



Прикладные задачи компьютерной лингвистики

Ефремова Наталья Эрнестовна
Грацианова Татьяна Юрьевна

Содержание



- Компьютерная лингвистика (КЛ)
- Особенности естественного языка (ЕЯ).
Единицы и уровни языка и текста
- Модели в компьютерной лингвистике
- Лингвистические процессоры (ЛП) и ресурсы
- Прикладные задачи КЛ и подходы к их
решению
- Показатели качества решения задач
- Организационная информация

Компьютерная лингвистика (1)



- ❖ Использование формальных моделей для описания ЕЯ
- ❖ Создание алгоритмов и прикладных программ для обработки языковой информации
- ❖ Междисциплинарное научное направление:
 - ❑ лингвистика (знания о языке)
 - ❑ математика (статистика, формальные языки)
 - ❑ информатика (построение программ)
 - ❑ искусственный интеллект (моделирование интеллектуальных процессов)
- ❖ Начало работ – 50-е годы, машинный перевод

Компьютерная лингвистика (2)



ЕЯ – средство общения людей →
автоматическая обработка информации на ЕЯ –
актуальное научное направление

- ❖ Основные направления:
 - автоматическая обработка текстов (включая оптическое распознавание символов)
 - обработка устной речи
 - построение лингвистических ресурсов
 - мультимодальные приложения, совмещающие обработку текста, звучащей речи и/или изображений
- ❖ Существуют готовые инструментальные системы, помогающие решать прикладные задачи КЛ

Особенности ЕЯ



- *Естественный язык* – система знаков, возникшая для обмена информацией в процессе человеческой деятельности
- Функции ЕЯ:
 - ▶ коммуникация (обмен информацией)
 - ▶ мышление
 - ▶ сохранение знаний
- Особенности ЕЯ как знаковой системы
 - большая: несколько сот тысяч знаков (единиц)
 - многоуровневая: подсистемы (уровни)

Знаки (единицы) ЕЯ



- В общем случае знаки языка (языковые единицы) обозначают понятия действительности
- Совокупности языковых единиц образуют уровни (подсистемы) языковой системы
- Уровни образуют взаимосвязанную иерархию →
- Единицы высшего уровня разложимы на единицы низшего уровня (возможно неоднозначное разложение)
- В рамках одного уровня единицы комбинируются друг с другом (даже для одного смысла возможно несколько комбинаций)
- На каждом уровне – свои правила сочетаемости

Единицы языка и текста



В лингвистике есть разграничение на

- язык в целом
- речь (устная, письменная: текст) – линейная последовательность единиц языка, строящаяся в соответствии с принятыми правилами

❖ Единицы языка:

- фонемы (звуки)
- морфемы
- слова (лексемы)

<a>, <o>, <л>

-бег-, -сь, раз-, -о-

алло, бегун, разбежись

❖ Единицы текста (речи):

- слова (словоформы)
- словосочетания
- предложения (фразы)

бегуны, пятиэтажный

прекрасный кварк

Я пошла бегать.

Уровни языка и текста



- **Фонологический:** звуки
Графематический: символы текста
- **Морфологический:** слова
 - ✓ подуровень морфем
- **Лексический:** множество лексем
- **Синтаксический:** предложения
 - ✓ подуровень словосочетаний
 - ✓ надуровень сверхфразовых единств
- **Семантический:** *семы* (слова *хороший* – *нехороший* различаются семой отрицания)
- **Дискурсивный:** схематические структуры текстов (*патентные формулы, деловые письма* и т.п.)

Естественные vs. искусственные языки



- Естественные языки: *русский, английский, хинди*
Искусственные языки: *программирования, логики*
 - Близки по функциям, но:
 - ◆ ЕЯ открыт и изменчив (на всех уровнях)
 - ◆ единицы ЕЯ сочетаются не стандартно
кидать – *кидает*, но *держать* – *держит*
стадо коров, но *табун лошадей*
 - ◆ ЕЯ избыточен, его единицы многозначны:
нет однозначного соответствия между
единицами и их смыслами
- ➔ Формально описать информацию о ЕЯ **сложно**

Избыточность и многозначность ЕЯ



Избыточность (*синонимия*) – совпадение по основному смыслу, различие в оттенках и стиле
горячий – жаркий, о лете – про лето, или – либо, воспретить – запретить

Многозначность

- ◆ *полисемия* – наличие у языковой единицы нескольких сходных по смыслу значений
земля – это *наша планета, суша, почва* и пр.
- ◆ *омонимия* – совпадение по форме разных по смыслу единиц
брак – супружество или дефект?
стали – существительное или глагол?
Декан назначен ректором – кто кем назначен?

Моделирование в КЛ



- Используемые в КЛ формальные модели
 - ▶ воспроизводят отдельные функции языка
 - ▶ ориентированы на конкретную прикладную задачу и при этом:
 - ✓ покрывают множество текстов
 - ✓ экспериментально обоснованы
 - ✓ опираются на *лингвистические ресурсы*
 - ▶ алгоритмизируемы
- Создаваемые модели используются при написании *лингвистических процессоров*

Виды моделей ЕЯ и текста



■ Статистические модели ЕЯ/текста

- опора на частоту употребления единиц (букв/слов/...) и их последовательностей
- модель – единицы и соответствующие им частоты

■ Структурные модели ЕЯ/текста

- учет лингвистических особенностей одного/нескольких уровней ЕЯ/текста
- модель – набор лингвистических правил

■ Признаковые модели текста

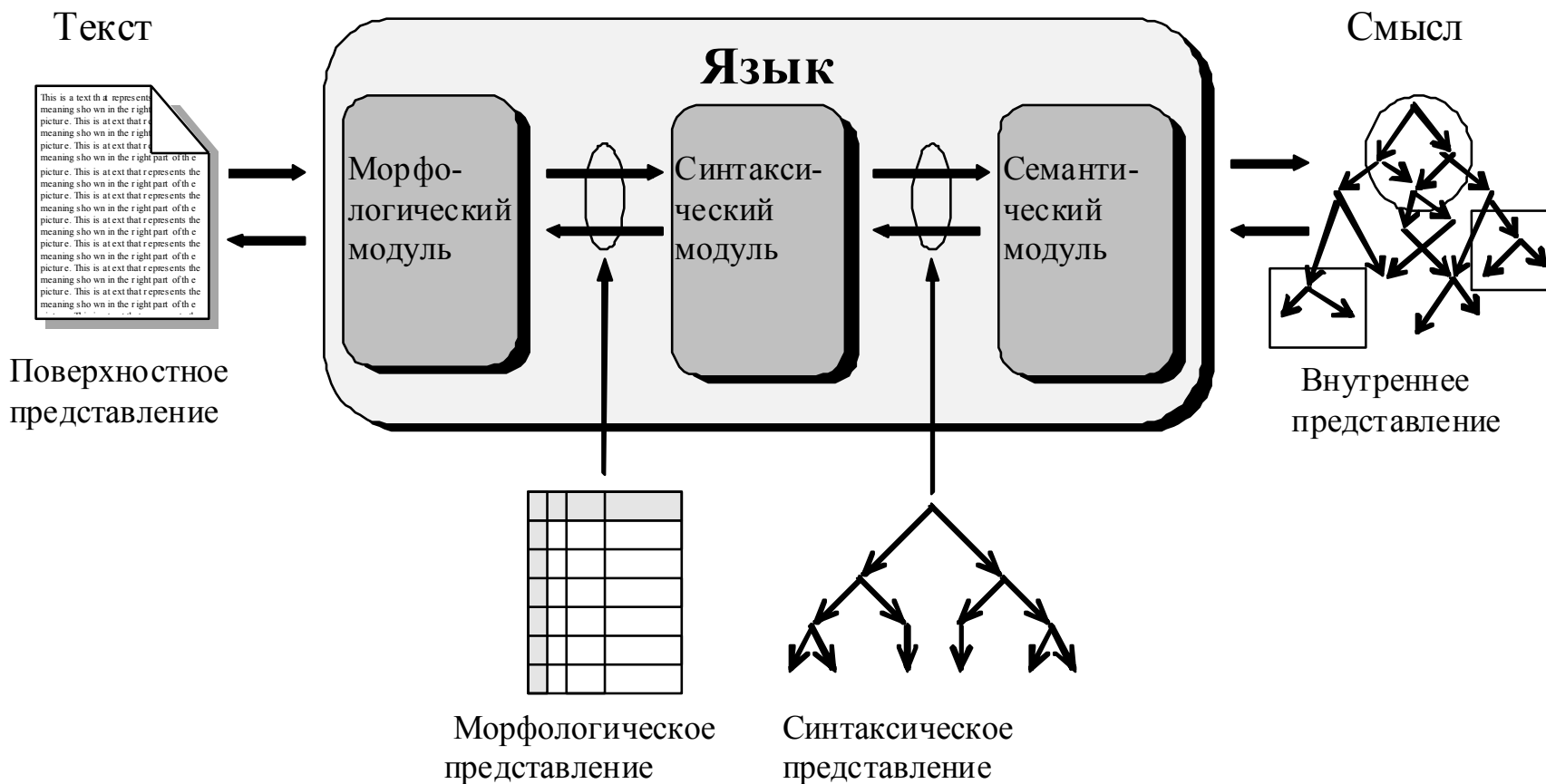
- учет лингвистических и статистических свойств (признаков) текста
- модель – множество признаков с весами

Лингвистические процессоры



- *Лингвистический процессор (ЛП)* – программа обработки текстов на ЕЯ в рамках конкретной прикладной задачи
- ЛП обеспечивает (в зависимости от задачи)
 - **анализ**: разбор и «понимание» поступившей на вход фразы на ЕЯ
 - **синтез**: построение фразы ЕЯ по формальному описанию ее смысла
- Основа ЛП – модель языка и/или текста → зависит от конкретного ЕЯ и прикладной задачи
- Опирается на *лингвистические ресурсы*

Пример: этапы обработки текста в структурной модели



Виды лингвистических ресурсов



Содержат необходимую для модели и задачи *лингвистическую информацию*

- Компьютерные словари *слов*, *морфем* и пр.
- Грамматики для описания *синтаксической структуры предложений* и др.
- Базы словосочетаний, отражающие сочетаемость слов ЕЯ
- Семантические словари, тезаурусы, онтологии, фиксирующие семантические связи слов
- Коллекции и корпуса текстов

Коллекции и корпусы текстов



- *Коллекция* – представительный набор текстов, собранных по определенному принципу
- *Корпус* – коллекция текстов с лингвистической разметкой (*морфологической*, *синтаксической*, *семантической*, *акцентологической* и пр.)
- Используются в лингвистических исследованиях и для машинного обучения моделей
- Созданием и использованием электронных корпусов текстов занимается **корпусная лингвистика**



Прикладные задачи КЛ

- Машинный перевод
- Информационный поиск
- Реферирование и аннотирование текстов
- Создание вопросно-ответных систем
- Организация диалога на ЕЯ
- Автоматизация создания и редактирования текстов

Задачи направления Text Mining

- Извлечение информации из текстов
- Выявление терминов и ключевых слов
- Классификация и кластеризация текстов
- Извлечение мнений и оценка тональности текстов
- Компьютерная текстология

Список не полон, есть еще другие задачи!

Подходы к решению задач



- **Статистический подход** – для статистической модели:
 - ✓ Вычисляются разные меры на основе частот единиц
 - + Прост в применении
 - Возможны одинаковые значения мер у различных для решаемой задачи единиц
- **Инженерный подход** – для структурной модели, применение лингвистических шаблонов и правил
- **Машинное обучение** – для признаковой модели
- **Комбинирование** подходов – рассмотрение различных моделей ЕЯ и текста

Лингвистические шаблоны



- *Лингвистический шаблон* – описание языковой конструкции, ее лексического состава и грамматических свойств
- Используются специальные языки записи шаблонов и поддерживающие их программные средства
- + Высокая точность извлечения информации
- + Легко локализовать и исправить ошибку
- Требуется работа эксперта
- Не переносимо на другую тематику текстов

$(^([0-9]+[.]+[0-9]+)|(0))\$$ – 0, 1.2, 12.345

A N – великий уравнитель, солнечное сплетение

Машинное обучение



Вход: ♦ множество ситуаций (C) и реакций (P)
♦ обучающая выборка: пары $C - P$

Задача: по обучающей выборке выявить зависимость между C и P

Алгоритм обучения пытается найти эту зависимость

Выход: функция, ставящая в соответствие C определенную P

- + Не требуется работа эксперта
- + Можно пробовать разные алгоритмы обучения
- Нужна большая, качественно размеченная обучающая выборка
- Сложно локализовать и исправить ошибку

Прикладные задачи: машинный перевод



- Джорджтаунский эксперимент, 1954 г.:
автоматический перевод с русского на английский
- Периодизация методов/систем:
 - 50-60 гг. — двуязычные системы, пословный и пословно-пооборотный перевод
 - 60-70 гг. — пофразный перевод
 - 70-80 гг. — многоязычные системы
 - 80-90 гг. — использование крупных лексических и терминологических ресурсов
 - 2000 гг. — статистический машинный перевод

Проблема полностью автоматического и при этом
качественного перевода текстов неразрешима

Информационный поиск



С 50-х гг. поиск в коллекциях текстовых документов:

- Вручную создавался *поисковый образ документа* (ключевые слова для отражения его содержания)
- Слова/словосочетания из запроса сопоставлялись с поисковыми образами документов
- Результат поиска – *релевантные документы*
- ❖ До сих пор применяется в современных корпоративных информационных системах

С 90-х гг. полнотекстовый поиск (в т.ч. в сети Интернет)

- *Автоматическое индексирование* текстов
- Применение модели «мешка слов»

Реферирование и аннотирование текстов



- Реферирование текста – построение краткого реферата для одного или нескольких текстов
 - ▶ основная стратегия – отбор наиболее значимых предложений
 - ▶ сложность: связность текста
- Аннотирование текста – построение аннотации документа
 - ▶ аннотация – вторичный документ, еще более краткий, чем реферат
 - ▶ в простейшем случае – перечень основных тем/ключевых слов документа

Создание вопросно-ответных систем



- Нужно получить не ссылку, документ или его фрагмент (сниппет), а ответ на вопрос
Как выглядит опёнок?, Сколько лет Кремс Л.О.?
- Примерная стратегия построения ответа:
 - определение типа вопроса и запрашиваемого понятия
 - построение запроса к источнику информации: интернет-поисковику, базе данных и т.п.
 - извлечение из найденных документов нужной информации
 - построение фразы ответа (синтез текста)
- Не предполагается дальнейшее ведение диалога

Организация диалога на ЕЯ



- Возможно общение посредством
 - текста (нужен анализ и/или синтез текста)
 - устной речи (нужен анализ и/или синтез речи)
- Основная часть – *диалоговый менеджер*: понимает когда, что и как отвечать, какова тема и структура беседы и т.д.
- *Чат-боты* – упрощенная диалоговая система
 - анализируют высказываний на наличие ключевых слов и речевых шаблонов
 - ответ выбирают из имеющихся вариантов
 - реакция – словесный ответ или действие
- Примеры систем: *Siri, OK, Google*

Автоматизация подготовки и редактирования текстов



- В идеале необходимо:
 - ✓ оформление текста: переносы, шрифты, поддержка структуры и т.д.
 - ✓ помощь при написании: подбор нужной лексики (синонимов, устойчивых словосочетаний и пр.)
 - ✓ проверка текста: орфографии, пунктуации, синтаксиса, стиля, структуры и др.
 - ✓ обработка текста: оценка сложности, выявление ключевых слов, составление реферата и т.п.
- Существующие системы (*Орфограммка*, *MS Word*) обеспечивают эти функции частично и не всегда удобно

Извлечение информации из текстов



- Значимая для текста/области/темы информация извлекается из текстов
 - ✓ конкретные объекты: *имена персон*
 - ✓ атрибуты объектов: для *персоны* – *место работы, должность, телефон, подразделение*
 - ✓ отношения (связи) между объектами: *быть частью, быть владельцем*
 - ✓ факты/события: *прошла встреча, выдан кредит*
 - Полученные данные структурируются, накапливаются в базах знаний, визуализируются
- Используется при *мониторинге новостных лент, текстовой аналитике* и пр.

Выявление терминов и ключевых слов



Разновидность задачи извлечения информации

- *Терминологические слова и словосочетания называют понятия специальной области знаний **экзопланета, красный гигант, взрыв сверхновой***
- ❖ *Выявление терминов нужно при построении **словарей, тезаурусов, онтологий, глоссариев***
- *Ключевые слова – слова и словосочетания, отражающие основной смысл текста*
- ❖ *Нужны для облегчения индексирования текста*
- ❖ *Ключевые слова необязательно являются терминами и присутствуют в тексте*

Классификация и кластеризация текстов



- *Классификация текстов* – отнесение текстов к классам с заданными свойствами/параметрами
- *Кластеризация текстов* – создание подмножеств близких (например, тематически) документов
- Используется, в частности:
 - ✓ для упорядочивания и навигации по набору документов (при поиске)
 - ✓ для фильтрации документов (фильтрация спама и дубликатов, подбор рекламы)

Извлечение мнений и оценка их тональности



- Мнение – общая эмоция, суждение или оценка автора по поводу некоторой сущности
- Тональность – эмоциональная оценка, выраженная по отношению:
 - ✓ к некоторой сущности в целом
 - ✓ ее аспектам (частям/свойствам/...)
- Виды оценок тональности: двоичная (+/–), троичная (+/–/0), по некоторой шкале

Области применения:

- ❖ выделение и суммирование мнений (о товарах, фильмах, партиях и пр.) в сети Интернет
- ❖ анализе и оценке документа: выделение цитат, определение тональности и субъективности

Компьютерная текстология



- Текстология изучает историю изменений текста и ищет причины этих изменений
- Цель – найти первоисточник или наиболее близкий к нему документ
- Возможные изменения в тексте – вставка, удаление, замена слов (при переписывании)
... государь пожаловал даде им грады и села ...
... государь пожаловал грады и села ...
- С 70-80 гг. основные этапы анализа текстов автоматизированы:
 - ✓ сравнение – поиск различий между текстами
 - ✓ кластеризация текстов на основании различий, построение дерева связей

Основные показатели качества работы систем КЛ



Результаты работы системы можно разделить

эксперт система	правильные (positive – P)	неправильные (negative – N)
правильные (P)	True P = TP	False P = FP
неправильные (N)	FN	TN

■ Точность (*Precision*)

$$P = TP / (TP + FP)$$

■ Полнота (*Recall*)

$$R = TP / (TP + FN)$$

■ F-мера – геометрическое среднее P и R

$$F = 2PR / (R + P)$$

Программа курса



- ❖ Модели КЛ:
 - ✓ этапы обработки текста в структурной модели
 - ✓ статистические модели ЕЯ и их применение
 - ✓ признаковые модели текста и их применение
- ❖ Лингвистические ресурсы:
 - ✓ корпуса текстов
 - ✓ методы КЛ для построения словарных ресурсов
- ❖ Прикладные задачи КЛ:
 - ✓ извлечение информации из текстов
 - ✓ выявление мнений и анализа тональности
 - ✓ реферирование и аннотирование текстов
 - ✓ машинный перевод и генерация текстов
 - ✓ создание вопросно-ответных систем и чат-ботов
 - ✓ распознавание и синтез речи

Литература (часть)



- ❖ Автоматическая обработка текстов на естественном языке и компьютерная лингвистика: учеб. пособие / Большакова Е.И. и др. – М.: МИЭМ, 2011.
- ❖ Автоматическая обработка текстов на естественном языке и анализ данных: учеб. пособие / Большакова Е.И. И др. – М.: Изд-во НИУ ВШЭ, 2017.
- ❖ Прикладная и компьютерная лингвистика / Николаев И.С. и др. (Ред.). – URSS, 2016.
- ❖ Jurafsky D., Martin J. Speech and Language Processing. An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition, Prentice Hall, 2000.

Организация учебного процесса



- ❖ На каждом занятии – самостоятельная работа и/или обсуждение лекционного материала (СР)
- ❖ После некоторых занятий – домашнее задание (ДЗ), сдача через систему Moodle
- ❖ В середине семестра – контрольная работа (КР)
- ❖ В течение семестра – поэтапное выполнение итогового задания (ИЗ)
- ❖ СР, ДЗ, КР и ИЗ оцениваются в баллах
- ❖ Действует балльно-рейтинговая система: оценка за курс выставляется по сумме баллов (автомат)
- ❖ При неудовлетворительной оценке – экзамен

Регистрация в Moodle



Необходимо отправить письмо на адрес

TGratsianova@cs.msu.su

- ❖ Тема письма: слово *магистр* и фамилия
- ❖ В письме строка: все через запятую без пробелов
логин, пароль, имя, фамилия, почтовый адрес
- ❖ Логин: 6-10 символов, производный от фамилии, маленькие латинские буквы и цифры
- ❖ Пароль: 9-12 символов, обязательно хотя бы одна цифра, маленькая и большая латинская буква, спецсимвол (#&@)
- ❖ Имя: полное, русскими буквами с большой буквы
- ❖ Фамилия: русскими буквами с большой буквы
- ❖ Пример: tgrats, Tata2017#, Татьяна, Грацианова, TGratsianova@cs.msu.su

Итоговое задание: постановка задачи



- ❖ Выбрать одну из прикладных задач КЛ (к 14.09.17)
- ❖ Найти информацию об инструментальных системах, решающих эту задачу (к 28.09.17)
- ❖ Написать программу, решающую выбранную задачу с помощью одной из систем (к 23.11.17)
- Форма работа над заданием: индивидуальная или парная
- В контрольные даты – обсуждение и представление полученных к данному моменту результатов

Итоговое задание: форма отчетности



- ❖ Отчет:
 - описание функциональности найденных систем
 - инструкция по использованию выбранной системы
 - описание полученной программы и результатов ее работы
- ❖ Презентация, представляющая функциональность найденных систем
- ❖ Демонстрация работы полученной программы



Спасибо за внимание!

Статистические модели: N-граммная модель



- Отвечает на вопрос, насколько данная фраза типична/правильна для языка
- Типичность/правильность задается с помощью вероятности
- Вероятность P вычисляется исходя из частоты F употребления и приписывается последовательностям из N слов
- Обычно $N = 1, 2, 3$ или 4

Для предложения *I want to eat* вероятность при $N=2$, будет вычисляться так:

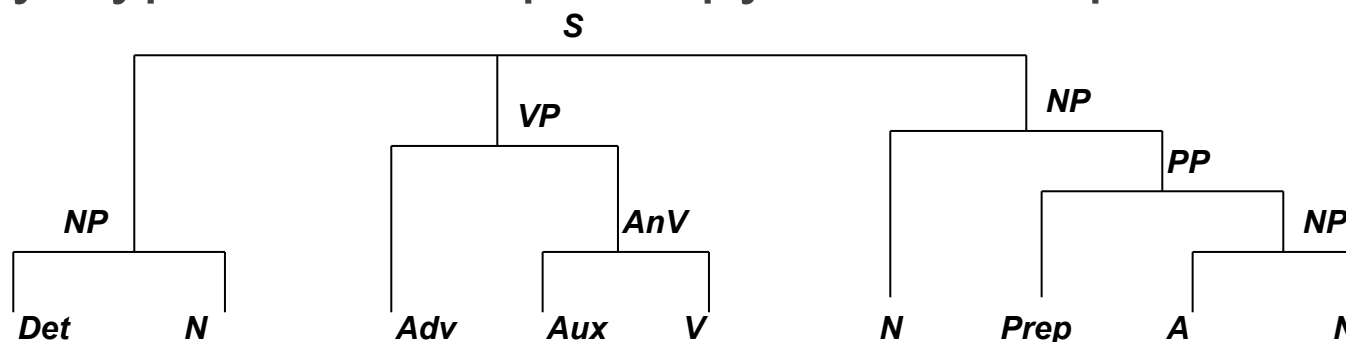
$$P(I \text{ want to eat}) = P(I) * P(want|I) * P(to|want) * P(eat|to)$$

где, например, $P(want|I) = F(I \text{ want}) / F(I)$

Структурные модели: деревья составляющих



- Является моделью синтаксиса
- Из отрезков текста (*составляющих*) для предложения строится *синтаксическое дерево*
- Составляющие могут вкладываться друг в друга, но не пересекаться
- Грамматически правильная синтаксическая структура обычно фиксируется КС-грамматикой



(Этот пес) (скоро (будет грызть)) (кость (в (красном углу)))

Признаковые модели: модель «мешок слов»



- Текст – множество его слов (значимых) → порядок слов в тексте и их грамматические формы не учитываются
- Чаще всего вес – частота употребления слова в тексте

t_1 : Карл¹ у Клары² украл³ кораллы⁴

t_2 : Клара² у Карла¹ украла³ кларнет⁵

t_3 : Клара² у Карла¹ украла³ кораллы⁴

t_4 : Простота⁶ – хуже⁷ воровства⁸

t_1 : {1,1,1,1,0,0,0,0}

t_2 : {1,1,1,0,1,0,0,0}

t_3 : {1,1,1,1,0,0,0,0}

t_4 : {0,0,0,0,0,1,1,1}