ЛАБОРАТОРНАЯ РАБОТА 5

Представление знаний на основе семантической сети

1. Цель работы

Целью работы является изучение способов представления семантической информации для реализации анализа связного текста с использованием неформальных стратегий.

2. Задание на лабораторную работу

Реализовать систему автоматического реферирования статей с входным параметром – коэффициентом сжатия исходной статьи (коэффициент задается в процентах).

Статья для реферирования должна представлять собой текст не менее чем из двадцати распространенных предложений на русском языке. В начале текста должен быть заголовок, отражающий смысл содержания всей статьи в целом.

Система автоматического реферирования должна иметь следующую структуру:

- текстовый редактор для ввода статьи и отображения полученного реферата по этой статье (результат работы системы). Необходимо предусмотреть возможность загрузки текста из файла;
- блок автоматического анализа содержания текста с помощью семантических сетей, который имеет следующую подструктуру:
 - а) модуль построения семантической сети для заголовка;
 - б) модуль построения списка ключевых слов на основе частотной встречаемости слов в тексте. При этом объем списка ключевых слов находится в зависимости от коэффициента сжатия текста (коэффициент сжатия задается в процентах как параметр системы реферирования);
 - в) модуль, непосредственно выполняющий анализ содержимого предложений статьи с помощью семантических сетей.

Связь между предложениями устанавливается при нахождении в них хотя бы одного общего слова или нескольких слов, связанных друг с другом отношениями «род–вид» или «целое–часть». Вначале строится семантическая сеть для первого предложения статьи и сравнивается с семантической сетью заголовка. Далее, при обнаружении семантической связи вышеуказанным способом, необходимо дополнить семантическую сеть заголовка информацией, содержащейся в анализируемом предложении (для этого нужно использовать [операции над семантическими сетями](#) [1, с. 99–119]), а также включить анализируемое предложение в генерируемый файл-реферат. При отсутствии семантической связи предложение в файл-реферат не включается и осуществляется переход к следующему предложению текста.

Процесс сравнения полученной семантической сети с сетью следующего предложения по тексту рекурсивно продолжается до конца статьи. Обнаружение ключевых слов в тексте реализовать с использованием семантических множителей.

В процессе работы блока семантического анализа текста система должна генерировать файл-протокол, содержащий следующую информацию:

- а) список выделенных ключевых слов;
- б) для каждого анализируемого предложения – семантическую сеть в произвольном формате представления;
- в) промежуточные семантические сети, получающиеся в результате [операций над семантическими сетями](#) обрабатываемых предложений и заголовка.

3. Некоторые понятия и определения

В компонентном анализе под семантическими множителями понимают элементарные составляющие (атомы смысла, семантические маркеры/параметры/компоненты), в терминах которых можно описывать неограниченное количество лексических единиц неограниченного количества естественных языков [2, с. 36].

Техника выделения семантических множителей состоит в рассмотрении слов и выделении некоторых признаков, разбивающих слова на разные семантические группы. Примерами таких признаков могут послужить «одушевлённость», «неодушевлённость», «взрослость», «мужской пол», «женский пол» и т. д. Далее эти признаки располагают по бинарному принципу: «одушевлённость – неодушевлённость», «мужской пол – женский пол» и т. д. Допускается выделение и более дифференцированных признаков (например, «рыбы», «животные», «живые организмы» и т. п.) с получением (в пределе) «атомов смысла». В итоге значение каждого слова представляется как множество семантических множителей (без учёта связей между самими множителями). Для многозначных слов строится многоуровневая система «значение слова – параметры значения». Недостаток: комбинаторный рост числа вариантов описания при попытке выразить смысл более крупных единиц текста (словосочетаний, предложений и т. д.).

В целях введения единой семантики в семантические сети рядом исследователей было предложено использовать процедурные семантические сети. В этом случае сеть строится на основе класса (понятия), а вершины, дуги (отношения) и процедуры представлены как объекты. Процедурами определяются основные действия (*операции*) над дугами (связями):

- установление связи;
- аннулирование связи;
- подсчёт числа вершин, соединённых заданной дугой;
- проверка наличия – отсутствия связи между заданными вершинами

Существуют также процедуры, определяющие основные операции над вершинами, например:

- определение экземпляра класса;
- аннулирование экземпляра;
- подсчёт числа экземпляров, принадлежащих к классу;
- проверка принадлежности экземпляра к некоторому классу.

Атрибуты класса разделяют на атрибуты определения и атрибуты свойства, причём атрибуты свойства отображаются в качестве отношений между классами и не наследуются классом нижнего уровня. Процедура наследования атрибутов между классами заключается в том, что наследование самого атрибута осуществляется с помощью связи экземпляра (*instance link*), а наследование значений – через отношения IS-A. Кроме того, наследование значений атрибутов свойств не производится. По этой причине для наследования некоторым концептом (вершиной класса x) специфических атрибутов определения из класса верхнего уровня необходимо, чтобы он имел атрибут, соответствующий метаклассу, в котором x является экземпляром.

4. Содержание отчёта по работе

В отчете должны быть представлены:

- описание формата представления семантических сетей в файле-протоколе;
- описание алгоритма выделения ключевых слов (здесь же указать зависимость размерности списка ключевых слов от коэффициента сжатия);
- описание алгоритма определения связности текста;
- тестовый набор данных с распечаткой протокола и комментариями.

Рекомендуемая литература

1. Представление и использование знаний [Электронный ресурс] / под ред. Х. Уэно, М. Исидзука. Режим доступа: <http://www.novsu.ru/file/1004921> (дата обращения: 28.08.2014).
2. Попов Э. В. Общение с ЭВМ на естественном языке [Электронный ресурс] / Э. В. Попов. Режим доступа: <http://www.novsu.ru/file/1004896> (дата обращения: 28.08.2014).
3. Михайлов Д. В. Функциональное программирование: курс лекций. Лекция 11: Модели представления знаний в интеллектуальных системах. Часть 2: Сетевая модель и её реализация на Лиспе [Электронный ресурс] / Д. В. Михайлов. Режим доступа: http://www.machinelearning.ru/wiki/images/f/ff/Lect_11_2_fp_mdv.pdf (дата обращения: 28.08.2014).
4. Михайлов Д. В. Системы искусственного интеллекта: курс лекций. Семантические сети и анализ связного текста [Электронный ресурс] / Д. В. Михайлов. Режим доступа: http://www.machinelearning.ru/wiki/images/f/f9/Semnets_mdv_pres.pdf (дата обращения: 12.08.2015).