Helping Our Own: Text Massaging for Computational Linguistics as a New Shared Task

Robert Dale
Centre for Language Technology
Macquarie University
Sydney, Australia
Robert.Dale@mq.edu.au

Adam Kilgarriff
Lexical Computing Ltd
Brighton
United Kingdom
adam@lexmasterclass.com

Abstract

In this paper, we propose a new shared task called HOO: Helping Our Own. The aim is to use tools and techniques developed in computational linguistics to help people writing about computational linguistics. We describe a text-to-text generation scenario that poses challenging research questions, and delivers practical outcomes that are useful in the first case to our own community and potentially much more widely. Two specific factors make us optimistic that this task will generate useful outcomes: one is the availability of the ACL Anthology, a large corpus of the target text type; the other is that CL researchers who are non-native speakers of English will be motivated to use prototype systems, providing informed and precise feedback in large quantity. We lay out our plans in detail and invite comment and critique with the aim of improving the nature of the planned exercise.

1 Introduction

A forbidding challenge for many scientists whose first language is not English is the writing of acceptable English prose. There is a concern—perhaps sometimes imagined, but real enough to be a worry—that papers submitted to conferences and journals may be rejected because the use of language is jarring and makes it harder for the reader to follow what the author intended. While this can be a problem for native speakers as well, non-native speakers typically face a greater obstacle.

The Association for Computational Linguistics' mentoring service is one part of a response.[1] A mentoring service can address a wider range of problems than those related purely to writing; but a key motivation behind such services is that an author's material should be judged on its research content, not on the author's skills in English.

This problem will surface in any discipline where authors are required to provide material in a language other than their mother tongue. However, as a discipline, computational linguistics holds a privileged position: as scientists, language (of different varieties) is our object of study, and as technologists, language tasks form our agenda. Many of the research problems we focus on could assist with writing problems. There is already existing work that addresses specific problems in this area (see, for example, (Tetreault and Chodorow, 2008)), but to be genuinely useful, we require a solution to the writing problem as a whole, integrating existing solutions to sub-problems with new solutions for problems as yet unexplored.

Our proposal, then, is to initiate a shared task that attempts to tackle the problem head-on; we want to 'help our own' by developing tools which can help non-native speakers of English (NNSs) (and maybe some native ones) write academic English prose of the kind that helps a paper get accepted.

The kinds of assistance we are concerned with here go beyond that which is provided by commonly-available spelling checkers and grammar checkers such as those found in Microsoft Word (Heidorn, 2000). The task can be simply expressed as a text-to-text generation exercise:

[1] See http://acl2010.org/mentoring.htm.

> Given a text, make edits to the text to improve the quality of the English it contains.

This simple characterisation masks a number of questions that must be answered in order to fully specify a task. We turn to these questions in Section 3, after first elaborating on why we think this task is likely to deliver useful results.

2 Why This Will Work

2.1 Potential Users

We believe this initiative has a strong chance of succeeding simply because there will be an abundance of committed, serious and well-informed users to give feedback on proposed solutions. A familiar problem for technological developments in academic research is that of capturing the time and interest of potential users of the technology, to obtain feedback about what works in a real world task setting, with an appropriate level of engagement.

It is very important to NNS researchers that their papers are not rejected because the English is not good or clear enough. They expect to invest large amounts of time in honing the linguistic aspects of their papers. One of us vividly recalls an explanation by a researcher that, prior to submitting a paper, he took his draft and submitted each sentence in turn, in quotation marks (to force exact matches only), to Google. If there were no Google hits, it was unlikely that the sentence was satisfactory English and it needed reworking; if there were hits, the hits needed checking to ascertain whether they appeared to be written by another non-native speaker.[2] To give that researcher a tool that improves on this situation should not be too great a challenge.

For HOO, we envisage that the researchers themselves, as well as their colleagues, will want to use the prototype systems when preparing their conference and journal submissions. They will have the skills and motivation to integrate the use of prototypes into their paper-writing.

2.2 The ACL Anthology

Over a number of years, the ACL has sponsored the ongoing development of the ACL Anthology, a large collection of papers in the domain of computational linguistics. This provides an excellent source for the construction of language models for the task described here. The more recently-prepared ACL Anthology Reference Corpus (Bird et al., 2008), in which 10,921 of the Anthology texts (around 40 million words) have been made available in plain text form, has also been made accessible via the Sketch Engine, a leading corpus query tool.[3]

The corpus is not perfect, of course: not everything in the ACL Anthology is written in flawless English; the ARC was prepared in 2007, so new topics, vocabulary and ideas in CL will not be represented; and the fact that the texts have been automatically extracted from PDF files means that there are errors from the conversion process.

3 The Task in More Detail

3.1 How Do We Measure Quality?

To be able to evaluate the performance of systems which attempt to improve the quality of a text, we require some means of measuring text quality. One approach would be to develop measures, or make use of existing measures, of characteristics of text quality such as well-formedness and readability (see, for example, (Dale and Chall, 1948; Flesch, 1948; McLaughlin, 1969; Coleman and Liau, 1975)). Given a text and a version of that text that had been subjected to rewriting, we could then compare both texts using these metrics. However, there is always a concern that the metrics may not really measure what they are intended to measure (see, for example, (Le Vie Jr, 2000)); readability metrics have often been criticised for not being good measures of actual readability. The measures also tend to be aggregate measures (for example, providing an average readability level across an entire text), when the kinds of changes that we are interested in evaluating are often very local in nature.

Given these concerns, we opt for a different route: for the initial pilot run of the proposed task, we intend to provide a set of development data consisting

[2] See the Microsoft ESL Assistant at `http://www.eslassistant.com` as an embodiment of a similar idea.

[3] See `http://sketchengine.co.uk/open`.

of 10 conference papers in two versions: an original version of the paper, and an improved version where errors in expression and language use have been corrected. We envisage that participants will focus on developing techniques that attempt to replicate the kinds of corrections found in the improved versions of the papers. For evaluation, we will provide a further ten papers in their original versions, and each participant's results will then be compared against a held-back set of corrected versions for these papers. We would expect the evaluation to assess the following:

- Has the existence of each error annotated in the manually revised versions been correctly identified?
- Have the spans or extents of the errors been accurately identified?
- Has the type of error, as marked in the annotations, been correctly identified?
- How close is the automatically-produced correction to the manually-produced correction?
- What corrections are proposed that do not correspond to errors identified in the manually-corrected text?

With respect to this last point: we anticipate looking closely at all such machine-proposed-errors, since some may indeed be legitimate. Either the human annotators may have missed them, or may not have considered them significant enough to be marked. If there are many such cases, we will need to review how we handle 'prima facie false positives' in the evaluation metrics.

Evaluation of the aspects described above can be achieved automatically; there is also scope, of course, for human evaluation of the overall relative quality of the system-generated texts, although this is of course labour intensive.

3.2 Where Does the Source Data Come From?

We have two candidates which we aim to explore as sources of data for the exercise. It is almost certain the first of these two options will yield material which is denser in errors, and closer to the kinds of source material that any practical application will have to work with; however, the pragmatics of the situation mean that we may have to fall back on our second option.

First, we intend to approach the Mentoring Chairs for the ACL conferences over the last few years with our proposal; then, with their permission, we approach the authors of papers that were submitted for mentoring. If these authors are willing, we use their initial submissions to the mentoring process as the original document set.

If this approach yields an insufficient number of papers (it may be that some authors are not willing to have their drafts made available in this way, and it would not be possible to make them anonymous) then we will source candidate papers from the ACL Anthology. The process we have in mind is this:

- Identify a paper whose authors are non-native English speakers.
- If a quick reading of the paper reveals a moderately high density of correctable errors with in the first page, that paper becomes a candidate for the data set; if it contains very few correctable errors, the paper is ruled as inappropriate.
- Repeat this process until we have a sufficiently large data set.

We then contact the authors to determine whether they are happy for their papers to be used in this exercise. If they are not, the paper is dropped and the next paper's author is asked.

3.3 Where do the Corrections Come From?

For the initial pilot, two copy-editors (who may or may not be the authors of this paper) hand-correct the papers in both the development and evaluation data sets. For a full-size exercise there should be more than two such annotators, just as there should be more than ten papers in each of the development and evaluation sets, but our priority here is to test the model before investing further in it.

The copy-editors will then compare corrections, and discuss differences. The possible cases are:

1. One annotator identifies a correction that the other does not.

2. Both annotators identify different corrections for the same input text fragment.

We propose to deal with instances of the first type as follows:

- The two annotators will confer to determine whether one has simply made a mistake—as many authors can testify, no proofreader will find *all* the errors in a text.
- If agreement on the presence or absence of an error cannot be reached, the instance will be dealt with as described below for cases of the second type, with absence of an error being considered a 'null correction'.

Instances of the second type will be handled as follows:

- If both annotators agree that both alternatives are acceptable, then both alternatives will be provided in the gold standard.
- If no agreement can be reached, then neither alternative will be provided in the gold standard (which effectively means that a null correction is recorded).

Other strategies, such as using a third annotator as a tie-breaker, can be utilised if the task generates a critical mass of interest and volunteer labour.

3.4 What Kinds of Corrections?

Papers can go through very significant changes and revisions during the course of their production: large portions of the material can be added or removed, the macro-structure can be re-organised substantially, arguments can be refined or recast. Ideally, a writing advisor might help with large-scale concerns such as these; however, we aim to start at a much simpler level, focussing on what is sometimes referred to as a 'light copy-edit'. This involves a range of phenomena which can be considered sentence-internal:

- domain- and genre-specific spelling errors, including casing errors;
- dispreferred or suboptimal lexical choices;
- basic grammatical errors, including common ESL problems like incorrect preposition and determiner usage;
- reduction of syntactic complexity;
- stylistic infelicities which, while not grammatically incorrect, are unwieldy and impact on fluency and ease of reading.

The above are all identifiable and correctable within the context of a single sentence; however, we also intend to correct inconsistencies across the document as whole:

- consistency of appropriate tense usage;
- spelling and hyphenation instances where there is no obvious correct answer, but a uniformity is required.

We envisage that the process of marking up the gold-standard texts will allow us to develop more formal guidelines and taxonomic descriptions for use subsequent to the pilot exercise. There are, of course, existing approaches to error markup that can provide a starting point here, in particular the schemes used in the large-scale exercises in learner error annotation undertaken at CECL, Louvain-la-Neuve (Dagneaux et al., 1996) and at Cambridge ESOL (Nicholls, 2003).

3.5 How Should the Task be Approached?

There are many ways in which the task could be addressed; it is open to both rule-based and statistical solutions. An obvious way to view the task is as a machine translation problem from poor English to better English; however, supervised machine learning approaches may be ruled out by the absence of an appropriately large training corpus, something we may not see until the task has generated significant momentum (or more volunteer annotators at an early stage!).

There is clearly a wealth of existing research on grammar and style checking that can be brought to bear. Although grammar and style checking has been in the commercial domain now for three decades, the task may provide a framework for the first comparative test of many of these applications.

Because the nature of errors is so diverse, this task offers the opportunity to exercise a broad range of approaches to the problem, and also allows for narrowly-focussed solutions that attempt to address specific problems with high accuracy.

4 Some Potential Problems

Our proposal is not without possible problems and detrimental side effects.

Clearly there are ethical issues that need to be considered carefully; even if an author is happy for their data to be used in this way, one might find retrospective embarrassment at eponynmous error descriptions entering the common vocabulary in the field—it's one thing to be acknowledged for Kneser-Ney smoothing, but perhaps less appealing to be famous for the Dale-Kilgarriff adjunct error.

Our suggestion that the ACL Anthology might be used as a source for language modelling brings its own downsides: in particular, if anything is likely to increase the oft-complained-about sameness of CL papers, this will! There is also an ethical issue around the fine line between what such systems will do and plagiarism; one might foresee the advent of a new scholastic crime labelled 'machine-assisted style plagiarism'.

There are no doubt other issues we have not yet considered; again, feedback on potential pitfalls is eagerly sought.

5 Next Steps

Our aim is to obtain feedback on this proposal from conference participants and others, with the aim of refining our plan in the coming months. If we sense that there is a reasonable degree of interest in the task, we would aim to publish the initial data set well before the end of the year, with a first evaluation taking place in 2011.

In the name of better writing, CLers of the world unite—you have nothing to lose but your worst sentences!

Acknowledgements

We thank the two anonymous reviewers for useful feedback on this proposal, and Anja Belz for encouraging us to develop the idea.

References

Steven Bird, Robert Dale, Bonnie Dorr, Bryan Gibson, Mark Joseph, Min-Yen Kan, Dongwon Lee, Brett Powley, Dragomir Radev, and Yee Fan Tan. 2008. The acl anthology reference corpus: A reference dataset for bibliographic research in computational linguistics. In *Proceedings of the Language Resources and Evaluation Conference (LREC 2008), location = Marrakesh, Morocco*.

Meri Coleman and T. L. Liau. 1975. A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, 60:283–284.

E Dagneaux, S Denness, S Granger, and F Meunier. 1996. Error tagging manual version 1.1. Technical report, Centre for English Corpus Linguistics, Université Catholique de Louvain.

Edgar Dale and Jeanne S. Chall. 1948. A formula for predicting readability. *Educational research bulletin*, 27:11–20.

Rudolph Flesch. 1948. A new readability yardstick. *Journal of Applied Psychology*, 32:221–233.

George Heidorn. 2000. Intelligent writing assistance. In R Dale, H Moisl, and H Somers, editors, *Handbook of Natural Language Processing*, pages 181–207. Marcel Dekker Inc.

Donald S. Le Vie Jr. 2000. Documentation metrics: What do you really want to measure? *Intercom*.

G. Harry McLaughlin. 1969. SMOG grading – a new readability formula. *Journal of Reading*, pages 639–646.

D Nicholls. 2003. The cambridge learner corpus: error coding and analysis for lexicography and ELT. In *Proceedings of the Corpus Linguistics 2003 Conference (CL 2003)*, page 572.

J R Tetreault and M S Chodorow. 2008. The ups and downs of preposition error detection in ESL writing. In *Proceedings of the 22nd International Conference on Computational Linguistics*.