An Empirical Investigation of Selected Spell Checkers and Correctors

Hsuan Lorraine Liang
Prof Bruce W. Watson
Prof Derrick G. Kourie

Fastar Research Group, http://www.fastar.org
University of Pretoria
University of South Africa

Background and Motivation

- In order to classify various spell checking/correcting packages, an empirical investigation is required to provide a comparison of these packages in terms of their functionalities, implementation strategies, and performance.
- Most spell checking/correcting packages follow a two-layer approach: error detection and error correction.
- In this research, only context-independent spell checking/correcting packages are investigated.

Introduction

- Experiments were conducted on English-based spell checking/correcting packages, a Northern Sotho spell checker/corrector, and a Chinese spell checker/corrector.
- Hardware: an Intel Pentium D at 2.8GHz with 512MB of RAM and 40GB of HD was used for the experiments.
- Performance was evaluated with respect to the recall rate of spelling error detection, suggestion accuracy (except Unix SPELL), and time efficiency.
- Controlled data sets were used.

English-based Spell Checking/Correcting Packages

- Spell checking/correcting packages selected:
  - Unix SPELL (hashing)
  - ASPELL (reverse edit distance and Metaphone)
  - AGREP (bitap)
  - FSA Package [Daciuk 1998] (finite-state automata and perfect hashing)
  - MS Word 2003
- Platforms:
  - Linux (Mandrakelinux 10.1): SPELL, ASPELL, AGREP, and FSA Package
  - Windows XP Professional 2002: MS Word 2003

English-based Spell Checking/Correcting Packages (cont.)

- Data sets:
  - Data set 1:
    - 15,000 most frequently-used words extracted (using the Oxford WordSmith Tools 4.0) from the British National Corpus (BNC).
    - 1,000 unique misspellings were extracted from the Birkbeck Spelling Error Corpus (BSEC). 742 single-letter errors, 201 2-letter errors, 44 3-letter errors, 9 4-letter errors, and 4 5-letter errors were identified.
  - Data set 2 (used as a sanity check for consistency):
    - 15,000 words extracted from the BNC.
    - 1,000 unique misspellings extracted from BSEC. 734 single-letter errors, 198 2-letter errors, 45 3-letter errors, 16 4-letter errors, and 7 5-letter errors.

English-based Spell Checking/Correcting Packages (cont.)

- Dictionaries used:
  - SPELL, ASPELL and MS Word were all equipped with their standard supplied dictionaries.
  - AGREP and FSA package made use of the dictionary supplied in SPELL (29,197 words).
- Experiments conducted:
  - Spell checking accuracy: SPELL, ASPELL, AGREP, FSA package and MS Word
  - Spell correcting accuracy: ASPELL, AGREP, FSA package and MS Word
  - Time efficiency: SPELL, ASPELL, AGREP and FSA package
- Experimental results:
  - Spell checking recall rates: SPELL/AGREP – MS Word – FSA package - ASPELL

English-based Spell Checking/Correcting Packages (cont.)

- Experimental results (cont.):
  - Spell correction accuracy: ASPELL – MS Word 2003 – FSA package - AGREP
  - Time performance: SPELL – ASPELL – AGREP – FSA package
- The performance of spell checking packages is directly related to the size and the content of the dictionaries used or supplied.
- Problems identified:
  - The results of the running time are not fine-grain enough. A clock-cycle timer needs to be employed.
  - Determine how morphological analysis influenced the spell checking performance of each package. New data sets need to be composed. More data sets are required for the measurement of the morphological intelligence:
    - Only root words, e.g. reach
    - Words in more complex inflected forms, e.g. unreachable

DAC Northern Sotho Spell Checker

- DAC Northern Sotho (DNS) Spell Checker made use of the TshwaneDJe Sesotho sa Leboa corpus. A list of 804 single-letter errors, 78 2-letter errors, and 26 3-letter errors were used for the experiments.
- Experimental results:
  - 882 out of 908 misspellings were detected which achieved a spell checking recall rate of 97.14%.
  - The precision rate was 86.73% (117 correctly-spelled words out of 882 words that were wrongly detected as misspellings).
  - The overall spell correction accuracy is 91.24%.

DAC Northern Sotho Spell Checker (cont.)

- Findings:
  - DNS Spell Checker is less mature than all the English-based spell checking/correcting packages (low recall rate although high correction rate) and requires further fine tuning in terms of the content of the dictionary (e.g. names of places) and adapt a more complex morphological analysis.
  - Unfortunately details on the error detection and spelling suggestion strategies are not available as DNS Spell Chekcer is merely a Windows installer.

CInsunSpell

- Each Chinese word can consist at least one character and up to four characters [Li, Sun & Wang 2002]. Thus, error detection was performed on the character level and sub-string level within each word.
- Test data:
  - It contains approximately 59,000 Chinese characters. It was obtained from People's Daily 1993, 1994 and 2000.
  - 595 single-word errors were identified.
- Experimental results:
  - The overall spell checking recall rate is 80.84% which compares roughly with that of DNS Spell Checker, but is far less than that of the English-based spell checking/correcting packages.

CInsunSpell (cont.)

- Experimental results (cont.):
  - It achieves a spelling correction accuracy of 69.44%.
- CInsunSpell performs worse than the English-based spell checking/correcting packages and DNS Spell Checker in both spell checking and correcting is partially attributable to the complexity of the language structure of concerned. It is also partially attributable to Chinese's input methods used for the corpus, e.g. Pinyin method.

Conclusion

- The performance of spell checking packages is directly related to the size and the content of the dictionaries used or supplied.
- The most studied and used technique is edit distance and probabilistic techniques are catching up. Similarity key and neural nets are not found in many spell checkers and correctors.
- Hybrid approach
- Phonetic spell checkers and correctors have slower running time but better correction rate.
- Morphological analysis or affix stripping scheme plays an important role in the size of the dictionary as well as the spell checking performance of various spell checking/correcting packages.

Work to Be Done

- Determine how morphological analysis influenced the performance of various spell checking/correcting packages.
- Fine tune experiments for time performance using a clock-cycle timer.
- Classification needs to be constructed according to functionalities, implementation strategies, and performance.
- Investigate Google Spell Checker.

Future Work

- Research and classification on context-dependent error correcting algorithms.
- Faster running time for phonetic spell checkers/correctors.
- Automatic spell correcting vs. interactive spell correcting.
- Spell checking and correcting for Chinese which has a comparatively more complex language structure and a much larger lexicon.