УДК 004.853

ГЕНЕРАЦИЯ ДВУХУРОВНЕВОЙ СЕМАНТИЧЕСКОЙ СЕТИ ДЛЯ ДИАЛОГОВОЙ ИНФОРМАЦИОННОЙ СИСТЕМЫ

Н. И. Гурин, Я. А. Жук

Белорусский государственный технологический университет, г. Минск, Республика Беларусь, root@belstu.by

Дистанционная обучающая технология находит все более широкое применение в современном мире. Однако по-прежнему вызывает опасения качество дистанционного обучения. С целью его повышения ведется разработка новых обучающих средств, интегрируемых в среды дистанционного обучения. Одним из направлений таких разработок является создание диалоговых информационных систем, приоритетной задачей которых является поиск точных ответов на вопросы обучающихся.

Для создания таких диалоговых систем необходимо выполнить генерацию базы знаний, которая будет в дальнейшем использоваться для поиска ответов. Структура нашей базы знаний информационной системы основана на наиболее общей модели представления знаний – семантической сети – в ее наиболее компактной форме списка дуг [1]. Отличительной особенностью списка дуг семантической сети от списка дуг обычного графа является наличие типов отношений, обозначаемых дугами.

В семантической сети были выделены два уровня. Первый уровень выражает отношения между конкретными информационными единицами. На данном уровне количество типов отношений ограничено только возможностями самого языка. Второй же уровень отражает отношения между типами связей первого уровня. Такими отношениями выступают синонимия и антонимия. Ключевым звеном, связывающим два уровня семантической сети между собой, являются типы отношений, выступающие характеристикой дуг первого уровня сети и узлами второго уровня сети.

Автоматическая генерация семантической сети по тексту, который является основным источником знаний в электронных учебных материалах, требует решения ряда задач компьютерной лингвистики. В первую очередь необходимо выполнить поиск слов, выполняющих роль сказуемых в предложениях, т.к. именно сказуемое выражает тип семантической связи. Для определения наиболее распространенных типов сказуемых был проведен анализ предложений текста электронного учебника по электрохимии. В ходе анализа рассматривались предложения первых двух глав исследуемого текста. Для слов определялась их роль в предложении, в частности выявлялись слова, выполняющие роль сказуемого.

На основании данного анализа было выявлено, что текстах научного стиля в роли сказуемых преимущественно выступают глаголы в формах третьего лица и краткие прилагательные. Реже встречаются глаголы в форме первого лица. Кроме того, причинно-следственные связи выражаются при помощи союзов «если», «поскольку» и «когда».

Для оценки общего числа семантических связей и расстановки приоритетов в их поиске был проведен анализ всего исследуемого текста на содержание псевдоокончиний, соответствующих наиболее распространенным средствам выражения семантических связей [2]. В качестве средства такого анализа использовались регулярные выражения, в состав которых включались псевдоокончания и обозначение конца слова «\b». Такой анализ позволил учесть, как слова, отделенные от следующего пробелом, так и слова, после которых стоял знак препинания.

Стоит отметить неточность данной методики подсчета, связанную с совпадением псевдоокончаний некоторых слов с псевдоокончаниями глаголов. Например, слово «атрибут» заканчивается на «ут», как и глаголы первого спряжения в форме третьего лица, множественного числа. Данный факт вносит незначительную погрешность при обнаружении глаголов в форме третьего лица (1 ложное совпадение из 38) и гораздо более заметную погрешность при выявлении глаголов в форме первого лица (160 ложных совпадений из 353). Совершенствование морфологического анализа возможно путем применения существующих в настоя-

щее время программных пакетов для решения задач компьютерной лингвистики. Выбор конкретных средств зависит от используемого при разработке генератора семантической сети языка программирования. Так, для языка программирования Python наиболее подходящим является программный пакет морфологического анализа pyMorphy, а для языка программирования C# – пакет морфологического и синтаксического анализа Solarix.

Данные о числе совпадений псевдоокончаний глаголов и кратких прилагательных со словами текста электронного учебника представлены в таблице 1. В связи с высоким числом ложных совпадений, обнаруженных при поиске глаголов в форме первого лица, статистика по ним не приводится.

Таблица 1 – Количество форм глаголов и кратких прилагательных в исследуемом тексте

Часть речи	Форма	Псевдоокончание	Число совпадений	% совпадений
глагол I спряжения	3 лицо, ед. число	-ет	417	24,62
глагол I спряжения	3 лицо, мн. число	-ут	38	2,24
глагол I спряжения	3 лицо, мн. число	-ют	109	6,43
глагол II спряжения	3 лицо, ед. число	-ит	88	5,19
глагол II спряжения	3 лицо, мн. число	-ат	39	2,30
глагол II спряжения	3 лицо, мн. число	-ят	15	0,89
глагол I спряжения	3 лицо, ед. число, возвратный	-ется	318	18,77
глагол I спряжения	3 лицо, мн. число, возвратный	-утся	8	0,47
глагол I спряжения	3 лицо, мн. число, возвратный	-ются	115	6,79
глагол II спряжения	3 лицо, ед. число, возвратный	-ится	52	3,07
глагол II спряжения	3 лицо, мн. число, возвратный	-атся	0	0,00
глагол II спряжения	3 лицо, мн. число, возвратный	-ятся	11	0,65
краткое прилагательное	муж. род, ед. число	-ан	29	1,71
краткое прилагательное	жен. род, ед. число	-ана	17	1,00
краткое прилагательное	ср. род, ед. число	-ано	19	1,12
краткое прилагательное	мн. число	-аны	17	1,00
краткое прилагательное	муж. род, ед. число	-ен	64	3,78
краткое прилагательное	жен. род, ед. число	-ена	34	2,01
краткое прилагательное	ср. род, ед. число	-ено	22	1,30
краткое прилагательное	мн. число	-ены	36	2,13
краткое прилагательное	муж. род, ед. число	-жен	13	0,77
краткое прилагательное	жен. род, ед. число	-жна	13	0,77
краткое прилагательное	ср. род, ед. число	-жно	88	5,19
краткое прилагательное	мн. число	-жны	10	0,59
союз		если	71	4,19
союз		когда	34	2,01
союз		поскольку	17	1,00
		Всего:	1694	100,00

Как видно из таблицы, из текста можно выделить около 1700 семантических связей. Наиболее распространенными средствами выражения семантических связей являются глагол первого спряжения в форме третьего лица, единственного лица и его возвратная форма. Таким образом, корректное выявление данных словоформ и разбор предложений с ними является приоритетной задачей при формировании семантической сети. Однако, поскольку большого числа ложных обнаружений других приведенных в таблице слов не было замечено, разработка функций выявления требуемых словоформ была выполнена с расчетом на все перечисленные в таблице формы слов.

Функция проверку принадлежности слов к указанным средствам выражения семантических связей выполнена на языке C# и представлена на рисунке 1:

```
private bool isVerb(string word)
{
    //массив нужных окончаний
    string[] endings = { "ет", "ут", "ют",
                         "ит", "ат", "ят",
                         "ется", "утся", "ются",
                         "ится", "атся", "ятся",
                         "ен", "ена", "ено", "ены",
                         "ан", "ана", "ано", "аны",
                         "жен", "жна", "жно", "жны"};
    //перебор окончаний
    for (int j = 0; j < endings.Length; j++)
    {
        //проверка, оканчивается ли i-ое слово на j-ое окончание
        if(word.EndsWith(endings[j])){
            return true;
        }
    }
    return false;
}
```

Рисунок 1 – Функция определения глаголов и кратких прилагательных

При помощи приведенной функции становится возможным выделить из текста только искомые слова после сегментации текста [3]. На рисунке 2 приведена функция сегментации текста с учетом знаков препинания и вывода найденных слов, выражающих семантические связи.

```
protected void Page_Load(object sender, EventArgs e)
{
    //знаки препинания
    char[] separators = "'\",.!?()[]\\/".ToCharArray();
    string txt = Request["txt"];
    //добавляем пробелы перед знаками препинания
    for (int i = 0; i < separators.Length; i++)
        txt = txt.Replace("" + separators[i], " " + separators[i]);
    //массив слов и знаков препинания
    string[] words = txt.Split();
    //перебор слов
    for (int i = 0; i < words.Length; i++)
    {
        //проверка, выбирающая нужные слова
        if(isVerb(words[i]))
            //если проверка прошла, выводим слово
            Response.Write(words[i] + "<br/>\n");
    }
}
```

Рисунок 2 – Функция определения глаголов и кратких прилагательных

Представленные функции используются в качестве одного из функциональных компонентов серверной части генератора семантической сети. Этап поиска средств выражения се-

мантических связей является вторым после сегментации. В настоящее время ведется отладка данной операции в отдельности от последующих этапов. На рисунке 3 показаны результаты работы данного компонента, выведенные в клиентской части. В текстовом поле в верхней части интерфейса виден исходный текст, а в нижней – найденные глаголы с замененными на специальные теги окончаниями.

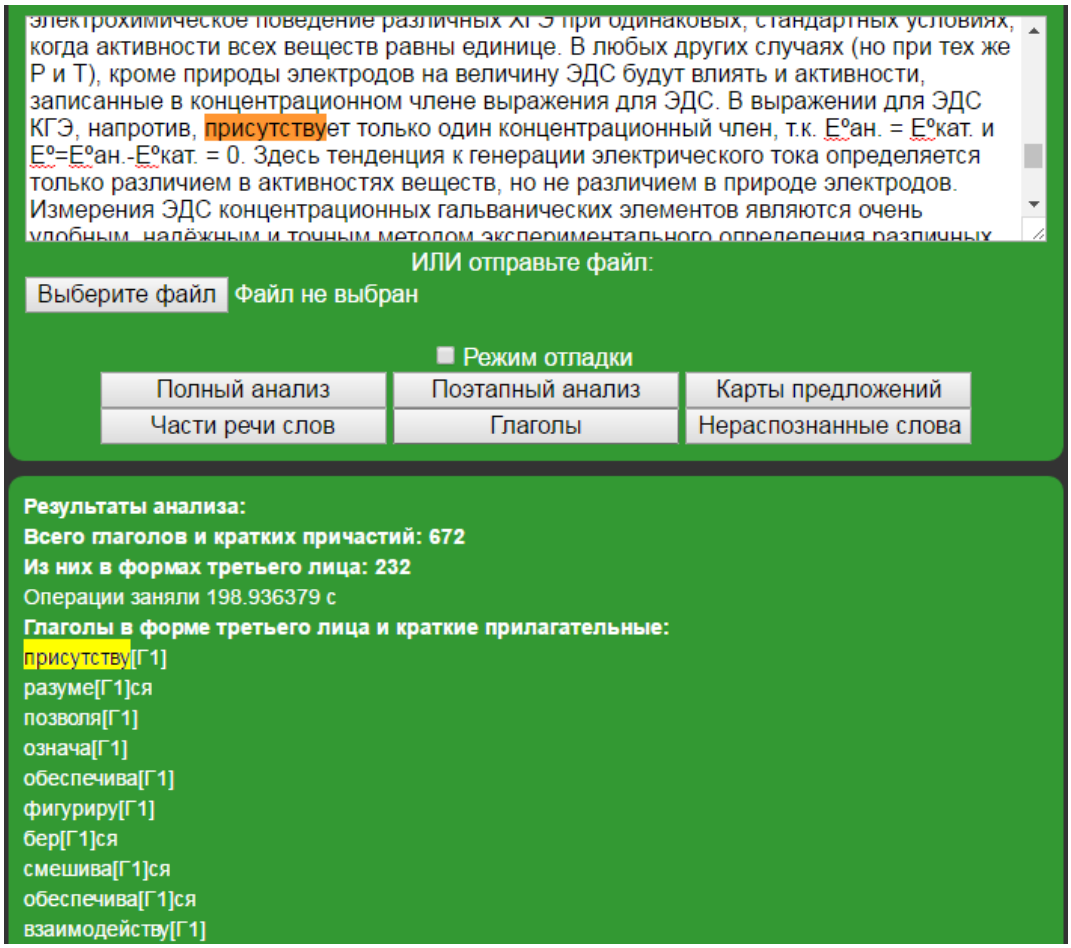

Рисунок 3 – Демонстрация работы модуля обнаружения глаголов

Как видно из рисунка, разработанные функции генератора семантической сети в первую очередь обнаружили наиболее распространенные средства выражения семантических связей – глаголы I спряжения в форме третьего лица. Данный набор средств выражения семантических связей позволяет создать пары из утвердительного и вопросительного шаблонов, которые затем будут использоваться при разборе предложений исходного текста и при распознавании задаваемых вопросов.

Список литературы

1. Жук Я.А., Гурин Н.И. Диалоговый модуль обработки семантической сети обучающей системы // Дистанционное обучение – образовательная среда XXI века: материалы VIII междунар. науч.-метод. конф. (Минск, 5–6 декабря 2013 года). – Минск: БГУИР, 2013. – С. 250–251.

2. Большаков И. А., Большакова Е. И. Автоматический морфоклассификатор русских именных групп // Компьютерная лингвистика и интеллектуальные технологии: материалы междунар. конф., Бекасово, 30 мая – 3 июня 2012 г. В 2 т. Т. 1: Основная программа конференции. / РГГУ. М., 2012. – С. 81–92.

3. Сегментация текста в проекте «Открытый корпус» / В. В. Бочаров [и др.] // Компьютерная лингвистика и интеллектуальные технологии: материалы междунар. конф., Бекасово, 30 мая – 3 июня 2012 г. В 2 т. Т. 1: Основная программа конференции. / РГГУ. М., 2012. – С. 51–60.