УДК 004.9

К АНАЛИЗУ ФОРМУЛ СЛОЖЕНИЯ ЛЕКСИЧЕСКИХ ПАР НА ОСНОВЕ КОЭФФИЦИЕНТОВ СЕМАНТИЧЕСКИХ ЕДИНИЦ ЯЗЫКА

Полякова О.С., Личаргин Д.В., Щурова А.В., Подлесный А.О.
научный руководитель канд. тех. наук Личаргин Д.В.
Сибирский федеральный университет

С каждым днем население развитых стран всё больше вовлечено в работу с программным обеспечением. Компьютерная техника проникла во все сферы человеческой деятельности. С помощью этой техники решаются все более сложные задачи. Лингвистические программные системы должны формировать корректный и осмысленный перевод текста. Отсюда вытекает проблема формального представления ассоциативных переходов между предложениями и фрагментами текстов на естественном языке. Здесь затрагиваются более частные проблемы таких наук, как философия, психология, физиология, информатика и других и, в первую очередь, лингвистики.

Лексические и стилистические особенности языка и речи давно и широко рассматриваются различными авторами, в частности Ю.Д. Левиным, А.В. Федоровым, В.С. Виноградовым, J. Catford, J. Holmes, P. Newmark, L. Kelly и другими. В работах М.А.К. Хэллидэя доказывается сопоставимость различных единиц двух языков и ставится вопрос об их формальной эквивалентности.

Однако вопрос об информационной обработке текстов, о получении осмысленного и полноценного текста требует дополнительных исследований в рамках семантического подхода.

Цель данной работы состоит в построении дерева генерации последовательностей предложений на основе выделения темы, ремы, связки, модальности и других уровней генерации осмысленных фраз естественного языка.

Задачи данной работы заключаются в:

1. Измерении метрического расстояния между парами слов;
2. Построении ключевых слов в виде дерева лексических пар.

Основная идея работы состоит в рассмотрении пар слов естественного языка с точки зрения пары векторов многомерного семантического пространства и в анализе проблемы измерения метрического расстояния между парами слов.

Новизна работы обусловлена важностью компонента осмысленности переходов между предложениями при генерации осмысленной в целом речи.

Важной составляющей проблемой при создании программной системы генерации осмысленных повествований является задача моделирования естественного языка. Решением проблемы может быть как анализ фраз, так и генерация самих фраз и текстов на основе учета пар слов как векторов признаков многомерного пространства слов – точек и предложений – функций на этом пространстве. А инструментом, позволяющим осуществить решение данной проблемы, является формализация семантики языка на основе соответствующих математических моделей.

Семантика изучает смысловое значение единиц языка. Особенностями семантики объясняются лингвистические сложности перевода компьютерными программами с одного языка на другой. Семантическая теория перевода должна позволить вычислять параметры единиц языка: переводческой эквивалентности, качества, адекватности и т.п.

Существуют различные типы пар слов и их отношений.

1. В частности, рассмотрим омонимические отношения между словами. Например: bank – 1) берег (реки), 2) банк (в котором хранятся деньги). В данном случае омонимы – это различные значения одного и того же слова (не связанные ассоциацией), которые обычно рассматриваются как абсолютно разные слова с общим написанием.

2. В следующем примере приведем понимание омофонии, то есть явления, когда слова звучат одинаково, но пишутся по-разному: meat [mi:t] – мясо, meet[mi:t] – встречать; whole – hole; knew – new.

3. Типы синонимии: мерзкий – отвратительный;

4. Антонимия также имеет место в отношениях между словами: хороший – плохой.

5. Эквонимы тесно связаны с явлениями гипонимов. Эквонимы определяются как слова одного уровня обобщения при общем гипериониме. Например: эквонимами по отношению к друг другу являются слова «бабушка» и «дедушка», «мать» и «отец».

6. Гипонимы – слова видового, более специального значения по отношению к слову более обобщенного смысла. Гиперонимы представляют собой имя родового понятия. Например: слова «мать» и «отец» в свою очередь являются гипонимами по отношению к слову «родитель».

7. В свою очередь «родитель» для этой же пары слов будет являться гиперонимом. Семантика гиперонимов является более объёмной, чем семантический план эквонимов, поэтому в рамках семантического плана гиперонимов объединяются значения двух или более самостоятельных слов: «родитель» – «мать», «компьютер» – «ноутбук».

В области семантики определен ряд задач, в частности, измерение семантических расстояний является одной из них. Измерить семантическое расстояние означает: оценить плотность семантических и ассоциативно-семантических связей между словами и понятиями словаря, между единицами текста и с точки зрения более сложных задач – между фрагментами текста. Для получения количественной оценки плотности семантической связи необходимы знания о природе отношений, о типах единиц – терминов отношений, а также выбор исследовательского инструментария, такого как OLAP технологии, инструментарий многомерных баз данных и векторного представления данных.

Возможно построение многомерной грамматической базы данных со следующими координатами вектора понятийного, что описано в работах Д.В. Личаргина:

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, 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, …}.

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. Необходимо численно задать расстояния в рамках многомерного пространства отношений между словами в рамках семантической классификации слов и понятий (точек семантического пространства) языка. Рассмотрим сложение лексических пар в пространстве семантических состояний с учетом метрики семантического пространства.

Таблица 1.

Распределение коэффициентов по отдельным типам слов языка

Тип отношения	**Значение в баллах от 0 до 1**	**Примеры**
Слово само с собой	1	To Jump – To Jump
«Радикальные» антонимы	0,95	All – No
«Умеренные» антонимы	0,9	Many – Few
Эквонимы	0,85	All – Some, Son – Daughter
Гиперонимы и гипонимы	0,8	Child – Son
Дефинонимы	0,75	Café – To Eat – Food – Eater
Другие близкие отношения: часть-целое и другие	0,7	Car – Transmission
Социально-обусловленные сближения значений	0,65	Love – Flower
Сдвиг по частям речи	0,6	To Eat – Eating – Meal
Сдвиг по членам предложения	0,5	To eat is to chew, I've bought it to eat.
Сдвиг по категориям	0,4	Café – Cafes.

Введем обозначение: MSSimilarity – Minimal Semantic Similarity, Минимальное семантическое расстояние.

MSSimilarity(0, Professor, Professor).

MSSimilarity(0, Number, Number).

MSSimilarity(0, ANYWORD1, ANYWORD1).

MSSimilarity(0.45, Professor, Number) ≥ MSSimilarity(0.6, Professor, Maths) + MSSimilarity(0.75, Maths, Number).

MSSimilarity(0.6, Professor, Maths) ≥ MSSimilarity(0.75, Professor, Science) + MSSimilarity(0.8, Science, Maths).

MSSimilarity(0.27, Many, Professor) ≥ MSSimilarity(0.6, Many, Number) + MSSimilarity(0.45, Professor, Number).

MSSimilarity(0.16, Multiple, Professor) ≥ MSSimilarity(0.6, Multiple, Many) + MSSimilarity(0.56, Professor, Academician) ≥ MSSimilarity(0.75, Professor, Science) + MSSimilarity(0.75, Academician, Science).

Возникает математическая задача – посчитать значения минимальных семантических расстояний для подобных слов и их цепочек для любых слов и цепочек слов. Постановка эксперимента предполагает оценку экспертами семантического расстояния в баллах и их сравнении с, принятыми на основе приводимой ниже таблицы, значениями. Далее, рассмотрим примеры деревьев семантико-грамматических пар слов.

Тема: «Мясо», Рема: «Повара», Связка: «Делает», Модальность: «По-разному»;

1. Тема: «Курица», Рема: «Мама», Связка: «Делает», Модальность: «С удовольствием»;

1.1. Тема: «Курица», Рема: «Мама», Связка: «Делает», Модальность: «Хорошо»;

1.2. Тема: «Курица», Рема: «Мама», Связка: «Училась делать», Модальность: «Часто»;

1.2.1. Подтема: «Блюда», Рема: «Мама», Связка: «Училась делать», Модальность: «Отлично»\ «Классно»;

2. Подтема: «Мясо», Рема: «Мама», Связка: «Делает», Модальность: «Отлично»\ «Классно\Вкусно»;

2.1. Тема: «Куропатка», Рема: «Я», Связка: «Есть», Модальность: «С удовольствием»;

2.2. Тема: «Куропатка», Рема: «Ресторан», «Связка»: Готовят: «Хуже»;

2.2.1. Тема: «Куропатка», Подрема: «Официант», Связка: «Сервирует», Модальность: «Хорошо»/ «Красиво»;

2.2.1.1. Подтема = Подрема: «Официант», Рема: «Вежливый», Связка: «Является», Модальность: «Необычно»;

2.2.1.1.1. Подтема = Рема: «Вежливость»/ «Галантность», Рема: «Доход», Связка: «Обеспечивает», Модальность: «Всегда»;

2.2.2. Тема = Рема: «Ресторан», Рема: «Музыка», Связка: «Играет в», Модальность: «Хорошая»/ «Спокойная»;

В результате, мы получаем обход сгенерированного дерева пар слов, отдельно по теме и реме фразы, также необходимо учитывать тип связи темы и ремы и модальность. Так, например, на основе приведенного выше дерева можно сгенерировать фразы вида: «Моя мама любит готовить курицу. Она научилась так классно готовить у бабушки. Это просто класс, как мама делает мясо. В ресторане значительно хуже готовят куропатку, чем это делает наша мама. В ресторанах очень вежливо справляются со своей работой официанты. Они блестяще сервируют куропатку. Но, мама делает более вкусно. Хотя рестораны знамениты своей спокойной музыкой. Мне нравится классика…». Такие фразы можно генерировать с привлечением сленга и художественных оборотов. Так, 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).

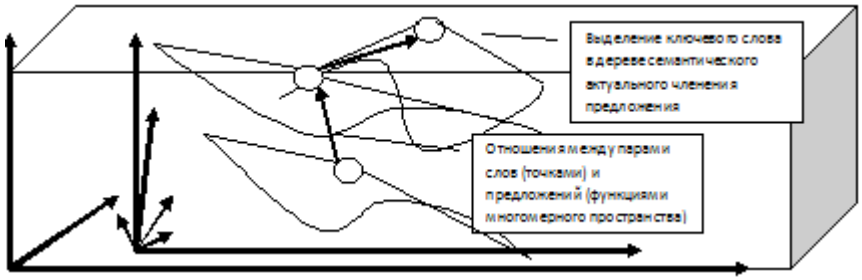

Рис. 2. Модель лексико-грамматического пространства

Парсинг позволяет сопоставить линейной последовательности слов естественного языка дерево разбора предложения на основе его формального лексико-грамматического представления. Предложение делится в контексте на исходную часть – тему (данное) и на то, что утверждается о ней – рему (новое). В некоторых случаях выделяется третий элемент – переходный элемент (связующий член). Часто он выражается глагольным сказуемым, содержащим временные и модальные показатели. Таким образом, траектория движения ключевых слов в предложениях определенного 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. Построение дерева генерации синонимичных предложений и измерение метрического расстояния между парами слов позволяет оптимизировать процесса построения итогового предложения.

В заключение необходимо отметить, что величина семантического расстояния играет большую роль при определении смысла с учетом контекста нескольких предложений. Вычисление этой величины является отправной точкой для моделирования языкового контекста, так как слово, употребленное в контексте определенного предложения, и его значение должно согласовываться со значениями слов, которые стоят рядом. Семантические значения слов в предложении должны создавать смысловое единство, поэтому значения концептов (и сем) слов, которые стоят рядом в предложении, должны быть семантически как можно ближе друг к другу.