Why people care about spelling

- Misspellings can cause misunderstandings
- Standard spelling makes it easy to organize words & text:
  - e.g., Without standard spelling, how would you look up things in a lexicon or thesaurus?
  - e.g., Optical character recognition software (OCR) can use knowledge about standard spelling to recognize scanned words even for hardly legible input
- Standard spelling makes it possible to provide a single text, accessible to a wide range of readers (different backgrounds, speaking different dialects, etc.)
- Using standard spelling can make a good impression in social interaction

Outline

Tasks are typically divided into:

- **Error detection** = simply find the misspelled words
- **Error correction** = correct the misspelled words
  - e.g., *ater* is a misspelled word, but what is the correct word? *water? later? after?*

We will consider three types of techniques:

- Non-word error detection
- Isolated-word error detection & correction
- Grammar correction (Context-dependent word error detection & correction)

Use of writers’ aids

How are spell checkers (and grammar checkers) used?

- **Interactive spelling checker**: spell checker detects errors as you type
  - It may or may not make suggestions for correction
  - It needs 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, and so we may not require 100% accuracy (especially with a list of choices)
- **Automatic spelling corrector**: spell checker runs on a whole document, finds errors, and corrects them
  - A more difficult task
  - A human may or may not proofread the results later

Spelling & grammar correction

We are all familiar with spelling & grammar correctors

- They are used to **improve document quality**
- They are not typically used to **provide feedback**

Typically designed for native speakers of a language

- Next unit (Language Tutoring Systems): feedback for non-native speakers

Non-word error detection

- **Word recognition**: split up “words” into true words and non-words
  - **Non-word error detection**: detect the non-words
- How is non-word error detection done?
  - Using a dictionary (construction and lookup)
  - *n*-gram analysis (more for OCR error detection)
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.

Two aspects:
- **Dictionary construction**: build the dictionary (what do you put in it?)
- **Dictionary lookup**: look up a potential word in the dictionary (how do you do this quickly?)

Challenges for spelling correction

**Tokenization**

Tokenization splits a sentence into its component words

Intuitively, a "word" is simply whatever is between two spaces, but this is not always so clear.
- **Contractions**: two words combined into one
  - e.g., can't, he's, John's [car] (vs. his car)
- **Multi-word expressions**: single term with space(s)
  - e.g., New York, in spite of, déjà vu
- **Hyphens** (ambiguous if a hyphen ends a line)
  - Some are always 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

**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

Challenges for spell checking:
- Exceptions to the rules: *mans, *runned
- Words which look like they have a given ending, but they don't: Hans, deed

N-gram analysis

**Idea**: use typical phonotactic patterns to identify words

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

<table>
<thead>
<tr>
<th>n</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1-gram (unigram)</td>
</tr>
<tr>
<td>2</td>
<td>2-gram (bigram)</td>
</tr>
<tr>
<td>3</td>
<td>3-gram (trigram)</td>
</tr>
<tr>
<td>4</td>
<td>4-gram</td>
</tr>
</tbody>
</table>

We can use this n-gram information to define what the possible strings in a language are.
- e.g., po is a possible English string, whereas kvl is not.

This is more useful to correct optical character recognition (OCR) output, but we'll still take a look.
What leads to errors? What properties do errors have?

Isolated-word error correction

- Having discussed how errors can be detected, we want to know how to correct these misspelled words:
  - **Isolated-word error correction**: correcting words without taking context into account
  - This technique can only handle errors resulting in non-words
  - Knowledge about what is a typical error helps in finding correct word
  - What leads to errors? What properties do errors have?

Types of errors

Phonetic errors

- Errors stemming from imperfect sound-letter correspondences
  - **Homophones**: two words which sound the same
    - e.g., red/read (past tense), cite/site/sight, they’re/their/there
  - Substitutions: replacing a letter (or sequence) with similar-sounding one
    - e.g., separate (for separate)

Positional bigram array

- To store information specific to the beginning, the end, or some other position in a word, use a **positionally 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:

<table>
<thead>
<tr>
<th>...</th>
<th>k</th>
<th>l</th>
<th>m</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>k</td>
<td>0</td>
<td>1</td>
<td>(elk)</td>
<td>1 (hall)</td>
</tr>
<tr>
<td>l</td>
<td>0</td>
<td>1</td>
<td>(hamlet)</td>
<td>1 (hammer)</td>
</tr>
<tr>
<td>m</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Types of errors

- **Run-on errors**: two separate words become one
  - e.g., the fuzz becomes thefuzz
- **Split errors**: one word becomes two separate items
  - e.g., equalization becomes equalization
  - The resulting items might still be words: e.g., a tollway becomes atoll way

Keyboard proximity

- e.g., program might become program since a and s are next to each other on a QWERTY keyboard

Space bar issues

- **Caveat emptor**
- **Grammar correction rules**
- **Syntax and Computing**
- **Probabilistic methods**
- **Minimum edit distance**
- **Similarity key techniques**
- **Rule-based methods**
- **N-gram analysis**
- **Dictionaries**
- **Introduction**
- **Non-word error detection**
- **Isolated-word error correction**
Describing typical errors

Errors can be examined under a more mechanistic lens:

Types of operations

- **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.

Typical error properties

- **Word length effects**: most misspellings are within two characters in length of original
  - When searching for 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

Isolated-word error correction methods

- Many different methods are used; we will briefly look at four methods:
  - Rule-based methods
  - Similarity key techniques
  - Probabilistic methods
  - Minimum edit distance

- The methods play a role in one of the three basic steps:
  1. Detection of an error (discussed above)
  2. Generation of candidate corrections
    - rule-based methods
    - similarity key techniques
  3. Ranking of candidate corrections
    - probabilistic methods
    - minimum edit distance (also usable for generation)

Rule-based methods

One can generate correct spellings by writing rules:

- **Common misspelling rewritten as correct word**: e.g., *hte* → *the*

  - **Rules**
    - based on inflections:
      - e.g., VClng → VCIng, where
        - V = letter representing vowel, basically the regular expression [aeiou]
        - C = letter representing consonant, basically [bcdfghjklmnpqrstvwxyz]
    - based on other common spelling errors (such as keyboard effects or common transpositions):
      - e.g., CsC → CaC
      - e.g., cie → cei

Similarity key techniques (SOUNDEX)

- Problem: How can we find a list of possible corrections?
  - Solution: Store words in different boxes in a way that puts the similar words together.

- Example:
  1. Start by storing words by their first letter (first letter effect),
     - e.g., punc starts with the code P.
  2. Then assign numbers to each letter
     - e.g., 0 for vowels, 1 for b, p, f, v (all bilabials), and so forth, e.g., punc → P052
  3. Then throw out all zeros and repeated letters,
     - e.g., P052 → P52.
  4. Look for real words within the same box, e.g., punk is also in the P52 box.

http://en.wikipedia.org/wiki/Soundex

Minimum edit distance

- In order to rank possible spelling corrections, it can be useful to calculate the **minimum edit distance** = minimum number of operations it would take to convert one word into another.

- For example, we can take the following five steps to convert *junk* to *haiku*:
  1. junk → junk (deletion)
  2. juk → huk (substitution)
  3. huk → hiku (transposition)
  4. hiku → haku (insertion)
  5. hiku → haiku (insertion)

- But is this the minimal number of steps needed?
Computing edit distances

To be able to compute the edit distance of two words at all, we need to ensure there is a finite number of steps.

This can be accomplished by:
- requiring that letters cannot be changed back and forth a potentially infinite number of times, i.e., we
- limit the number of changes to the size of the material we are presented with, the two words.

Idea: Never deal with a character in either word more than once.

Result:
- We could delete each character in the first word and then insert each character of the second word.
- Thus, we will never have a distance greater than length(word1) + length(word2)

To calculate minimum edit distance, we set up a directed, acyclic graph, a set of nodes (circles) and arcs (arrows).

Horizontal arcs correspond to deletions, vertical arcs correspond to insertions, and diagonal arcs correspond to substitutions (a letter can be “substituted” for itself).

The simple but dumb way of doing it:
- Follow every path from start (A) to finish (T) and see how many changes we have to make.
- But this is very inefficient! There are many different paths to check.
Computing edit distances
The smart way to compute the least cost

- The smart way to compute the least cost uses **dynamic programming**: process designed to make use of results computed earlier
  - We follow the topological ordering & calculate the least cost for each node:
    - We add the cost of an arc to the cost of reaching the node this arc originates from.
    - We take the minimum of the costs calculated for all arcs pointing to a node and store it for that node.
  - The key point is that we are storing partial results along the way, instead of recalculating everything, every time we compute a new path.

---

Probabilistic methods

When converting from one word to another, a lot of words will be the same distance.

- e.g., for the misspelling *wil*, all of the following are one edit distance away:
  - *will*
  - *wild*
  - *wilt*
  - *nil*

Probabilities will help to tell them apart

---

The Noisy Channel Model

Probabilities can be modeled with the **noisy channel model**

- Hypothesized Language: *X*
  - Noisy Channel: *X → Y*
  - Actual Language: *Y*

Goal: Recover *X* from *Y*

- The noisy channel model has been very popular in speech recognition, among other fields

(Thanks to Mike White for the slides on the Noisy Channel Model)

---

Noisy Channel Spelling Correction

Goal: Recover correct spelling *X* from misspelling *Y*

- Noisy word: *Y* = observation (incorrect spelling)
- We want to find the word (*X*) which maximizes: \( P(X|Y) \), *i.e.*, the probability of *X*, given that *Y* has been seen

---

Example

Goal: Recover correct spelling *swam* from misspelling *sawm* (i.e., \( P(swam|sawm) \))

---

Conditional probability
(Reminder)

- \( p(x|y) \) is the probability of *x* given *y*
  - Let’s say that it rains appears 20 times in a span of 40 days
    - \( p(\text{rain}) = 20/40 = 0.5 \)
  - Now, let’s say I bring an umbrella to work on 18 of the 40 days, and it rains on 2 of those days
    - \( p(\text{rain} | \text{umbrella}) = 2/18 = 0.1111 \)

Note: there is no causation implied here; we are simply counting things
Bayes Rule

With \( X \) as the correct word and \( Y \) as the misspelling ...

\[ P(Y|X) \] is impossible to calculate directly, so we use:

\[ P(Y|X) = \text{the probability of the observed misspelling given the correct word} \]

\[ P(X) = \text{the probability of the (correct) word occurring anywhere in the text} \]

Bayes Rule allows us to calculate \( P(X|Y) \) in terms of \( P(Y|X) \):

1. **Bayes Rule**: \( P(X|Y) = \frac{P(Y|X)P(X)}{P(Y)} \)

The Noisy Channel and Bayes Rule

We can directly relate Bayes Rule to the Noisy Channel:

\[
\frac{\text{Posterior}}{\text{Prior}} = \frac{P(Y|X)}{P(Y) \times \text{Normalization}}
\]

Goal: for a given \( y \), find \( x \)

\[
\text{arg max}_x \quad P(y|x) \times P(x)
\]

The denominator is ignored because it’s the same for all possible corrections, i.e., the observed word \( (y) \) doesn’t change

Finding the Correct Spelling

Goal: for a given misspelling \( y \), find correct spelling \( x \)

\[
\text{Error Model} \quad \text{Language Model}
\]

arg max \( _x \quad Pr(y|x) \times Pr(x) \)

1. List “all” possible candidate corrections, i.e., all words with one insertion, deletion, substitution, or transposition
2. Rank them by their probabilities

Example: calculate for swam

\[
Pr(\text{sawm}|\text{swam})Pr(\text{swam})
\]

and see if this value is higher than for any other possible correction.

Obtaining probabilities

Confusion probabilities

- It is impossible to fully investigate all possible error causes and how they interact, but we can learn from watching how often people make errors and why.
- One way is to build a **confusion matrix** = a table indicating how often one letter is mistyped for another correct

\[
\begin{array}{c|ccc}
\text{...} & r & s & t & \text{...} \\
\hline
r & \text{n/a} & 12 & 22 & \\
s & 14 & \text{n/a} & 15 & \\
t & 11 & 37 & \text{n/a} & \\
\vdots
\end{array}
\]

(cf. Kernighan et al 1999)

Obtaining probabilities

Using a spelling error-annotated corpus:

- These matrices are calculated by counting how often, e.g., \( ab \) was typed instead of \( a \) in the case of insertion

To get \( P(Y|X) \), then, we find the probability of this kind of typo in this context. For insertion, for example \( X_n \) is the \( p^{th} \) character of \( X \):

\[
(2) \quad P(Y|X) = \frac{\text{inst}(X_{n-1}, X_n)}{\text{count}(X_{n-1})}
\]
Some resources ...

Want to try these some of these things for yourself?
- How to Write a Spelling Corrector by Peter Norvig: http://norvig.com/spell-correct.html
  - 21 lines of Python code (other programming languages also available)
- Birkbeck spelling error corpus: http://www.ota.ox.ac.uk/headers/0643.xml

Spelling correction for web queries

A nice little side topic ...

Spelling correction for web queries is hard because it must handle:
- Proper names, new terms, etc. (blog, shrek, nsync)
- Frequent and severe spelling errors
- Very short contexts

Algorithm

Main Idea (Cuerozan and Brill (EMNLP-04))
- Iteratively transform the query into more likely queries
- Use query logs to determine likelihood
  - Despite the fact that many of these are misspelled!
  - Assumptions: the less wrong a misspelling is, the more frequent it is; and correct > incorrect

Example:
- anol scwartegger
  - arnold schwartnegger
  - arnold schwarznegger
  - arnold schwarzenegger

Examples

Context Sensitivity
- power crd → power cord
- video crd → video card
- platnuin rings → platinum rings

Known Words
- golf war → gulf war
- sap opera → soap opera

Spelling correction for web queries

Algorithm (2)

- Compute the set of all close alternatives for each word in the query
  - Look at word unigrams and bigrams from the logs; this handles concatenation and splitting of words
  - Use weighted edit distance to determine closeness
- Search sequence of alternatives for best alternative string, using a noisy channel model

Constraint:
- No two adjacent in-vocabulary words can change simultaneously

Examples (2)

Tokenization
- chat inspanich → chat in spanish
- ditroitigers → detroit tigers
- brittenetspear inconcert → britney spears in concert

Constraints
- log wood → log wood (not dog food)
**Context-dependent word correction**

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

- This will handle errors which are real words, just not the right one or not in the right form.
- This is very similar to a **grammar checker** = a mechanism which tells a user if their grammar is wrong.

**Grammar correction—what does it correct?**

- Syntactic errors = errors in how words are put together in a sentence: the order or form of words is incorrect, i.e., ungrammatical.
  - **Local** syntactic errors: 1-2 words away
    - e.g., *The study was conducted mainly by 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.*
    - Agreement error between subject kids and verb is.

**More on grammar correction**

- Semantic errors = errors where the sentence structure sounds okay, but it doesn’t really mean anything.
  - e.g., *They are leaving in about fifteen minutes to go to her house.*
  - ⇒ *minutes and minutes* are both plural nouns, but only one makes sense here

There are many different ways in which grammar correctors work, two of which we’ll focus on:

- **N-gram model**
- **Rule-based model**

**N-gram grammar correctors**

Remember that bigrams & trigrams model the probability of sequences

- **Question n-grams address**: Given the previous word (or two words), what is the probability of the current word?
  - Use of n-grams: compare different candidates:
    - e.g., given *these*, we have a lower chance of seeing *report* than of seeing *reports*
    - Since a confusable word (reports) can be put in the same context, resulting in a higher probability, we flag *report* as a potential error

But there’s a major problem: we may hardly ever see these reports, so we won’t know its probability.

- **Some possible solutions**:
  - use bigrams/trigrams of parts of speech
  - use massive amounts of data and only flag errors when you have enough data to back it up

**Rule-based grammar correctors**

We can target specific error patterns. For example:

- **To a certain extend**, we have achieved our goal.
  1. Match the pattern some or certain followed by *extend*, which can be done using the regular expression `some|certain extend`
  2. We’ll discuss regular expressions with searching; for now, think of them as short ways to write patterns or templates
  2. Change the occurrence of *extend* in the pattern to *extent*


**Beyond regular expressions**

- But what about correcting the following:
  - A baseball teams were successful.
  - We should see that A is incorrect, but a simple pattern doesn’t work because we don’t know where the word teams might show up.
    - A wildly overpaid, horrendous baseball teams were successful. (Five words later; change needed.)
    - A player on both my teams was successful. (Five words later; no change needed.)

We need to look at how the sentence is constructed in order to build a better rule.
**Syntax**

- **Syntax** = the study of the way that sentences are constructed from smaller units.
- There cannot be a “dictionary” for sentences since there is an infinite number of possible sentences:
  1. Linear order
  2. Hierarchical structure (Constituency)

There are two basic principles of sentence organization:

- Linear order
- Hierarchical structure (Constituency)

**Constituency**

- What are the “meaningful units” of a sentence like *Most of the ducks play extremely fun games*?
  1. Most of the ducks
  2. of the ducks
  3. extremely fun
  4. play extremely fun games

- We refer to these meaningful groupings as **constituents** of a sentence.

**Linear order**

- **Linear order** = the order of words in a sentence.
  - A sentence can have different meanings, based on its linear order:
    1. John loves Mary.
  - Languages vary as to what extent this is true, but linear order in general is used as a guiding principle for organizing words into meaningful sentences.
  - Simple linear order as such is not sufficient to determine sentence organization, though.
  - e.g., we can’t simply say “The verb is the second word in the sentence.”
  1. I eat at really fancy restaurants.
  2. Many executives eat at really fancy restaurants.

**Hierarchical structure**

- Constituents can appear within other constituents
- Constituents shown through brackets:
  - [[Most [of the ducks]]] [play [[extremely fun] games]]

  Constituents displayed as a **syntactic tree**:

  ```
  a
  \b\ c \ e
  d g
  Most play of the ducks extremely fun games
  ```

**Lexical categories**

- **Lexical categories** are simply word classes, or what you may have heard as *parts of speech*. The main ones are:
  1. verbs: eat, drink, sleep, ...
  2. nouns: gas, food, lodging, ...
  3. adjectives: quick, happy, brown, ...
  4. adverbs: quickly, happily, well, westward
  5. prepositions: on, in, at, to, into, of, ...
  6. determiners/articles: a, an, the, this, these, some, much, ...

- We would also like some way to say that
  1. *the ducks*, and
  2. *extremely fun games*

- are the same type of grouping, or constituent, whereas
  1. *of the ducks*

- seems to be something else.

- For this, we will talk about different **categories**
  1. Lexical
  2. Phrasal
Determining lexical categories

How do we determine which category a word belongs to?

- **Distribution**: Where can these kinds of words appear in a sentence?
  - e.g., Nouns like *mouse* can appear after articles ("determiners") like *some*, while a verb like *eat* cannot.
- **Morphology**: What kinds of word prefixes/suffixes can a word take?
  - e.g., Verbs like *walk* can take an *ed* ending to mark them as past tense. A noun like *mouse* cannot.

(We'll discuss this more with Language Tutoring Systems)

Phrase Structure Rules

- We can give rules for building these phrases.
  - We want a way to say that a determiner and a noun make up a noun phrase, but a verb and an adverb do not.
- **Phrase structure rules** are a way to build larger constituents from smaller ones.
  - e.g., \( S \rightarrow NP \ VP \)
  - This says:
    - A sentence \( (S) \) constituent is composed of a noun phrase \( (NP) \) constituent and a verb phrase \( (VP) \) constituent. (hierarchy)
    - The NP must precede the VP. (linear order)

Phrase Structure Rules and Trees

With every phrase structure rule, you can draw a tree for it.

**Lexicon**:

- \( Vt \rightarrow \text{saw} \)
- \( \text{Det} \rightarrow \text{the} \)
- \( \text{Det} \rightarrow \text{a} \)
- \( \text{N} \rightarrow \text{dragon} \)
- \( \text{N} \rightarrow \text{boy} \)
- \( \text{Adj} \rightarrow \text{young} \)

**Syntax rules**:

- \( S \rightarrow \text{NP} \ VP \)
- \( \text{VP} \rightarrow \text{Vt} \ NP \)
- \( \text{NP} \rightarrow \text{Det} \ N \)
- \( \text{N} \rightarrow \text{Adj} \ N \)
Some Properties of Phrase Structure Rules

- Potentially (structurally) ambiguous = have more than one analysis

(10) We need more intelligent leaders.
(11) Paraphrases:
   a. We need leaders who are more intelligent.
   b. Intelligent leaders? We need more of them!

- Recursive = property allowing for a rule to be reapplied (within its hierarchical structure).
  - e.g., NP → NP PP
  - The property of recursion means that the set of potential sentences in a language is infinite.

Parsing

Using these phrase structure rules, we can get a computer to parse a sentence = assign a structure to a sentence.

There are many, many parsing techniques out there.

- **Top-down**: build a tree by starting at the top (i.e. S → NP VP) and working down the tree.
- **Bottom-up**: build a tree by starting with the words at the bottom and working up to the top.

Trace of a top-down parse

S
  NP₂
    Det₃
    Adj₆
    N₅
    N₆
    N₇
    adj₁₂
    a₁₅
    dragon₁₇

Trace of a bottom-up parse

S₁₇
  NP₈
    Det₂
    Adj₄
    N₆
    N₈
    N₉
    a₁₁
    dragon₁₃

More finely articulated rules

In practice, one actually works with rules like:

- S → NPₙ₇ VPₙ₇
- Or uses features & variables like:
  - S → NPNUM=X VPNUM=X

It can get very complicated (& fun) very quickly:

- S:TENSE=Z → NPNUM=X,PER=Y VPNUM=X,PER=Y,TENSE=Z

Writing grammar correction rules

So, with our rules, we can now write some correction rules, which we will just sketch here.

- **A baseball teams were successful.**
  - A followed by PLURAL NP: change A → The
  - i.e., one looks for a tree like: NP → Det₉ NP₉
  - We'll talk about this more with mal-rules in Language Tutoring Systems

- **John at the pizza.**
  - The structure of this sentence is NP PP, but that doesn't make up a whole sentence.
  - We need a verb somewhere.
The more we depend on spelling correctors, do we try less to correct things on our own?

- But spell checkers are not 100%
- One (older) study found that students made more errors (in proofreading) when using a spell checker!

<table>
<thead>
<tr>
<th></th>
<th>high SAT scores</th>
<th>low SAT scores</th>
</tr>
</thead>
<tbody>
<tr>
<td>use checker</td>
<td>16 errors</td>
<td>17 errors</td>
</tr>
<tr>
<td>no checker</td>
<td>5 errors</td>
<td>12.3 errors</td>
</tr>
</tbody>
</table>

(cf., http://www.wired.com/news/business/0,1367,58058,00.html)

Caveat emptor

Dangers of spelling and grammar correction


Candidate for a Pullet Surprise

("The Spell-Checker Poem")

by Mark Eckman and Jerrold H. Zar
http://grammar.about.com/od/spelling/a/spellcheck.htm

I have a spelling checker,
It came with my PC.
It plane lee marks four my revue
Miss steaks aye can knot sea.

Eye ran this poem threw it,
Your sure reel glad two no.
Its vary polished in it's weigh.
My checker tolled me sew.

...