

Spelling & Grammar Correction

Linguistics 384
Winter 2004

Much taken from Karen Kukich (1992)
Techniques for Automatically Correcting Words in Text

Spelling and Grammar Checkers

- Why do we care?
- What causes errors?
- What makes spelling correction difficult?
- What is the task (exactly)?
- How is spelling correction done?
- The dangers of spelling and grammar correction

2

Why do we care?

Aoccdrnig to a rscheearch at Cmabrigde Uinervtisy, it deosn't mttar in waht oredr the ltteers in a wrod are, the olny iprmoetnt tihng is taht the frist and lsat ltteer be at the rghit pclae. The rset can be a toatl mses and you can sitll raed it wouthit porbelm. Tihs is bcuseae the huamn mnid deos not raed ervey lteter by istlef, but the wrod as a wlohe.

⇒ Spelling shouldn't even matter to us, right?

3

Why we care

- People want to appear to be educated.

http://www.canoe.ca/CNEWSWeirdNews0202/13_three-ap.html

February 13, 2002

Spelling errors dominate education fight

MADRID, Spain (AP) -- The government couldn't help but chuckle last week when students opposed to reforms aimed at raising education standards released a flier calling for demonstrations with a glaring spelling mistake.

Now it's the government's turn to blush.

4

A letter written in Catalan, signed by Environment Minister Jaume Matas and sent to tens of thousands of homes in northeastern Spain, contained 13 spelling errors and two geographical errors.

The letter defends a controversial hydrological project in which water is to be diverted from the Ebro River, which flows through the Aragon and Catalonia regions, to the Mediterranean coast.

Education and Culture Minister Pilar del Castillo, the architect of the education overhaul, reacted to the flier by saying, "Students who call demonstrations are the ones who get the worst grades." She has said repeatedly in recent days that Spain's schools are churning out uneducated young people.

Responding to minister Matas' multiple slip-ups, a Socialist Party leader in Catalonia, Jaume Antich, said: "Are you trying to prove Pilar del Castillo right when she talks about low cultural levels? In view of this letter, I don't know if you'd pass the exam she wants to reinstate."

5

Why we care (cont.)

- Misspellings can cause problems or misunderstandings.
e.g. Jane bought a god yesterday; it's a big golden retriever.
e.g. This will be a fee [free] concert.
- Typos can be dangerous: 1991 Bell Atlantic & Pacific Bell telephone network outages partly caused by typographical errors.
A *b* in a line of computer code was supposed to be a *D*. "That one error caused the equipment and software to fail under an avalanche of computer-generated messages." (Wall Street Journal, Nov. 25, 1991)
- If you're reading in a hand-written document from a scanner and want to convert it into a text file, poorly written, unrecognizable characters can be corrected with spell checking technology

Note: the spelling claim supposedly from Cambridge University isn't quite true. See: <http://www.mrc-cbu.cam.ac.uk/personal/matt.davis/Cmabrigde/>

6

A note on language learners

People learning a language (e.g. English) often make spelling and grammar errors as they learn the language

- Can use their errors to help design spell checking systems
- Can use correction systems to improve their English

7

What causes errors?

- Keyboard mistypings
- Phonetic errors
- Knowledge problems

8

Keyboard mistypings

- Space bar issues
 - **run-on** errors = two separate words become one “word”
e.g. *the fuzz* becomes *thefuzz*
 - **split** errors = one word becomes two separate words
e.g. *equalization* becomes *equali zation*
- ⇒ Resulting items might still be words
e.g. *a tollway* becomes *atoll way*
- Keyboard proximity:
e.g. *Jack* becomes *Hack* since *h* and *j* are next to each other on the keyboard
- Physical similarity, if, e.g., a person is typing something they handwrote.
e.g. *tight* for *fight*

9

Phonetic errors

phonetic errors = errors based on the sounds of a language (not necessarily on the letters)

⇒ Letters and sounds do not always match up – i.e. writing and speaking are two different things

e.g. *c* can sound like a *k* or an *s*

Phonetic errors tend to be more distorted than typographical errors.

10

Kinds of phonetic errors

- **homophones** = two words which sound the same
e.g. *red/read, cite/site/sight, they're/their/there*
- spelling like it sounds
e.g. *sikologee*
 - letter substitution: replacing a letter (or sequence of letters) with a similar-sounding one
e.g. *John kracked his nuckles*
 - word replacement: replacing one word with some similar-sounding word
e.g. *John battled me on the back*
- **Spoonerisms** = switching two letters/sounds around
e.g. *It's a tavy grain with biscuit wheels.*

11

Knowledge problems

- not knowing a word and guessing its spelling (can be phonetic)
e.g. *sientist*
- not knowing a rule and guessing it
e.g. Do we double a consonant for *ing* words? *jog* → *joging*

12

What kinds of errors are there?

- **insertion** = a letter is added to a word
- **deletion** = a letter is deleted from a word
- **substitution** = a letter is put in place of another one
- **transposition** = two adjacent letters are switched

Note that the first two alter the length of the word, whereas the second two maintain the same length

- **single-error misspellings** = only one instance of an error
- **multi-error misspellings** = multiple instances of errors (harder to identify)

13

What makes spelling correction difficult?

- **Tokenization**: what is a word?
- **Inflection**: how are some words related?
- **Productivity** of language: how many words are there?

14

Tokenization

At first it may seem intuitive that a “word” is simply something between two spaces, but this is not always so clear.

- **contractions** = two words combined into one
e.g. *can't*, *Chris's* (vs. *Chris'*)
- **multi-token words** = (arguably) a single word with a space in it
e.g. *New York*, *in spite of*, *deja vu*
- **hyphens** (Note: often on a single line, but ambiguous if a hyphen ends a line.)
 - Some are (only) a single word: *e-mail*, *co-operate*
 - Others are two words combined into one: *Columbus-based*, *sound-change*
- **Abbreviations**: may stand for multiple words
e.g. *etc.* = *et cetera*, *ATM* = *Automated Teller Machine*

15

Inflection

A word in English may appear in various guises due to word **inflections** = word endings which are fairly systematic for a given part of speech

- plural noun ending: *the boy* + *s* → *the boys*
- past tense verb ending: *walk* + *ed* → *walked*

This can make spell-checking hard:

- There are exceptions to the rules: *mans*, *runned*
- There are words which look like they have a given ending, but they don't: *Hans*, *deed*

16

Productivity

- part of speech change: nouns can be verbified
e.g. *basketing* is not a word, but *chairing* is (and *basketing* could easily become a word). *emailed* is now a word.■
- morphological productivity: prefixes and suffixes can be added
e.g. I can speak of a *post-chair-tari-an*, one who insists on sitting behind a chair.■
- words entering and exiting the lexicon: *spleet* 'split' (*Hamlet III.2.10*) leaves, while *d'oh* comes in

17

What do we expect from spell checkers?

Two ways to run spell-checkers:

- **interactive spelling checkers** = spell checker detects errors as you type.■
 - It may or may not make suggestions for correction.■
 - Requires a “real-time” response (i.e. must be fast)■
 - It is up to the human to decide if the spell checker is right or wrong.■
 - If there are a list of choices, we may not require 100% accuracy in the corrected word■
- **automatic spelling correctors** = spell checker runs on a whole document, finds errors, and corrects them■
 - A much more difficult task.■
 - A human may or may not proofread the results later.

18

Detection vs. Correction

There are two very different tasks:

- **error detection** = simply find the misspelled words
- **error correction** = correct the misspelled words■

e.g. It might be easy to tell that *ater* is a misspelled word, but what is the correct word? *water?* *later?* *after?*■

⇒ Depends on what we want to do with our results as to what we want to do.
Note, though, that detection is a prerequisite for correction.

19

What is the task?

Three main ways of viewing the spell checking task:

- Nonword error detection
- Isolated-word error correction
- Context-dependent word correction

20

Nonword error detection

nonword error detection is essentially the same thing as **word recognition** = splitting up “words” into true words and nonwords. ■

How is nonword error detection done?

- n-gram analysis
- dictionary (lookup and construction)

21

N-gram analysis

An **n-gram** is a string of n letters.

a 1-gram (unigram)
at 2-gram (bigram)
ate 3-gram (trigram)
late 4-gram

We can use this n-gram information to define what are possible strings in a language.

e.g. *po* is a possible string in English, while *kvt* is not.

22

How do we store n-gram information?

We could have a list of all possible (or impossible) n-grams like so (1 = possible, 0 = impossible)

<i>po</i>	1
<i>kvt</i>	0
<i>police</i>	1
<i>asdf</i>	0

Any word which has a 0 for any substring is a misspelled word ■

- Requires a lot of computer storage space ■
- Inefficient (slow) when looking up every string ■
- Information is repeated (*po* is in *police*)

23

Bigram array

Instead, we can define a bigram **array** = information stored in a rectangular, organized fashion (columns and rows)

...	k	l	m	...
k	0	1 (<i>tackle</i>)	1 (<i>Hackman</i>)	...
l	1 (<i>elk</i>)	1 (<i>hello</i>)	1 (<i>alms</i>)	...
m	0	0	1 (<i>hammer</i>)	...

This is a **nonpositional bigram array** = the array 1's and 0's apply for a string found anywhere within a word (beginning, 4th character, ending, etc.).

24

Positional bigram array

More useful might be a **positional bigram array** = the array only applies for a given position in a word.█

Here's the same array as before, but now only applied to word endings:

...	k	l	m	...
k	0	0	0	...
l	1 (<i>elk</i>)	1 (<i>hall</i>)	1 (<i>elm</i>)	...
m	0	0	0	...

25

Dictionaries

Intuition: have a complete list of words and check the input words against this list. If it's not in the dictionary, it's not a word.█

- **Dictionary lookup** = lookup a potential word in the dictionary (how do you do this quickly?)
- **Dictionary construction** = build the dictionary (what do you put in it?)

26

Dictionary lookup

- Inflections: must strip the word of prefixes and suffixes before looking it up█
- Have to make lookup fast by using efficient lookup techniques, such as a hash table (which we discussed with web searching)

27

Dictionary construction

- Want the dictionary to have only the word relevant for the user → **domain-specificity**
e.g. For most people *memoize* is a misspelled word, but in computer science this is a technical term and spelled correctly.█
- foreign words, hyphenations, derived words, proper nouns, and new words will always be problems for dictionaries since we cannot predict these words until humans have made them words█
- Dictionary should probably also be dialectally consistent, e.g. include only *color* or *colour* but not both

28

Isolated-word error correction

- Have looked at error detection (n-grams and dictionaries)
- Now, we want to know how to correct these misspelled words: **Isolated-word error correction** = correcting words without taking context into account.
⇒ Only handles errors that result in nonwords

Knowledge about how errors are made helps

29

Some general techniques

We will look at a few techniques that can be added to a spell checking system.

These issues affect how one builds a dictionary, how one writes rules, and so on.

- word length effects: most misspellings are within two characters in length of original
→ when finding the correct spelling, we do not usually need to look at words with greater length differences.
- first-position error effects: the first letter of a word is rarely erroneous
→ when searching for the correct spelling, the process is sped up by being able to look only at words with the same first letter.

30

Some general techniques (cont.)

- keyboard effects: many misspellings are the result of typing a nearby letter
→ we can build a **confusion matrix** = a table that indicates how often one letter is mistyped for another

...	k	l	m	...
k	n/a	98	14	...
l	104	n/a	0	...
m	4	12	n/a	...

31

Isolated-word error correction methods

There are many different ways to correct; we will briefly look at three methods:

- rule-based methods
- minimum edit distance
- probabilistic methods

The basic steps:

1. Detection of an error
2. Generation of candidate corrections
3. Ranking of candidate corrections

We will talk mostly about generation of candidates.

32

Rule-based methods

Can generate possible correct spellings by writing rules.

- Common misspelling rewritten as correct word: *hte* → *the*
- Rules based on inflections: $V+CC+ing \rightarrow V+CC+ing$ (where V = vowel and C = consonant)
- Rules based on other common spelling errors (such as keyboard effects or common transpositions): $CsC \rightarrow CaC$, $cie \rightarrow cei$

33

Minimum edit distance

In order to rank possible correct spellings, it sometimes helps to calculate the **minimum edit distance** = shortest (minimum) number of operations it would take to convert one "word" into another.

1. *junk* → *juk* (deletion)
2. *juk* → *huk* (substitution)
3. *huk* → *hku* (transposition)
4. *hku* → *hiku* (insertion)
5. *hiku* → *haiku* (insertion)

34

Probabilistic techniques

Two main probabilities are taken into account:

- **transition probabilities** = probability (chance) of going from one letter to the next.
e.g. What is the chance that *a* will follow *p* in English? That *u* will follow *q*?
- **confusion probabilities** = probability of one letter being mistaken (substituted) for another (can be derived from a confusion matrix)
e.g. What is the chance that *q* will be confused with *p*?

Useful to combine probabilistic techniques with dictionary methods

35

Context-dependent word correction

Context-dependent word correction = correcting words based on the surrounding context.

- This will handle errors which are actually words
- Essentially a fancier name for a **grammar checker** = a mechanism which tells a user if their grammar is wrong

36

What do grammar correctors correct?

- Syntactic errors = errors in how words are put together in a sentence: different parts of speech (noun, verb, etc.) are in the wrong places, or something isn't "grammatical."■
 - **Local** syntactic errors: 1-2 words away
e.g. *The study was conducted mainly be John Black.*
⇒ a verb is where a preposition should be.■
 - **Long-distance** syntactic errors: (roughly) 3 or more words away
e.g. *The kids who are most upset by the little totem is going home early.*
⇒ *A kid is* vs. *Kids are*■
- Semantic errors = errors where the sentence structure sounds okay, but it doesn't really mean anything.
e.g. *They are leaving in about fifteen minuets to go to her house.*
⇒ *minuets* and *minutes* are both plural nouns, but only one makes sense here

37

How do grammar correctors correct?

Again, there are many different ways, and we will look at two:

- Rule-based
- Bigram model (bigrams of words)

38

Rule-based grammar correctors

Example: *These report need to be checked.*■

- We know that *report* should be plural (*reports*), but what tells us that?■
- Answer: *These* + the fact that *report* is a noun.■
- Rule: *these/those* + NOUN + PLURAL-VERB
→ *these/those* + NOUNs + PLURAL-VERB

39

Bigram grammar correctors

We could also look at **bigrams**: now we are talking about bigrams of words, i.e. two words that appear next to each other.■

- **Question:** Given the previous word, what is the probability of the current word?■
- e.g. given *these*, we have a 5% chance of seeing *reports* and a 0.001% chance of seeing *report* (*these report cards*).■
- Thus, we will change *report* to *reports*

But there's one major problem: we may hardly ever see *these reports* and so won't know the probability of that bigram.■

(Partial) Solution: use bigrams of **parts of speech**. e.g. What is the probability of a noun given that the previous word was an adjective?

40

Is this really how spell checkers work?

Yes and no.

- Some spell checkers are secret; we don't know how they work exactly
- Others, such as ispell and aspell, are **open source** spell checkers, meaning that anyone can see how they work. And they do use similar methods

Advantage of open source:

- Users can add to the dictionaries, or remove words.
- Users can understand what kinds of errors will be caught and what kinds won't, and so know what they need to look for on their own.

41

The dangers of spelling and grammar correction

The more we depend on spelling correctors, the less we try to correct things on our own. But spell checkers are not 100%

A study at the University of Pittsburgh found that students made **more** errors when using a spell checker!

	high SAT scores	low SAT scores
no spell checker	5 errors	12.3 errors
use spell checker	16 errors	17 errors

<http://www.wired.com/news/business/0,1367,58058,00.html>

42

A Poem on the Dangers of Spell Checkers

Michael Livingston (<http://www.courses.rochester.edu/livingston/guide/phonix.html>)

Eye halve a spelling chequer
It came with my pea sea.
It plainly marques four my revue
Miss steaks eye kin knot sea.
Eye strike a key and type a word
And weight four it two say
Weather eye am wrong oar write
It shows me strait a weigh.
As soon as a mist ache is maid
It nose bee fore two long
And eye can put the error rite
Its rare lea ever wrong.
Eye have run this poem threw it
I am shore your pleased two no
Its letter perfect awl the weigh
My chequer tolled me sew.

43