УДК 004.822

МОДЕЛЬ ПОСТРОЕНИЯ СЕМАНТИЧЕСКОЙ СЕТИ НАУЧНОГО ТЕКСТА

Аюшеева Н.Н., Диких А.Ю.

ФГБОУ ВО «Восточно-Сибирский государственный университет технологий и управления», Улан-Удэ, e-mail: anndonir@gmail.com, arkh1991@gmail.com

Настоящая статья посвящена описанию модели построения семантической сети научного текста на основе использования знаний онтологии предметной области. Система построения семантической сети научного текста состоит из модуля лексического и морфологического анализа, модуля выделения терминов и отношений, модуля вычисления весовых коэффициентов и модуля визуализации семантической сети. Модуль лексического и морфологического анализа разбивает весь входной поток предложений текста на лексемы и определяет их морфологические характеристики. Модуль выделения терминов и отношений вначале формирует на основе полученных лексем список терминов, а затем выполняет поиск функциональных и нефункциональных отношений между ними. Функциональные отношения определяются на основе анализа взаимного расположения терминов в пределах одного предложения. Для определения нефункциональных отношений терминам текста сопоставляются фреймы онтологии, а затем производится поиск связи между фреймами в самой онтологии. На основе частотных характеристик, а также места расположения терминов в тексте происходит вычисление весовых коэффициентов для терминов и отношений. На последнем шаге на основе всей полученной информации формируется взвешенный граф, вершинам которого соответствуют термины, дугам – отношения.

Ключевые слова: семантика текста, семантическая сеть, морфологический анализ, лексический анализ, онтология, научный текст

MODEL OF CONSTRUCTING A SEMANTIC NETWORK OF SCIENTIFIC TEXT

Ayusheeva N.N., Dikikh A.Yu.

East Siberian State University of Technology and Management, Ulan-Ude, e-mail: anndonir@gmail.com, arkh1991@gmail.com

This paper is devoted to description the model of a scientific text semantic network construction on the basis of using a domain ontology knowledge. The system of a scientific text semantic network construction consists of a module for lexical and morphological analysis, a module for identification the terms and relationships, a module for calculating the weight coefficients, and a module for visualizing the semantic network. The module for lexical and morphological analysis divides the input stream of text sentences into lexemes and determines their morphological characteristics. The module for identification the terms and relationships, in the first place, forms a list of terms based on the obtained lexemes, and then search a functional and non-functional relations between them. The functional relations are determined after analysis of the relative disposition of terms in the one sentence. The frames of ontology are compared a text terms for the determination of the non-functional relations. Then the relations between the frames of ontology are found. Based on the frequency characteristics, as well as on the location of terms in the text, the weight coefficients of the terms and relations are calculated. At the last step, the weighted graph is formed on the basis of an all received information. In graph nodes correspond to terms and arcs – to relations.

Keywords: text semantic, semantic network, morphological analysis, lexical analysis, ontology, scientific text

Семантические сети были разработаны в качестве общего аппарата представления знаний. С момента их разработки они активно использовались в системах обработки естественного языка и оказались одним из самых наглядных способов представления семантики высказываний на естественном языке. Семантические сети являются инструментом представления сложных совокупностей объектов и отношений между ними, которые, в свою очередь, выступают в сети элементами знания [1]. Выводы в семантической сети основаны на анализе отношений между объектами. Модели семантической сети в значительной степени универсальны и являются легко настраиваемыми на любую конкретную предметную область. Задачи, связанные с извлечением знаний из текстов, обучающие системы, информационный поиск, реферирование, проверка корректности терминологических словарей и определений – это далеко не полный список задач, для решения которых успешно используются модели семантических сетей. Для создания семантической сети необходимо провести комплексный анализ текста, который позволит представить взаимосвязь объектов, их свойства и атрибуты, а также определить важность терминов и отношений текста, что даёт возможность сделать выводы о его содержании, наиболее и наименее важных фактах в рамках данного текста и их зависимостей друг от друга.

В настоящее время происходит рост объёмов обрабатываемой информации, в связи с чем растёт потребность в интеллектуальных системах. Следовательно, задачи семантического анализа графики, звука, текста приобретают всё большую

актуальность. Наибольшее применение в решении прикладных задач находят методы, базирующиеся на анализе факторных (статистических) характеристик слов и словосочетаний исследуемого текста. Существенной проблемой данных методов является невозможность отображения в полной мере содержания или смысла анализируемого объекта, например текста. Кроме того, при попытке извлечения знаний из построенной семантической сети могут произойти сложности с верной интерпретацией содержания представленных текстовых данных. В связи с вышесказанным для корректного построения семантической сети текста необходимо производить интеллектуальную обработку анализируемых данных, дающую возможность совместно с использованием нечеткой системы определять композиционную структуру текста, а также при помощи предметных онтологий выделять термины и отношения между ними. Предложенная в работе [2] технология построения семантической сети позволяет проследить последовательность действий такой интеллектуальной обработки. В данной работе предпринята попытка формализованного описания предложенной технологии с позиций морфологического представления системы построения семантической сети.

Описание модели системы построения семантической сети

Семантическая сеть, построенная для текстового документа, должна достаточно точно отображать смысл и семантику текста. Поэтому основной задачей при построении семантической сети текста является извлечение его семантики, которая будет максимально отображать смысловой аспект текста. Для решения данной задачи предлагается использовать систему, структурная схема которой представлена на рисунке.

Система построения семантической сети научного текста представляет собой четверку

$$D = (Ml, Mtr, Mw, Ms),$$

где *Ml* – модуль лексического и морфологического анализа;
Mtr – модуль выделения терминов и отношений;
Mw – модуль вычисления весовых коэффициентов;
Ms – модуль визуализации семантической сети.

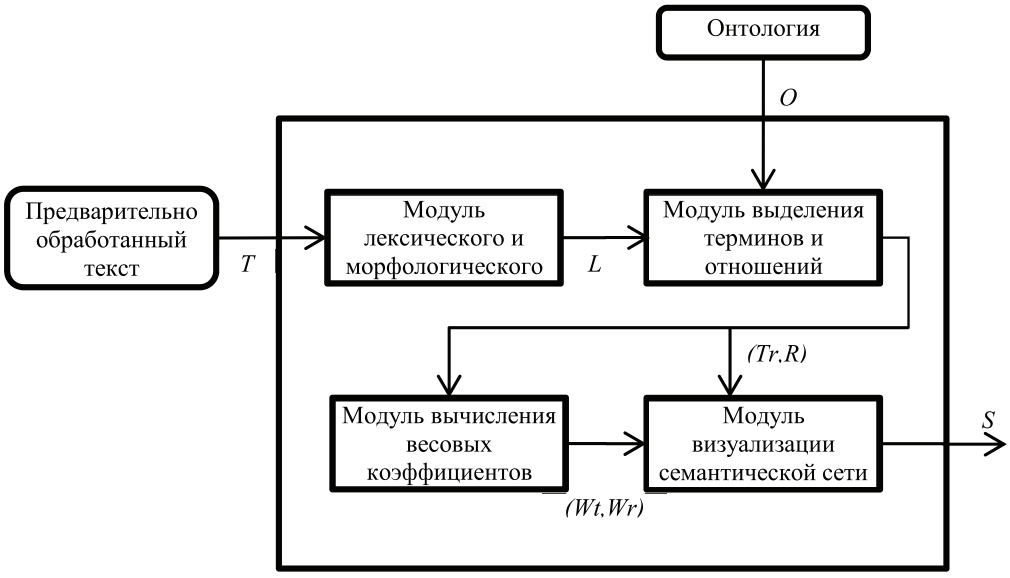
На вход данной системы поступает предварительно обработанный текст $T = \{p_i\}$, $i = 1..n$, где *p* – простое предложение, не имеющее однородных членов предложения, личных, указательных, относительных местоимений. На входе кроме этого мы имеем онтологию предметной области *O*.

Модуль лексического и морфологического анализа *Ml* на основе *T* формирует множество описаний лексем

$$L = \{Lx_i\}, i = 1..m,$$

$$Lx = (form, textForm, cid, vid, position, ftp),$$

где *form* – нормальная форма лексемы;
textForm – форма лексемы в тексте;



Структурная схема системы построения семантической сети

cid – код набора постоянных характеристик лексемы (например, часть речи);
vid – код набора временных характеристик (например, число, падеж);
position – порядковый номер лексемы в тексте;
ftp – код формального текстового признака.

Модуль выделения терминов и отношений *Mtr* на основе *L* и с помощью онтологии предметной области *O* формирует пару (*Tr*, *R*), где множество терминов $Tr = \{Trm_i\}$, $i = 1..t$, $Trm_j = \{Lx_j\}$, $j = 1..l$, и множество отношений $R = \{Rl_k\}$, $k = 1..r$, $Rl_i = (Trm_l, Rel, Trm_r)$, в котором тип отношения $Rel = Lx$ в случае наличия функционального отношения между Trm_l и Trm_r или $Rel \in$ {«class», «kind», «whole», «part»} в случае нефункционального отношения.

Модуль вычисления весовых коэффициентов *Mw* для пары (*Tr*, *R*) создаёт соответствующую ей пару (*Wt*, *Wr*), где $Wt = \{Wtr_i\}$ – вес Trm_i, $Wr = \{Wrl_k\}$ вес Rl_k.

Модуль визуализации семантической сети *Ms*, являясь агрегатором, собирает результаты работы *Mtr* и *Mw* и генерирует семантическую сеть *S* в виде взвешенного графа (*V*, *T*), где множество вершин $V = Tr$, множество отношений $T = R$, вес вершины $V_i = Wtr_i$, а вес отношения $R_k = Wrl_k$ [3].

Далее рассмотрим каждый модуль подробнее.

Модуль лексического и морфологического анализа

На этапе лексического анализа происходит определение содержательно-смысловых блоков текста.

В научном тексте лингвисты определяют четыре логически выделенных содержательных блока:

1) проблема – блок, в котором рассказывается о постановке и понимании проблемы;

2) опыт – блок, в котором перечислен опыт предшественников по данному направлению исследования;

3) решение – блок, в котором предлагается и обосновывается способ решения проблемы, его доказательства и аргументы;

4) итог – блок, в котором происходит обобщение всех полученных данных, а также подводится общий итог.

Для идентификации содержательно-смысловых блоков используется метод поиска формальных текстовых признаков, которые употребляются в том или ином блоке [3].

Для каждого блока существует своё множество текстовых признаков. Так, для блока «Проблема» множество текстовых признаков M_p = {«в данной статье», «в данной работе», «в статье», «рассмотрим», …}, для блока «Опыт» – M_e = {«в своих работах», «исходя из опыта ученых», «вспомним», «утверждает, что», …}, для блока «Решение» M_d = {«анализ показал», «исследование показало», «заметим, что», «можно выделить», «нужно заметить», …}, для блока «Итог» M_t = {«можем сделать вывод», «в итоге», «отсюда», «таким образом»,…}. Множество $M = \{M_p, M_e, M_d, M_t\}$ есть множество формальных текстовых признаков, определяющих различные блоки текста. Так как один и тот же формальный текстовый признак может принадлежать разным блокам, для определения значения *ftp* для всех Lx_i, выделенных в данном абзаце текста, используется нечёткий регулятор [3].

Лексический анализ необходим для разбиения текста документа *T* на последовательность лексем *L*. Символы входной последовательности могут принадлежать каким-либо лексемам – *A* = {«а», «б», «в», «г», «д», «е», «ё», «ж», «з», «и», «й», «к», «л», «м», «н», «о», «п», «р», «с», «т», «у», «ф», «х», «ц», «ч», «ш», «щ», «ъ», «ы», «ь», «э», «ю», «я», «А», «Б», «В», «Г», «Д», «Е», «Ё», «Ж», «З», «И», «Й», «К», «Л», «М», «Н», «О», «П», «Р», «С», «Т», «У», «Ф», «Х», «Ц», «Ч», «Ш», «Щ», «Ъ», «Ы», «Ь», «Э», «Ю», «Я», «0», «1», «2», «3», «4», «5», «6», «7», «8», «9», «№»}, а могут являться символами-разделителями – *Dv* = {«!», «?», «.», «,», « », «:» «;»}. Только в редких случаях между лексемами не бывает разделителей. *Ml* представляет каждое $p_i \in T$ в виде последовательности символов *X*. На основе анализа вхождения очередного $x_i \in X$ в множества *A* и *Dv* можно определить значение *textForm* и *position* для очередной лексемы $Lx_i \in L$.

Для проведения морфологического анализа используется функционал библиотеки Mcr.dll, реализующей доступ к словарю на основе *n*-грамм [4], который позволяет для лексемы Lx_i с определённым значением *textForm* определить значения *form*, *cid*, *vid*.

Модуль выделения терминов и отношений

В данной работе выделение терминов базируется на поиске субстантивных именных словосочетаний, представляющихся моделью *согласуемое слово + существительное*. Согласно этой модели основой словосочетания является *существительное*, а в качестве зависимого выступает *согласуемое слово*, которое, как правило, представляется в виде имени существительного или имени прилагательного. Состав именного словосочетания не ограничивается только существительным и прилагательным. Так же в словосочета-

нии можно встретить наречия, предлоги и сочинительные союзы.

Для выделения терминологических словосочетаний используется КС-грамматика, описанная в работе [2]. Грамматика позволяет определять семантическую связность подряд идущих лексем и определить очередной $Trm_i \in T_r$. Данная грамматика базируется на принципе согласованности лексем, т.е. словоформа в словосочетании должна быть связана с другими словоформами и подходить к ним по временным характеристикам, таким как род, число и падеж.

Стоит учесть, что данные правила не позволяют выделять словосочетания, которые однозначно являются терминологическими. Как правило, качественные и притяжательные прилагательные не входят в их состав (*большой дом*, *широкое распространение*) за некоторым исключением. Для определения разряда прилагательного можно успешно использовать Национальный корпус русского языка (ruscorpora.ru).

В научном тексте для построения семантической сети между терминами выделяются квалитативные и квантитативные отношения [5]. Квалитативными отношениями являются отношения иерархии – «род – вид», отношения агрегации – «целое – часть» и функциональные отношения, отражающие прагматику предметной области и имеющие вид «объект действия – действие – субъект действия». Из квантитативных отношений в рамках данной работы рассматривалось только отношение тождества – «синоним».

Выделение очередного функционального отношения $Rl_i \in R$ происходит с помощью поиска в тексте Lx_j со значением *cid*, определяющим часть речи – глагол и находящуюся между *Trmi* и *Trmi* + 1. В таком случае, $Rel = Lx_j$.

Для выделения нефункциональных отношений происходит анализ онтологии O на предмет наличия в них описания терминов Tr [6]. Таким образом, Trm_i может быть сопоставлен фрейм O_j. Если фрейм O_i имеет ссылку, определяющую квалитативное отношение, на фрейм O_j, то следовательно можно утверждать, что между соответствующим фреймам Trm_i и Trm_j также будет квалитативное отношение. В таком случае *Rel* определяется как «class», «kind», «whole» или «part» в зависимости от типа обнаруженного квалитативного отношения.

Модуль вычисления весовых коэффициентов

В качестве вершин семантической сети научного текста используются термины. Дугами же являются отношения между ними. Весовой коэффициент позволяет количественно определить значимость термина и отношения. Определение весовых коэффициентов терминов и отношений текста является одной из приоритетных задач, так как данные характеристики позволяют определить важнейшие идеи и основную суть всего текста. Под «значимостью» подразумевается как «наличие смысла, значения», так и отношение данного термина к другим терминам в рамках текста. Определение значимости термина происходит на основе анализа критериев значимости:

– частота встречаемости w_1: если термин часто встречается в тексте, он образует большее количество отношений;

– категория текста w_2: термины, соответствующие тематике текста, являются более значимыми, чем остальные;

– содержательно-смысловой блок w_3: термины, встречающиеся в основных содержательно-смысловых блоках, например «итог» или «проблема», являются более значимыми для определения смысла текста.

Для термина *Trmi* w_1 рассчитывается по формуле

$$w_1 = 1 - \log_{\max(r)} r, \quad (1)$$

где r – ранг частоты термина.

Вес w_2 принято считать равным 1, в случае, если Trm_i отражает тему текста, 0 в противном случае.

Вес w_3 принимает различные значения в зависимости от типа содержательно-смыслового блока, в котором встречается Trm_i. Для блока «Проблема» $w_3 = 0{,}3$, для блока «Опыт» $w_3 = 0{,}15$, для блока «Решение» $w_3 = 0{,}25$, для блока «Итог» $w_3 = 0{,}3$. В случае, если Trm_i встречается в нескольких блоках, то w_3 рассчитывается как сумма значений w_3 в соответствующих блоках.

Общий вес Wtr_i для Trm_i рассчитывается по формуле

$$Wtr_i = \sum_{j=1}^{3} k_j w_j, \quad (2)$$

где значения $k_1 = 0{,}309$, $k_2 = 0{,}406$, $k_3 = 0{,}285$, вычислены согласно процедуре взвешивания, предложенной в работе [7].

Значимость отношений семантической сети зависит от двух критериев:

– частота совместной встречаемости терминов в данном отношении w_1;

– принадлежность отношения к содержательно-смысловому блоку w_2.

Значение w_1 рассчитывается по формуле 1, где r – ранг частоты встречаемости отношения.

w_2 принимает различные значения в зависимости от типа содержательно-смыслового блока, в котором встречается Rl_i.

Для блока «Проблема» $w_2 = 0{,}17$, для блока «Опыт» $w_2 = 0{,}09$, для блока «Решение» $w_2 = 0{,}35$, для блока «Итог» $w_2 = 0{,}39$. В случае, если Trm_i встречается в нескольких блоках, то w_2 рассчитывается как сумма значений w_2 в соответствующих блоках.

Общий вес Wrl_i для Rl_i рассчитывается по формуле

$$Wrl_i = w_1 * k_1 + w_2 * k_2, \qquad (3)$$

где значения $k_1 = 0{,}425$, $k_2 = 0{,}575$ также вычислены согласно процедуре взвешивания, предложенной в работе [7].

Модуль визуализации семантической сети

Данный модуль предназначен для представления результатов вычислений в удобном для исследователя виде. Исходными данными модуля являются список терминов *Tr*, список отношений *R*, а также веса терминов *Wtr* и веса отношений *Wr* .

Выходными данными является список смежности графа семантической сети $G = \{L_i\}$, где $i = 1..t$, $Ls_i = (Trm_i, Wtr_i, Links_i)$, $Links_i = \{Link_j\}$, где $j = 1..p$ и p – количество терминов Trm_j, с которыми Trm_i находится в отношении. $Link_i = (Rel_j, Ls_j, Wr_j)$ является описанием связи Trm_i с Trm_j.

Заключение

Модель построения семантической сети представляет собой систему, включающую четыре взаимосвязанных модуля. Модуль лексического и морфологического анализа использует словарь на основе *n*-грамм из библиотеки Mcr.dll, что обеспечивает динамичное пополнение словаря. Модуль выделения терминов основан на использовании контекстно-свободной грамматики, которая позволяет расширить количество структур терминологических словосочетаний. При вычислении весовых коэффициентов терминов и отношений учитывается композиционная структура текста. Модуль визуализации обеспечивает проведение экспериментов. Предложенная модель соответствует технологии построения семантической сети, описание которой приведено в ранних работах.

Список литературы

1. Бессмертный И.А. Визуализация знаний на основе семантической сети / И.А. Бессмертный // Программирование. – 2010. – № 4. – С. 16–24.

2. Аюшеева Н.Н., Диких А.Ю. Технология построения семантической сети научного текста на основе анализа онтологии предметной области // Знания – онтологии – теории: материалы VI Всероссийской научно-технической конференции с международным участием. – Новосибирск: Изд-во ИМ СО РАН, 2017. – С. 38–46.

3. Аюшеева Н.Н., Гомбожапова Т.Н., Диких А.Ю. Определение содержательно-смыслового блока терминов и отношений научного текста // Теоретические и прикладные вопросы современных информационных технологий: материалы XII Всероссийской научно-технической конференции. – Улан-Удэ: Изд-во ВСГУТУ, 2015. – С. 88–95.

4. Выдрин Д.В., Поляков В.Н. Реализация электронного словаря с использованием n-грамм // Искусственный интеллект. – 2002. – № 4. – С. 180–183.

5. Найханова Л.В. Технология создания методов автоматического построения онтологий с применением генетического и автоматного программирования: монография. – Улан-Удэ: Изд-во ВСГУТУ, 2013. – 268 с.

6. Лукашевич Н.В. Тезаурусы в задачах информационного поиска / Н.В. Лукашевич. – М.: Изд-во МГУ, 2011. – 89 с.

7. Балацкий Е.В. Методы диагностики социального самочувствия населения / Е.В. Балацкий // Мониторинг общественного мнения. – 2005. – № 3 (75). – С. 47–53.