УДК 004.9

АНАЛИЗ ЛЕКСИЧЕСКИХ ПАР ДЛЯ РЕШЕНИЯ ПРОБЛЕМЫ АВТОМАТИЧЕСКОЙ ГЕНЕРАЦИИ ТЕМАТИЧЕСКИ СВЯЗНОЙ РЕЧИ

Васильев К.О., Личаргин Д.В., Щурова А.В.
научный руководитель канд. тех. наук Личаргин Д.В.
Сибирский федеральный университет

В работе рассматривается проблема формирования корректного и осмысленного текста посредством использования программных систем, а именно проблема формального представления ассоциативных переходов между предложениями и фрагментами текстов на естественном языке. Основной идеей решения этой проблемы является формализация и оценка расстояния между парами слов естественного языка как пар векторов многомерного семантического пространства слов естественного языка. Определяются семантические координаты лексического и грамматического пространства слов, пар слов и предложений естественного языка. Приводятся примеры различных типов лексико-грамматических отношений между словами естественного языка. Рассматривается дерево генерации синонимичных предложений на основе выделения темы, ремы, связки, модальности и других уровней генерации осмысленных фраз естественного языка.

На сегодняшний день порождение (синтез) речи компьютером является, безусловно, важной проблемой. В данной области широко распространены и разрабатываются разнообразные системы формирования высказываний и обработки естественного языка, а также языковых баз данных различными программными системами: экспертными системами, программами электронного перевода, «ботами» (системами диалога с пользователем), синонимизаторами, программами генерации текстов по тематике «прогноз погоды», «технический справочник» и т.п.

Проблема генерации осмысленной речи исследуется со времен появления вычислительной техники и широко исследуется различными авторами, в частности Э. Кодда, А. Хомского, А.С.Нариньяни, Т. Винограда, М.В. Никитина, К. Шеннона, и даже задолго до появления компьютерной техники (Машина Луллия и др.).

Важными проблемами являются проблемы перевода, машинного перевода, построения экспертных систем, естественно-языковых интерфейсов и др. Для решения этих проблем используются различные средства и методы: метод резолюций, мультииерархические системы параллельного разбора (грамматики, семантики, морфологии, фонетического членения предложения и других единиц языка), реляционные, многомерные и иерархические базы данных и многие другие.

Решение задач семантики, дискретной математики, лингвистики и искусственного интеллекта направлены на прохождение теста Тьюринга с все более жесткими условиями, включающими в себя широкий набор слов, конструкций, фактов и эмуляции отношения к предмету разговора со стороны собеседника или выступающего.

Рассмотрим многомерное пространство объектов естественного языка: слов и выражений. Многие словосочетания могут быть сформированы правильно относительно грамматики, но не иметь семантического смысла. Допустим, фраза «See I» грамматически построена неверно, а фраза «I eat a hat» грамматически корректна, но не имеет семантического смысла, а фраза «I eat a pear» верна и в грамматическом, и в семантическом смысле.

Возможно построение многомерной грамматической базы данных со следующими координатами вектора понятийного описания:

G_1 = Части речи {«Артикль», «Прилагательное», «Существительное»,«Глагол», ...};

G_2 = Члены предложения {«Определитель», «Определение», «Подлежащее», «Сказуемое», …};

$G_{3,3,1}$ = Лица {«1-ое», «2-ое», «3-ее», «Не определено»};

$G_{3,3,2}$ = Аспект {«Неопределенный», «Продолженный», «Совершенный», «Совершенный продолженный», «Не определен»};

$G_{3,1,1}$, $v_{3,1,2}$, … – Другие размерности, выраженные грамматическими категориями.

Далее, определим лексическое пространство языка (лексический куб) со следующими координатами:

S_1 = Порядок слов {Исполнитель, Действие, Реципиент, Получатель, Место, Время, Инструмент, Метод};

S_2 = Тема {Еда, одежда, тело, здание, группа людей, транспорт, ...};

S_3 = Варианты замены слов в предложении {to cook, to boil, to roast, to fry, to bake, …, to eat, to chew, …} (см. рисунок 1).

Все грамматические конструкции располагаются в ячейках многомерного массива данных – многомерного пространства слов языка. Координаты вектора, такие как, например, V[Глагол / Признак / Совершенный, ...], определяют ячейку с грамматической конструкцией «having + ГЛАГОЛ + -(e)d». Вектор V[Прилагательное / Предикат / Первое лицо, Превосходная степень, длинное прилагательное, ...] определяет конструкцию «am the most + ПРИЛАГАТЕЛЬНОЕ». Реляционные таблицы, как часть этого многомерного массива, представлены в лингвистике в форме традиционных грамматических парадигм.

В отличие от популярной в традиционной дисциплине «обработка естественного языка» статистической модели языка, в которой вероятность языковых выражений определяется на основе Марковских процессов и других вероятностных и статистических методов и их применения к анализу корпусов текстов на естественном языке, рассматриваемая модель определяет язык как векторизованное пространство, или иначе, векторизованные классификации. Приведем несколько примеров такого подхода, составляющего общий контекст исследования отношений между парами слов естественного языка. M(«модель естественного языка»)[L(«уровень предложения»), S(«лексика»), G(«грамматика») [O(«порядок слов и члены предложения») {субъект, предикат, объект}, T(«объекты по тематике изучения») {идеи {науки, представления, чувства …}, предметы {одежда, еда, части тела, здания, транспорт, …}, существа, …}, V(«варианты подстановок слов в предложение») {позитивное{обожать, любить,…}, негативное {не любить, ненавидеть,…},…}], N(«функции предложения над точками слов»)].

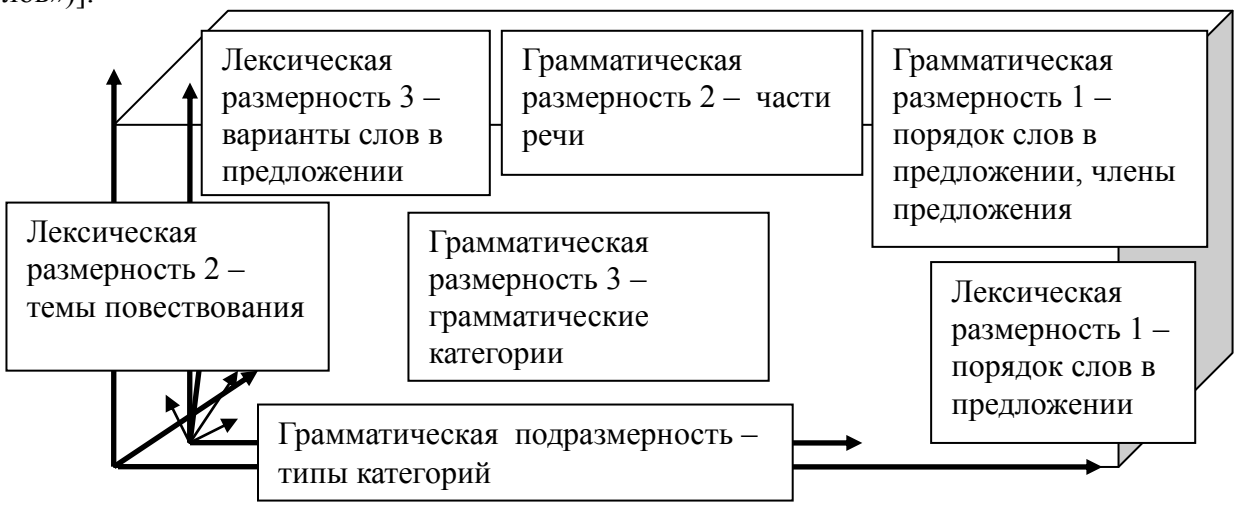

Рис. 1. Координаты многомерного лексико-грамматического подпространства леса данных естественного языка

Такое многомерное пространство включает комбинаторно сочетающиеся группы слов, например, группа слов {носить, одевать, снимать, гладить, шить, …} относится к ячейке многомерного пространства M(«модель языка»)[G(«грамматика»)[«отношение-существо- объект предмет», «одежда»; «глагол», «предикат», «неопределенная форма»]]. Пример, подстановочной таблицы, как среза многомерного понятийного пространства слов естественного языка, приводится ниже.

Для решения проблемы анализа отношений пар слов и предложений рассматриваются следующие разделы модели естественного языка на основе леса классификаций. M(«модель естественного языка»)[L(«уровень пар слов»), S(«семантика»)[«объект», «одежда»]+ S(«семантика»)[«объект», «устройство» ; «действие», «над одеждой»]] ⊇ {«кепка – стиральная машина», «свитер – швейная машина», «кофта – утюг»}. Важно отметить, что рассматриваемое трехмерное лексико-семантическое пространство слов общей мультииерархической модели языка и его различные отображения на трехмерное грамматическое пространство слов той же модели дают возможность не просто выявлять осмысленные синтагматические отношения между словами, но и выявлять различного рода ассоциативные отношения между словами и их цепочки.

Рассмотрим принцип сведения переходов между предложениями к переходам между словами на основе парсинга в форме дерева актуального членения предложения.

Традиционно актуальное членение предложений включает в себя деление на тему и рему, рема является ключевым словом в предложении, а тема относится ко всему тексту или его фрагменту. На вершине дерева актуального членения предложения имеет место ключевое слово (рема); на втором уровне – тема и рема; на третьем имеет место четверка: тема, связка, рема, модальность; на четвертом уровне добавляются обстоятельства, имеющие важную уточняющую функцию. На пятом уровне имеют место очевидные, понятные из контекста обстоятельства и конкретизация; на шестом – полупустые слова, уточняющие аспекты слов, указанных выше в дереве разбора.

Например,

0. Тема повествования: «суп»;
1. Ключевое слово: «вкуснятина» = «вкусный»;
2. Тема-Рема: «суп – вкуснятина» = «суп – вкусный»;
3. Тема-Рема-Связка-Модальность: «суп-вкусным-вышел-классно (очень хорошо)»;
4. Важная конкретизация: «…вкусным и профессиональным»;
5. Контекстуальная конкретизация: «суп, который готовила Аня, …»;
6. Аспекты понятий: «впечатление от супа, …, это просто восторг от вкусняшки, профессиональной штуки…»;
7. Различные эквивалентные преобразования, например, двойное отрицание.

Таким образом, одну и ту же мысль, что суп вкусный, можно выразить астрономическим количеством более частных по смыслу и по форме фраз.

Таким образом, от модели траекторий в виде цепочек пар слов естественного языка, как точек многомерного пространства, можно перейти к соответствующей траектории ключевых слов как вершин деревьев генерации каждого из вариантов синонимичных фраз языка (см. рис. 2). Парсинг актуального членения предложения дает возможность выделить в предложении ключевое слово, тему и рему, тему-рему-связку-модальность и другие уровни. Данный парсинг отличается от грамматического парсинга и семантического анализа предложения. В связи с развитием электронного обучения, важным остается аспект применения генерации речи в обучающих системах.

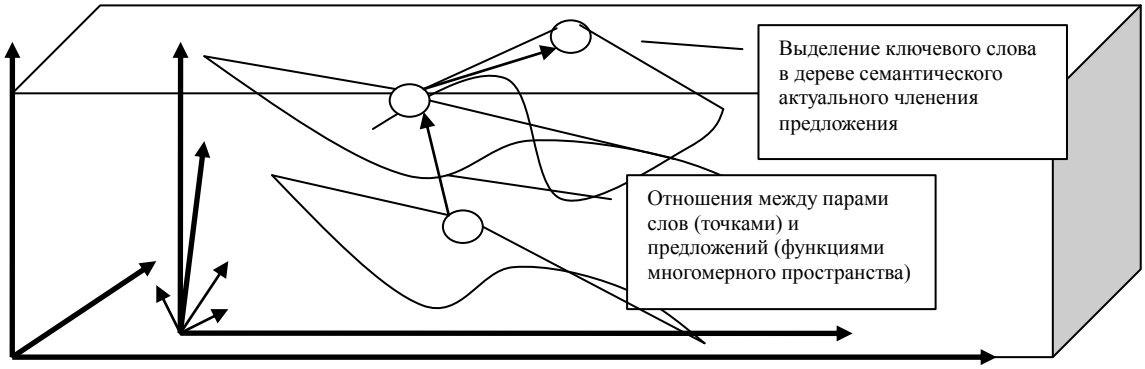

Рис. 2. Модель лексико-грамматического пространства

Таким образом, траектория движения ключевых слов в предложениях некоего текста может соответствовать цепочкам пар слов и соответствующих векторов слов естественного языка в многомерном семантическом пространстве, что дает возможность осуществлять генерацию повествований с «тематическим скольжением» на основе классификации пар ассоциативно связных слов языка.

В заключении необходимо отметить, что анализ пар слов естественного языка, как пар векторов многомерного семантического пространства дает возможность улучшить качество генерации текстов на основе, например, статистических методов порождения естественного языка.