Введение в компьютерную и корпусную лингвистику

Селегей В.П.

20 февраля 2016 г.

Это всё будет обновляться и дизайниться по мере хода лекций (надеюсь) и по мере отсутствия лени у меня (верю). Буду рад обсудить разные фишечки live-TeXинга с теми, кто практикует.
AQ

0 Предисловие

Все предметы на КЛ делятся на 4 блока:

- База Физтеха: теория графов, автоматов, формальные грамматики.
- Лингвистика: сначала введение (в прошлые года в бакалавриате не было), в магистратуре будет уже по темам: морфология, синтаксис, семантика.
- Обзор моделей для лингвистики: различные open-source прикладные проекты
- Задачи: машинный перевод, распознавание и т.д.

Вот здесь будет второй пункт – базовые понятия лингвистики.

Первые 2 лекции будут совсем вводные в лингвистику и компьютерную лингвистику, дальше уже по сути пойдём.

1 Лекция 1. 18.02.2016

1.1 Чем занимается лингвистика?

Итак, чем же занимается лингвистика?

Обычно под "лингвистом"подразумевается человек, знающий языки. Здесь определяем более узкий срез понятия:

Определение 1. Лингвистика теоретическая – наука о естественном человеческом языке вообще и о всех языках мира как индивидуальных его представителях.

Сейчас современная наука (напр. на конференциях) более тяготеет к разнообразию в исследуемых языках, как следствие, необходимо развивать общие формальные модели.

Всё это ведёт к методологическому переходу от субъективной интроспекции (самоосознание внутри одного языка) к объективным корпусным и нейролингвистическим методам исследований.

3 кита теоретической лингвистики: типологический подход, формальные представления, объективные связи.

Прикладная же лингвистика занимается NLP (Natural Language Processing).

1.2 А "компьютерная"лингвистика (КоЛинг)?

Суть КоЛинг – автоматическая обработка ЕЯ (естественного языка) в научных или прикланых целях.

КоЛинг как часть лингвистики:

- Теория: создание формальных моделей языка (*аналогия: модель расширяющейся Вселенной*)
- Практика: применение компьютеров в лингвистических исследованиях (*аналогия: компьютерная медицина*)
- Инженерная деятельность (наша проекция практики): решения с помощью компов (*парсеры, корпуса, лингворесурсы*) различных задач обработки ЕЯ (*маш. перевод звучащей речи, обработка запросов на ЕЯ, анализ новостей, генерация репортажа по видео, голосовое управление аппаратом и т.д.*)

В результате получается эдакий "Лингвистический пылесос".

В идеале между Линг и КоЛинг будет связь: модели языка от гуманитариев соединяются с методами получения языковых данных от прогеров (*или наоборот, я что-то запутался*), что ведёт к общей пользе ура.

1.3 Окей, как это смоделировать?

Определение 2. Язык – это устройство для кодирования значений с помощью системы специальных средств в целях коммуникации.

А что такое "значение"? Мысль? Идея? Её нельзя положить на стол, препарировать и распознать.

Есть различные структуры-попытки формализовать всё это.

// Теория Роджера Шенка (wiki)

// Compreno-модель

В принципе, можно пытаться снять энцефалограмму мозга при различных мыслях, но по понятным причинам это слишком сложно, чтобы работать. Пока что люди до анализа таких данных не доросли.

1.4 Естественный язык: basics

Основоположником всего этого был, конечно, Ноам Хомский (Chomsky).

Итак, язык – это кодирующая система.

Якобсон пытался формализовать весь процесс коммуникации (адресант, адресат, код, котнакт, сообщение и т.д.)

Само кодирование смысла получается через различные средства: на самом верхнем уровне они делятся на лексические и грамматические средства.

Далее, если высказыание на ЕЯ – это фраза, предложение, то описывать (иными словами, пытаться формализовать) их мы можем на поверхностном (например, в виде синтаксического анализа, довольно легко) и глубинном (что-то типа семантического анализа, гораздо сложнее) уровнях.

Пример 1. Мы можем сказать, что "в аудитории 16 человек"великим множеством способов ==> язык как средство выражения чудовищно избыточен.

Кроме того, надо понять, что смысл передаётся не только языковыми средствами (мимика, жесты, интонирование). Все их надо учитывать.

1.5 Средства описания значения

Как мы можем представлять фразу математически? (простейший случай – bag of words)

Из проги можем вспомнить деревья-графы, мб даже семантические сети.

В таком случае, какая глубина деревьев и вообще как их настраивать?

В итоге:

- глубина описаний
- язык описания значений
- связь с неязыковыми структурами.

1.6 Грамматические средства кодирования

Попробуем перечислить все уровни выражения мыслей. Начинаем с кирпичиков-слов.

Идеи: Оттуда берутся словоформы, куда затем можно добавить служебные слова. Потом организуется связь с артикуляцией через пунктуацию.

Систематизируем:

1 Самое главное и образующее свойство - линейный порядок слов. "пять человек"и "человек пять разница в примерности, которую даёт лишь порядок слов

   Ещё: "Mary gave John an apple"vs "John gave Mary an apple".

2 Словоформы и словообразование.

"Отец купил сыну..."и "Отцу купил сын"

"прыгнуть-подпрыгнуть-перепрыгнуть"

"выпить – выпивать"

3 Вспомогательные слова

4 Пунктуация

5 Просодика: акценты, паузы, интонация

6 Новые средства кодирования: смайлы, зачёркивания, гипертекст, шрифт и прочая графика.

1.7 Простые примеры

Предложение:

Этот парень делает любую вещь, к которой прикоснётся, неисправной.

Вообще, школьный синтаксический разбор суть обыкновенное дерево разбора. А что тогда корень дерева? Подлежащее или сказуемое? В школе думают о первом, но это чушь.

Главное слово должно объяснять много о устройстве предложения. В данном случае "парень"не даёт почти никакой информации, это может быть кто(что) угодно без ущерба смыслу. Значит, лучше думать о глаголе.

Итак, "делает вершина дерева.

Делает – кто? Парень (именительный падеж) А детектим его через число+падеж (изменяемые, берём окончание –> зависит от конкретного предложения и определяет предложение)+род+одушевлённость (классифицирующие: отсюда берём и смысл)

Вещь – определяем падеж по наличию зависимости "любую"(неодуш, ...).

Неисправной: падеж берётся от глагола, а всё остальное – от "вещи". ещё одно такое слово – "(к) которой у них по 2 родителя де-факто. Т.е. чисто древесная структура не работает.

Де-факто тут есть эллипсис. Это когда мы неполно опускаем смысловые слова (мб повторяющиеся и т.д.)

1.8 Система уровней анализа языка

0 Лексический анализ (слова, знаки препинания, цифры и т.д.)

1 Морфологический анализ (все возможные грамматические характеристики выделенных лексем)

2 Синтаксический анализ (установление базовых связей между словами)

Тут начался интерпретационный уровень

3 Семантический анализ (понятие структуры, связанной со смыслом, но при этом мы остаёмся в границах языка)

4 Прагматический анализ (интерпретация семантики в контексте конкретной ситуации, отнологии (общих знаний об устройстве мира))

картинка про треугольник перевода от Селегея

1.9 Задачи и уровни их решения

1) Проверка орфографии
2) Машинный переносы