СНЯТИЕ ГРАММАТИЧЕСКОЙ ОМОНИМИИ В ТЕКСТЕ С ПОМОЩЬЮ СТАТИСТИЧЕСКИХ МЕТОДОВ

Максим КРЫГИН, Владислав ШКУРКО,
Ольга АФАНАСЬЕВА, Киев, Украина
Украинский языково-информационный фонд НАН Украины

Аннотация. *Работа посвящена статистическому подходу к снятию грамматической омонимии в текстах.*
Ключевые слова: *грамматическая омонимия, учебная выборка, цепь Маркова, вероятность перехода, грамматическая разметка текста.*

Abstract. *The article devoted to statistical approach to grammatical disambiguation in the texts.*
Keywords: *grammatical homonymy, training sample, Markov chain, transition probability, grammatical markup of text*

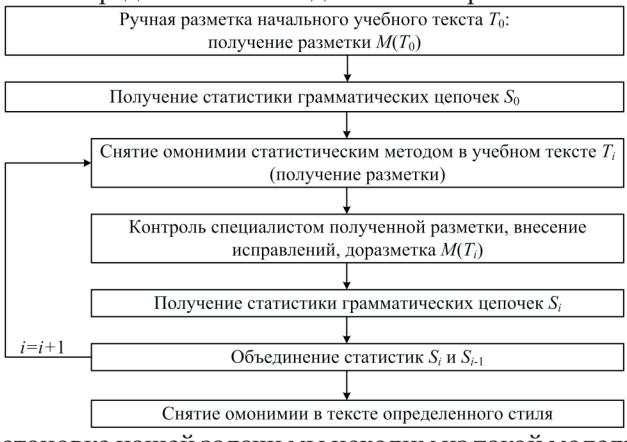
Современные информационные технологии все больше ориентируются на естественный язык, что предусматривает развитие методов и формализованных средств, работающих с процессами понимания и генерации связного текста. Одним из существенных препятствий на пути создания таких средств является присущая языковой системе разнотипная многозначность языковых единиц и особенно – явление грамматической омонимии. Поэтому в современной прикладной лингвистике и лингвистической технологии чрезвычайно актуализируется исследование языковых неоднозначностей и разработка эффективных формализованных методов их анализа и устранения.

Целью нашей работы является разработка методологии автоматического определения грамматических характеристик словоформ в тексте, которые служат маркерами грамматических неоднозначностей и в том числе – грамматической омонимии. Понятие омонимии широко освещено в работах лингвистов. Воспользуемся определением омонимов, которое приведено в энциклопедии «Украинский язык» [2]: омонимы – слова или их отдельные грамматические формы, а также устойчивые словосочетания, морфемы, синтаксические конструкции, которые при одинаковом звучании (или написани) имеют абсолютно разные значения (в отличие от полисемии). Для решения ряда практических задач нашей работы мы рассматриваем грамматические омонимы, к которым относим разные грамматические формы одного и того же слова или разных слов, которые имеют одинаковое написание.

Именно грамматическая омонимия является серьезным препятствием на пути создания грамматически размеченного корпуса текс-

тов. Например украинское слово *мати* имеет одинаковое написание в формах имени существительного женского рода (*мать*), имени существительного женского рода единственного числа, мужского рода множественного числа и неопределенной формы глагола; слово *зображення* – в формах существительного среднего рода единственного числа, именительного, родительного или винительного падежа, а также именительного и винительного падежей множественного числа; слово *прав* – в формах существительного среднего рода родительного падежа множественного числа (от *право*), глагола несовершенного вида повелительного наклонения второго лица единственного числа (от *править*) и глагола несовершенного вида прошедшего времени мужского рода единственного числа (от *прати*).

Схематически предлагаемый метод можно изобразить таким образом:



При постановке нашей задачи мы исходим из такой модели текста:

$$T_i=\{(w_1)r_1(w_2)r_2(w_3)\ldots(w_N)\},$$

где w_i – словоформы, r_i – разделители между словоформами – знаки пунктуации, пробелы и др., N – количество словоформ в тексте T_i.

Каждое слово текста T мы представляем в параметрическом виде $w_i = (v_i, g_i)$, где v_i задает часть речи для словоформы w_i, а g_i – грамматическое значение, т. е. текст приобретает размеченный вид: $M(T)=\{(v_1, g_1)\ (v_2, g_2)\ (v_3, g_3)\ldots(v_N, g_N)\}$. Преобразование $M : T \quad M(T)$ будем называть разметкой текста T.

Пример: *T = Людина повинна мати у серці велику любов до світу.*

w_1=Людина	w_2=повинна	w_3=мати	w_4=у	w_5=серці	w_6=велику	w_7=любов	w_8=до	w_9=світу
v_1=*іменник,* g_1= *одина, жіночий рід, називний відмінок;*	v_2= *прикметник* g_2=*жіночий рід, однина*	v_3=*дієслово недоконаного виду,* g_3=*інфінітив;*	v_4=*прийменник* g_4=*незм.*	v_5=*іменник,* g_5=*однина, середній рід, місцевий відмінок*	v_6=*прикметник,* g_6=*однина, жіночий рід, знахідний відмінок*	v_7=*іменник,* g_7= *однина, жіночий рід, знахідний відмінок*)	v_8=*прийменник;* g_8=*незм.*	v_9=*іменник,* g_9= *однина, чоловічий рід, родовий відмінок;*

Предложенный нами метод снятия грамматической омонимии предусматривает морфологический анализ текста алгоритмом лематизации [3], в результате применения которого для каждого слова получаем грамматические характеристики, которые вследствие явления грамматической омонимии являются однозначными далеко не для всех слов, т.е. в общем случае каждое слово *w* характеризуется вектором грамматических состояний (*v*, *g*):

w_1=Людина	w_2=повинна	w_3=мати	w_4=у	w_5=серці	w_6=велику	w_7=любов	w_8=до	w_9=світу
$(v_1, g_1)=\{(v_1^1, g_1^1), (v_1^1, g_1^1)\}$	$(v_2^1, g_2^1)=(v_2^1, g_2^1)$	$(v_3^1, g_3^1)=\{(v_3^1, g_3^1), (v_3^2, g_3^1), (v_3^2, g_3^3), (v_3^3, g_3^3)\}$	$(v_4^1, g_4^1)=(v_4^1, g_4^1)$	$(v_5^1, g_5^1)=(v_5^1, g_5^1)$	$(v_6^1, g_6^1)=(v_6^1, g_6^1)$	$(v_7^1, g_7^1)=\{(v_7^1, g_7^1), (v_7^1, g_7^1)\}$	$(v_8^1, g_8^1)=\{(v_8^1, g_8^1), (v_8^2, g_8^1)\}$	$(v_9^1, g_9^1)=\{(v_9^1, g_9^1), (v_9^1, g_9^2)\}$
v_1^1=*іменник*; g_1^1=*однина, жіночий рід, називний відмінок;* v_1^2=*прикметник, присвійний,* g_1^1=*однина, жіночий рід, називний відмінок*)	v_2^1=*прикметник* g_2^1=*жіночий рід, однина*	v_3^1=*дієслово недоконаного виду,* g_3^1=*інфінітив;* v_3^2=*іменник;* g_3^1=*однина, жіночий рід, називний відмінок;* g_3^3=*іменник, чоловічий рід, множина, називний відмінок;* g_3^3=*іменник, чоловічий рід, множина, знахідний відмінок*)	v_4^1=*прийменник* g_4^1=*незм.*	v_5^1=*іменник;* g_5^1=*однина, середній рід, місцевий відмінок*	v_6^1=*прикметник,* g_6^1=*однина, жіночий рід, знахідний відмінок*	v_7^1=*іменник;* g_7^1=*однина, жіночий рід, називний відмінок;* g_7^2=*однина, жіночий рід, знахідний відмінок*)	v_8^1=*прийменник;* g_8^1=*незм.* v_8^2=*іменник;* g_8^1=*незм.*	v_9^1=*іменник;* g_9^1=*однина, чоловічий рід, родовий відмінок;* g_9^2=*однина, чоловічий рід, давальний відмінок*

Морфологический анализ текста является отображением M': T $M'(T)$. Разметка $M'(T)$ является неоднозначной и содержит информацию про все теоретически возможные грамматические значения словоформы. Нашей задачей является максимально возможное приближение разметки $M'(T)$ до $M(T)$.

Параметры разметки текста $M(T)$ образуют n-связную цепь Маркова (ЦМ) [1], элементами которой выступают грамматические состояния словоформ (*v*, *g*). Для изучения поведения этой цепи и нахожде-

ния вероятностей перехода мы создаем учебную выборку полностью грамматически размеченных текстов. Затем тексты этой выборки подвергаются статистическому анализу, в результате которого получаем вероятности перехода для грамматических состояний словоформ: $\{p([v_{i+2}=v^1, g_{i+2}=g^1])/([v_i=v^2, g_i=g^2],[v_{i+1}=v^3, g_{i+1}=g^3])$, v^1, v^2, v^3 – пробегают все множество частей речи, а g^1, g^2, g^3 – все множество возможных грамматических состояний для конкретной части речи$\}$.

$$i:(v_i,g_i)(v_{i\ 1},g_{i\ 1})(v_{i\ 2},g_{i\ 2}) \qquad \begin{matrix}(v_i^1,g_i^1)(v_{i\ 1}^1,g_{i\ 1}^1)(v_{i\ 2}^1,g_{i\ 2}^1)\\(v_i^1,g_i^1)(v_{i\ 1}^1,g_{i\ 1}^1)(v_{i\ 2}^1,g_{i\ 2}^2)\\(v_i^1,g_i^1)(v_{i\ 1}^1,g_{i\ 1}^1)(v_{i\ 2}^2,g_{i\ 2}^1)\\(v_i^1,g_i^1)(v_{i\ 1}^1,g_{i\ 1}^2)(v_{i\ 2}^1,g_{i\ 2}^1)\\ \dots \quad \dots \quad \dots\\(v_i^m,g_i^n)(v_{i\ 1}^l,g_{i\ 1}^q)(v_{i\ 2}^e,g_{i\ 2}^r)\end{matrix}$$

выбирается та последовательность грамматических состояний, вероятность перехода для которой больше. Результаты применения нашего метода к конкретным текстам представлены в таблице:

Текст	Учебная выборка	Количество словоформ в тексте			Распознано омонимичных словоформ			Нераспознано омонимичных словоформ
		Всего уникальных	Омонимичных	Однозначных	Всего	Неверно	Верно	
Конституция Украины	-	14131	10263 (72,63%)	3868 (27,37%)	-	-	-	-
Хозяйственный кодекс	Конституция Украины	51982	35161 (67,64%)	16821 (32,36%)	526 (1,78%)	321 (61%)	205 (39%)	34,635 (98,22%)
Семейный кодекс	Конституция Украины + Семейный кодекс	23524	15819 (67,25%)	7705 (32,75%)	11273 (71%)	1853 (16,44%)	9420 (83,56%)	4546 (28,74%)

Таблица. Результаты автоматического снятия омонимии.

Эксперименты по созданию учебной выборки и снятию омонимии показали положительную динамику результатов, что позволяет рассчитывать на высокий процент точности снятия грамматической омонимии при условии накопления адекватной учебной выборки для разных стилей текстов.

ЛИТЕРАТУРА

1. *Вентцель Е. С.* Теория случайных процессов и ее инженерные приложения [Текст] / Вентцель, Е.С., Овчаров Л.А. – М.: Наука, 1991. – С. 98–127.
2. «Українська мова». Енциклопедія / Редкол.: Русанівський В.М., Тараненко О.О. (співголови), М.П.Зяблюк та ін. – К.: Укр. енцикл., 2000. – 752 с.
3. *Шевченко И. В.* Электронный грамматический словарь украинского языка [Текст] / Шевченко И.В., Рабулец А.Г., Широков В.А. // Труды Международной конференции «Megaling'2005. Прикладная лингвистика в поиске новых путей». 27 июня – 2 июля 2005 года. – Меганом, Крым, Украина. – 2005. – С. 124–129.