1 Advanced Language Modelling

Exercise 1.1: Kneser-Ney Language Models

a) We know that,

$$\sum_{v \epsilon V} P_{KN}(w3 = v \mid w1w2) = 1 \text{ and}$$

$$\sum_{v \epsilon V} P_{KN}(w3 = v \mid w2) = 1, \text{ where } N(w1w2) > 0$$

$$P_{KN}(w3 \mid w1w2) = \frac{\max\{N(w1w2w3) - d, 0\}}{N(w1w2)} + \lambda(w1w2)\, P_{KN}(w3 \mid w2) \text{ [From eqn. (2)]}$$

Plugging the value of $\sum_{v \epsilon V} P_{KN}(w3 = v \mid w1w2)$ *in eqn.* (2):

$$1 = \frac{\max\{N(w1w2w3) - d, 0\}}{N(w1w2)} + \lambda(w1w2)\, P_{KN}(w3 \mid w2)$$

$$1 - \frac{\max\{N(w1w2w3) - d, 0\}}{N(w1w2)} = \lambda(w1w2)\, P_{KN}(w3 \mid w2)$$

So the back-off factor, $\lambda(w1w2) = \dfrac{\left(1 - \frac{\max\{N(w1w2w3) - d, 0\}}{N(w1w2)}\right)}{P_{KN}(w3 \mid w2)}$

$$P_{KN}(w3 \mid w2) = \frac{\max\{N_{1+}(\bullet\, w2w3) - d, 0\}}{N_{1+}(\bullet\, w2\, \bullet)} + \lambda(w2)\, P_{KN}(w3) \text{ [From eqn. (4)]}$$

Plugging the value of $\sum_{v \epsilon V} P_{KN}(w3 = v \mid w2)$ in eqn. (4):

$$1 = \frac{\max\{N_{1+}(\bullet\, w2w3) - d, 0\}}{N_{1+}(\bullet\, w2\, \bullet)} + \lambda(w2)\, P_{KN}(w3)$$

$$1 - \frac{\max\{N_{1+}(\bullet\, w2w3) - d, 0\}}{N_{1+}(\bullet\, w2\, \bullet)} = \lambda(w2)\, P_{KN}(w3)$$

So the back-off factor, $\lambda(w2) = \dfrac{\left(1 - \frac{\max\{N_{1+}(\bullet\, w2w3) - d, 0\}}{N_{1+}(\bullet\, w2\, \bullet)}\right)}{P_{KN}(w3)}$

b)

	w = 'york'	w = 'matter'
N(w)	2374	2367
$N_{1+}(\bullet\, w)$	9	270
$-\log_2 P_{ML}(w)$	12.258374283571046	12.262634512478337
$-\log_2 P_{Lids}(w)$	12.268770360977008	12.273028793453731
$-\log_2 P_{KN}(w)$	17.964084485236466	13.057193889627946

Table 1

From the table above we can see that the word ‘york’ and the word ‘matter’ almost have the same frequency. However, by observing their bigrams we can observe that ‘york’ comes after only 9 words whereas, ‘matter’ comes after 270 different words. So, in a unknown context the probability of the next word being ‘matter’ should be more than the ‘york’ which comes after some certain words. The maximum likelihood estimation and the lidstone smoothing technique give almost same probability which is not accurate in such kind of scenarios. However, Kneser-Ney smoothing can do the job way better by checking the history (bigram) and gives a better prediction of the next word in an unknown context. Since, ‘matter’ is preceded by more different words, it should have more probability than the word ‘york’ in an unknown context and Kneser-Ney smoothing is doing exactly the same by calculating ‘york’ as 17.964084485236466 and ‘matter’ 13.057193889627946. As it is expressed as – log 2 base the lower the value the better the probability of getting selected as a next word in an unknown context.

c) Among the various smoothing techniques Kneser-Ney smoothing is the most precise and gives better estimation. It is usually applied in interpolated form and is developed on the idea of absolute discounting interpolation that makes the use of both higher and lower order LM. The process is done by moving probability mass from higher to lower order language model.

The Kneser-Ney model uses technique of absolute discount smoothing interpolation as a backoff model. The Kneser-Ney design retains the first term of absolute discounting interpolation, but rewrites the second term. Whereas absolute discounting interpolation in a bigram model would simply default to a unigram model in the second term, Kneser-Ney depends upon the idea of a continuation probability associated with each unigram.

If we recall the maximum likelihood estimation of unigram language model we get the following:

$$P_{ML}\ (\mathrm{w}) = \frac{c(w)}{\sum_i c(w_i)}$$

However, from the equations of a) we can clearly see that Kneser-Ney smoothing replace raw counts with the count of histories. Instead of determining how likely a word w_i is to appear in unigram model, the Kneser-Ney smoothing evaluates how likely a word w_i is to appear in an unfamiliar bigram context.

2 Text Classification

2.1 Author indefication using LMs

Dickens				Doyle				Twain			
Unigrams		Bigrams		Unigrams		Bigrams		Unigrams		Bigrams	
'the'	65440	'$< s >$', 'i'	7392	'the'	72028	'$< s >$', 'i'	10863	'the'	67253	'$< s >$', 'i'	7105
'and'	50579	('of', 'the'	6559	'of'	34862	'of', 'the'	8869	'and'	59974	'of', 'the'	7029
'to'	36773	'in', 'the'	6135	'and'	34405	'in', 'the'	6003	'a'	34965	'in', 'the'	5607
'of'	36288	'$< s >$', 'the'	4212	'i'	31920	'$< s >$', 'it'	5271	'to'	33284	'$< s >$', 'the'	4997
'a'	28874	'to', 'the'	3495	'a'	31080	'$< s >$', 'the'	5181	'of'	30456	'$< s >$', 'he'	4062
'i'	26157	'$< s >$', 'he'	3188	'to'	30829	'$< s >$', 'he'	4855	'i'	26541	('it', 'was'	3687
'in'	24205	'to', 'be'	2981	'that'	21884	'it', 'was'	4068	'it'	25226	'it', $'< /s >'$	3259
'that'	18512	'it', 'was'	2615	'it'	21448	'to', 'the'	3589	'was'	20767	'and', 'the'	3214
'was'	17260)	on', 'the'	2550	'in'	21173	'it', 'is'	3290	'in'	19669	'$< s >$', 'it'	3178
'it'	16969	'$< s >$', 'it'	2528	'was'	18735	'$< s >$', 'you'	3117	'he'	18982	'to', 'the'	3142
'he'	16635	'and', 'the'	2441	'he'	18161	'i', 'have'	2767	'that'	17877	'he', 'was'	2459
'his'	15938	'in', 'a'	2255	'you'	17336	'at', 'the'	2714	'you'	13064	'don', 't'	2379
'her'	12926	'$< s >$', 'but'	2194	'his'	14661	'it', $'< /s >'$	2600	't'	11101	'$< s >$', 'but'	2349
'with'	12582	'with', 'a'	2136	'is'	12663	'$< s >$', 'but'	2483	'his'	10703	'was', 'a'	2227
'you'	12313	'of', 'his'	2071	'had'	11265	'that', 'i'	2348	'but'	10442	'$< s >$', 'and'	2170)

c)

f) Unigrams:

	Dickens	Doyle	Twain
DickensLM	**963**	**965**	830
DoyleLM	1100	669	865
TwainLM	1033	1001	**513**

g) Bigrams:

	Dickens	Doyle	Twain
DickensLM	**1039**	809	919
DoyleLM	1281	**320**	1017
TwainLM	1402	1003	**286**

i) The Lm work quite well,with only one mistake in the unigram (see table).This is most likely caused by the fact that the authors have very different writing style. For example different words to begin the sentences, like you can see in the table of ex 2.2. Important to say that the bigram have a way better result than the unigrams. This is caused by the fact that they have a similar vocabulary in their text and only the combination

of the words or the beginning of the sentences show realy big differences between the authors.

2.2 Feature Selection

Using the point-wise mutual information (PMI) measure, we explored the most discriminating n-grams for each author: Charles Dickens, Arthur Doyle, and Mark Twain.

c) A table of the top 10 unigram features for each author (highest positive PMI value):

	Dickens	Doyle	Twain
1	('her',)	('is',)	('t',)
2	('with',)	('you',)	('and',)
3	('his',)	('had',)	('but',)
4	('in',)	('i',)	('it',)
5	('to',)	('that',)	('a',)
6	('of',)	('his',)	('was',)
7	('and',)	('the',)	('he',)
8	('had',)	('of',)	('to',)
9	('the',)	('he')	('the',)
10	('that',)	('it',)	('i',)

Unigram features for all three authors very similar; there are many common features. Therefore, they cannot give any information for author identification or for characterization of author style. Though there are miserable differences: each author has at least one unique unigram: Dickens – (‘her’), Doyle – (‘you’), Twain - ('t'), ('but'), ('was',), however it cannot provide strong assurance.

d) A table of the top 10 bigram features for each author (highest positive PMI value):

	Dickens	Doyle	Twain
1	('of', 'his')	('it', 'is')	('<s>', 'and')
2	('to', 'be')	('<s>', 'you')	('<s>', 'the')
3	('that', 'i')	('i', 'have')	(('it', '</s>')
4	('i', 'have')	('that', 'i')	('<s>', 'he')
5	('<s>', 'and')	('<s>', 'it')	('in', 'a')
6	('in', 'a')	('on', 'the')	('he', 'was')
7	('<s>', 'but')	('at', 'the')	('<s>', 'but')
8	('he', 'was')	('and', 'the')	('was', 'a')
9	('don', 't')	('it', 'was')	('in', 'the')
10	('with', 'a')	('of', 'the')	('of', 'the')

Bigram features provide more information for characterization of author style or author identification. Each author has specific features that could point out at characteristic structures in their works:

- Dickens: ('<s>', 'but'), ('don', 't') – frequently used negation and imperative mood.

- Doyle: ('it', 'is'), ('it', 'was') – explanations of what happened; ('<s>', 'you') – a lot of dialogs.
- Twain: ('he', 'was'), ('was', 'a') – narration in the past tense; ('<s>', 'he'), ('he', 'was') – male heroes.