Preface

François YVON
LIMSI-CNRS & Université Paris Sud, Orsay

Robert Dale
Macquarie University, Sydney

Spelling error correction is one of the most ancient applications of Natural Language Processing. It was also the first to make its way to end-users, in the form of automatic spell checkers in text processing software. Such tools are nowadays embedded in most text input interfaces and are routinely used by millions of users. As it turns out, however, the problem of automatic error correction is far from solved and existing systems still fail to detect and fix many errors. On the one hand, many error types remain out of reach of existing spelling and grammar checkers, which rely on only very local contexts to make their decisions. On the other hand, spell checkers often lack the morphological and lexicographical expertise that is required to distinguish a legitimate neologism from a typo. And computer tools that could fix more subtle kinds of errors at the grammatical, stylistic, semantic, or discourse level are still in their infancy.

Automatic correction tools are however more needed than ever, as more and more users depend on written exchanges for their professional and private activities: in emails, on forums, to query the Web, in messaging systems, in chatrooms, on recommendation sites, in social networks, etc—the occasions to type and to typo are many, all the more so as writers often try to adhere to norms, or to use idioms and languages they only imperfectly master. Not only would such tools help to liberate writers from the fear of making mistakes, they could also greatly help improve many other NLP applications (such as sentiment analysis, machine translation, or information distillation, to name a few), the ultimate performance of which often depends on the correctness (or the predictability) of their input text.

If human are pretty good at generating errors, so are most machines that try to output language: indeed, optical character recognition and automatic speech transcription systems, like statistical machine translation engines, have long made themselves famous for their poor sense of spelling and grammar. Post-processing the output of these systems with error-correction tools has been attempted by many, with little success so far. ASR and MT engines are not the only NLP components to make errors, and it seems fair to say that no NLP system is completely perfect in its behavior: after 20 years of intense research, the best POS taggers still continue to commit a significant number of errors, even on clean text inputs; and the situation only worsens as one tries to perform more complex analysis tasks such as parsing, co-reference resolution, word-sense disambiguation, you name it. As often noted, these errors often tend to have cascading effects down the processing pipeline—but principled solutions for limitating these cascading effects have yet to be designed and implemented.

At both ends of our NLP pipelines we thus find errors. This is certainly no news: as humans, we are experts at coping with and dealing with errors in language production; but as the goals of developing error-free software slowly vanish, we are more and more forced to take this sad reality into account. This raises several important research issues, such as: Can automatic error correction still improve? How do we go beyond simple spelling/grammatical errors? How do we design error correction tools for non-standard spelling? It is also fair to assume that not all errors are equally bad: How do we automatically assess the gravity of errors? How do we make systems more robust to other components' mistakes?

Following a renewed interest in automatic error correction (such as the "Helping Our Own" shared task on massaging non-native scientific texts in English[1] or the "Bing speller challenge",[2] on fixing Web queries), this special issue of the *TAL journal* aimed to revisit the old problem of automatic error detection, correction and evaluation in the light of these new contexts and applications. For this special issue, 11 long paper submissions were received (seven in English), covering a wide spectrum of topics. The final selection of five papers offers a nice overview of the current state-of-the-art in the development of NLP systems capable of detecting, fixing or, more generally, handling errors in real-world applications.

The paper by **Torsten Zesch** studies the automatic correction of *real world errors*, in his case, errors found in Wikipedia articles. By carefully mining revisions, *i.e.* differences in successive versions of the same article, Torsten Zesch has managed to identify several hundred naturally occurring grammatical errors that should be hard to fix for any spell-checking systems. Using this challenging test set, he then compares the corrective power of several sets of contextual cues. His results suggest that, on this data, even hybrid approaches, combining the best of statistical-based and of knowledge-based cues, will struggle to reliably discriminate between true and false negatives. Another very positive outcome of this work is an open source package that will help other teams working on grammatical errors replicate these findings.

Fabrizio Gotti and his co-authors also consider realistic errors: the false alarms of a commercial level spell-checking system. The focus of this joint academic-industrial work is not so much on fixing the human, but on fixing the machine trying to fix the humans. In this study, statistical error detection and repair takes place at the end of a (mostly symbolic) text processing chain. The challenge here is to improve the performance of a very precise, state-of-the-art, system. The authors show how, in this situation, machine learning techniques and statistical learning can actually contribute to improve the overall precision of the spell-checker. As explained in great

1. `http://clt.mq.edu.au/research/projects/hoo/`
2. `http://web-ngram.research.microsoft.com/spellerchallenge/Default.aspx`

detail, this success was however not easily obtained: indeed, reducing the overdetection rate required the design of a linguistically-oriented typology of these false alarms and the training of error-specific learners using error-specific features for each error type. Even then, it appears that some false alarms remain difficult to spot during post-processing, suggesting that a more integrated process might be required.

Michael Flor, from the *Educational Testing Service*, looks at student essays from international English language assessments. Automatically fixing the misspellings these texts contain is a first step towards implementing automatic grading procedures. As shown by the author, these short texts contain very high misspelling rates as compared to text produced by native speakers, most of them corresponding to non-word errors, leading to the study of the optimal combination of contextual cues that could help select the right correction. As it turns out, in this specific context, the simple-minded, non-contextual baseline is hard to improve upon, even when using sophisticated statistical cues. This might be because non-native errors are unpredictable and can be quite distant from the correct word, thereby requiring the spell-checking procedure to explore a larger search space than is typically done. A conclusion of this exhaustive study is that, in the absence of a proper error model, the problem of fixing non-native non-word errors is far from solved.

The paper by **Patrick Juola, John Noecker Jr and Michael Ryan** tries to answer a different question. Unavoidably, real-world texts—in their study, OCRed documents—will be noisy: before trying to fix them, which can be challenging, it might be worth trying to evaluate whether these errors are actually hurting the performance of subsequent processing tools. As it turns out, for the task of stylometric authorship detection considered in this study, the conclusion might encourage a 'laisser-faire' policy: indeed, artificially corrupting the input texts has only a moderate impact on system performance. This finding might not be so surprising, given the robustness of text classification tools used for authorship attribution; yet, it is of obvious practical significance.

Finally, the work by **Anne Bonneau** and her co-authors focuses on another type of error: errors in the production of speech. In this work, the normative references against which errors are projected is provided by the targeted application: Computer Assisted Language Learning. For such applications, it is important to check that the learner (i) has actually pronounced the words she was expected to read and (ii) has done so with the right intonation. Providing accurate feedback to the user therefore requires the implementation of a complex speech processing pipeline aimed at analyzing and segmenting the learner's input and comparing it with the reference, both at the phonetic and the prosodic levels. This paper describes and diagnoses several modules in this pipeline, aimed respectively at checking the correctness of the input phonemic string, then at assessing the precision of fine-grained segmentations in phonemes and the impact of these segmentation errors. An especially interesting aspect of this work, which resonates well with Flor's paper on typical foreigner misspellings, is the account of non-native (mis)pronunciations: indeed, foreign language learners cannot expect to perfectly master pronunciation rules and provision has to be made for such inconsistencies.

Acknowledgements

We would like to express our gratitude to the whole scientific committee of the *TAL journal*, and even more so to the members of the specific scientific committee that was set up for this special issue:

– Martine Adda, LPL/CNRS, Paris (France)
– Delphine Bernhard, LiLPa, Université de Strasbourg (France)
– Simon Charest, Druide informatique, Montréal (Canada)
– Anne Dister, Facultés Universitaires Saint-Louis, Bruxelles (Belgique)
– Yannick Estève, LIUM, Université du Maine, Le Mans (France)
– Thierry Fontenelle, Centre de traduction des organes de l'Union européenne, Luxembourg (Luxembourg)
– Iñaki Alegría, University of the Basque Country (Spain)
– Diana Inkpen, Université d'Ottawa (Canada)
– Marie-José Hamel, Université d'Ottawa (Canada)
– David Langlois, LORIA, Université de Lorraine, Nancy, (France)
– Alessandro Lenci, Università di Pisa (Italy)
– Pierre Nugues, University of Lund (Sweden)
– Martin Raynaert, Tilburg University (The Nederlands)
– Alla Rozovskaya, University of Illinois at Urbana-Champaign (USA)
– Benoit Sagot, ALPAGE/INRIA, Paris (France)
– Michel Simard, NRC, Ottawa (Canada)
– Khaled Shaalan, The British University in Dubai (Dubai)
– Serge Sharroff, University of Leeds (United Kingdom)
– Eric Wehrli, LATL, Université de Genève (Suisse)