Контрольная по АОЕЯ

*Леша Виняр*

Часть 1.

1. Задачи – определение микро- (анафора, кореферентность и тд) и макроструктуры дискурса. Высокоуровневыми задачами, для которых требуется «дискурсивный модуль», является извлечение фактов и сбор и анализ мнений.

2. Расстояние Левенштейна – «емкость» операций вставки, замены или удаления символа, отличающих одну строку от другой (грубо говоря – количество таких операций).

Между мошка и мускат – расстояние 2 (заменить о на у и вставить в конце т).

3. Нормализация – это необходимый этап обработки текста, при котором входные данные приводятся к некоторому выбранному стандарту, в котором данные о словоформе отражены либо в виде леммы (начальной формы) – лемматизация, либо в виде «псевдоосновы» - стемминг (выбирается какой-то вариант основы, от которого удобно строить грамматические формы) и грамматических тэгов. Как уже сказано выше, есть два типа – стемминг и лемматизация.

Стемминг дает неудовлетворительные результаты в том случае, когда у нас между «основой» и «окончанием» происходят сложные морфонологические чередования, сильно осложняющие простую «аглюттинативную» схему. Например – беглая гласная – не совсем просто, как ее включать или нет в «основу».

4. Лексическая вероятность – вероятность самой по себе формы из пары «омонимичных» встретиться в этом значении (в самом широком смысле – от грамматического до семантического). Когда мы используем лексическую вероятность, то допускаем, что она у одного элемента одинакова в разных текстах (грубо говоря – не учитываем контекст). Вклад лексической вероятности очень высок в том случае, когда один из «омонимов» очень частотен, а другой – очень редок. К примеру, форма «кругом» гораздо чаще является наречием, а не существительным в инструменталисе (одного этого различия может хватить чтобы принять решение).

5.

Существуют конструкции, в которых критерии вершинности показывают разные результаты (к примеру – русские констуркции с числителньыми).

Требование единственности вершины иногда заставляет нас испытывать трудности и вводить нули (Петя был мальчиком, который видел только одним глазом) – видел поверхностно является как бы зависимым и «мальчика», и «который»

Непроективность

Часть 2.

7.

Используя подкорпус из 6 003 398 слов (омонимия снята), можно это посчитать. На запрос «существительное на расстоянии 1 после наречия» было найдено 15 388 вхождений, на запрос «существительное на расстоянии 1 после прилагательного» было найдено 344 134. Теперь посчитаем условную вероятность – насколько выше вероятность увидеть перед существительным (условие) прилагательное (событие А), чем увидеть наречие (событие Б). Поскольку нас интересуют только эти исходы – представим, что в принципе нам встречаются перед существительными только наречия или прилагательные.

Тогда получаем примерную вероятность:

(Adv & Noun) = 0,0428

(Adj & Noun) = 0,9571

8.

|  |  |  |
| --- | --- | --- |
| S, VP (S -> VP; VP -> V NP) |  |  |
| 0 (!VP -> V Det) | NP (NP -> Det N) |  |
| V | Det | N, NP |
| Build | a | house |

Построим табличку, пронумеровав строки и столбцы начиная с самого левого и самого нижнего. Теперь двигаемся слева напаво, «увеличивая» с каждой строкой число n в n-граммах, и смотря, можно ли их получить, применяя правила. В клетках мы записали, какими правилами получаются или не получаются n-граммы. С помощью этих правил можно получить такое предложение.

КС-грамматика:

|  |  |
| --- | --- |
| S → NP VP  S → VP  NP → Det N  NP → N  NP → Pron  VP → V NP  VP → V Part  VP → V NP NP  VP → Aux V PP  PP → Prep NP | Det → a / the  N → John / house / Mary / roses / curse  Pron → me  V → built / ran /gave / come / build / run  Part → away  Aux → has  Prep → upon |