**УДК 004.9**

# ОЦЕНКА РАССТОЯНИЯ В СЕМАНТИЧЕСКОМ И ГРАММАТИЧЕСКОМ ПРОСТРАНСТВЕ МЕЖДУ ОТДЕЛЬНЫМИ ЯЗЫКОВЫМИ ЕДИНИЦАМИ И ФРАГМЕНТАМИ ТЕКСТОВ

**Личаргин Д.В.[1], Полякова О.С.[1], Подлесный А.О.[1], Кравченко М.В.[1]**

*[1]ФГАОУ ВПО «Сибирский федеральный университет», Красноярск, Россия (660074, Красноярск, ул. Киренского, 26), e-mail: lichdv@hotmail.ru*

**В работе рассматривается проблема оценки семантического и грамматического расстояния между словами и другими единицами осмысленного языка. Данная проблема является подпроблемой генерации семантически осмысленного и грамматически корректного текста на естественном языке. В качестве инструмента оценки семантических расстояний между единицами языка используется многомерная грамматическая база данных с координатами понятийного вектора. В этой базе грамматические конструкции определены в ячейках многомерного массива данных – многомерного пространства слов языка. Решение вопроса об измерении метрического расстояния между парами слов позволяет определить степень мощности множества сем, общих для рассматриваемого семантического отношения между словами. Численно заданы расстояния в рамках многомерного пространства отношений между словами в рамках семантической классификации слов и понятий языка. Рассмотрены примеры сложений лексических пар в семантическом пространстве с учетом метрики семантического пространства. Также в работе приводятся подходы по определению семантического расстояния между фрагментами текста на основе расстояния между ключевыми словами естественного языка.**

Ключевые слова: компьютерная лингвистика, семантически осмысленный текст, семантика, семантическое расстояние, математическая семантика, расстояние между парами слов языка.

# ESTIMATION OF THE DISTANCES IN SEMANTIC AND GRAMMATICAL SPACE BETWEEN INDIVIDUAL LANGUAGE UNITS AND TEXT FRAGMENTS

**Lichargin D.V.[1], Polyakova O.S.[1], Podlesniy A.O.[1], Kravchenko M.V. [1]**

*[1] FSAEI HPE "Siberian Federal University", Krasnoyarsk, Russia (660074, Krasnoyarsk, ul. Kirenskogo, 26), e-mail:lichdv@hotmail.ru*

**This paper considers the problem of estimating the semantic and grammatical distance between words and other units of the meaningful language. This problem is a sub-problem of generating semantically meaningful and grammatically correct text in a natural language. As a tool for estimating the semantic distance between the units of the language, a grammatical multidimensional database with the coordinates of the notional vector is applied. In this database grammatical structures are defined in the cells of a multidimensional database - a multidimensional space of words of the language. Solving the task of estimating the metric distance between pairs of words allows determining the power of semes set, common for the considered semantic relations between words. The distance within the multidimensional space of relations between words within the semantic classification of words and notions of the language is numerically determined. The examples of adding lexical unit pairs in the semantic space are viewed, taking into account the metrics of the semantic space. Besides, in the paper approaches for determining the semantic distance between the fragments of the text is present based on the distance between the keywords of the natural language.**

Keywords: computational linguistics, semantically meaningful text, semantics, semantic distance, mathematical semantics, the distance between pairs of words in the language.

На сегодняшний день широко распространены и разрабатываются разнообразные системы машинного перевода, поддержки диалога с пользователем, естественно-языковые интерфейсы, системы автоматического реферирования и аннотирования, экспертные системы и т.п. Однако такие программы еще не в состоянии полностью описать, породить, обеспечить анализ и синтез естественного языка с учетом потенциального мультилингвизма. В работах [1-4] показано, что проблема перевода с одного естественного языка на другой

может сводиться к генерации осмысленного множества языка и эвристическому поиску в этом потенциально генерируемом множестве.

Проблема генерации предложений, повествований и текстов на основе семантически осмысленного и грамматически корректного перехода от предложения к предложению является актуальной в связи с необходимостью правильного определения семантической связи единиц языка при переводе и учете ближайшего и дальнего контекста.

Определение коэффициентов семантической связности отношений между единицами языка дает возможность оценки соответствия найденного фрагмента перевода и составляющих его фраз (или функциям) точкам многомерного пространства потенциально сгенерированного упорядоченного множества – классификации фраз естественного языка. Оценка расстояния между единицами языка применима и в других задачах, при построении компьютерных тезаурусов, при автоматическом построении компьютером связного и осмысленного текста, работе экспертных систем. На основе оценки семантической близости можно посчитать суммарную семантическую близость с учетом частотности списка слов, характерных для искомого текста, на основе такой численной оценки можно определить тематику текста. Различные типы отношений между словами языка, такие как омонимы, синонимы, гиперонимы, антонимы, эквонимы и другие должны получить точную численную оценку на основе единообразной скалярной величины.

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.

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. оценке 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 речи или некоторого приближения к множеству осмысленной речи из надмножества осмысленных фраз с семантическим шумом.

Важной проблемой для создания программной системы генерации осмысленных фраз естественного языка является задача моделирования естественного языка. Решением проблемы может служить как построение модели учета пар слов как пар векторов признаков многомерного пространства слов. При этом инструментом, позволяющим осуществить решение данной проблемы, может служить формальная модель семантики языка на основе описания семантических массивов слов в многомерном семантически упорядоченном пространстве данных.

Семантика и математическая семантика являются науками, рассматривающими смысловое значение единиц языка, таких как слова, предложения, повествования и тексты. Особенностями семантики объясняются лингвистические сложности перевода компьютерными программами с одного языка на другой. Семантическая теория перевода призвана описать модели вычисления параметров для единиц разного уровня языка в целях оценки переводческой эквивалентности, качества перевода, 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). Эти слова являются омонимами, то есть рассматриваются как абсолютно разные слова с общим написанием.

2. В следующем примере приведем понимание омофонии, что соответствует одинаковому звучанию, но разному написанию слов: «meat» [mi:t] – «meet» [mi:t]; «whole» [houl] – «hole» [houl]; «knew» [nju:] – «new» [nju:].

3. Типы синонимии определяются степенью близости значений слов: «fine» - «nice», «awful» - «terrible» - «terrific» и другие.

4. Антонимия определяется отношениями между словами: «good» – «bad», «large» – «small». Необходимо учитывать также множественные противопоставления: «live» м «be born» – «die» – «revive», а также смысловые ряды: «huge» – «large» – «medium» – «small» – «tiny».

5. Эквонимы тесно связаны с явлениями гипонимов. Эквонимы определяются как слова одного уровня обобщения, относящиеся к общему гипериониму. Например: эквонимами по отношению к друг другу являются слова «grandmother» и «grandfather», «mother» и «father», «computer» и «calculator».

6. Гипонимы – слова видового, более специального значения по отношению к слову более обобщенного семантического значения. Гиперонимы представляют собой имя родового понятия. Например: слова «mother» и «father» в свою очередь являются гипонимами по отношению к слову «parent».

7. В свою очередь «parent» 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: «parent» – «mother», «computer» – «notebook».

Одной из задач математической семантики является измерение семантических расстояний между словами естественного языка. Оценка семантического расстояния позволяет оценить плотность семантических и ассоциативно-семантических связей между словами и понятиями словаря, между единицами текста и, в рамках более сложных задач, между фрагментами текста. Правила сложения расстояний между смысловыми единицами языка A и C определяют нижнюю границу, связанную с суперпозицией расстояний между словами A и B, а также словами B и C. Величина измеренного семантического расстояния между любыми единицами текста показывает степень семантической, и, соответственно, тематической и контекстуальной близости отношений между смысловыми элементами текста.

Для получения количественной оценки плотности семантической связи необходимы знания о природе отношений, о типах единиц: терминов, слов, их смысловых значений. Также возможно привлечение 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 пространства, описанного в работах [1, 4]:

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

$G_2$ = Члены предложения {«Определитель», «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).

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. |

Для пояснения вышеприведенной таблицы приведем пример. Возьмем два предложения:

1) «я хочу пожарить свежую курицу»;
2) «я хочу пожарить свежие ножки буша».

Гипонимичными по отношению друг к другу будут являться фразы, отличающиеся лишь словами: «курица» и «ножки буша». Коэффициент между такими словосочетаниями составит 0,8. Таким образом, мы можем определить, что расстояние между первыми словами равно «1», так как фраза «я хочу пожарить свежее» относится к такому типу отношений, как «отношение слова с самим собой». С другой стороны, предложение, отличающееся от аналогичного по двум позициям, например, «father» и «mother», «sociable» и «reserved» может называться эквонимично-антонимичное. Значит, эти два предложения являются двумя пересекающимися функциями в многомерном семантическом пространстве слов языка.

При употреблении фразы «я хочу пожарить свежую утку» расстояние между словами «курица» и «утка» составит 0,85, потому что они являются эквонимами, но еще не антонимами, в виду отсутствия достаточно резкого противопоставления, как, например, для слов «часто» – «редко».

Измеряя коэффициенты между единицами языка с помощью определения их типов отношений, можно вычислять результирующее расстояние между предложениями в целом.

Введем обозначение: MSSimilarity – Minimal Semantic Similarity, минимальное семантическое расстояние. Его вычисление позволяет определить степень согласованности значений слова в определенном контексте и степень их смысловой нагрузки на предложение или абзац при использовании именно данного варианта языковой единицы.

Расстояние между одними и теми же словами равно 1, так как отношение А-А является характеристикой отношения тождества. Ниже приведены примеры таких отношений:

1) *MSSimilarity(1, Professor, Professor);*
2) *MSSimilarity(1, Number, Number);*
3) *MSSimilarity(1, ANYWORD1, ANYWORD1).*

В следующих примерах показано, что такое отношение расстояния обратно пропорционально отношения семантической близости / семантического подобия. В случае необходимости перемножения значений коэффициентов пары слов производится следующая оценка. Близость пары «Professor, Number» больше, чем двух пар «Professor, Maths» и «Maths, Number». Отношение неравенства вида «больше» означает возможность других путей в графе пар слов между данными словами.

1) *MSSimilarity (0.45, Professor, Number) ≥ MSSimilarity (0.6, Professor, Maths) + MSSimilarity (0.75, Maths, Number).*
2) *MSSimilarity (0.6, Professor, Maths) ≥ MSSimilarity (0.75, Professor, Science) + MSSimilarity (0.8, Science, Maths).*
3) *MSSimilarity (0.27, Many, Professor) ≥ MSSimilarity (0.6, Many, Number) + MSSimilarity (0.45, Professor, Number).*
4) *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).*

Имеет место задача оценки значения минимальных семантических расстояний для подобных слов и их цепочек. Постановка эксперимента по верификации данных результатов 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»;

Таким образом, имеет место обход сгенерированного дерева пар слов отдельно по теме и реме фразы, также необходимо учитывать тип связи темы и ремы и других смысловых центров предложения. 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».

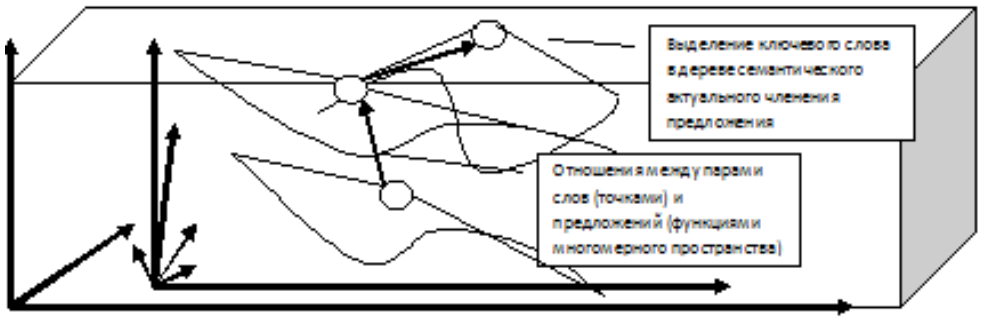


Рис. 1. Модель лексико-грамматического пространства

Такие фразы можно генерировать с привлечением сленга и художественных оборотов с учетом выбранного стиля изложения. Так, модель траекторий в виде цепочек пар слов естественного языка строится на основе представления о многомерном пространстве слов языка с учетом векторов грамматических и семантических значений. Допустим переход к соответствующей траектории ключевых слов как вершин деревьев актуального членения предложения в целях генерации каждого из вариантов синонимичных фраз языка (см. рис. 1).

Парсинг позволяет сопоставить каждой линейной последовательности слов естественного языка одно 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 процесс построения итогового предложения.

**В заключение** необходимо отметить, что величина семантического расстояния играет большую роль при определении смысла с учетом контекста нескольких предложений. Семантические значения слов в предложении должны создавать смысловое единство, поэтому значения концептов (и сем) слов, которые стоят рядом в предложении, должны находиться в оптимальном диапазоне семантической близости.

## Список литературы


1. Личаргин Д.В. «Методы и средства порождения семантических конструкций естественно языкового интерфейса программных систем». Диссертация. Кандидат технических наук: 05.13.17. / Д.В. Личаргин. Защищена 05.07.2004, Утв. 10.12.2004; №137428. – Красноярск, 2004. – 154 с.
2. Личаргин Д.В., Маглинец А.Ю., Рыбков М.В., Бачурина Е.П. Построение алгоритма преобразования деревьев иерархических систем как элементов порождаемой классификации // Журнал Информатизация образования и науки. – 2014. – 15 с.
3. Личаргин Д.В., Сафонов К.В., Егорушкин О.И., Бачурина Е.П. К вопросу об упорядочения многоуровневой семантической сети на иерархии семантической классификации / Вестник Сибирского государственного аэрокосмического университета. – 2014. – 8 с.
4. Личаргин Д.В., Щурова А.В., Курбатова Е.А., Колбасина И.В. Анализ лексических пар для автоматической генерации диалогической и монологической речи // Вестник Сибирского государственного аэрокосмического университета. – 2013. – С. 47-51.
5. Личаргин Д.В. Порождение дерева состояний на основе порождающих грамматик над деревьями строк // Вестник Сибирского государственного аэрокосмического университета: сб. научн. трудов. – Красноярск: СибГАУ. - №1 (27). – 2010. – С. 57-59.


**Рецензенты:**

Ченцов С.В., д.т.н., профессор, заведующий кафедрой Системы автоматики, автоматизированного управления и проектирования Сибирского федерального университета, г. Красноярск;

Бронов С.А., д.т.н., профессор, руководитель научно-учебной лаборатории Систем автоматизированного проектирования кафедры Системы искусственного интеллекта Сибирского федерального университета, г. Красноярск.