



A Mathematical Model for Spell Checking/Correcting

Hsuan Lorraine Liang

Prof Bruce W. Watson

Prof Derrick G. Kourie

Fastar Research Group, <http://fastar.org>



Motivation

There exist many commercial and non-commercial spell checkers/correctors. However, no formal taxonomy has been constructed for these spell checking/correcting algorithms. The existing techniques are limited in their scope and accuracy. The aim of this mathematical model is to provide a thorough understanding of the basics which will form part of the foundation of the taxonomy construction and the design and implementation of the toolkit and its domain-specific language.



Spell Checking/Correcting: Definitions

- History: Research began in the 1960s.
 - *spell*
 - *grope*
- The automatic spell correcting research has focused on three main problems [Kukich 1992]:
 - nonword error detection
 - isolated-word error correction
 - context-dependent word correction
- We are making the assumption that all words are morphologically-complete.



Spell Checking/Correcting: Definitions (cont.)

- **Spell checking:**

Strings in a text, encoded by some encoding, that do not appear in a given word list, dictionary, or lexicon are detected, i.e. words that are invalid in a language.

- $f(T, D) \rightarrow \{M\}$, where a set of misspelled words will be returned.

- **Spell correction:**

One or more words are suggested from the chosen word list, dictionary, or lexicon for each misspelled word as the correct spelling.

- $f(\{M\}, D) \rightarrow \{M, \{S\}\}$, where a set of valid words will be suggested for each misspelled word.



Spell Checking/Correcting: Basic Formulation

- U denotes the set of characters in some encoding.
- $\mathbf{L} = \{L_i \mid i \in LD\}$,
where L_i denotes the set of valid words in language i as represented in the encoding U and LD denotes the set of names that describe the language under consideration.
- V_{L_i} denotes the set of characters used in language i to construct words.

Thus, we can write a set of valid characters in some language as follows:

$$\forall L \in \mathbf{L} : (V_L \subseteq U) \wedge (L \subseteq V_L^+ \subseteq U^+)$$



Spell Checking/Correcting: Mathematical Model: *Smash*

- *Smash* is an operation on a string or a sequence of symbols that will return all the valid and invalid words in a given language, i .
- Subsequences that are obviously nonwords or misspelled words should be stripped out.
- This operation returns a set of strings, each of which is in the set V_{Li}^+ .
- A text can be viewed as some strings out of the set U^+ .



Spell Checking/Correcting: Mathematical Model: *Smash* (cont.)

- *Smash* has the following signature:

$$Smash_i: U^+ \rightarrow \mathbf{P}(V_{Li}^+)$$

- For example,

$$Smash_{English}(\pi = 3.1416 \text{ is an equation}) = \{\text{is, an, equation}\}$$

$$Smash_{English}(\text{the blu flower im the garden}) = \{\text{the, blu, flower, im, garden}\}$$



Spell Checking/Correcting: Mathematical Model: *Misses*

- *Misses* is an operation to removing all valid words in language i so that what remains is the set of misspelled words, $(V_{Li}^+ - L_i)$.
- Given a text, T encoded in U , for language i . *Misses* can be stated as follows:

$$Misses_i(Smash_i(T))$$

- The signature of *Misses* is as follows:

$$Misses_i : \mathbf{P} (V_{Li}^+) \rightarrow \mathbf{P} (V_{Li}^+ - L_i)$$



Spell Checking/Correcting: Mathematical Model: *Misses* (cont.)

- For example,

$$\text{Misses}_{\text{English}}(\{\text{the, blu, flower, im, garden}\}) = \{\text{blu, im}\}$$



Spell Checking/Correcting: Mathematical Model: *Suggest*

- *Suggest* provides a set of possible alternative spellings in language i for each misspelled word in the set delivered by *Misses*.
- Given a text T , encoded in U , for language i . *Suggest* can be stated as follows:

$$Suggest_i(Misses_i(Smash_i(T)))$$

- *Suggest* returns a set of pairs which is synonymous with a function. This function therefore maps each element from the set $(V_{L_i}^+ - L_i)$ to a subset of L_i and has the signature $(V_{L_i}^+ - L_i) \rightarrow \mathbf{P}(L_i)$



Spell Checking/Correcting: Mathematical Model: Smash (cont.)

- Hence, the signature of $Suggest_i$ is:

$$Suggest_i: \mathbf{P}(V_{Li}^+ - L_i) \rightarrow ((V_{Li}^+ - L_i) \rightarrow \mathbf{P}(L_i))$$

- For example,

$$Suggest_{English}(\{\text{blu}, \text{im}\}) \rightarrow \{(\text{blu}, \{\text{blue}, \text{blew}, \text{blues}\}), (\text{im}, \{\text{in}, \text{on}, \text{am}\})\}.$$



Conclusion and Future Work

- Literature survey
- Taxonomy construction
 - Feature diagrams
- Toolkit design
- Domain-specific language (DSL) design
- Toolkit implementation
- Benchmarking
- DSL implementation