УДК 004.89:004.4+004.912+004.82

Парамонов А.И.,
Ломонос Я.Г.,
Каргин А.А., д-р техн. наук

ТЕХНОЛОГИЯ ПОСТРОЕНИЯ КОНТЕКСТА ЕЯ-ТЕКСТОВ НА ОСНОВЕ НЕЧЕТКОЙ ЛОГИКИ

Донецкий национальный университет

Рассмотрена проблема моделирования понимания естественно-языкового текста. Определены и изложены основные принципы построения контекста естественного языка. Предложен и описан подход к решению данной задачи на основе нечеткой логики и концептуальных моделях, построенных на сведениях из когнитивной психологии. Описаны алгоритмы, используемые для преобразования входного текста в элементы нечеткой концептуальной модели

Введение

Одним из основных средств передачи информации на сегодняшний день является текст. Роль текста особенно возрастает с увеличением объемов цифровой информации. Развиваются формы представления и методы передачи, в том числе всемирная сеть Интернет. Тексты на естественном языке (ЕЯ) составляют большую часть существующих электронных документов. Столь стремительное развитие информационного потока приводит к необходимости создания инструментов для обработки цифровой информации. С каждым этапом развития перед разработчиками инструментария возникают новые проблемы, ставятся новые задачи. существующие на сегодняшний день программные средства не удовлетворяют полной мере всем предъявляемым требованиям. Преобладание текстовой информации на ЕЯ в неоднородном представлении делает особенно востребованными системы, обеспечивающие новый качественный уровень работы с ЕЯ-текстом.

Задачи обработки ЕЯ-текстов, такие как поиск требуемой информации, категоризация и аннотирование документов, требуют принципиально новых подходов, учитывающих смысловую нагрузку текстов и уникальные особенности языка.

Проблема понимания языковых сообщений не нова для современной науки. Однако её решение по-разному осуществляется в различных отраслях знания. Разнообразие подходов объясняется как известной изолированностью различных наук друг от друга, так и, безусловно, сложностью самого объекта описания.

Над этой проблемой работают многие ученные как в области ИИ, так и когнитивные психологи и лингвисты.

Задача интерпретации текста на естественном языке требует комплексного решения ряда подзадач. Авторы статьи предлагают объединить механизмы нечеткой логики и модели, основанные на сведениях из когнитивной психологии. Это позволит решить проблемы, связанные с разработкой моделей, методов и алгоритмов выделения семантических элементов текста, оценки нечеткости элементов текста, обработки ошибок в символьном представлении текста и т.д. Таким образом, разработка моделей синтаксического, морфологического и терминологического анализа ЕЯ-текстов, основанных на нечетких моделях обработки лингвистической информации и учитывающих когнитивные аспекты обработки текста человеком очень актуальна. Эта проблема и рассматривается в статье.

В предлагаемом подходе задача сводится к преобразованию входного текста, который представлен в виде уровней конкретизации смысловой нагрузки, в элементы нечеткой концептуальной модели.

Для построения контекста в статье предлагается три смысловых уровня единиц текста: первичный текст, нечеткие термины, контекст. Переход между ними осуществляется в два этапа.

Под первичным текстом понимается конечная последовательность базовых элементов электронного текста – символов. Первый этап подразумевает под собой обобщение смысловой информации по уровням символ (каждый элемент в отдельности не несет смысловой нагрузки) – морфема (каждый элемент несет смысловую нагрузку либо общего, либо вспомогательного содержания) – термин (каждый элемент несет конкретную нагрузку в рамках текста). Второй этап основан на разработанной гибридной нечеткой модели (ГНМ) представления концептуальных знаний [1].

Терминологическая разметка текста

Этап терминологического анали за текста с учетом возможности ошибок предполагает составление нечетких характеристик текста и подробно описан в работе [2].

Для того, чтобы учесть различные варианты ошибок в тексте, происходящие при вводе текста оператором ПК или при интерпретации аудио или графических документов, как то замена символа, вставка лишнего символа, выпадение символа или их комбинаций, необходимо построение трех нечетких характеристик текста: символьной, морфемной и терминологической.

Построение нечеткой характеристики текста уровня символов требует минимальной информации о том, каким образом был создан электронный текст. Исходя из этой информации, выбирается наиболее подходящая нечеткая характеристика базового алфавита, которая определяет, насколько один символ алфавита похож на другой с позиции метода создания электронного текста. Характеристика похожести задается в виде нечеткого множества на базовом множестве фактора уверенности [3] – действительного отрезка [–1;+1] и представляет собой обобщенные экспертные данные. Так, например, при условии, что текст набирался оператором ПК, символ «а» очень похож на символ «А» (разные регистры) и похож на символ «в» (располагается рядом на клавиатуре).

В каждой нечеткой характеристике уровня символов позиции текста содержится вектор уверенностей системы в том, что в этой позиции находится некоторый символ алфавита для всех символов алфавита. Уверенность также задается в виде нечеткого множества на базовом множестве фактора уверенности [3]. По сути, каждая позиция нечеткой характеристики текста уровня символов содержит похожесть встреченного в этой позиции символа первичного текста на все другие символы используемого алфавита, включая себя. Такая форма представления текста позволяет учесть ошибки в тексте, связанные с заменой символа.

Построение нечеткой характеристики текста уровня морфем происходит путем последовательного анализа каждой позиции нечеткой характеристике текста уровня символов так, как это показано в работах [2, 4]. В нечеткой характеристике конечной подпоследовательности позиций текста уровня символов выбираются уверенности тех символов соответствующих символам морфемы. Объединение этих уверенностей формирует уверенность системы в том, что в этой позиции в первичном тексте присутствует морфема, для которой проводился анализ. Алгоритм выбора нечеткой характеристике подпоследовательности позиций текста и алгоритм объединения уверенностей построены таким образом, чтобы учитывать возможность выпадения или добавление лишнего символа.

Нечеткая характеристика текста уровня морфов представляет собой последовательность нечетких характеристик уровня морфов каждой позиции первичного текста. В каждой нечеткой характеристике уровня морфов позиции текста содержатся уверенности системы для тех морфем, которые, по мнению системы, могут присутствовать в первичном тексте и заканчиваться в этой позиции. Терминологический анализ начинается со структуры термина. В данной работе предполагается, что термин – это конечная последовательность слов не менее чем из одного слова (фраза), несущая конкретную смысловую нагрузку.

Причем, порядок слов может изменяться. Каждое слово – конечная последовательность морфем, содержащая только одну корневую морфему, которая несет основной смыл слова.

Терминологический анализ предполагает параллельный механизм анализа нечеткой характеристики текста уровня морфем, описанный в [2, 4]. Этот алгоритм выделяет термины словаря, присутствующие в первичном тексте независимо от порядка следования слов. Уверенность термина строится на основе уверенностей морфем из нечеткой характеристики текста уровня морфем и расстояния между позициями, в которых эти морфемы имеют уверенность. Для выделения термина и построения его нечеткой характеристики – уверенности необходимы уверенности всех корневых морфем, но уверенности вспомогательных морфем (префиксов, суффиксов, инфиксов) необязательно.

Результатом терминологического анализа является нечеткая характеристика текста уровня терминов подобная нечеткой характеристике текста уровня морфем. Нечеткая характеристика текста уровня терминов отражает уверенность системы о том, какие термины, в какой позиции первичного текста и с какой уверенностью присутствует.

Построение контекста

Согласно предлагаемой технологи формирование контекста ЕЯ-текстов базируется на гибридной нечеткой модели представления знаний [1].

Структура гибридной модели (ГМ) разработана на основе комплексного подхода к решению поставленной задачи.

Анализ когнитивного подхода позволил выделить базовые единицы модели: объекты, действия и события [1, 5]. Элементы модели объединяются в классификационные структуры: семантические и пропозициональные сети [6]. В модели учтены особенности индивидуального восприятия окружающего мира, представленные набором индивидуальных знаний о мире в виде прототипов гибридной модели [6, 7]. Прототипы формируются в соответствии с классификационными структурами и разделяются на схемы и скрипты [6]. Таким образом, множество элементов модели и связи между ними представляют систему знаний.

Но данное представление не может быть полным, потому что человеческое мышление представляет собой нечёткий механизм [8]. Таким образом, полная модель знаний должна быть представлена нечеткой ГМ (НГМ), в которой каждой составляющей ГМ приписывается фактор уверенности (*CF*).

Исходя из принятого в [1] описания, объекты и действия есть наборы признаков, формируемые из множества всех возможных признаков при построении базы знаний (БЗ). Для каждой БЗ задается разная степень принадлежности определённого признака конкретному понятию. Это даёт возможность определить каждый РЭГМ как подмножество признаков, обладающее нечеткой характеристикой.

Фактор уверенности для элементов множества событий задаётся уверенностью в составляющих этого события. Согласно [1] событие состоит из элементов множеств объектов и действий, а также дополнительных аргументов, которые влияют на отношения между событиями, но не являются составляющими уверенности в самом событии.

Фактор уверенности классификационных структур НГМ есть множество нечётких характеристик связей в этих структурах. Построенная с учётом заданных факторов уверенности НГМ описывает индивидуальные знания о предметной области и формирует БЗ, используемую при построении контекста.

Контекстом будем считать цепочку событий (последовательность узлов пропозициональной сети), субъектами которых выступают «главные действующие объекты» (подмножество узлов семантической сети объектов, элементы которого наиболее часто представлены в обрабатываемом тексте).

Таким образом, задача выделения контекстных знаний сводится к нахождению подграфа НГМ, который формируется при обработке входных нечетких терминов, полученных на этапе терминологической разметки.

В основе процесса обработки знаний в НГМ используется модель активизации сетей, разработанная на базе модели логогена Мортона [3].

В общем виде алгоритм обработки фрагмента текста, предложенный в работе, представляет собой последовательность действий:

- анализ *j*-й нечеткой фразы (множество нечетких терминов);
- сопоставление фразы с узлом сети;
- вычисление активности узла, поступившей после обработки *j*-й нечеткой фразы;
- расчет активности узла полученной при обработке предыдущих фраз;
- расчет активности узла, поступившей от смежных узлов;
- распространение активности от узла на смежные с ним узлы;
- пересчет активности узла с учетом всех составляющих.

Процесс обработки входного фрагмента текста представляет собой дискретный во времени и непрерывный по состоянию процесс. Каждое изменение временного шага связано с обработкой очередного события фрагмента текста.

В процессе интерпретации входных знаний у определенных узлов сети значения активности изменяются. При этом перерасчет происходит при обработке каждого события фрагмента текста.

Выводы

Представленная модель контекста может быть использована для полноценной категоризации и аннотации текстов, а полученные результаты для построения интеллектуальной поисковой системы.

Описанные алгоритмы были проверены ходе компьютерных экспериментов. Предложенная технология построения контекста использована при построении интеллектуальной системы категоризации и интерпретации текстовой информации "*Text-Term-Concept*" [9], а также в интеллектуальной поисковой системе "*EnewsAnalyzer*" [10].

Список литературы

1. *Парамонов А.И.* Представление знаний гибридной моделью для систем интеллектуального поиска // Вестник Донецкого университета. Донецк, Серия А, №1, 2005. – С. 404–409.

2. *Ломонос Я.Г.* Терминологическая разметка текста в автоматизированной системе интеллектуальной обработки текстовой информации // Штучний Інтелект.

– ІПШІ МОН і НАН України «Наука і освіта», 2006. – №3/2006. – С. 537–547.

3. *Люггер Дж. Ф.* Искусственный интеллект: стратегии и методы решения сложных проблем // М.: Изд. Дом «Вильямс», 2003.

4. *Ломонос Я.Г.* Использование механизма логогена Мортона для терминологического анализа электронных документов // Наукові праці ДонНТУ. – 2007. – №13(121). – С. 145 – 152.

5. *Schank R.C.* Scripts, Plans, Goals and Understanding: An Inquiry into Human Knowledge Structures // Hillsdale, NJ: Lawrence Erlbaum Associates, 1977. – 248 p.

6. *Солсо Р.* Когнитивная психология // СПб.: Питер, 2002. – 592 с. – (Серия «Мастера психологии»).

7. *Величковский Б.М.* Современная когнитивная психология // М.: МГУ, 1982. 8. *Кофман А.* Введение в теорию нечетких множеств // пер. с франц. – М.: Радио и связь, 1982. – 432 с.

9. *Каргин А.А., Парамонов А.И., Ломонос Я.Г.* Интеллектуальная систе атора, мірі інтер полянта і параметрів процедури проріджування даних.