
Advanced Natural Language Processing

Lecture 7

Spelling Correction and the EM Algorithm

Philipp Koehn

6 October 2011



A Text

The Communists disdain to conceal their views and aims. They openly declare that their ends can be attained only by the forcible overthrow of all existing social conditions. Let the ruling classes tremble at a communist revolution. The proletarians have nothing to lose but their chains. They have a world to win.

Spelling Errors

The **Comunists** disdain to conceal their views and aims. They openly declare that their ends can be **attaimed** only by the forcible overthrow of all existing social **conditionns**. Let the ruling classes **trebmle** at a communist revolution. The proletarians have nothing to lose but their chains. They have a world to win.

Types of Errors

Deletion dropping a letter

Communist → Comunist

Insertion adding a letter

conditions → conditionns

Substitution replacing a letter

attained → attaimed

Transposition swapping two letters

tremble → trebmle

Many Possible Corrections

Even if allowing only one spelling error per word:

Error	Correction	Transformation	Type
acress	actress	t→	deletion
acress	ccess	→a	insertion
acress	caress	ca→ac	transposition
acress	access	c→r	substitution
acress	across	o→e	substitution
acress	acres	→s	insertion

Finding Corrections

Task: find the minimal number corrections between `acress` and `across`

		a	c	r	e	s	s
	0						
a							
c							
r							
o							
s							
s							

Finding Corrections

Apply corrections or matches at the beginning of the word and its misspelling

		a	c	r	e	s	s
	0	←1					
a	↑1	↖0					
c							
r							
o							
s							
s							

Finding Corrections

Fill in the matrix further down

		a	c	r	e	s	s
	0	←1	←2	←3	←4	←5	←6
a	↑1	↖0	←1	←2	←3	←4	←5
c	↑2	↑1	↖0	←1	←2	←3	←4
r	↑3	↑2	↑1	↖0	←1	←2	←3
o	↑4	↑3	↑2	↑1	↖1	↖←2	↖←3
s	↑5	↑4	↑3	↑2	↖↑2	↖1	←2
s	↑6	↑5	↑4	↑3	↖↑3	↑2	↖1

Finding Corrections

Backtrack pointers along the minimal correction path

		a	c	r	e	s	s
	0	←1	←2	←3	←4	←5	←6
a	↑1	↖ 0	←1	←2	←3	←4	←5
c	↑2	↑1	↖ 0	←1	←2	←3	←4
r	↑3	↑2	↑1	↖ 0	←1	←2	←3
o	↑4	↑3	↑2	↑1	↖ 1	↖←2	↖←3
s	↑5	↑4	↑3	↑2	↖↑2	↖ 1	←2
s	↑6	↑5	↑4	↑3	↖↑3	↑2	↖ 1

Spelling Correction Model

- Given a text with misspellings t , we want to find the most likely corrections c

$$\operatorname{argmax}_c p(c|t)$$

- Noisy channel model

$$\operatorname{argmax}_c p(c|t) = \operatorname{argmax}_c p(t|c) p(c)$$

- Two components
 - language model $p(c)$
 - misspelling model $p(t|c)$

Modeling $p(t|c)$

Given the correct word c , how likely will it be misspelled as t ?

$$p(t|c) = \begin{cases} \frac{\text{del}(c_p)}{\text{count}(c_p)} & \text{if deletion} \\ \frac{\text{ins}(t_p)}{\text{count}(\epsilon)} & \text{if insertion} \\ \frac{\text{sub}(c_p, t_p)}{\text{count}(c_p)} & \text{if substitution} \\ \frac{\text{trans}(c_{p+1}, c_p)}{\text{count}(c_p, c_{p+1})} & \text{if transposition} \end{cases}$$

Supervised Approach

- Create a training corpus
 - find misspelled text
 - correct all errors
 - keep track of all corrections
- Estimate Model, for instance
 - letter **n** occurs 54,211 times in correct text
 - letter **n** is misspelled as **m** 12 times

$$\frac{\text{sub}(c_p, t_p)}{\text{count}(c_p)} = \frac{\text{sub}(\mathbf{n}, \mathbf{m})}{\text{count}(\mathbf{n})} = \frac{12}{54,211} = 0.00022$$

Unsupervised Approach

- Hand-correcting a large corpus is a lot of work
 - What if we only have a corpus with spelling errors and an English dictionary
- Can we still learn a spelling correction model?

Chicken and Egg Problem

- If we had a spelling correction **model**
 - then we could make all the **corrections** in the text.
- If we had all the **corrections**
 - then we could learn a spelling correction **model**.
- But we have neither

Incomplete Data

- We have some of the data
 - text with misspellings
 - English dictionary
 - We are missing some of the data
 - corrections for misspellings
 - We want to learn a model
- A problem for the EM algorithm

EM Algorithm

1. Initialize the model
 - for instance uniformly: all misspellings are equally likely
2. Expectation step: Apply the model to the data
 - consider all possible corrections, and score them
3. Maximization step: Estimate the model from the data
 - given the corrections labeled in the data, learn a model
 - using weighted counts based on likelihood of correction
4. Iterate: go back to step 2 until convergence

Example

- Sentence with misspellings

The acress played an important role im the theare
actress important in theatre
acres import is
across

- Collect counts for correction model
 - actress → acress: del(t) += 0.333
 - acres → acress: ins(s) += 0.333
 - across → acress: sub(o,e) += 0.333
 - important → important: sub(n,m) += 0.5
 - in → im: sub(n,m) += 0.5

Next Iteration

- Apply improved model

The acress played an important role im the theare
actress important in theatre
acres import is

- Collect counts for correction model

- actress \rightarrow **acress**: del(c,t) += 0.8
- acres \rightarrow **acress**: ins(c,t) += 0.1
- across \rightarrow **acress**: sub(o,e) += 0.8
- important \rightarrow **important**: sub(n,m) += 0.9
- in \rightarrow **im**: sub(n,m) += 0.9

Convergence

- EM algorithm discovers the right corrections

The acress played an important role im the theare
actress important in theatre
acres import is
across

- Correction model can be estimated