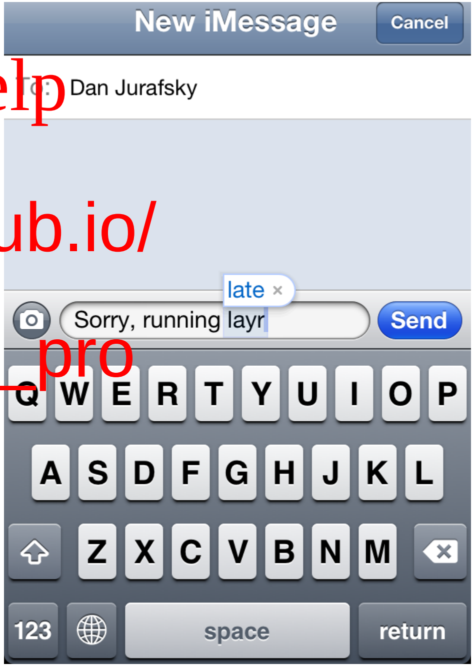
Spelling Correction and the Noisy Channel

pelling Correction Task

Applications for spelling correction

Word processing

Phones

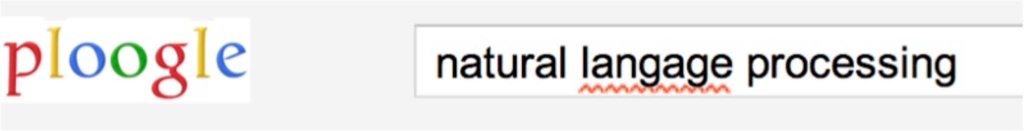


To: Dan Jurafsky

late ×

Sorry, running layr

Web search



natural langage processing

Showing results for natural ***language*** processing

Search instead for natural langage processing

Spelling Tasks

- Spelling Error Detection
- Spelling Error
  - Autocorrect
    - hte→the
  - Suggest a correction
  - Suggestion lists

Types of spelling errors

- Non-word Errors
  - *graffe* → *giraffe*
- Real-word Errors
  - Typographical err
    - *three* → *there*
  - Cognitive Errors (homophones)
    - *piece*→*peace*,
    - *too* → *two*

Rates of spelling errors

26%: Web queries Wang *et al.* 2003

13%: Retyping, *t al.* English&German

7%: Words corre e-sized organizer

2%: Words uncorrected on or oreff &MacKenzie 2003

1-2%: Retyping: Kane and Wobbrock 2007, Gruden et al. 1983

Non-word spelling errors

- Non-word spelling error detection:
  - Any word not in a ***dictionary*** is an error
  - The larger the dic
- Non-word spellin
  - Generate ***candidates***: real words tha error
  - Choose the one which is best:
    - Shortest weighted edit distance
    - Highest noisy channel probability

Real word spelling errors

- For each word *w*, generate candidate set:
  - Find candidate words with similar ***pronunciations***
  - Find candidate w
  - Include *w* in cand
- Choose best candidate
  - Noisy Channel
  - Classifier

Spelling Correction and the Noisy Channel

pelling Correction Task

Spelling Correction and the Noisy Channel

y Channel Model of Spelling

Noisy Channel Intuition

Noisy Channel

- We see an observation x of a misspelled word
- Find the correct word w

$$\hat{w} \qquad {}_{w \in V}$$

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

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

History: Noisy channel for spelling proposed around 1990

- **IBM**

  - Mays, Eric, Fre L. Mercer. 1991. Context based *ation Processing and Management,*

- **AT&T Bell Labs**
  - Kernighan, Mark D., Kenneth W. Church, and William A. Gale. 1990. A spelling correction program based on a noisy channel model. Proceedings of COLING 1990, 205-210

Non-word spelling error example

acr

Candidate generation

- Words with similar spelling
  - Small edit dist
- Words with si
  - Small edit distance of pronunc or

Damerau-Levenshtein edit distance

- Minimal edit distance between two strings, where edits are:
  - Insertion
  - Deletion
  - Substitution
  - Transposition of two adjacent

Words within 1 of `acress`

Error	Candidate Correction	Correct Letter	Error Letter	Type
acress	a			deletion
acress	c			insertion
acress	caress	ca		nsposition
acress	access	c		bstitution
acress	across	o	e	substitution
acress	acres	–	s	insertion
acress	acres	–	s	insertion

Candidate generation

- 80% of errors are within edit distance 1

- Almost all erro 2

- Also allow insertion of **space**
  - `thisidea` → `this idea`
  - `inlaw` → `in-law`

Language Model

- Use any of the language modeling algorithms we’ve learned
- Unigram, bigram, trigram
- Web-scale spellin
    - Stupid backoff

Unigram Prior probability

Counts from 404,253,213 words in Corpus of Contemporary English (COCA)

word	Fre	
actress	[illegible]	
cress	220	42
caress	686	69
access	37,038	.0000916207
across	120,844	.0002989314
acres	12,874	.0000318463

Channel model probability

- **Error model probability, Edit probability**
- *Kernighan, Church, Gale 1990*

- *Misspelled word* $x = x_1, x_2, x_3\ldots, x_m$
- *Correct word* $w = w_1, w_2, w_3, \ldots, w_n$

- $P(x|w)$ = probability of the edit
  - (deletion/insertion/substitution/transposition)

Computing error probability: confusion matrix

```
del[x,y]:     count(xy typed as x)
ins[x,y]:     count(x typed as xy)
sub[x,y]:
trans[x,y]:   count(xy ty           )
```

Insertion and deletion conditioned on previous character

Confusion matrix for spelling errors

Generating the confusion matrix

- Peter Norvig's list of errors
- Peter Norvig's list of counts of single-edit errors

Channel model

Kernighan, Church, Gale 1990

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

Channel model for acress

Candidate Correction	Correct Letter	Error Letter	x|w	P(x|word)
actress	t	-	c|ct	.000117
cress	-	a	[illegible]	[illegible]
caress	ca	ac	ac|ca	.00
access	c	r	r|c	.00
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|word)	P(word)	10^9 *P(x|w)P(w)
actress	t	-	c|ct	.000117	.0000231	2.7
cress	-	a	[illegible]	[illegible]	000000544	.00078
caress	ca	ac	ac|ca	.00	000170	.0028
access	c	r	r|c	.00	00916	.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|word)	P(word)	10^9 *P(x|w)P(w)
actress	t	-	c|ct	.000117	.0000231	2.7
cress	-	a			000000544	.00078
caress	ca	ac	ac|ca	.00	000170	.0028
access	c	r	r|c	.00	00916	.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

Using a bigram language model

- “a stellar and versatile **acress** whose combination of sass and glamour…”
- Counts from the [illegible] merican English with add-1 smoothing
- P(actress|versatile)=.00002 [illegible] 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}

Using a bigram language model

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

- Counts from the merican English with add-1 smoothing

- `P(actress|versatile)=.00002 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}

Evaluation

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

Spelling Correction and the Noisy Channel

y Channel Model of Spelling

Spelling Correction and the Noisy Channel

rd 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***
- The study was [illegible]ohn Black.

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

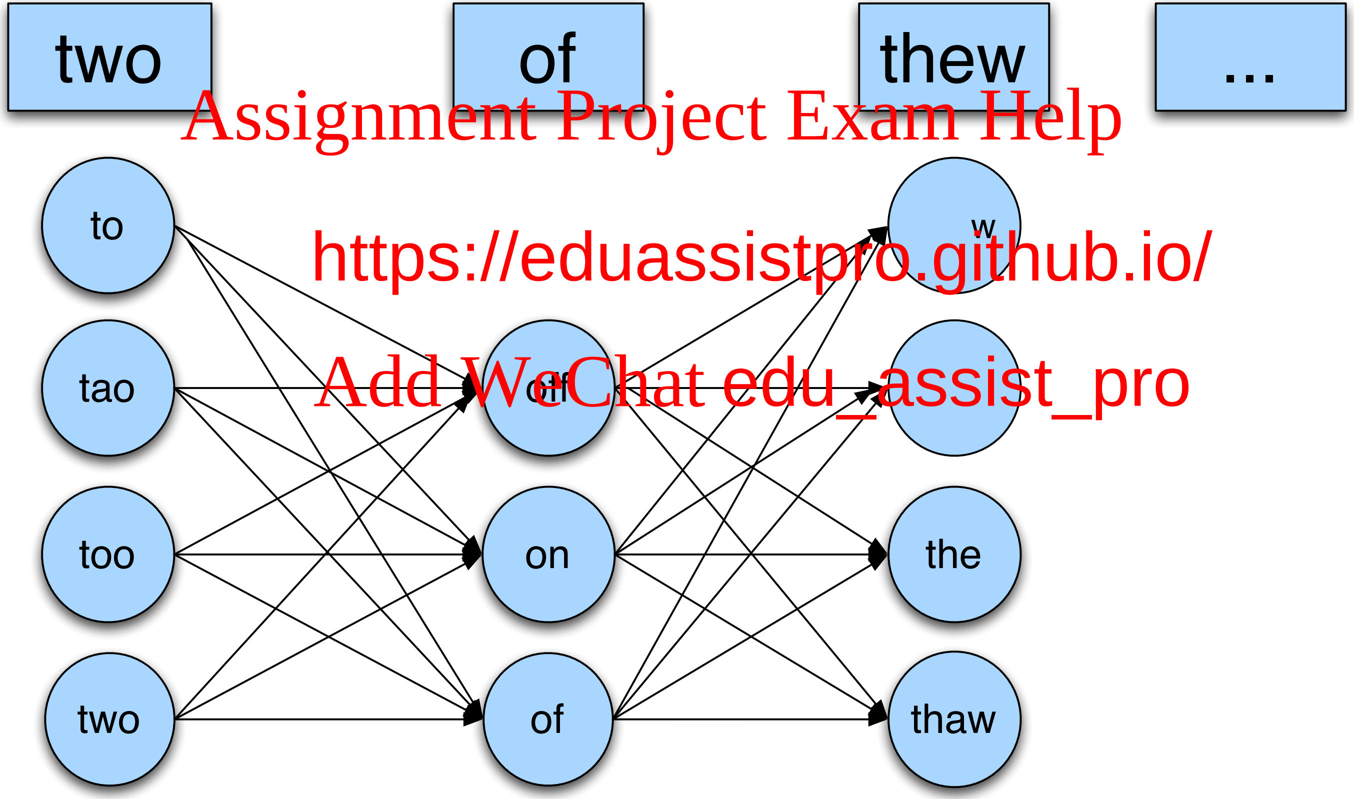
Solving real-world spelling errors

- For each word in sentence
  - Generate *cand*
    - the word its
    - all single-letter edits that ar rds
    - words that are homophone
- Choose best candidates
    - Noisy channel model
    - Task-specific classifier

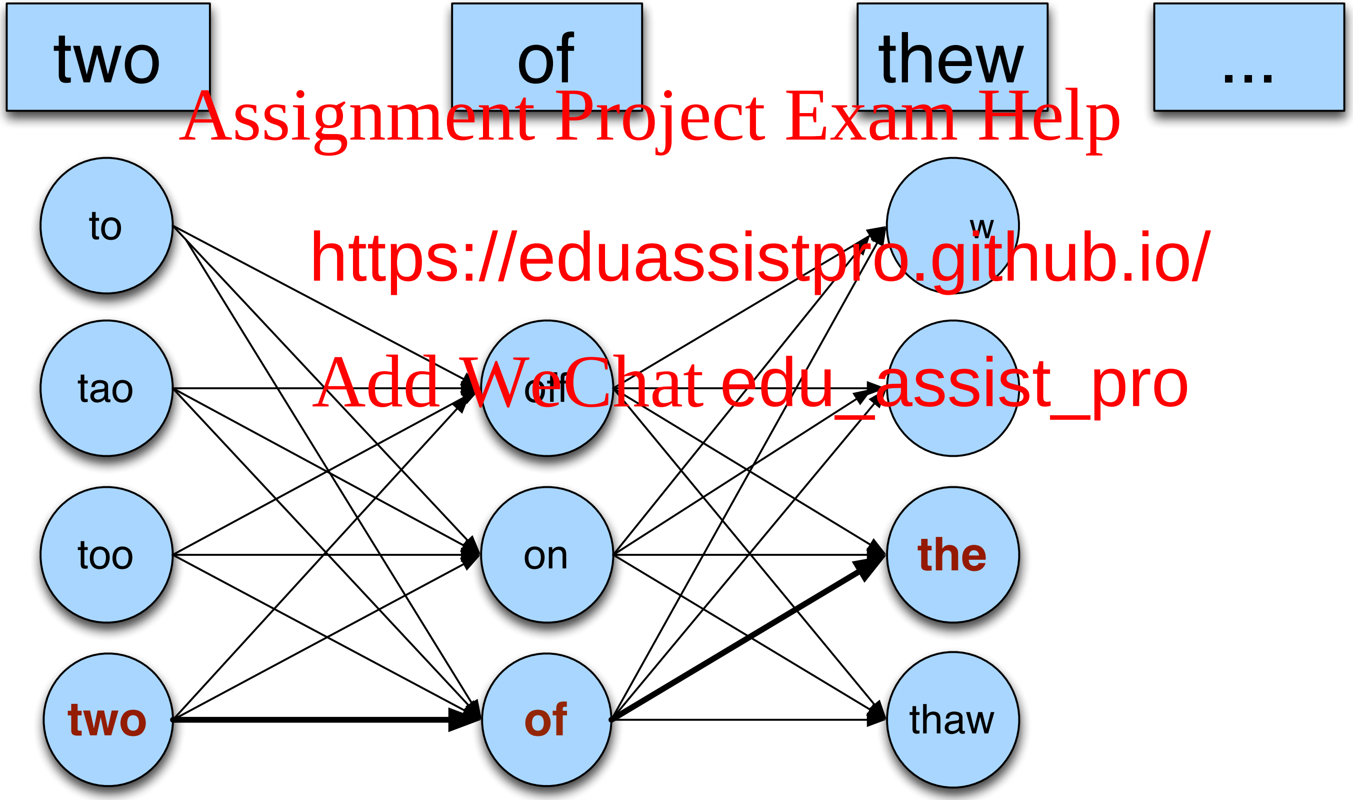
Noisy channel for real-word spell correction

- Given a sentence $w_1,w_2,w_3,…,w_n$
- Generate a set of candidates for each word w_i
  - Candidate(w_1) = {
  - Candidate(w_2) = {
  - Candidate(w_n) = {w_n, w'_n , w''_n , w'''_n
- Choose the sequence W that ma

Noisy channel for real-word spell correction



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
  - $\mathbf{w'''_1}$,w_2,w_3,w_4
  - …
- Choose the sequence W that ma

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")

- Obviously this depends on the ap
  - .90 (1 error in 10 words)
  - .95 (1 error in 20 words)
  - .99 (1 error in 100 words)
  - .995 (1 error in 200 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			144
thew	thew				90
thew	thaw	e|a	0.001		0.7
thew	threw	h|hr	0.000008	0.000004	0.03
thew	thwe	ew|we	0.000003	0.00000004	0.0001