Московский Государственный Университет имени М.В. Ломоносова
Факультет вычислительной математики и кибернетики

Задание на сглаживание в языковых статистических моделях в рамках курса «Автоматическое порождение графов знаний»

ОТЧЕТ

Работу выполнил:
Саховский Андрей Сергеевич
Группа 618/1

Рассмотренные модели

Были рассмотрены следующие N-граммные языковые статистические модели:

- Униграммная модель без сглаживания;
- Биграммная модель без сглаживания;
- Униграммная модель с аддиктивным сглаживанием;
- Биграммная модель с аддиктивным сглаживанием;
- Простая линейная интерполяционная модель на основе униграммной и биграммной моделей со сглаживанием.

Униграммные модели

В таблице 1 приведены частоты и вероятности слов в униграммных моделях без сглаживанием и со сглаживанием Лапласа с параметром λ, равным 1. Частоты и вероятности посчитаны по тренировочному тексту для слов, присутствующих в тренировочном и тестовом текстах.

Таблица 1: Частоты и вероятности слов в униграмных моделях без сглаживания и со сглаживанием Лапласа.

Слово	**Частота**		**Вероятность**	
	Без сглаживания	**Сглаживание Лапласа (+1)**	**Без сглаживания**	**Сглаживание Лапласа (+1)**
а	2	3	0.053	0.048
в	4	5	0.105	0.079
весёлая	1	2	0.026	0.032
ворует	1	2	0.026	0.032
вот	1	2	0.026	0.032
джек	3	4	0.079	0.063
дом	1	2	0.026	0.032
доме	2	3	0.053	0.048
и	0	1	0.000	0.016
кот	0	1	0.000	0.016
которая	3	4	0.079	0.063
который	3	4	0.079	0.063
ловит	0	1	0.000	0.016
построил	3	4	0.079	0.063
птица	1	2	0.026	0.032
пугает	0	1	0.000	0.016
пшеница	1	2	0.026	0.032
пшеницу	1	2	0.026	0.032
синица	1	2	0.026	0.032
синицу	0	1	0.000	0.016
тёмном	2	3	0.053	0.048
хранится	2	3	0.053	0.048
часто	1	2	0.026	0.032
чулане	2	3	0.053	0.048
это	2	3	0.053	0.048
Σ	37	62	1.0	1.0

Перплексия униграммных и биграммных моделей

На рис. 1 представлены значения перплексии униграммных и биграммных моделей с аддиктивным сглаживанием при различных значениях параметра сглаживания λ, лежащим в интервале $[0; 1]$ с шагом 0.01. Построение моделей осущетвлялось

на основе тренировочного текста, перплексия была посчитана по тестовому тексту. Наилучшие значения λ составили 0.01 и 1 для униграммной и биграммной моделей соответственно. При значении $\lambda = 0$ модели вырождаются в соответствующие модели без сглаживания. По построенным графикам можно заметить, что при добавлении сглаживания униграммная, и биграммная модели начинают демонстрировать существенно более низкую (лучшую) перплексию за счёт исчезновения нулевых значений вероятности на этапе подсчёта вероятности текста и подсчёта перплексии. С ростом λ униграммная модель начинает показывать более низкую перплексию, а для биграммной модель перплексия увеличивается.

Линейная интерполяционная модель

На основе лучших построенных униграммной и биграммных моделей (с параметрами сглаживания 0.01 и 1 соответственно) была построена простая линейная интерполяционная модель:

$$P_{li}(w_n|w_{n-1}) = \mu P(w_n) + (1 - \mu)P(w_n|w_{n-1})$$

Результаты подбора веса униграммной и биграммной моделей (параметр μ) представлены на рис. 2. Результаты эксперимента показали, что наилучшая перплексия на тестовом тексте достигается при значении $\mu = 0$, то есть при использовании биграммной модели без униграммной модели.

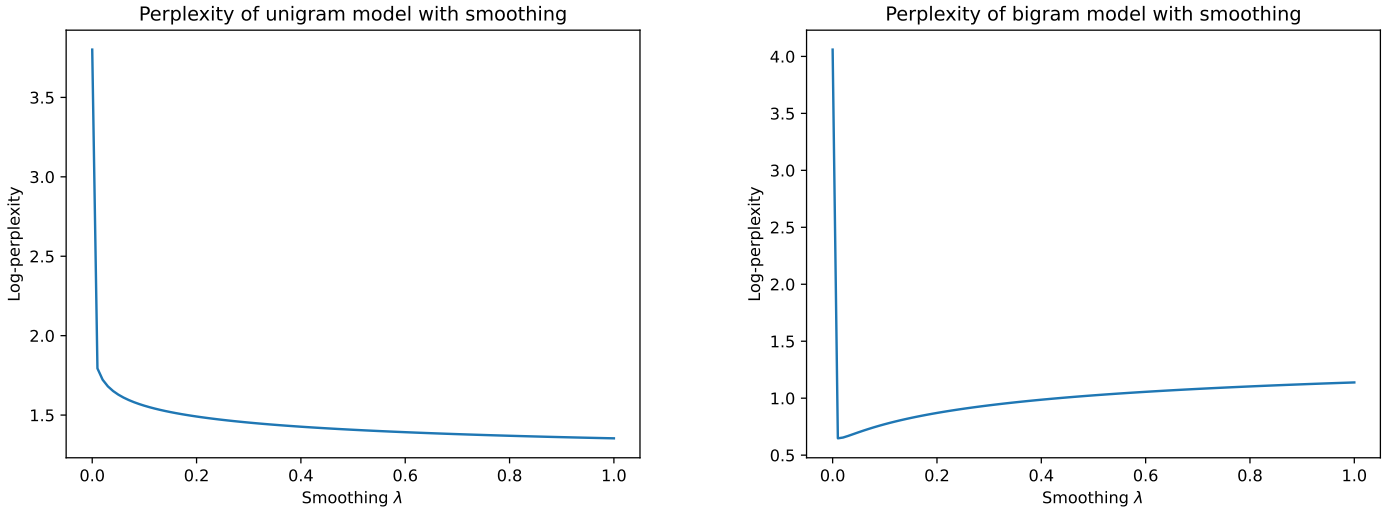

Рис. 1: Перплексия обученных на тренировочном тексте униграммной и биграммной моделей с различным параметром сглаживания λ.

Результаты сравнения рассмотренных моделей приведены в таблице 2. Результаты экспериментов показали неэффективность униграммных и биграммных моделей без сглаживания. Их неэффективность объясняется наличием нулевых вероятностей при подсчёте перплексии. Наилучших значений перплексии достигла биграммная модель с аддитивным сглаживанием при $\lambda = 0.01$. Использование линейной интерполяционной модели не привело к улучшению перплексии относительно биграммной модели.

Таблица 2: Частоты и вероятности слов в униграмных моделях без сглаживания и со сглаживанием Лапласа.

Модель	**Перплексия**
Униграммная, без сглаживания	6323.6
Биграммная, без сглаживания	11475.6
Униграммная, со сглаживанием (+1)	25.4
Биграммная, со сглаживанием (+1)	13.7
Биграммная, со сглаживанием ($\lambda = 0.01$)	4.4
Линейная интерполяционная, со сглаживанием	4.4

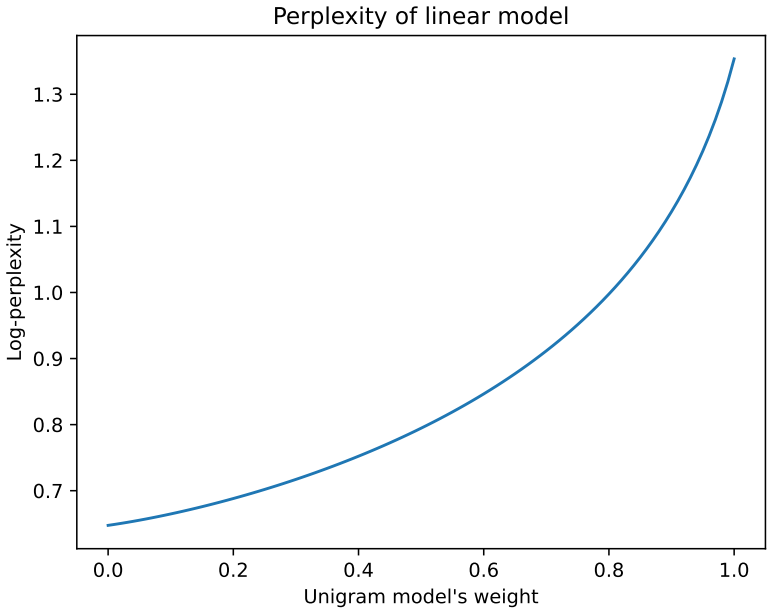

Рис. 2: Перплексия обученной на тренировочном тексте простой линейной интерполяционной модели на основе униграммной и биграммной моделей с различным весом униграммной/биграммной моделей.

Заключение

Были рассмотрены униграммные и биграммные языковые статистические модели. Наилучшие значения перплексии показала биграммная модель с аддитивным сглаживанием. Было обнаружено, что использование аддитивного сглаживание в униграммных и биграммных моделей позволяет существенно улучшить значения перплексия за счёт избавления от нулевых вероятностей униграмм и биграмм. Эксперименты с простой линейной интерполяционной модели не позволили улучшить значения перплексии относительно биграммной аддитивной модели и было обнаружено, что наилучшие значения перплексии на тестовом тексте достигаются при вырождении линейной модели в биграммную.