Monsoon 2022

Spelling Correction

- Independent Word Spelling Correction

Dr. Rajendra Prasath

Indian Institute of Information Technology Sri City, Chittoor

06th September 2022 (rajendra.2power3.com)

> Topics to be covered

- Recap:
  - Phrase Queries
  - Proximity Search
  - Permuterm Index
  - Bi-gram Indexes

- Spelling Correction
  - Independent Word Spelling Correction
    - Spelling Detection
    - Specific tasks in Spelling Correction
    - Spelling Suggestion

- More topics to come up … Stay tuned …!!

Overview

Recap: Information Retrieval

- **Information Retrieval (IR)** is finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers).

- During these days, we frequently think first of web search, but there are many other cases:
  - E-mail search
  - Searching your laptop
  - Corporate knowledge bases
  - Legal information retrieval
  - Images / Patent / Transportation Related Searches
    - and so on . . .

Recap: Wild-card queries: *

- **mon*:** find all docs containing any word beginning with "mon".

- Easy with binary tree (or B-tree) dictionary: retrieve all words in range: **mon ≤ w < moo**

- ***mon:** find words ending in "mon": harder
  - Maintain an additional B-tree for terms backwards.

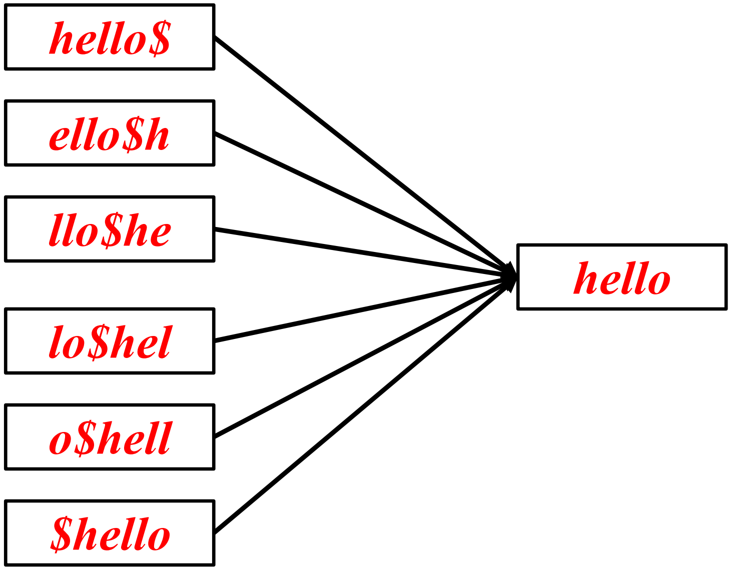
  Can retrieve all words in range: **nom ≤ w < non**

From this, how can we enumerate all terms meeting the wild-card query ***pro*cent*** ?

Recap: Permuterm index

- Add a $ to the end of each term
- Rotate the resulting term and index them in a B-tree
- For term hello, index under:
  - **hello$, ello$h, llohe, lohel, o$hell, $hello**

  where $ is a special symbol.



Empirically, dictionary quadruples in size

Spelling Correction

Non-word Spelling Error - An Example

acress

Words within 1 of acress

Error	Candidate Correction	Correct Letter	Error Letter	Type
acress	actress	t	-	deletion
acress	cress	-	a	insertion
acress	caress	ca	ac	transposition
acress	access	c	r	substitution
acress	across	o	e	substitution
acress	acres	-	s	Insertion / deletion

How to we perform Channel Modeling?

Channel model

$$P(x|w) = \begin{cases} \dfrac{\text{del}[w_{i-1}, w_i]}{\text{count}[w_{i-1} w_i]}, & \text{if deletion} \\ \dfrac{\text{ins}[w_{i-1}, x_i]}{\text{count}[w_{i-1}]}, & \text{if insertion} \\ \dfrac{\text{sub}[x_i, w_i]}{\text{count}[w_i]}, & \text{if substitution} \\ \dfrac{\text{trans}[w_i, w_{i+1}]}{\text{count}[w_i w_{i+1}]}, & \text{if transposition} \end{cases}$$

Kernighan, Church, Gale 1990

Smoothing probabilities: Add-1 smoothing

- But if we use the confusion matrix example, unseen errors are impossible!
- They'll make the overall probability 0. That seems too harsh
  - e.g., in Kernighan's chart q➔a and a➔q are both 0, even though they're adjacent on the keyboard!
- A simple solution is to add 1 to all counts and then if there is a |A| character alphabet, to normalize appropriately:

$$\text{If substitution, } P(x \mid w) = \frac{\text{sub}[x,w]+1}{\text{count}[w]+A}$$

Channel model for acress

Candidate Correction	Correct Letter	Error Letter	*x|w*	*P(x|w)*
actress	t	–	c|ct	.000117
cress	–	a	a|#	.00000144
caress	ca	ac	ac|ca	.00000164
access	c	r	r|c	.000000209
across	o	e	e|o	.0000093
acres	–	s	es|e	.0000321
acres	–	s	ss|s	.0000342

Noisy channel probability for acress

Candidate Correction	Correct Letter	Error Letter	*x|w*	*P(x|w)*	*P(w)*	10^9 * *P(x|w)** *P(w)*
actress	t	-	c|ct	.000117	.0000231	2.7
cress	-	a	a|#	.00000144	.000000544	.00078
caress	ca	ac	ac|ca	.00000164	.00000170	.0028
access	c	r	r|c	.000000209	.0000916	.019
across	o	e	e|o	.0000093	.000299	2.8
acres	-	s	es|e	.0000321	.0000318	1.0
acres	-	s	ss|s	.0000342	.0000318	1.0

Noisy channel probability for acress

Candidate Correction	Correct Letter	Error Letter	x|w	P(x|w)	P(w)	10^9 *P(x|w)P(w)
actress	t	-	c|ct	.000117	.0000231	2.7
cress	-	a	a|#	.00000144	.000000544	.00078
caress	ca	ac	ac|ca	.00000164	.00000170	.0028
access	c	r	r|c	.000000209	.0000916	.019
across	**o**	**e**	**e|o**	**.0000093**	**.000299**	**2.8**
acres	-	s	es|e	.0000321	.0000318	1.0
acres	-	s	ss|s	.0000342	.0000318	1.0

Evaluation

- Some spelling error test sets
  - Wikipedia's list of common English misspelling
  - Aspell filtered version of that list
  - Birkbeck spelling error corpus
  - Peter Norvig's list of errors (includes Wikipedia and Birkbeck, for training or testing)

SPELLING CORRECTION WITH THE NOISY CHANNEL

Context-Sensitive Spelling Correction

Real-word spelling errors

- … leaving in about fifteen ***minuets*** to go to her house.
- The design ***an*** construction of the system…
- Can they ***lave*** him my messages?
- The study was conducted mainly ***be*** John Black.

- 25-40% of spelling errors are real words(Kukich 1992)

Context-sensitive spelling error fixing

- For each word in sentence (phrase, query …)
  - Generate *candidate set*
    - the word itself
    - all single-letter edits that are English words
    - words that are homophones
    - (all of this can be pre-computed!)
- Choose best candidates
    - Noisy channel model

Noisy channel for real-word spell correction

- Given a sentence $x_1,x_2,x_3,\ldots,x_n$
- Generate a set of candidates for each word x_i
  - Candidate(x_1) = {x_1, w_1 , w'_1 , w''_1 ,…}
  - Candidate(x_2) = {x_2, w_2 , w'_2 , w''_2 ,…}
  - Candidate(x_n) = {x_n, w_n , w'_n , w''_n ,…}
- Choose the sequence W that maximizes P(W|$x_1,\ldots,x_n$)

$$\hat{w} = \underset{w \in V}{\operatorname{argmax}} P(w \mid x)$$

$$= \underset{w \in V}{\operatorname{argmax}} P(x \mid w) P(w)$$

Incorporating context words: Context-sensitive spelling correction

- Determining whether **actress** or **across** is appropriate will require looking at the context of use

- We can do this with a better **language model**

- A **bigram language model** conditions the probability of a word on (just) the previous word

$$P(w_1 \ldots w_n) = P(w_1)P(w_2 \mid w_1) \ldots P(w_n \mid w_{n-1})$$

Incorporating context words

- For unigram counts, P(w) is always non-zero
  - if our dictionary is derived from the document collection
- This won't be true of $P(w_k \mid w_{k-1})$. We need to smooth
- We could use add-1 smoothing on this conditional distribution
- But here's a better way – interpolate a unigram and a bigram:
  - $Pl_i(w_k \mid w_{k-1}) = \lambda P_{uni}(w_k) + (1-\lambda) Pb_i(w_k \mid w_{k-1})$
    - $Pb_i(w_k \mid w_{k-1}) = C(w_{k-1}, w_k) \,/\, C(w_{k-1})$

All Important Points

- Note that we have several probability distributions for words
  - Keep them straight!
    - You might want/need to work with log probabilities:
      - log P(w1…wn) = log P(w1) + log P(w2|w1) + … + log P(wn|wn−1)
  - Otherwise, be very careful about floating point underflow
- Our query may be words anywhere in a document
  - We'll start the bigram estimate of a sequence with a unigram estimate
  - Often, people instead condition on a start-of-sequence symbol, but not good here
  - Because of this, the unigram and bigram counts have different totals – not a problem

Using a bigram language model

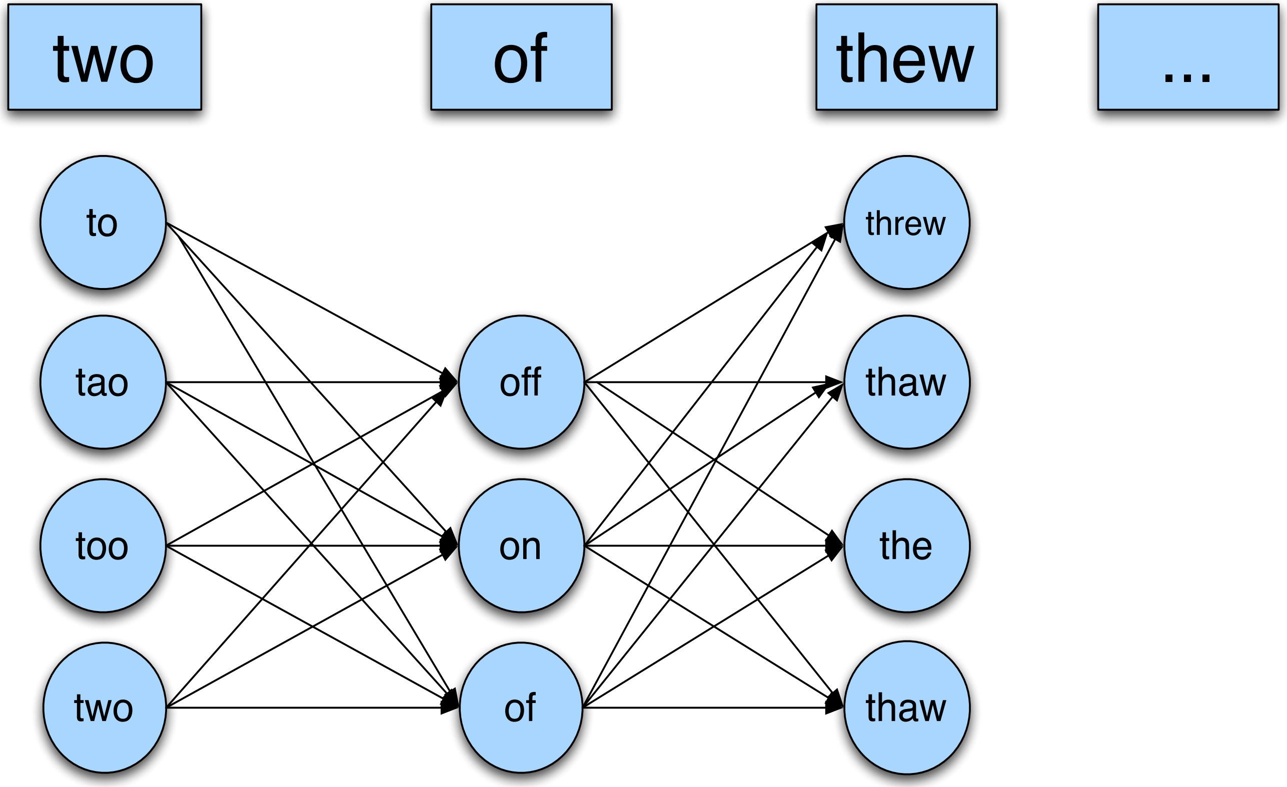
- “a stellar and versatile acress whose combination of sass and glamour…”

- Counts from the Corpus of Contemporary American English with add-1 smoothing

- P(actress | versatile)=.000021 P(whose | actress) = .0010
- P(across | versatile) =.000021 P(whose | across) = .000006

- P(“versatile actress whose”) = .000021*.0010 = 210 x10-10
- P(“versatile across whose”) = .000021*.000006 = 1 x10-10

Noisy channel for real-word spell correction



Simplification: One error per sentence

- Out of all possible sentences with one word replaced
  - w_1, $\mathbf{w''_2}$,w_3,w_4 two **off** thew
  - w_1,w_2,$\mathbf{w'_3}$,w_4 two of **the**
  - $\mathbf{w'''_1}$,w_2,w_3,w_4 **too** of thew
  - …
- Choose the sequence W that maximizes P(W)

Where to get the probabilities?

- **Language model**
  - Unigram
  - Bigram
  - etc

- **Channel model**
  - Same as for non-word spelling correction
  - Plus need probability for no error, $P(w|w)$

Probability of no error

- What is the channel probability for a correctly typed word?
- P("the" | "the")
  - If you have a big corpus, you can estimate this percent correct

- But this value depends strongly on the application
  - .90 (1 error in 10 words)
  - .95 (1 error in 20 words)
  - .99 (1 error in 100 words)

Peter Norvig’s “thew” example

x	w	x|w	P(x|w)	P(w)	10^9 P(x|w)P(w)
thew	the	ew|e	0.000007	0.02	144
thew	thew		0.95	0.00000009	90
thew	thaw	e|a	0.001	0.0000007	0.7
thew	threw	h|hr	0.000008	0.000004	0.03
thew	thwe	ew|we	0.000003	0.00000004	0.0001

State of the art noisy channel

- We never just multiply the prior and the error model
- Independence assumptions→probabilities not commensurate
- Instead: Weight them

$$\hat{w} = \underset{w \in V}{\operatorname{argmax}} P(x \mid w) P(w)^{\lambda}$$

- Learn λ from a development test set

Improvements to channel model

- Allow richer edits (Brill and Moore 2000)
    - ent→ant
    - ph→f
    - le→al
- Incorporate pronunciation into channel (Toutanova and Moore 2002)
- Incorporate device into channel
    - Not all Android phones need have the same error model
    - But spell correction may be done at the system level

Summary

In this class, we focused on:

(a) Recap: Positional Indexes

i. Positional Index Size

ii. Wild card Queries

iii. Permuterm index

(b) Spelling Correction

i. Types of Spelling Correction

ii. Noisy Channel modelling for Spell Correction

iii. Spelling Suggestions

Acknowledgements

Thanks to ALL RESEARCHERS:

- Modern Information Retrieval Baeza-Yates and Ribeiro-Neto, Addison Wesley, 1999.

- **Introduction to Information Retrieval Manning, Raghavan and Schutze, Cambridge University Press, 2008.**

- Search Engines Information Retrieval in Practice W. Bruce Croft, D. Metzler, T. Strohman, Pearson, 2009.

- Information Retrieval Implementing and Evaluating Search Engines Stefan Büttcher, Charles L. A. Clarke and Gordon V. Cormack, MIT Press, 2010.

- Many Authors who contributed to SIGIR / WWW / KDD / ECIR / CIKM / WSDM and other top tier conferences

- **Prof. Mandar Mitra, Indian Statistical Institute, Kolkatata (https://www.isical.ac.in/~mandar/)**

Assistance

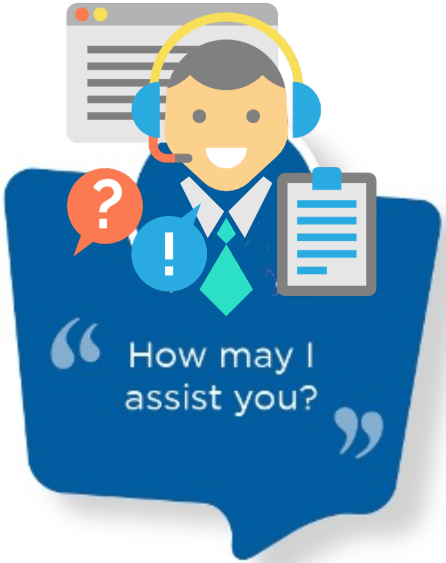
- You may post your questions to me at any time
- You may meet me in person on available time or with an appointment
- You may ask for one-to-one meeting

Best Approach

- You may leave me an email any time
  (email is the best way to reach me faster)

Questions

It’s Your Time



Contact Information

Dr. Rajendra Prasath
IIIT Sri City, Chittoor