Лекция 6. Компьютерная лингвистика

2022

План

- 1. Машинный перевод
- 2. Информационный поиск
- 3. Извлечение информации
- 4. Диалоги и чат-боты
- 5. Анализ тональности
- 6. Квантитативная лингвистика

Тема

- **1. Машинный перевод**
- 2. Информационный поиск
- 3. Извлечение информации
- 4. Диалоги и чат-боты
- 5. Анализ тональности
- 6. Квантитативная лингвистика

Появление научного перевода

- Письмо американского математика Уоррена Уивера Норберту Винеру: «Когда я вижу текст на русском языке, я говорю себе, что на самом деле он написан по-английски и зашифрован при помощи странных знаков. Мне нужно его просто расшифровать» (4.03.1947)

Перевод как дешифровка

- Подсчитывается частота взаимной встречаемости элементов текста. Статистически значимые отклонения от случайности позволяют найти ключ к дешифровке текста.
- Эти методы стали активно использоваться 50 лет спустя.

Основные подходы к машинному переводу

- 1. Перевод на основе **правил** (rule-based machine translation – **RBMT**) работает с грамматиками и словарями.
- 2. Статистический машинный перевод (statistical machine translation – **SMT**) – работает на основе методов машинного обучения, анализируя частоту совместной встречаемости слов в большом количестве пар «предложение + его перевод».
- 3. Гибридный перевод (hybrid machine translation – **HMT**) – наиболее современный подход, комбинирующий правила и статистику.

Автоматизированный перевод

- computer-aided translation – **CAT**
- Текст переводится человеком, использующим разные компьютерные технологии

Гибридный перевод

1 этап – перевод при помощи словарей и грамматик

Time flies like an arrow

Время летит как стрела (1)

Мухи времени любят стрелу (2)

2 этап – сравнение частоты встречаемости сочетаний *время летит* и *мухи времени*.

Модель постредактирования

- PROMT: корпус состоит из предложений, переведённых системой с помощью правил, в соответствие которым поставлены эти же предложения, исправленные носителями языка.

Методы оценки качества перевода

- 1. Экспертная оценка
- 2. Автоматическая оценка

Экспертная оценка

- 1. Не менее 4 экспертов оценивают перевод каждого предложения по **полноте** (точности) и **гладкости** (правильность с точки зрения носителя). По каждому из этих параметров каждый эксперт ставит оценки в соответствии с заранее заданной шкалой.
- 2. **Ранжирование** вариантов перевода.
- 3. Оценка **трудозатрат на редактирование** перевода.

Автоматическая оценка

- Сравнение с эталоном, выполненным или отредактированным вручную: **совпадение n-грамм**.
- Метрики автоматической оценки: BLEU, NIST, MERT, METEOR, TER
- http://asiya.lsi.upc.edu/demo/asiya_online.php - оценка статистического перевода

Некоторые системы машинного перевода

- Systran (США, Франция, Корея)
- Logos, OpenLogos (США, Германия)
- PROMT (Россия)
- Linguatec (Германия)
- IdiomaX (Швейцария, Италия)
- Babylon (Израиль)
- Apertium (Испания)
- Google Translate (США)
- Bing (США)
- Яндекс, Переводчик (Россия)
- ABBYY Compreno (Россия)

Тема

- 1. Машинный перевод
- **2. Информационный поиск**
- 3. Извлечение информации
- 4. Диалоги и чат-боты
- 5. Анализ тональности
- 6. Квантитативная лингвистика

Информационная потребность

- представление пользователя о том, что он хочет найти

Поисковый запрос

- формулировка информационной потребности.
- Информация для поиска представлена в **коллекции документов**. Совпадающие части запроса и документа называют **терминами (дескрипторами)**.

Классический алгоритм поиска

- 1. **Обработка текста документа**. Морфологический анализатор, синтаксический анализатор, получение последовательности графов – деревьев зависимостей для предложений в документе. Семантический анализатор строит на их базе семантическое представление документа.
- 2. **Обработка текста запроса**. С помощью тех же операций строится семантическое представление запроса.
- 3. Сравнение по **индексу**.

Индекс

Слова	Номера документов
а	1, 2, 3, 4, 5, …
Абакан	172, 198
…	
ящур	11

Проблемы информационного поиска

- Семантико-синтаксический анализатор, распознающий анафору, эллипсис и т.п.
- Распознавание цели запроса
- Анализ текстов запросов
- [дорога владимир николаев]

Виды запросов

- **Информационные** (*расстояние до Марса, всё о кроликах*)
- **Навигационные** (*оф сайт фк зенит*)
- **Транзакционные** (цель – выполнение задачи: билет плацкарт воронеж 6 августа)

Критерии качества поисковой системы

- **Релевантность**: документы, нужные пользователю
- **Точность** – доля релевантных документов в числе всех найденных
- **Полнота** – доля найденных документов в числе всех релевантных документов коллекции
- **Ранжированная поисковая система**: получение в первую очередь наиболее релевантных документов

Факторы ранжирования

- Способы численного представления характеристик документа и запроса, важных для качества поиска.
- **Текстовые** (доля слов запроса, встретившихся в документе; доля биграмм запроса, встретившихся в документе; доля слов запроса, встретившихся в документе в той же форме, в какой они представлены в запросе)
- **Ссылочные** (частота встречаемости слов запроса в ссылках на документ)
- **Поведенческие** (количество просмотренных документов, время просматривания документа, переформулирование запроса).
- Используется порядка 1000 факторов.

Алгоритм ранжирования

- машинное обучение на основании экспертной оценки по шкале релевантности документов, полученных по запросу

Стандартные лингвистические модули

- 1. **Лемматизатор**. Распознавание языка. Сведение словоформ к лексеме, обработка имён собственных.
- 2. **Модуль исправления опечаток**. Работа с контекстом ([тстер] – тестер/тостер? [цифровой тстер]). Автозамена, подсказки, смешанные результаты поиска.
- 3. **Модуль диакритики**. Например, в таких языках, как турецкий или венгерский, вариант без диакритики встречается в запросах чаще, чем с диакритикой, что создаёт проблему для статистических алгоритмов.

Модули расширения

- **Синонимы.** [купить картошку недорого]/[купить картофель дешево], но [пирожное картошка]/[пирожное картофель].
- **Классы условной эквивалентности**:
- Словообразовательные [законы физики]/[физические законы]
- Транслиты [Bosch]/[Бош]
- Аббревиатуры [ИП]/[индивидуальный предприниматель]
- Склейка-разрезание [автокредит]/[авто кредит]

Построение модулей расширения

- Тезаурус
- Лингвистические модели (дериватемы, алгоритмы транслитерации и т.п.)
- Статистические модели (встречаемость в одном документе, замена в переформулированном запросе: [айфон 10]/[iphone 10] и т.п.)

Фильтры расширения

- **Контекст**. [hugo] = только [хьюго] в [hugo boss]/но = [хьюго]/[гюго] в [victor hugo]
- **Регион**. [МГУ]=[Московский государственный университет] в Москве или Подмосковье + [Мордовский государственный университет] в Саранске

Фильтры расстояния

- [Владимир Даль]/[Владимир Иванович Даль]
- [Владимир всматривался в даль]

Генерация динамических сниппетов

- построение с учётом запроса короткой аннотации документа, чтобы пользователь мог решить, стоит ли открывать ссылку на найденный документ

Тема

- 1. Машинный перевод
- 2. Информационный поиск
- **3. Извлечение информации**
- 4. Диалоги и чат-боты
- 5. Анализ тональности
- 6. Квантитативная лингвистика

Задачи извлечения

- Связаны с получением конкретных ответов на вопросы и включают определение
- 1) **именованных сущностей** (*В каком году основан петербургский университет/университет в петербурге?*)
- 2) **отношений между сущностями** (*является частью, основан в, в браке с, является владельцем, работал в*).

Событие

- фиксированный набор сущностей и отношений между ними, может иметь несколько синонимичных шаблонов:
  - *Яндекс купил Кинопоиск за 80 млн долларов в октябре 2013 года.*
  - *Осенью 2013 года Кинопоиск был приобретён Яндексом за 80 млн долларов.*
  - *Яндекс стал владельцем Кинопоиска в октябре 2013 года, заплатив $ 80 млн.*

Задача распознавания именованных сущностей

- 1) нахождение в тексте упоминания сущности;
- 2) однозначное указание на объект или лицо;
- 3) приписывание категории.

Извлечение информации из фрагмента текста

- Современный [СПбГУ] в [России] – преемник [Академического университета], который был учреждён одновременно с [Академией наук] указом [Петра I] от [28 января 1724 года], в частности, в [1758 – 1765] годах ректором [Академического университета] был [М.В. Ломоносов].

Сущности и категории

Сущности	Возможные категории
СпбГУ Академический университет Академия наук	**Организация**, образовательное учреждение, вуз Организация, образовательное учреждение, вуз Организация, научная организация, академия
Россия	**Место**, страна, государство
Пётр I М.В. Ломоносов	**Человек**, исторический деятель, политик, правитель Человек, учёный, химик, писатель, философ, художник
28 января 1724 года 1758 – 1765	Время (дата) Время (отрезок)

Зависимость категории от контекста

- **Россия** отказалась от американского мяса. Россельхознадзор вводит временные ограничения на поставки продукции птицеводства США в **Россию**.

Неоднозначность идентификации

- – **Толстому** подражаете, – сказал Рудольфи.
- – Кому именно из Толстых? – спросил я. – Их было много… Алексею ли Константиновичу, известному писателю, Петру ли Андреевичу, поймавшему за границей царевича Алексея, нумизмату ли Ивану Ивановичу или Льву Николаевичу?

Анафора и кореферентность

- Грамоте обучил [Михайла Ломоносова] [дьячок местной Дмитровской церкви С.Н Сабельников]. «Вратами учёности», по **его** собственному выражению, для **него** делаются «Грамматика» [Мелетия Смотрицкого], «Арифметика» [Л.Ф. Магницкого], «Стихотворная Псалтырь» [Симеона Полоцкого]. В четырнадцать лет **юный помор** грамотно и чётко писал.
- По заголовку и теме текста именованным сущностям может приписываться вес.

Знания о мире

- Аня подарила Маше конфеты, потому что у **неё** был день рождения.
- Аня подарила Маше конфеты, потому что у **неё** было две коробки.

Идентификаторы для разрешения кореферентности

- «Евгений Онегин» стал одним из самых значительных произведений **А.С. Пушкина.**
- Евгений Онегин – молодой дворянин, отправляющийся в самом начале романа к умирающему дяде.
- «Евгений Онегин» состоит из трёх **действий** и семи картин.

Тема

- 1. Машинный перевод
- 2. Информационный поиск
- 3. Извлечение информации
- **4. Диалоги и чат-боты**
- 5. Анализ тональности
- 6. Квантитативная лингвистика

Тест Тьюринга

- Английский математик Алан Тьюринг в 1950 году предположил, что к 2000 году качество имитации человеческого диалога компьютером достигнет такого уровня, что в 30% случаев после 5 минут общения человек не сможет различить живого собеседника и компьютер.
- В 1990 году учреждена премия Лёбнера – ежегодное соревнование чат-ботов в прохождении теста Тьюринга.
- В 2014 году в г. Рединг (Великобритания) бот Женя Густман прошёл тест Тьюринга (33% судей).

Моделирование диалога (интеракционная социолингвистика)

- Порядок обмена репликами
- Общий контекст для собеседников
- Структура диалога (установление, поддержание, прерывание контакта)
- Инициатива в диалоге (смешанная, односторонняя)

Модули диалоговых систем

- Распознавание речи
- Понимание языка
- Диалоговый менеджмент
- Генерация естественного языка
- Синтез речи

Модуль понимания естественного языка

- **Задача**: семантическое представление входного текста
- **Знания о мире:** базы знаний, пополняемые алгоритмами извлечения информации из текстов
- **Знания об участниках диалога**: статусы, роли, предпочтения и др. сведения

Диалоговый менеджер

- центральная составляющая диалоговых систем, которая координирует деятельность других компонентов.
- Задачи:
- обновление контекста диалога на основании проинтерпретированного общения;
- представление контекстно-зависимых интерпретаций сигналов;
- работа с базами знаний;
- распознавание речевых актов;
- координирование диалогового и недиалогового поведения.

Модуль генерации естественного языка

- Планирование документа
- Микропланирование
- Поверхностная реализация.

Планирование документа

- Определение содержания
- Структурирование дискурса

Микропланирование

- Лексикализация
- Аггрегация (определение информации для одного предложения)
- Генерация отсылочных выражений.

Поверхностная реализация

- Построение грамматически правильных предложений
- Конвертация текста в запрашиваемый формат

Чат-боты

- Siri (Apple)
- Maluuba (Android)
- Robin (Android)
- Iris (Android)
- Vlingo (Android)
- Skyvi (Android)
- Voice Mate (LG)
- S-Voice (Samsung)
- Google Now
- Cortana (Microsoft)

Artificial Intelligence Markup Language (AIML)

- <aiml> тег, который начинает и заканчивает документ
- <category> тег, обозначающий элемент в базе знаний
- <pattern> содержит простой шаблон: что пользователь может сказать чат-боту
- <template> содержит ответ чат-бота пользователю
- 20 тегов для уточнения шаблонов и сохранения контекста беседы.

Вопросно-ответные системы

- **IBM Watson** – медицинское консультирование
- Модуль контентной аналитики DEEPQA с машинным обучением на основе нейронных сетей

Тема

- 1. Машинный перевод
- 2. Информационный поиск
- 3. Извлечение информации
- 4. Диалоги и чат-боты
- **5. Анализ тональности**
- 6. Квантитативная лингвистика

Анализ тональности

- определение эмоциональной окраски сообщений.
- **Sentiment analysis** – сентимент-анализ, анализ мнений, анализ эмоциональной составляющей сообщений.

Корпус текстов

- Блоги, социальные сети, твиты, отзывы в интернет-магазинах (UGC – User Generated Content).
- Webometric Analyst (программа сбора текстов по заданным параметрам)

Анализ тональности

- 1) субъект тональности (кто? – турист)
- 2) объект тональности (о чём? – отель)
- 3) аспект тональности (местоположение отеля)
- 4) тональная оценка (сообщение о свойствах «очень милый персонал»)

Подходы к анализу тональности

- 1) правила (русский язык)
- 2) машинное обучение (английский язык)

Правила

- Используются шаблоны, описывающие предметную область
- По этим шаблонам из текстов извлекаются n-граммы
- Пример правила: Если цепочка содержит глагол из списка 1 (любить, нравиться, обожать и др.) и не содержит глагол из списка 2 (ужасать, отвращать и др.) или отрицания, то её тональность положительная.
- Механизмы комбинации правил: насколько часто используется, на каких позициях и т.п. (**отличный** фильм для **страдающих** бессонницей).

NRC Word-Emotion Association Lexicon

Слово	Эмоция или тональная оценка	Значение (1 – есть соответствие, 2 – нет соответствия)
frank	anger	0
frank	anticipation	0
frank	disgust	0
frank	fear	0
frank	joy	0
frank	negative	0
frank	positive	1
frank	sadness	0
frank	surprise	0
frank	trust	1

NRC Hashtag Sentiment Lexicon

Слово/ словосочетание	Значение тональной оценки	Частота позитивной связи (хештег, эмотикон и др.)	Частота негативной связи
elegant	5, 665	537	3
excellent movie	5	7	0
kindness	1,006	39	23
sinister	-3,12	7	256

Разработка словарей

- НКРЯ (ev: posit, ev: neg)
- Перевод списков слов с другого языка,
- Пополнение списков при помощи правил (если слово есть в списке, а другое присоединено к нему союзом И, то оно тоже включается в список; меры совместной встречаемости с положительно окрашенной лексикой)

Вычисление тональности слова (SO – sentiment orientation)

- PMI = log2 P(слово A около слова B)/P (слово A)*P(слово B)

- SO (A) = PMI (хорошо или хороший, слово A) – PMI (плохо или плохой, слово A)

Тезаурусы с разметкой эмоциональной составляющей

- SenticNet
- SentiWordNet
- WordNet-Affect
- RussNet

Программы определения тональности текста

- Stanford Live Demo
- SentiStrength
- LIWC

Оценка качества работы алгоритмов

- **Полнота** – отношение верно приписанных тональностей к общему числу тональностей (приписанных и не приписанных)
- **Точность** – отношение верно определённых тональностей ко всем определённым системой тональностям
- **F-мера** – отношение удвоенного произведения полноты и точности к их сумме.

Тема

- 1. Машинный перевод
- 2. Информационный поиск
- 3. Извлечение информации
- 4. Диалоги и чат-боты
- 5. Анализ тональности
- **6. Квантитативная лингвистика**

Принцип квантитативной лингвистики

- **Экспонент** – означающее
- **Денотат** – означаемое
- Денотат «дерево» – экспоненты рус. *дерево*, англ. *Tree*
- Фонемы имеют только экспонент, не имеют денотата.
- На **подсчёте экспонентов** единиц языка и их сочетаний основаны алгоритмы квантитативной лингвистики.

Методика определения языка, на котором написан текст

- Зная частотность букв для каждого языка, мы можем определить, на каком языке написан текст, по частотности букв в тексте.
- Скорость и точность определения возрастает, если считать не отдельные буквы, а сочетания по 2, 3, 5 и т.д.

Проблема дешифровки текста на неизвестном языке

- 1) статистика букв
- 2) система письма
- 3) языковые структуры
- 4) сведения о культуре и образе жизни авторов текста
- 5) письменные памятники соседних народов (имена правителей и названия городов)

Типологические индексы Дж. Гринберга

- 1. **Индекс синтеза**. Сколько в среднем морфем в слове данного языка. Syn=M/W, где M – количество морфем в тексте, W – количество слов в тексте
- 2. **Индекс деривации**. Насколько широко используется словообразование при помощи морфем. Der=D/W, где D – количество деривационных морфем в тексте, W – количество слов в тексте.
- 3. **Индекс префиксации**. Насколько часто используются приставки.
- Pref=P/W
- 4. **Индекс суффиксации**. Suf=S/W

Языки разных морфологических типов

Индексы	Русский	Английский	Якутский	Вьетнамский	Эскимосский
Синтез	2,33	1,68	2,17	1,06	3,72
Деривация	0,37	0,15	0,35	0,00	1,25
Префиксация	0,17	0,04	0,00	0,00	0,00
Суффиксация	1, 15	0,64	1,15	0,00	2,72

Стилеметрия

- количественное исследование стилей текстов, написанных разными писателями в разных жанрах.

Предсказание популярности новых книг и сценариев

- Университет Стоуни Брук (США)
- 1) статистика скачивания книг разных жанров на сайте электронной библиотеки Проект Гутенберг
- 2) 50 самых популярных и самых непопулярных текстов в каждом жанре
- 3) обучающая выборка
- 4) обучение на основе 1000 первых текстов с учётом лингвистических параметров

Лингвистические параметры

- 1) лексика: униграммы и биграммы
- 2) части речи: распределение слов в текстах по частям речи
- 3) простые грамматические характеристики: распределение в текстах некоторых простых синтаксических структур
- 4) сложные грамматические характеристики: распределение в текстах некоторых сложных синтаксических структур
- 5) тональность и коннотации: слова, обозначающие чувства, и слова, имеющие дополнительные эмоциональные или оценочные значения

Результат

- 84% - максимальная популярность жанра «Приключения».
- Алгоритм может быть доработан для оценки и прогнозирования успешности научных статей.

Глоттохронология

- Два языка развиваются из праязыка независимо друг от друга.
- Можно вычислить долю совпадающих слов в основных списках (ОС) этих языков и определить время, прошедшее с момента их разделения.

Доля совпадения между языками

Корневая глоттохронология

Этимологическая статистика

Лексикостатистическая классификация

Частотные словари

- **Лемматизация** словоформ
- **Общая частота** – число употреблений на млн слов корпуса
- **Ранг** леммы или словоформы по частотности позволяет составлять лексические минимумы для изучения языков, их разных функциональных стилей

Квантитативная морфология

Корпус	Им.	Род.	Дат.	Вин.	Тв.	Предл.
НКРЯ	27,06	29,23	5,98	18,66	8,44	10,63
ХАНКО	24,30	32,62	5,50	17,73	8,08	11,78
Josselson	38,80	16,80	4,70	26,30	6,50	6,90
Steinfeldt	33,60	24,60	5,10	19,50	7,80	9,40

Выводы

- 1) квантитативные исследования позволяют выяснить, как язык используется в разных сферах коммуникации
- 2) частотные характеристики отличаются в разные периоды времени
- 3) частота использования связана со структурными свойствами языка (usage-based grammar)

Благодарю за внимание!

- Вопросы?