Automated Writing Assistance: Grammar Checking and Beyond
Topic 2: Spell Checking

Robert Dale
Centre for Language Technology
Macquarie University

Spell Checking

- What's a Spelling Error?
- Non-Word Error Detection
- Error Correction
- Real-Word Error Detection

What is a Spelling Error?

- How many spelling errors are there here?
  - Wot color is the dawg?
  - C u l8er
- A definition:
  - A spelling error is a word which is not spelled as it should be

Execution vs Intention

- Orthographic errors
- Typographic errors
- Examples:
  - performence → performance
  - teh → the
  - thier → their

Use Cases for Spell Checking

- Correcting spelling errors in text
- Fixing OCR output
- Correcting spelling errors in search queries
- Some solutions allow interaction, others require machine autonomy

Spell Checking

- What's a Spelling Error?
- Non-Word Error Detection
- Error Correction
- Real-Word Error Detection

Unix Spell

```
$ spell
This is the storry abuot an event that went from
baad to wurse
abuot
baad
storry
wurse
$
```

Storage Issues

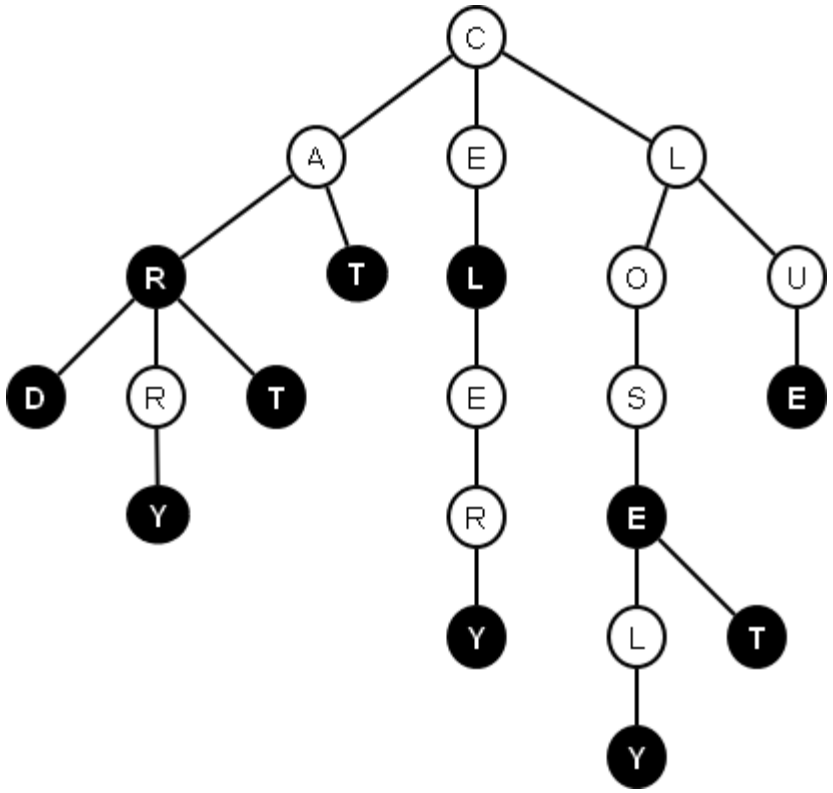
1981: The original PC's maximum memory using IBM parts was 256 KB: 64 KB on the motherboard and three 64 KB expansion cards.

A word list of100k words occupies around 500KB.

Peterson's Three Levels of Storage

- Small dictionary of frequently used words [100–200 words]
- Document-specific words [1000–2000 words]
- Larger secondary storage [10k–100k words]

Dictionary Storage via Tries



Problems with Word Lists

- False Positives
  - A valid word may be flagged as a spelling error because it is not in the list
- False Negatives
  - A misspelled word may not be flagged as a spelling error because it is orthographically identical to some other valid word

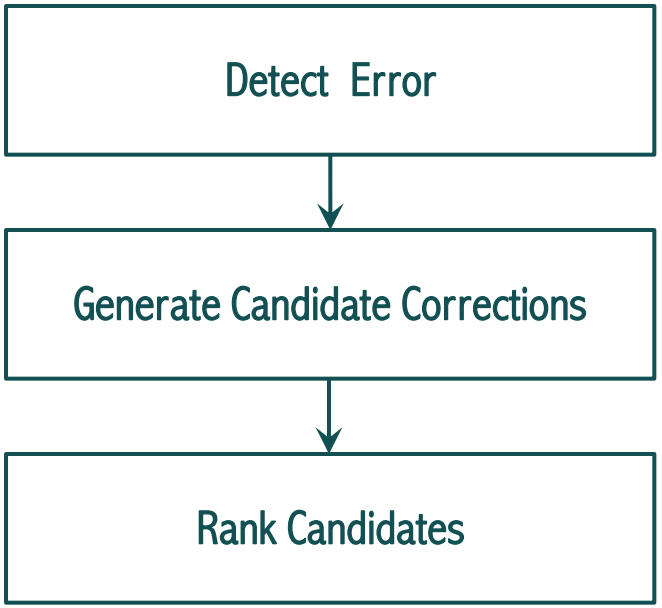
Spell Checking

- What's a Spelling Error?
- Non-Word Error Detection
- Error Correction
- Real-Word Error Detection

The Task

- Given a word which is assumed to be misspelled, find the word that the author intended to type

Spell Checking



Finding Candidate Corrections

- Look for ‘nearby’ real words
- Edit distance:
  - An edit = a deletion, an insertion, a transposition or a substitution
  - Each edit adds 1 to the edit distance between strings
- Damerau 1980: 80% of spelling errors are 1 edit from the correct string

Edit Distance

- Deletion:
  - continuous → continous
- Insertion:
  - explanation → explaination
- Substitution
  - anybody → anyboby
- Transposition:
  - automatically → autoamtically

Using Edit Distance

- For a hypothesized misspelled word:
  - Generate all strings within an edit distance of 1
  - Filter non-words out of the list

teh	→	tea	→	tea
		teb		~~teb~~
		…		…
		the		the

Potential Problems with Edit Distance

- For a string of n characters from an alphabet of size k, number of strings within edit distance 1:

$$k(2n + 1) + n - 1$$

- Peterson [1980]: an average of 200 dictionary accesses for each misspelling
- Also: words > edit distance 1 are ignored

Approaches to Spelling Correction

- Angell [1983]: Trigram Analysis
- Yannakoudakis and Fawthrop [1983]: Error Patterns
- van Berkel and De Smedt [1988]: Triphone Analysis
- Kernighan, Church and Gale [1990]: The Noisy Channel Model
- Agirre et al [1998]: Using Context
- Brill and Moore [2000]: String-to-String Edits
- Toutanova and Moore [2002]: Pronunciation Modeling

Angell's [1983] ACUTE: Trigram Analysis

- An alternative means of finding candidate replacements:
  - Find the closest dictionary word based on number of shared trigrams

Angell's [1983] ACUTE: An Example

- CONSUMING =
  $$C, $CO, CON, ONS, NSU, SUM, UMI, MIN, ING, NG$, G$$
- CONSUMMING =
  $$C, $CO, CON, ONS, NSU, SUM, UMM, MMI, MIN, ING, NG$, G$$

Angell's [1983] ACUTE: Similarity Measurement

- The DICE coefficient:

$$\frac{2c}{n + n'}$$

- where:

  c = number of shared trigrams

  n and n′ = number of trigrams in each of the two words

- In our example: similarity = 20/23 = 0.87

Angell's [1983] ACUTE: Performance

- Given:
  - A dictionary of 64,636 words
  - A corpus of 1544 misspellings
- Results:
  - 72.6% successfully corrected
  - 5.1% more than one best match
  - 9.0% correct spelling ranked second or third
  - 9.7% correct spelling not identified

Angell's [1983] ACUTE: Performance in Terms of Edit Distance

Error Type	N	Correct	Joint	2nd/3rd	Wrong
Omission	570	92.4	3.5	2.6	1.4
Insertion	267	88.4	4.5	4.5	2.6
Substitution	354	71.4	7.6	15.9	5.1
Transposition	136	36.0	6.6	19.1	38.2
Multiple	217	54.8	3.2	13.4	26.6

What Causes Spelling Errors?

- Typing errors (typographic errors, errors of execution)

  the → teh

  spell → speel

- Cognitive errors (orthographic errors, errors of intention)

  receive → recieve

  conspiracy → conspiricy

  abyss → abiss

  naturally → nacherly

Approaches to Spelling Correction

- Angell [1983]: Trigram Analysis
- Yannakoudakis and Fawthrop [1983]: Error Patterns
- van Berkel and De Smedt [1988]: Triphone Analysis
- Kernighan, Church and Gale [1990]: The Noisy Channel Model
- Agirre et al [1998]: Using Context
- Brill and Moore [2000]: String-to-String Edits
- Toutanova and Moore [2002]: Pronunciation Modeling

Yannakoudakis and Fawthrop [1983]: Error Patterns

- Problem Statement:
  - Given a non-word error, generate a ranked list of candidate replacements based on common error patterns
- Background assumption:
  - Many errors are due to phonetic confusion
  - But conversion into a phonetic coding assumes a dialect

Yannakoudakis and Fawthrop [1983]: The Approach

- Analysed a corpus of 1377 spelling errors
- Divide each word into spelling elements – a bit like vowel and consonant clusters, but oriented towards typical confusions in spelling:
  - F-OR-EI-GN
  - D-I-PH-TH-ER-IA
  - F-A-V-OUR-A-B-L-E

Yannakoudakis and Fawthrop [1983]: Error Rules

- A 'vocabulary' of 299 spelling elements
- Very large space of possible element-to-element replacements
- Constrained by observed patterns:
  - Doubling or singling of characters
  - Errors involving specific characters
  - Errors involving related phonemes
  - …

→3079 error rules

Yannakoudakis and Fawthrop [1983]: Other Heuristics

- The most frequent length of an error form is one character less than the dictionary form
- Typing errors are caused by hitting an adjacent key to the one intended or by hitting the correct key and its neighbour
- Short error forms do not contain more than one error
- If the error form is short, only dictionary words differing in length by one character from the error form are examined

Yannakoudakis and Fawthrop [1983]: Examples

- F-ILIPIN-OE-S → PH-ILIPIN-O-S
- CA-PH-EINE → CA-FF-EINE
- When there's more than one possible correction, choice is made via 'subjective Bayesian probabilities' on the dictionary words and the error rules

Yannakoudakis and Fawthrop [1983]: Performance

- Corrected 90% of 1153 error forms
  - In 95% of these corrections one word was identified
  - In 5% a choice of between 2 and 4 words was offered
- Mean time to correct an error was 22 seconds, with a minimum of five seconds and a maximum of 50 seconds

Approaches to Spelling Correction

- Angell [1983]: Trigram Analysis
- Yannakoudakis and Fawthrop [1983]: Error Patterns
- van Berkel and De Smedt [1988]: Triphone Analysis
- Kernighan, Church and Gale [1990]: The Noisy Channel Model
- Agirre et al [1998]: Using Context
- Brill and Moore [2000]: String-to-String Edits
- Toutanova and Moore [2002]: Pronunciation Modeling

van Berkel and De Smedt [1988]: Triphone Analysis for Phonetic Errors

- Statistical methods are best suited to typographic errors
- Linguistic methods are more appropriate for orthographic errors
- Assumption: orthographic errors are more important
  - They are more persistent
  - They leave a worse impression

van Berkel and De Smedt [1988]: Approach

- Use grapheme to phoneme conversion to generate all phonological variants of the misspelled word
- Split phoneme string into triphones
- Find dictionary words containing low frequency (ie more informative) triphones
- Choose most similar word found

van Berkel and De Smedt [1988]: Performance

- Evaluated on 188 misspelled Dutch surnames and a dictionary of 254 names

System	First Choice	2nd or 3rd	Not Found
SPELL	58.5	1.1	40.4
ACUTE	89.9	6.9	3.2
TRIPHONE ANALYSIS	94.1	5.9	0.0

Approaches to Spelling Correction

- Angell [1983]: Trigram Analysis
- Yannakoudakis and Fawthrop [1983]: Error Patterns
- van Berkel and De Smedt [1988]: Triphone Analysis
- Kernighan, Church and Gale [1990]: The Noisy Channel Model
- Agirre et al [1998]: Using Context
- Brill and Moore [2000]: String-to-String Edits
- Toutanova and Moore [2002]: Pronunciation Modeling

Kernighan, Church and Gale [1990]: Using the Noisy Channel Model

- The problem:
  - Given a word in error, find the most likely word intended by the author
- Approach:
  - Find all words within edit distance of 1
  - Determine the probability of each possible edit from a corpus
  - Use these probabilities to order the list of candidates

Kernighan, Church and Gale [1990]: Using the Noisy Channel Model

- We want to find the most likely correction c given a misspelling t
- By Bayes Rule, this means finding the c that maximizes

$$Pr(c).Pr(t|c)$$

Prior model of word probabilities

The channel (or error) model

Kernighan, Church and Gale [1990]: An Example: Candidate Corrections

Typo	Correction	Transformation	
acress	actress	@ t 2	deletion
acress	cress	a # 0	insertion
acress	caress	ac ca 0	reversal
acress	access	r c 2	substitution
acress	across	e o 3	substitution
acress	acres	s # 4	insertion
acress	acres	s # 5	insertion

Kernighan, Church and Gale [1990]: Prior Probabilities

- Pr(c) is estimated by:

$$\frac{freq(c) + 0.5}{N}$$

- where freq(c) is the number of times that the word c appears in the 1988 AP corpus (N = 44 million words)

Kernighan, Church and Gale [1990]: Conditional Probabilities

$$Pr(t|c) \approx \begin{cases} \dfrac{del[c_{p-1}, c_p]}{chars[c_{p-1}, c_p]}, & \text{if deletion} \\ \dfrac{add[c_{p-1}, t_p]}{chars[c_{p-1}]}, & \text{if insertion} \\ \dfrac{sub[t_p, c_p]}{chars[c_p]}, & \text{if substitution} \\ \dfrac{rev[c_p, c_{p+1}]}{chars[c_p, c_{p+1}]}, & \text{if reversal} \end{cases}$$

- ***del***, ***add***, ***sub*** and ***rev*** are derived from confusion matrices
- ***chars*** are occurrence counts derived from the corpus

Kernighan, Church and Gale [1990]: Confusion Matrices

sub[X, Y] = Substitution of X (incorrect) for Y (correct)

X \ Y (correct)	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y	z
a	0	0	7	1	342	0	0	2	118	0	1	0	0	3	76	0	0	1	35	9	9	0	1	0	5	0
b	0	0	9	9	2	2	3	1	0	0	0	5	11	5	0	10	0	0	2	1	0	0	8	0	0	0
c	6	5	0	16	0	9	5	0	0	0	1	0	7	9	1	10	2	5	39	40	1	3	7	1	1	0
d	1	10	13	0	12	0	5	5	0	0	2	3	7	3	0	1	0	43	30	22	0	0	4	0	2	0
e	388	0	3	11	0	2	2	0	89	0	0	3	0	5	93	0	0	14	12	6	15	0	1	0	18	0
f	0	15	0	3	1	0	5	2	0	0	0	3	4	1	0	0	0	6	4	12	0	0	2	0	0	0
g	4	1	11	11	9	2	0	0	0	1	1	3	0	0	2	1	3	5	13	21	0	0	1	0	3	0
h	1	8	0	3	0	0	0	0	0	0	2	0	12	14	2	3	0	3	1	11	0	0	2	0	0	0
i	103	0	0	0	146	0	1	0	0	0	0	6	0	0	49	0	0	0	2	1	47	0	2	1	15	0
j	0	1	1	9	0	0	1	0	0	0	0	2	1	0	0	0	0	0	5	0	0	0	0	0	0	0
k	1	2	8	4	1	1	2	5	0	0	0	0	5	0	2	0	0	0	6	0	0	0	4	0	0	3
l	2	10	1	4	0	4	5	6	13	0	1	0	0	14	2	5	0	11	10	2	0	0	0	0	0	0
m	1	3	7	8	0	2	0	6	0	0	4	4	0	180	0	6	0	0	9	15	13	3	2	2	3	0
n	2	7	6	5	3	0	1	19	1	0	4	35	78	0	0	7	0	28	5	7	0	0	1	2	0	2
o	91	1	1	3	116	0	0	0	25	0	2	0	0	0	0	14	0	2	4	14	39	0	0	0	18	0
p	0	11	1	2	0	6	5	0	2	9	0	2	7	6	15	0	0	1	3	6	0	4	1	0	0	0
q	0	0	1	0	0	0	27	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
r	0	14	0	30	12	2	2	8	2	0	5	8	4	20	1	14	0	0	12	22	4	0	0	1	0	0
s	11	8	27	33	35	4	0	1	0	1	0	27	0	6	1	7	0	14	0	15	0	0	5	3	20	1
t	3	4	9	42	7	5	19	5	0	1	0	14	9	5	5	6	0	11	37	0	0	2	19	0	7	6
u	20	0	0	0	44	0	0	0	64	0	0	0	0	2	43	0	0	4	0	0	0	0	2	0	8	0
v	0	0	7	0	0	3	0	0	0	0	0	1	0	0	1	0	0	0	8	3	0	0	0	0	0	0
w	2	2	1	0	1	0	0	2	0	0	1	0	0	0	0	7	0	6	3	3	1	0	0	0	0	0
x	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	9	0	0	0	0	0	0	0
y	0	0	2	0	15	0	1	7	15	0	0	0	2	0	6	1	0	7	36	8	5	0	0	1	0	0
z	0	0	0	7	0	0	0	0	0	0	0	7	5	0	0	0	0	2	21	3	0	0	0	0	3	0

Kernighan, Church and Gale [1990]: The Example: Scoring the Candidates

Correction	Score	Raw	freq(c)	Pr(t | c)
actress	37%	.157	1343	55/470,000
cress	0%	.000	0	46/32,000,000
caress	0%	.000	4	0.95/580,000
access	0%	.000	2280	0.98/4,700,000
across	18%	.077	8436	93/10,000,000
acres	21%	.092	2879	417/13,000,000
acres	23%	.098	2879	205/6,000,000

Kernighan, Church and Gale [1990]: The Example in Context

… was called a "stellar and versatile acress whose combination of sass and glamour has defined her ….

Kernighan, Church and Gale [1990]: Performance

- Test sample of 329 misspelled words with two candidate corrections
- Program agrees with majority of judges in 87% of cases

Approaches to Spelling Correction

- Angell [1983]: Trigram Analysis
- Yannakoudakis and Fawthrop [1983]: Error Patterns
- van Berkel and De Smedt [1988]: Triphone Analysis
- Kernighan, Church and Gale [1990]: The Noisy Channel Model
- Agirre et al [1998]: Using Context
- Brill and Moore [2000]: String-to-String Edits
- Toutanova and Moore [2002]: Pronunciation Modeling

Agirre et al [1998]: Using Context

- The Goal:
  - Given a non-word error, use the context to determine the most likely correction (the 'single proposal')

Agirre et al [1998]: Sources of Knowledge

- Syntactic:
  - Constraint Grammar (CG) used to rule out candidate corrections that are grammatically unacceptable
- Semantic:
  - Use distance in WordNet (CD) to choose the candidate noun correction that is closest to the words in the context
- Statistical:
  - Brown Corpus (BF) and document (DF) word frequencies

Agirre et al [1998]: Performance

- A large number of combinations tried on artificially generated error data
- Best performing combinations tested on real error data
- Main findings:
  - Combination of syntax and document frequencies works best
    - But effect of DF impacted by small documents
  - Brown Corpus frequencies and conceptual density not useful

Approaches to Spelling Correction

- Angell [1983]: Trigram Analysis
- Yannakoudakis and Fawthrop [1983]: Error Patterns
- van Berkel and De Smedt [1988]: Triphone Analysis
- Kernighan, Church and Gale [1990]: The Noisy Channel Model
- Agirre et al [1998]: Using Context
- Brill and Moore [2000]: String-to-String Edits
- Toutanova and Moore [2002]: Pronunciation Modeling

Brill and Moore [2000]: Improving the Noisy Channel Model

- The Approach:
  - Given a word assumed to be in error, use a noisy channel model based on string to string edits to determine candidate corrections

Brill and Moore [2000]: Approach

- Generalise the error model to permit generic string to string edits

  $\Pr(\alpha \rightarrow \beta)$ is the probability that the user types β when they meant α

- Edits are conditioned on position in the string:
  - $\Pr(\alpha \rightarrow \beta \mid \text{PSN})$ where PSN = start, middle, or end of word
- Observation:
  - $P(e \mid a)$ does not vary by location
  - $P(\text{ent} \mid \text{ant})$ does

Brill and Moore [2000]: Example

- Spelling error:
  - physical → fisikle
- Conceptually, the user picks a word; partitions it into substrings; generates each partition, perhaps erroneously
  - ph+y+s+i+c+al → f+i+s+i+k+le
- Probability of generating the error is then:
  - P(f | ph) . P(i | y) . P(s | s) . P(i | i) . P(k | c) . P(le | al)

Brill and Moore [2000]: Learning the Model

- String to string edits are derived from mismatches in aligned ⟨spelling error, correction⟩ pairs:

- Edits derived:

  $c \rightarrow k$, $ac \rightarrow ak$, $c \rightarrow kg$, $ac \rightarrow akg$, $ct \rightarrow kgs$

Brill and Moore [2000]: Testing

- 10000 word corpus of spelling errors + corrections
- 200k word dictionary
- Language model assigns uniform probabilities to all words

Brill and Moore [2000]: Performance

Without positional information:

Max Window	1-Best	2-Best	3-Best
0	87.0	93.9	95.9
Church and Gale	89.5	94.9	96.5
1	90.9	95.6	96.8
2	92.9	97.1	98.1
3	93.6	97.4	98.5
4	93.6	97.4	98.5

Brill and Moore [2000]: Performance

With positional information:

Max Window	1-Best	2-Best	3-Best
0	88.7	95.1	96.6
1	92.8	96.5	97.4
2	94.6	98.0	98.7
3	95.0	98.0	98.8
4	95.0	98.0	98.8
5	95.1	98.0	98.8

Approaches to Spelling Correction

- Angell [1983]: Trigram Analysis
- Yannakoudakis and Fawthrop [1983]: Error Patterns
- van Berkel and De Smedt [1988]: Triphone Analysis
- Kernighan, Church and Gale [1990]: The Noisy Channel Model
- Agirre et al [1998]: Using Context
- Brill and Moore [2000]: String-to-String Edits
- Toutanova and Moore [2002]: Pronunciation Modeling

Toutanova and Moore [2002]: Pronunciation Modeling

- Observation:
  - Many errors in Brill and Moore [2000] are due to word pronunciation

Misspelling	Correct Word	B+M Proposal
edelvise	edelweiss	advice
bouncie	bouncy	bounce
latecks	latex	lacks

Toutanova and Moore [2002]: Approach

- Build two error models:
  - The Brill and Moore [2000] model
  - A phone-sequence to phone-sequence error model
- Uses machine-learned letter-to-phone conversion
- At classification time, the two models are combined using a log linear model

Toutanova and Moore [2002]: Performance

Model	1-Best	2-Best	3-Best	4-Best
B+M	94.21	98.18	98.90	99.06
Phoneme	86.36	93.65	95.69	96.63
Combined	95.58	98.90	99.34	99.5
Error Reduction	23.8	39.6	40	46.8

Toutanova and Moore [2002]: Examples

Misspelling	Correct	LTR Guess
bouncie	**bouncy**	bounce
edelvise	**edelweiss**	advise
grissel	**gristle**	grizzle
latecks	**latex**	lacks
neut	**newt**	nut
rench	**wrench**	ranch
saing	**saying**	sang
stail	**stale**	stall

Spell Checking

- What's a Spelling Error?
- Non-Word Error Detection
- Error Correction
- Real-Word Error Detection

Real Word Errors are a Real World Problem

- Peterson:
  - 10% of typing errors are undetected when using a 50k word dictionary
  - 15% are undetected when using a 350k word dictionary
- Two Main Approaches in the Literature:
  1. Try to determine from contextual evidence whether a word is a real-word error
  2. Given a potential real-word error, determine the most likely correction

Mays, Damerau and Mercer [1991]: Using Trigrams to Detect Real-Word Errors

- The Goal:
  - Given a text, determine presence of real-word errors and propose candidate corrections
- Basic Idea:
  - If the trigram-derived probability of an observed sentence is lower than that of any sentence obtained by replacing one of the words with a spelling variation, then hypothesize that the original is an error and the variation is what the user intended.

Mays, Damerau and Mercer [1991]: The Idea

- Example:
  - I saw the man it the park
- Syntax can be used:
  - to determine that an error is present
  - to determine whether candidate corrections result in grammatical strings
- But we don't have 100% reliable parsers, so try something else: a trigram language model

Mays, Damerau and Mercer [1991]: The Key Insights

- A low-probability word sequence can be considered evidence for a real-word error
- High-probability sequences can be used to rank correction candidates

Mays, Damerau and Mercer [1991]: The Data

- Restricted to edit distance 1 errors, and one misspelled word per sentence
- Given a set of 100 randomly selected sentences:
  - For each sentence, generate all possible sentences where each word is subjected to edit distance 1 transformations

$\rightarrow$ 8628 misspelled sentences

Mays, Damerau and Mercer [1991]: The Noisy Channel Model

- We want to find the most likely correction w given a misspelling x
- By Bayes Rule, this means finding the w that maximizes

$$Pr(w).Pr(x|w)$$

The channel model

Prior model of word probabilities, approximated using the trigram model

Mays, Damerau and Mercer [1991]: The Noisy Channel Model

- The channel model:

$$P(x|w) = \begin{cases} \alpha & \text{if } x = w \\ (1-\alpha)/|SV(w)| & \text{if } x \in SV(w) \\ 0 & \text{otherwise} \end{cases}$$

- SV(w) is the set of spelling variations of w; all are considered equally likely
- The challenge: find the optimal value for α, the a priori belief that the observed input word is correct

Mays, Damerau and Mercer [1991]: Performance

α	Original	Changed	Correct	Composite
0.9000	15.0	94.4	78.7	74.4
0.9900	3.0	86.9	90.9	79.0
0.9990	1.0	76.7	95.4	73.2
0.9999	0.0	63.7	97.0	61.8

- original = %age of original input sentences changed to some other sentence
- changed = %age of misspelled sentences changed to some other sentence
- correct = %age of changed misspelled sentences that were changed correctly
- composite = %age of misspelled sentences that were changed correctly

Mays, Damerau and Mercer [1991]: Observations

- As α increases the correctness of the changes increases
- As α increases the percentage of misspelled sentences changed to some other sentence decreases
- A reasonable value for α lies in the range 0.99–0.999

See Wilcox-O'Hearn, Hirst and Budanitsky [2008] for a rational reconstruction and proposals for improvements

Hirst and Budanitsky [2005]: Lexical Cohesion for Real-Word Error Correction

- The Goal:
  - Determine real-word errors on the basis of their semantic incompatibility with the rest of the text
- Basic idea:
  - Words which are semantically unrelated to the context, but whose spelling variations are related to the context, are possible real-world spelling errors

Hirst and Budanitsky [2005]: Syntax Doesn't Always Help

- It is my sincere hole [hope] that you will recover swiftly.
- The committee is now [not] prepared to grant your request.

Hirst and Budanitsky [2005]: The Underlying Observation

- Linguistic cohesion is maintained by lexical chains: words linked by lexical and semantic relationships
  - literal repetition
  - coreference
  - synonymy
  - hyponymy

Hirst and Budanitsky [2005]: Key Assumptions

- A real-word spelling error is unlikely to be semantically related to the text.
- Usually, the writer's intended word will be semantically related to nearby words.
- It is unlikely that an intended word that is semantically unrelated to all those nearby will have a spelling variation that is related.
- So: detect tokens that fit into no lexical chain in the text and replace them with words for which they are plausible mistypings that do fit into a lexical chain.

Hirst and Budanitsky [2005]: Requirements

- A mechanism for generating candidate spelling variations
  - For example, all real words in edit distance 1
- A mechanism for determining whether two words are semantically related
  - For example, distance measures in WordNet

Hirst and Budanitsky [2005]: The Approach

- Ignore words not in the lexicon, closed class words, and elements of a list of non-topical words (eg know, find, world)
- For any remaining suspect:
  - Determine if it is semantically related to another word in the text
  - If not, then look for positive evidence: is any spelling variation a better fit?

Hirst and Budanitsky [2005]: Performance

Scope	Detection		
	P_D	R_D	F_D
1	0.184	0.498	0.254
3	0.205	0.372	0.245
5	0.219	0.322	0.243
MAX	0.247	0.231	0.211
Chance	0.0129	0.0129	0.0129

But: Wilcox-O'Hearn et al [2008] show that the Mays, Damerau, and Mercer model performs better.

Whitelaw et al [2009]: The Web as a Corpus for Spelling Correction

- Basic idea:
  - Use the web as a large noisy corpus to infer knowledge about misspellings and word usage
  - Avoid using any manually-annotated resources or explicit dictionaries
- Important feature: easily ported to other languages

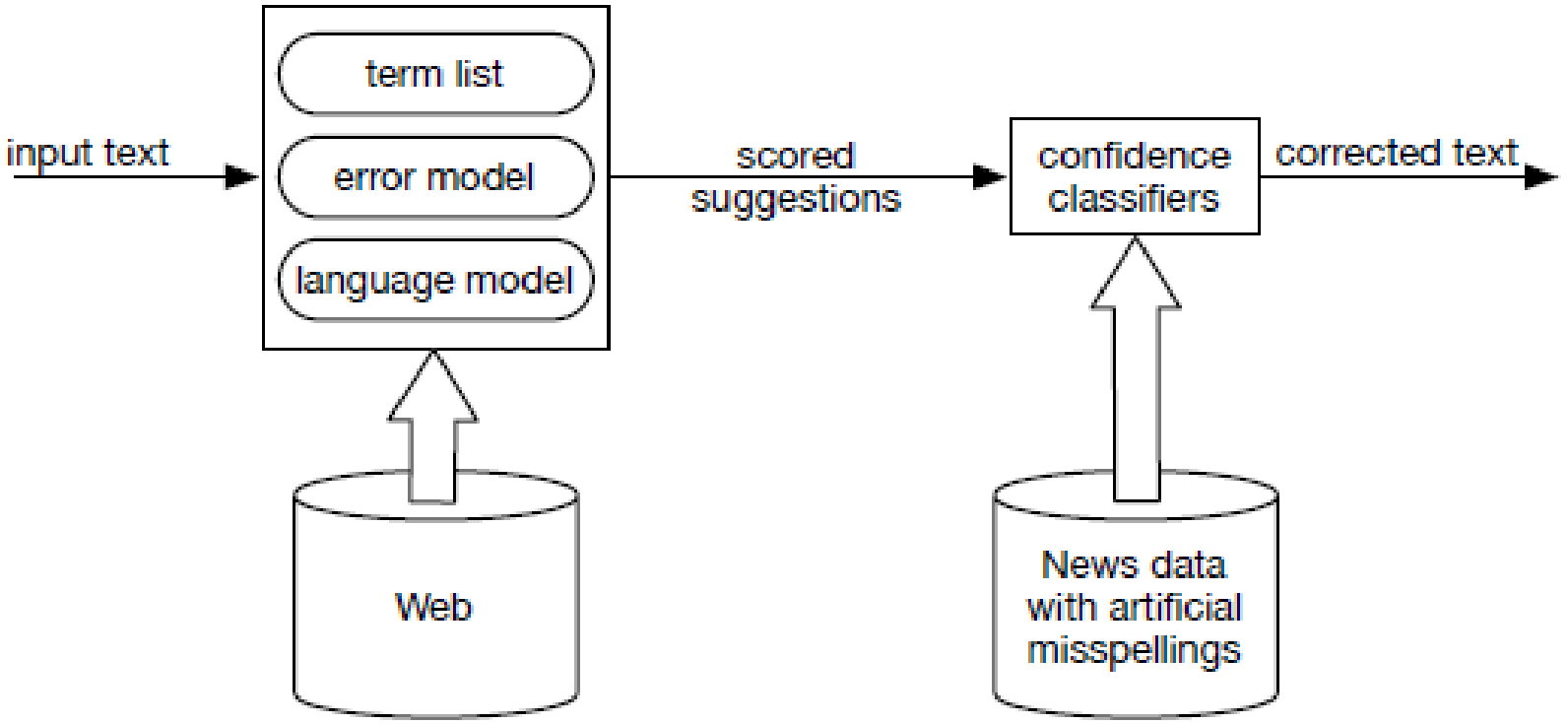
Whitelaw et al [2009]: Approach

- Infer information about misspellings from term usage observed on the Web, and use this to build an error model
- The most frequently observed terms are taken as a noisy list of potential candidate corrections
- Token n-grams are used to build a language model which is used to make context-appropriate corrections

Whitelaw et al [2009]: Key Feature

- Given error and LM scores, confidence classifiers determine the thresholds for spelling error detection and auto-correction
- Classifiers are trained on clean news data injected with artificial misspellings

Whitelaw et al [2009]: System Architecture



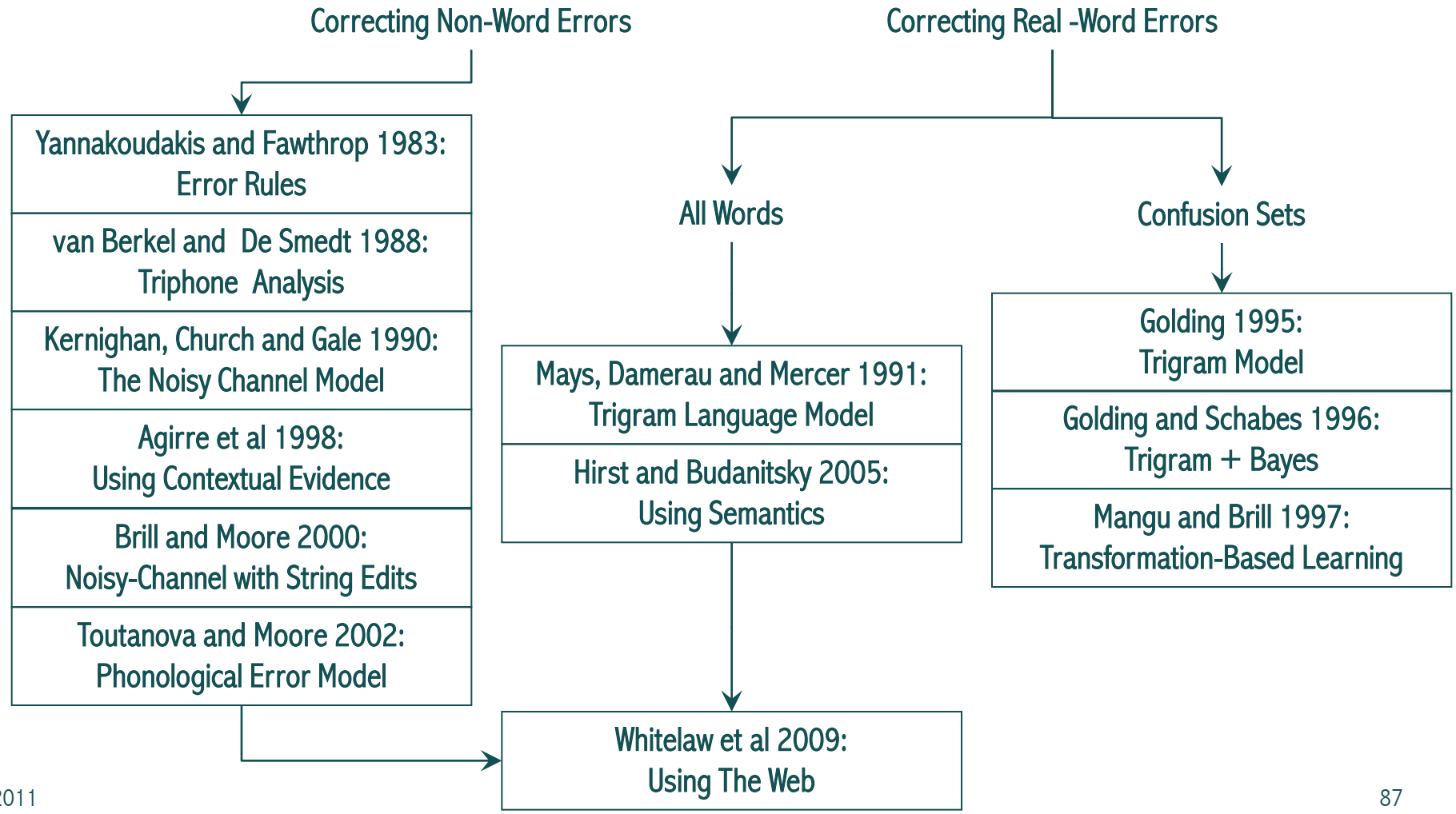
Whitelaw et al [2009]: Candidate Corrections

- The Term List:
  - The 10 million most frequently occurring tokens from a > 1 billion sample of web pages (so it's noisy)
- The Error Model:
  - A substring model like Brill and Moore's
  - Built using ⟨intended word, misspelling⟩ pairs inferred from the web
- The Language Model:
  - Derived from the web, of course

Whitelaw et al [2009]: Performance

- Total error rate for best configuration reduces the error of the best aspell system from 4.83% to 2.62% on artificial data
- Total error rate reduces the error of the best aspell system from 4.58% to 3.80% on human English data
- Total error rate reduces the error of the best aspell system from 14.09% to 9.80% on human German data

A Road Map


Correcting Non-Word Errors
Correcting Real -Word Errors
Yannakoudakis and Fawthrop 1983: Error Rules
van Berkel and De Smedt 1988: Triphone Analysis
Kernighan, Church and Gale 1990: The Noisy Channel Model
Agirre et al 1998: Using Contextual Evidence
Brill and Moore 2000: Noisy-Channel with String Edits
Toutanova and Moore 2002: Phonological Error Model
All Words
Confusion Sets
Mays, Damerau and Mercer 1991: Trigram Language Model
Hirst and Budanitsky 2005: Using Semantics
Golding 1995: Trigram Model
Golding and Schabes 1996: Trigram + Bayes
Mangu and Brill 1997: Transformation-Based Learning
Whitelaw et al 2009: Using The Web

The Bottom Line

- Methods for generating candidate corrections for a word known to be in error are now very sophisticated
  - The noisy channel model is a good fit
  - Lots of scope for refinement in the language model and the error model
- Determining when a word has been misspelled as another word is an AI-hard problem …
- … but Google-scale language modelling does surprisingly well

Are We There Yet?

- Don't know.
  - We are still hampered by a lack of shared data for evaluation.
  - We also lack a real understanding of how the different use cases for spelling correction are related.