Statistical Machine Translation for Grammatical Error Correction

Arvind Srinivasan
Columbia University
vs2371@columbia.edu

Louis Cialdella
Columbia University
lmc2179@columbia.edu

Abstract

This document contains the instructions for preparing a camera-ready manuscript for the proceedings of NAACL HLT 2012. The document itself conforms to its own specifications, and is therefore an example of what your manuscript should look like. Authors are asked to conform to all the directions reported in this document. Also included are instructions for submission to the conference, which are similar to the camera-ready instructions but remove author-identifying information.

1 Background

The following instructions are directed to authors of papers accepted for publication in the NAACL HLT 2012 proceedings. All authors are required to adhere to these specifications. Authors are required to provide a Portable Document Format (PDF) version of their papers. The proceedings will be printed on US-Letter paper. Authors from countries in which access to word-processing systems is limited should contact the publication chairs as soon as possible.

2 Task + Corpus

This project was initiated as a submission to the ConLL 2013 Shared Task: Grammatical Error Correction [1]. As a result, all training and test data was from a provided corpus, the National University Singapore Corpus of Learner English, or NUCLE, and trained systems were scored with the provided NUS MaxMatch, or M^2 scorer.

[1] http://www.comp.nus.edu.sg/ nlp/conll13st.html

The corpus itself is comprised of 1400 annotated essays written by english-second-language Singaporean students. The annotations provide an error type and correction, from the below types of errors:

tag	category
Vt	Verb tense
Vm	Verb modal
V0	Missing verb
Vform	Verb form
SVA	Subject-verb-agreement
ArtOrDet	Article or Determiner
Nn	Noun number
Npos	Noun possesive
Pform	Pronoun form
Pref	Pronoun reference
Prep	Preposition
Wci	Wrong collocation/idiom
Wa	Acronyms
Wform	Word form
Wtone	Tone
Srun	Runons, comma splice
Smod	Dangling modifier
Spar	Parallelism
Sfrag	Fragment
Ssub	Subordinate clause
WOinc	Incorrect sentence form
WOadv	Adverb/adjective position
Trans	Link word/phrases
Mec	Punctuation, capitalization, spelling, typos
Rloc	Local redundancy
Cit	Citation
Others	Other errors
Um	Unclear meaning (cannot be corrected)

For the initial approach to this task, however, we

looked to train a system that would identify a set of five errors: SVA, ArtOrDet, Nn, VForm, and Prep. The total corpus amounted to 67372 sentences, of which 15,000 were noisy (references, urls, etc). Of that, 11288 sentences had annotations with those tags, leaving a fairly small dataset on which to train. After the time of this writing, ConLL has released an additional dataset - the blind test data with the gold references, so we anticipate that this dataset will grow significantly.

3 Baseline

Our baseline system used Moses to translate between a test corpus of original essays and the counterpart essays with all the annotations applied. This flattened corpus was split into 58000 training sentences, 2000 sentences for tuning, and 8000 test sentences.

Moses used Giza++ to align the sentences, with default heuristics and reordering models (grow-diag-final, msd-bidirectional-fe). The tuning used MERT with default features.

To score the baseline, we use BLEU as well as the M^2 scorer. The M^2 scorer scores precision, accuracy, and F1 against the annotations, rather than the text itself. Though the conference version of M^2 is case-sensitive, we use a case-insensitive version for simplicity - recasing in English is a trivial task. Though both of these scorers support multi-reference evaluation, the corpus itself only has a single-reference gold reference file. Clearly, there are multiple equally fluent machine-generated correction candidates, even within the phrase table generated by the small training corpus, so we think this is an area that needs to be explored in terms of test corpus augmentation.

In reality, the optimality of a correction would be gauged by fluency and grammatical cohesion of the final generation, so neither Bleu nor M^2 adequately capture the ideal result - an ideal result would likely be modeled by a combination of the two, with multiple references.

3.1 Issues with Baseline

4 Corpus Cleaning + Datasets

The first immediate issue with the baseline phrase table was that references and urls generated a large amount of noise in the correction data. Since we used the wordpunct NLTK tokenization scheme, urls in particular were inconsistently treated as multiple words, creating noisy alignments as well. Thus, the first cleaning step was to eliminate references completely, re-adding them after the correction. We accomplished this with a simple reegex substitution.

Additionally, to experiment on the domain sensitivity of our system, we split the corpus by essay topic. This was particularly necessary because one of the challenges of the shared task was to submit system output for topics both inside and outside the training data. However, since the topics were not available a priori, we used an online version of the Latent Dirichlet Allocation (LDA) algorithm to generate a topic model, and then a K-Means clustering implementation to split the documents by those topics. We were then able to generate a held out dataset with 21572 training sentences (no references) with all but one topic, and 8695 test sentences (no references) with the remaining topic. This experiment showed that the NUCLE essay topics were largely similar in content, and that domain sensitivity is still a concern that ought to be tested by a heldout test set with a topic further removed from the training corpus.

To experiment with the impact of zero-annotation sentences in the training corpus, we also generated datasets without correct sentences. This left a dataset of 11288 sentences, of which we used 10000 to train and 288 to tune.

To run stemming experiments, we used the stemmers bundled with NLTK. We stemmed and lemmatized the two corpora mentioned above with the Lancaster and Snowball stemmers and the WordNet lemmatizer to experiment with various degrees of stemming aggressiveness.

4.1 Layout

Format manuscripts two columns to a page, in the manner these instructions are formatted. The exact dimensions for a page on US-letter paper are:

- Left and right margins: 1in
- Top margin:1in
- Bottom margin: 1in

- Column width: 3.15in
- Column height: 9in
- Gap between columns: 0.2in

Papers should not be submitted on any other paper size. Exceptionally, authors for whom it is *impossible* to format on US-Letter paper, may format for *A4* paper. In this case, they should keep the *top* and *left* margins as given above, use the same column width, height and gap, and modify the bottom and right margins as necessary. Note that the text will no longer be centered.

4.2 The First Page

Center the title, author's name(s) and affiliation(s) across both columns (or, in the case of initial submission, space for the names). Do not use footnotes for affiliations. Do not include the paper ID number assigned during the submission process. Use the two-column format only when you begin the abstract.

Title: Place the title centered at the top of the first page, in a 15 point bold font. (For a complete guide to font sizes and styles, see Table 1.) Long title should be typed on two lines without a blank line intervening. Approximately, put the title at 1in from the top of the page, followed by a blank line, then the author's names(s), and the affiliation on the following line. Do not use only initials for given names (middle initials are allowed). Do not format surnames in all capitals (e.g., "Bangalore," not "BANGALORE"). The affiliation should contain the author's complete address, and if possible an electronic mail address. Leave about 0.75in between the affiliation and the body of the first page.

Abstract: Type the abstract at the beginning of the first column. The width of the abstract text should be smaller than the width of the columns for the text in the body of the paper by about 0.25in on each side. Center the word **Abstract** in a 12 point bold font above the body of the abstract. The abstract should be a concise summary of the general thesis and conclusions of the paper. It should be no longer than 200 words. The abstract text should be in 10 point font.

Text: Begin typing the main body of the text immediately after the abstract, observing the two-column format as shown in the present document. Do not include page numbers.

Indent when starting a new paragraph. For reasons of uniformity, use Adobe's **Times Roman** fonts, with 11 points for text and subsection headings, 12 points for section headings and 15 points for the title. If Times Roman is unavailable, use **Computer Modern Roman** (LaTeX2e's default; see section 3.1 above). Note that the latter is about 10% less dense than Adobe's Times Roman font.

4.3 Sections

Headings: Type and label section and subsection headings in the style shown on the present document. Use numbered sections (Arabic numerals) in order to facilitate cross references. Number subsections with the section number and the subsection number separated by a dot, in Arabic numerals.

Citations: Citations within the text appear in parentheses as (Gusfield, 1997) or, if the author's name appears in the text itself, as Gusfield (1997). Append lowercase letters to the year in cases of ambiguities. Treat double authors as in (Aho and Ullman, 1972), but write as in (Chandra et al., 1981) when more than two authors are involved. Collapse multiple citations as in (Gusfield, 1997; Aho and Ullman, 1972).

References: Gather the full set of references together under the heading **References**; place the section before any Appendices, unless they contain references. Arrange the references alphabetically by first author, rather than by order of occurrence in the text. Provide as complete a citation as possible, using a consistent format, such as the one for *Computational Linguistics* or the one in the *Publication Manual of the American Psychological Association* (American Psychological Association, 1983). Use of full names for authors rather than initials is preferred. A list of abbreviations for common computer science journals can be found in the ACM *Computing Reviews* (Association for Computing Machinery, 1983).

The LaTeX and BibTeX style files provided roughly fit the American Psychological Association format, allowing regular citations, short citations and multiple citations as described above.

Appendices: Appendices, if any, directly follow the text and the references (but see above). Letter

Type of Text	Font Size	Style
paper title	15 pt	bold
author names	12 pt	bold
author affiliation	12 pt	
the word “Abstract”	12 pt	bold
section titles	12 pt	bold
document text	11 pt	
abstract text	10 pt	
captions	10 pt	
bibliography	10 pt	
footnotes	9 pt	

Table 1: Font guide.

them in sequence and provide an informative title: **Appendix A. Title of Appendix**.

Acknowledgment sections should go as a last (unnumbered) section immediately before the references.

4.4 Footnotes

Footnotes: Put footnotes at the bottom of the page. They may be numbered or referred to by asterisks or other symbols.[2] Footnotes should be separated from the text by a line.[3] Footnotes should be in 9 point font.

4.5 Graphics

Illustrations: Place figures, tables, and photographs in the paper near where they are first discussed, rather than at the end, if possible. Wide illustrations may run across both columns and should be placed at the top of a page. Color illustrations are discouraged, unless you have verified that they will be understandable when printed in black ink.

Captions: Provide a caption for every illustration; number each one sequentially in the form: “Figure 1. Caption of the Figure.” “Table 1. Caption of the Table.” Type the captions of the figures and tables below the body, using 10 point text.

5 Length of Submission

The NAACL HLT 2012 main conference accepts submissions of long papers and short papers. The maximum length of a long paper manuscript is eight (8) pages of content and two (2) additional pages of references *only* (appendices count against the eight pages, not the additional two pages). The maximum length of a short paper manuscript is four (4) pages including references. For both long and short papers, all illustrations, references, and appendices must be accommodated within these page limits, observing the formatting instructions given in the present document. Papers that do not conform to the specified length and formatting requirements are subject to be rejected without review.

6 Double-blind review process

As the reviewing will be blind, the paper must not include the authors' names and affiliations. Furthermore, self-references that reveal the author's identity, e.g., ”We previously showed (Smith, 1991) ...” must be avoided. Instead, use citations such as ”Smith previously showed (Smith, 1991) ...” Papers that do not conform to these requirements will be rejected without review. In addition, please do not post your submissions on the web until after the review process is complete.

Acknowledgments

Do not number the acknowledgment section.

References

Alfred V. Aho and Jeffrey D. Ullman. 1972. *The Theory of Parsing, Translation and Compiling*, volume 1. Prentice-Hall, Englewood Cliffs, NJ.

American Psychological Association. 1983. *Publications Manual*. American Psychological Association, Washington, DC.

Association for Computing Machinery. 1983. *Computing Reviews*, 24(11):503–512.

Ashok K. Chandra, Dexter C. Kozen, and Larry J. Stockmeyer. 1981. Alternation. *Journal of the Association for Computing Machinery*, 28(1):114–133.

Dan Gusfield. 1997. *Algorithms on Strings, Trees and Sequences*. Cambridge University Press, Cambridge, UK.

[2]This is how a footnote should appear.

[3]Note the line separating the footnotes from the text.