народов
ponomar_elena@mail.ru

Использование прецедентных текстов в последнее время получило широкое распространение в средствах массовой информации. Текст, в котором присутствует хотя бы один из прецедентных феноменов, изначально экспрессивен, так как «порожденная двуплановость или многоплановость, “включенный текст” служит целям языковой игры разного рода: способствует поэтизации текста, создает поэтический намек, подтекст, рождает загадку, создает ироническое, саркастическое, гротескное, трагическое или иное звучание, способствует иерархизации смысла, – придает бытовой фразе смысл иносказания – политического, поэтического, философского или какого-либо иного, иногда просто рождает непритязательную шутку» [Земская, 2004, с. 563].