УДК 004.9

АНАЛИЗ ЛЕКСИЧЕСКИХ ПАР ДЛЯ АВТОМАТИЧЕСКОЙ ГЕНЕРАЦИИ ДИАЛОГИЧЕСКОЙ И МОНОЛОГИЧЕСКОЙ РЕЧИ*

Д. В. Личаргин[1], А. В. Щурова[1], Е. А. Курбатова[1], И. В. Колбасина[2]

[1]Сибирский федеральный университет
Россия, 660074, Красноярск, ул. Академика Киренского, 28
[2]Сибирский государственный аэрокосмический университет имени академика М. Ф. Решетнева
Россия, 660014, Красноярск, просп. им. газ. «Красноярский рабочий», 31. E-mail: lichdv@hotmail.ru

Рассматривается подпроблема проблемы 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 подпроблемы является формализация и оценка расстояния между парами слов естественного языка как между парами 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.

Ключевые слова: искусственный интеллект, обработка естественного языка, тест Тьюринга, генерация осмысленных текстов, математическая семантика.

LEXICAL PAIRS ANALYSIS FOR AUTOMATIC GENERATION OF DIALOGUE AND MONOLOGUE SPEECH

D. V. Lichargin[1], A. V. Shchurova[1], E. A. Kurbatova[1], I. V. Kolbasina[2]

[1]Siberian Federal University
28 Kirenskiy st., Krasnoyarsk, 660074, Russia
[1]Siberian State Airspace University named after academician M. F. Reshetnev
31 “Krasnoyarskiy Rabochiy” prosp., Krasnoyarsk, 660014, Russia. E-mail: lichdv@hotmail.ru

In the paper the sub-problem of the problem of correct and meaningful text formation by means of software application, and namely the problem of formal presentation of association transfers between sentences and natural language texts fragments is considered. The main idea in solving the sub-problem is to formalize and estimate the distance between natural language word pairs as pairs of vectors of multidimensional semantic space of natural language words. The semantic coordinates of lexical and grammatical space of words, word pairs and sentences of the natural language are determined. Examples of different types of lexical and grammatical relations between natural language words are offered. The tree of synonymic sentences generation is considered based on the determination of the theme, the rheme, the link, the modality and other levels of natural language meaningful phrases generation.

Keywords: Artificial Intelligence, Natural Language Processing, Turing Test, Meaningful Texts Generation, Mathematical Semantics.

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 является актуальной в связи с важностью развития систем взаимодействия человека и компьютера на основе естественного языка (естественно-языкового интерфейса) и потребностью в формировании заданного множества осмысленных текстов различного рода с использованием соответствующих программных приложений.

Проблема генерации осмысленной речи исследуется со времен появления вычислительной техники и широко исследуется различными авторами, в частности Э. Кодда, А. Хомским, А. С. Нариньяни, М. В. Никитиным, К. Шенноном, А. И. Пиотровским, и даже задолго до появления компьютерной техники (машина Луллия и др.).

*Исследование поддержано Министерством образования и науки Российской Федерации, соглашение 14.B37.21.1010.

Важными проблемами являются проблемы перевода [1; 2], машинного перевода, построения экспертных систем, естественно-языковых интерфейсов и др. Для решения этих проблем используются различные средства и методы: метод резолюций, древесный парсинг предложения, мультииерархические системы параллельного разбора (грамматики, семантики, морфологии, фонетического членения предложения и других единиц языка), объектное представление и фреймы, реляционные, многомерные и иерархические базы данных, онтологии, семантические классификации, семантические сети и многие другие.

Кроме того, особого рассмотрения требует проблема анализа семантических пар слов языка, что может позволить генерировать ассоциативно связанные диалоги и монологи на естественном языке.

Цель данной работы состоит в том, чтобы дать анализ лексических пар (слов и предложений) для генерации диалогической и монологической речи.

Задачи данной работы заключаются:

– в анализе классификации слов и смысловых понятий английского языка для ее последующего использования в качестве основы для генерации осмысленного подмножества языка;

– анализе взаимосвязей между словами и выражениями в английском языке: их пар как векторов многомерного пространства слов языка и траекторий слов и предложений как цепочки или системы векторов.

Основная идея работы состоит в построении модели естественного языка на основе многомерного представления слов и пар слов языка и в применении этой модели для решения проблемы генерации ассоциативных переходов в диалогической и монологической речи.

Новизна данной работы состоит в формальном представлении ассоциативных переходов между словами и предложениями как траекторий (функций) в многомерном семантическом пространстве [3], заданном векторами признаков семантической классификации.

Современный уровень разработки в этой области характеризуется многими в определенном смысле не до конца успешными попытками создания систем генерации осмысленной речи на более или менее широких подмножествах естественного языка. В частности, на основе корпусов текстов, данных социальных сетей и отладки семантических сетей с логическими переходами вида «вопросы – ответы», а также «вопрос 1 – вопрос 2 – …», имеются различные реализации решения этой проблемы в некоторых приближениях (программа Alice, обучающая система «Робот Джордж» и др.).

Решение задач семантики, дискретной математики, лингвистики и искусственного интеллекта направлено на прохождение теста Тьюринга со 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 при этом не иметь семантического смысла. Допустим, фраза «See I» грамматически построена неверно, 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.

Ниже приводится пример учета комбинаторики слов естественного языка, представленного в форме подстановочной таблицы, способной генерировать осмысленные фразы на английском языке (табл.1).

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-е», «2-е», «3-е», «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 предложениям {to cook, to boil, to roast, to fry, to bake, …, to eat, to chew, …} (рис. 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.

Таблица 1

Принцип генерации осмысленных фраз естественного языка методом подстановки

the ... этот …	of the ... этого …	is over закончится	now сейчас
series серия	game игра	is left осталось	at the present moment в настоящем
season сезон	tournament турнир	starts начинается	today сегодня
cycle цикл	Olimpic games Олимпийские игры	goes on продолжается	this week на этой неделе

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

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 представляет язык как векторизованное пространство векторизованных классификаций.

Приведем несколько примеров такого подхода [3–5], 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»)].

Такое многомерное пространство включает в себя 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.

В качестве пояснения места анализа отношений пар слов в общей модели естественного языка как леса классификаций рассмотрим группу слов {кофта, носки, куртка, майка, фартук, …}, которая относится к ячейке многомерного массива M(«модель естественного языка»)[L(«уровень слова»), G(«грамматика»)S («семантика»)[«объект», «одежда», «существительное», «субъект», «единственное число»]]. Обе группы слов образуют синтагматические пары вида M(«модель естественного языка»)[L(«синтагмы»), G(«грамматика»)S («семантика»)[«объект», «одежда»; «существительное», «субъект», «единственное число»] + [«действие с объектом», «одежда»; «глагол», «предикат», «неопределенная форма»], N(«функция двух аргументов»)]: «носить кофту», «гладить фартук», «шить носки», «снимать куртку» и т. п.

Грамматический порядок слов получает в соответствие семантические групп слов, в результате чего данное пространство становится критерием семантической и грамматической осмысленности речи. Функции определенного вида, определенной геометрии над данными группами слов с хорошей вероятностью образуют осмысленные фразы. Фрагменты этих функций представляют собой предложения осмысленного естественного языка. Функции предложений соответствуют гнездящимся деревьям уровня предложения. 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»]] $\supseteq$ {«кепка – стиральная машина», «свитер – швейная машина», «кофта – утюг»}.

Важно отметить, что рассматриваемое трехмерное лексико-семантическое пространство слов общей мультииерархической модели языка, т. е. модели на основе векторизованного леса лингвистических данных, и его различные отображения на трехмерное грамматическое пространство слов той же модели дают возможность выявлять не просто осмысленные синтагматические отношения между словами, но и различного рода ассоциативные отношения между словами и их цепочки (табл. 2).

Представленные обозначения F.G.L.Ex.S соответствуют уровням дерева шкалы вариантов подстановки слов в семантическом пространстве слов и предложений, а именно:

– Группы вариантов;
– Группа слов;
– Список слов по изменяющемуся признаку;
– Экземпляр списка;
– Стилистический вариант.

Тогда как обозначение D.A.P означает вложение уровней дерева в виде цепочек именных групп вида:

– Позиция объекта в предельно полном предложении;
– Атрибут объекта;
– Часть объекта;
– Атрибут части объекта.

Рассмотрим принцип сведения переходов между предложениями к переходам между словами на основе парсинга предложения в форме дерева актуального членения предложения с одним ключевым словом на вершине дерева парсинга.

Таблица 2

Возможные отношения между словами со стороны шестимерного лексико-грамматического пространства

Название лексического и грамматического отношения	Вектор многомерного пространства для слова 1	Вектор многомерного пространства для слова 2	Пример отношения
Различие в частях речи	G[«Verb», B, C] + S[D, E, F]	G [«Noun», B, C] + S[D, E, F]	Love – to love
Различие в грамматической категории	G [A, B, «Singular»] + S[D, E, F]	G [A, B, «Plural»] + S[D, E, F]	Fan's – fans'
Различие в теме	G [A, B, C] + S[D, E1 = «Food», F = «Make»]	G [A, B, C] + S[D, E2 = «Clothes», F = «Make»]	Cook – sew
Различие в объекте	G[A, B, C] + S[D, E, F.G.H]	G [A, B, C] + S[D, E, F.G.¬H]	Start > launch
Антонимы	G [A, B, C] + S[D, E, F.G1.L1.Ex1]	G [A, B, C] + S[D, E, F.G1.L1.Ex2]	To be born – to live – to die – to revive
Гиперонимы	G [A, B, C] + S[D, E, F…G….Ex]	G [A, B, C] + S[D, E, F…G]	Mother – Parent
Гипонимы	G [A, B, C] + S[D, E, F…G]	G [A, B, C] + S[D, E, F…G….Ex]	Parent – Mother
Дефиноним	G [A, B, C] + S[D1.0.0, E, F.G.L...]	G[A, B, C] + S[D2.0.0, E, F.G.L…]	Driver – Vehicle, Driver – To Drive, Driver – Route или A Cook – To Cook – Dish – Cooked – Recipe
Эмотивный синоним	G[A, B, C] + S[D, E, F.G.L.Ex.S]	G [A, B, C] + S[D, E, F.G.L.Ex. ¬ S]	Лицо – Ряха – Лик – Харя или Waste – Spend – Have (Time)
Аспект / часть	G[A, B, C] + S[D, E, F…]	G [A, B, C] + S[D.0.P, E, F…]	Traffic – Car – Wheel – Tyre
Аспект / атрибут	G[A, B, C] + S[D, E, F…]	G [A, B, C] + S[D.A, E, F…]	Car – Old – 15 years old, Car – Powerful – 100 house-powers

Традиционно актуальное членение предложений включает в себя деление на тему и рему, при этом рема является ключевым словом в предложении, а тема относится ко всему тексту или его фрагменту. Таким образом, на вершине дерева актуального членения предложения имеет место ключевое слово (рема); на втором уровне дерева парсинга имеют место тема и рема; на третьем уровне 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.

Приведем дополнительный пример генерации дерева синонимичных по контексту фраз. Например:

0. Тема повествования: «автомобиль».

1. Ключевое слово: «надежность».

2. Тема–Рема: «автомобиль – надежность» = «автомобиль – надежный».

3. Тема–Рема–Связка–Модальность: «автомобиль–надежным–сконструировали–профессионалы (хорошо)».

4. Важная конкретизация: «…надежным и функциональным».

5. Контекстуальная конкретизация: «автомобиль, который купил Петр, …».

6. Аспекты понятий: «оценка автомобиля, …, это является идеалом надежности, комфортабельного дизайна…».

7. Различные эквивалентные преобразования, например двойное отрицание: «…нисколько не опасен», «нельзя не заметить…».

Приведем дополнительные примеры: генерации последовательностей фраз на естественном языке.

1. Тема: Овощи; Рема: Разговор. Генерация предложения: Говорить об овощах → я говорю об овощах → я хочу сказать об овощах → овощи, это – то, о чем я хочу сказать (Первое предложение).

2. Тема: Овощи; Рема: Вкусно. Генерация предложения: Овощи вкусные → … → Присутствие свежих овощей завораживает отличным вкусом (Второе предложение).

3. Тема: Овощи; Рема: Пять часов. Генерация предложения: Овощи были в пять часов → Овощи съели в пять часов → Овощи исчезли в пять часов → Овощи исчезли с тарелок в пять часов → Ерунда, что овощи не исчезли с тарелок в пять часов (Третье предложение).

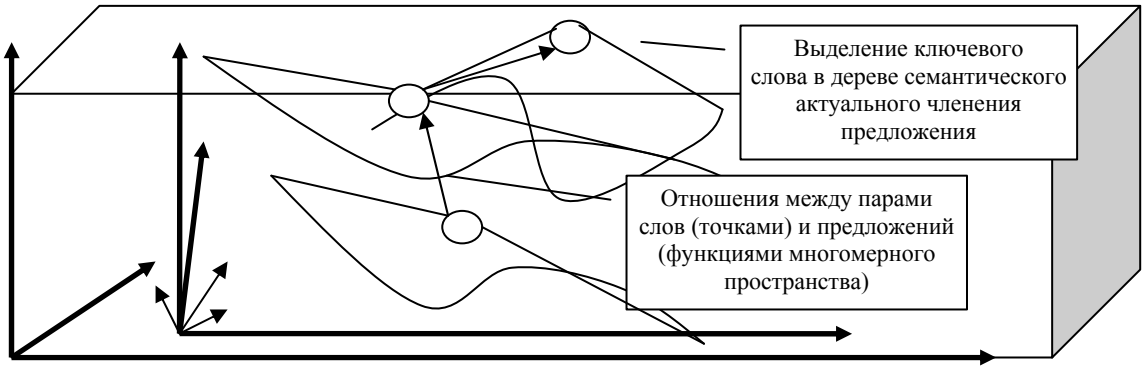

Рис. 2. Модель лексико-грамматического пространства

4. Тема: Кухня, Рема: Красивая. Генерация предложения: Кухня – красивая → Кухня приятна для еды → Сегодня кухня особенно приятна для еды → Сегодня кухня особенно приятна для «пожевать». (Четвертое предложение с элементами сленговых оборотов).

5. Тема: Повар, Рема: Хороший/профессиональный. Генерация предложений: Повар – профессионал → → Повар, конечно, профессионал → Я уверен, что повар, конечно, профессионал → Я уверен, что повар, конечно, профи → Я уверен, что повар, конечно, суперпрофи (Пятое предложения с фонетическим шумом сленговой стилистики языка).

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 [6] важным остается аспект применения генерации речи обучающих системах.

Таким образом, необходимо отметить, что проблема генерации логико-грамматических переходов между парами предложений нуждается в дальнейшем исследовании. Метод аналогии между переходами в виде пар слов и переходом между предложениями в виде дерева с одним ключевым словом на корне дерева актуального членения предложения является эффективным и нуждается в дальнейшем развитии.

Библиографические ссылки

1. Сдобников В. В. Новый взгляд на стратегию перевода: коммуникативно-функциональный подход // Журн. Сиб. федер. ун-та. 2011. Т. 4, № 10. С. 1444–1453.

2. Алексеева И. С. Текст как доминанта перевода // Журн. Сиб. федер. ун-та. 2011. Т. 4, № 10. С. 1375–1384.

3. Сафонов К. В., Личаргин Д. В. Elaboration of a Vector-Based Semantic Classification over the Words and Notions of the Natural Language // Вестник СибГАУ. 2009. Вып. 5 (26). С. 52–56.

4. Сафонов К. В., Личаргин Д. В. Разработка векторизованной семантической классификации над словами и понятиями естественного языка // Вестник СибГАУ. 2010. Вып. 4 (31). С. 33–37.

5. Сафонов К. В., Личаргин Д. В. Некоторые принципы автоматической генерации учебных материалов на основе баз знаний и лингвистической классификации // Вестник СибГАУ. 2012. Вып. 2 (42). С. 72–77.

6. Обучение математике в среде Moodle на примере электронного обучающего курса / Т. В. Зыкова, Т. В. Кытманов, Г. М. Цибульский, В. А. Шершнева // Вестн. Краснояр. гос. пед. ун-та им. В. П. Астафьева. 2012. № 1. С. 60–63.

References

1. Sdobnikov V. V. *The Journal of SibFU*, 2011, vol. 4, № 10, pp. 1444–1453.

2. Alekseeva I. S. *The Journal of SibFU.* 2011, vol. 4, № 10, pp. 1375–1384.

3. Safonov K. V., Lichargin D. V. *Vestnik SibGAU*, 2009, № 5 (26), pp. 52–56.

4. Safonov K. V., Lichargin D. V. *Vestnik SibGAU*, 2010, № 4 (30), pp. 33–37.

5. Safonov K. V., Lichargin D. V. *Vestnik SibGAU*, 2012, № 2 (42), pp. 72–77.

6. Zykova T. V., Kytmanov T. V., Tsibulskyi G. M., Shershneva V. A. *Vestnik KSPU*, 2012, № 1, pp. 60–63.

© Личаргин Д. В., Щурова А. В.,
Курбатова Е. А., Колбасина И. В., 2013