Automated Writing Assistance: Grammar Checking and Beyond

Robert Dale

September 8, 2011

1 What This Document is About

This document provides bibliographic details for the various materials cited in the summer school course on *Automated Writing Assistance: Grammar Checking and Beyond*, held in Tarragona, Spain as part of the International Summer School in Language and Speech Technologies (SSLST 2011) on 31st August and 1st September 2011. The class material was broken down into five topics; the references here are organized under that same structure.

2 The Nature of the Problem

We set the scene using the **conduit metaphor** as introduced by Reddy (1979), and adopted a definition of technical writing due to Britton (1965). The example of a **stage model** of the writing process was due to Rohman (1965); the more sophisticated **cognitive process model** is from Flower and Hayes (1981). We also examined a taxonomy of revision processes proposed by Faigley and Witte (1981).

We discussed an extensive analysis of student writing errors carried out by Connors and Lunsford (1988), and briefly reviewed a number of **taxonomies of error** (Douglas and Dale, 1991; Becker et al., 2003; Busta et al., 2009).

3 Spell Checking

The now somewhat mundane concerns of early spelling checkers are discussed by Peterson (1980). We discussed a range of approaches to correcting **non-word spelling errors**:

- Angell, Freund, and Willett (1983) on trigram analysis;
- Yannakoudakis and Fawthrop (1983) on error patterns;
- van Berkel and de Smedt (1988) on triphone analysis;
- Kernighan, Church, and Gale (1990) on the noisy channel model;
- Agirre et al. (1998) on the use of context;
- Brill and Moore (2000) on integrating a model of string-to-string edits into the noisy channel model; and
- Toutanova and Moore (2002) on integrating pronunciation modeling into the noisy channel model.

We also discussed Mays et al.'s (1991) approach to using trigrams to detect and correct **real-word errors**, and Hirst and Budanitsky's (2005) attempt to use semantics for the same problem. Finally, we described the work of Whitelaw et al. (2009) on using the web as a source of information for spelling correction.

We also alluded to, but did not discuss, a number of approaches to real-word error correction based on the idea of **confusion sets**: (Golding, 1995; Golding and Schabes, 1996; Mangu and Brill, 1997).

4 Grammar Checking

The range of different grammatical errors we discussed at the outset were drawn from (Douglas and Dale, 1991). The Unix Writer's Workbench is described in (Macdonald, 1983); Atwell's approach to Constituent-Likelihood Error Detection is described in (Atwell, 1987).

The description of EPISTLE/CRITIQUE was drawn from (Heidorn et al., 1982; Jensen et al., 1983). A number of relevant papers are collected together in (Jensen, Heidorn, and Richardson, 1993). The Microsoft Word grammar checker is described in some detail by Heidorn (2000).

In terms of specific techniques for grammar checking, an early example of **relaxation** is presented by Weischedel and Black (1980). Douglas and Dale (1992) provide a description of relaxation in the PATR grammatical framework; Schneider and McCoy (1998) provides a description of the **mal-rules** approach.

We also mentioned the following other techniques:

- fitted parsing (Jensen et al., 1983);
- mixed bottom-up and top-down parsing (Mellish, 1989); and
- minimum edit distance parsing (Lee et al., 1995).

Kohut and Gorman (1995) provide an evaluation of five commercial grammar-checking packages that were available in the mid-1990s.

5 Handling ESL Errors

Bolt (1992) tested seven grammar-checking programs of the time against 35 sentences containing ESL errors. Donahue (2001) provides an analysis of ESL errors that is contrasted with the findings of Connors and Lunsford (1988) for native speakers. The counts of errors in the Cambridge Learners Corpus were taken from Leacock et al. (2010), as were a number of other tabulations of data used in this lecture. The 'Helping Our Own' (HOO) task is described in (Dale and Kilgarriff, 2010).

The three approaches to article errors we discussed are (Knight and Chander, 1994; Han, Chodorow, and Leacock, 2006; De Felice and Pulman, 2008). Pelletier's universal grinder and universal packager are introduced in (Pelletier, 1975).

The papers cited in the tabular summary of work on preposition errors are as follows:

- Papers which address preposition selection in well-formed text: (Lee and Seneff, 2008; Chodorow, Tetreault, and Han, 2007; De Felice and Pulman, 2007; De Felice and Pulman, 2008; Tetreault and Chodorow, 2008a; Gamon et al., 2008; Bergsma, Lin, and Goebel, 2009);
- Papers which address preposition error detection on learner data: (Eeg-Olofsson and Knuttson, 2003; Tetreault and Chodorow, 2008b; De Felice and Pulman, 2009; Hermet, Dsilets, and Szpakowicz, 2008; Tetreault and Chodorow, 2009; Gamon, 2010; Han et al., 2010).

We also briefly mentioned Lee and Seneff's (2008) work on verb form errors. Other supporting tools that were cited in passing were Collin's parser (Collins, 1999) and Lin's work on automatic thesaurus construction (Lin, 1999).

6 Beyond the Sentence

We revisited Flower and Hayes' (1981) cognitive process model of writing and Faigley and Witte's (1981) taxonomy of revision operations to motivate looking at other aspects of the writing process where assistance might be provided.

Our discussion of architectures for natural language generation was drawn from (Reiter and Dale, 2000); the primary reference on Rhetorical Structure Theory is (Mann and Thompson, 1988).

References

Agirre, Eneko, Koldo Gojenola, Kepa Sarasola, and Atro Voutilainen. 1998. Towards a single proposal in spelling correction. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and the 17th International Conference on Computational Linguistics*, pages 22–28, Montreal, Canada.

Angell, R. C., G. E. Freund, and P. Willett. 1983. Automatic spelling correction using a trigram similarity measure. *Information Processing and Management*, 19:255–261.

Atwell, Eric. 1987. How to detect grammatical errors in a text without parsing it. In *Proceedings of the Third Conference of the European Association for Computational Linguistics*, pages 38–45, Copenhagen, Denmark.

Becker, M., A. Bredenkamp, B. Crysmann, and J. Klein. 2003. Annotation of error types for a german newsgroup corpus. In A. Abeille, editor, *Treebanks: Building and Using Parsed Corpora*. Kluwer, Dordrecht, chapter 6, pages 89–100.

Bergsma, Shane, Dekang Lin, and Randy Goebel. 2009. Web-scale n-gram models for lexical disambiguation. In *Proceedings of the 21st International Joint Conference on Artifical Intelligence*, pages 1507–1512, Pasadena.

Bolt, Philip. 1992. An evaluation of grammar-checking programs as self-help learning aids for learners of english as a foreign language. *Computer Assisted Language Learning*, 5(1):49–91.

Brill, Eric and Robert C. Moore. 2000. An improved error model for noisy channel spelling correction. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, pages 286–293, Hong Kong.

Britton, W. Earl. 1965. What is technical writing? *College Composition and Communication*, 16(2):113–116, May.

Busta, Jan, Dana Hlavackova, Milos Jakubicek, and Karel Pala. 2009. Classification of errors in text. In Petr Sojka and Ales Horak, editors, *Proceedings of Recent Advances in Slavonic Natural Language Processing*, pages 109–119, Masaryk University, Brno.

Chodorow, Martin, Joel Tetreault, and Na-Rae Han. 2007. Detection of grammatical errors involving prepositions. In *Proceedings of the Fourth ACL-SIGSEM Workshop on Prepositions*, pages 25–30, Prague, Czech Republic.

Collins, Michael J. 1999. *Head-driven Statistical Models for Natural Language Parsing*. PhD Thesis, University of Pennsylvania, Philadelphia, Pennsylvania.

Connors, Robert J. and Andrea A. Lunsford. 1988. Frequency of formal errors in current college writing, or ma and pa kettle do research. *College Composition and Communication*, 39(4):395–409.

Dale, Robert and Adam Kilgarriff. 2010. Helping our own: Text massaging for computational linguistics as a new shared task. In *Proceedings of the 6th International Natural Language Generation Conference*, pages 261–265, Dublin, Ireland, 7th–9th July 2010.

De Felice, Rachele and Stephen G. Pulman. 2007. Automatically acquiring models of preposition use. In *Proceedings of the Fourth ACL-SIGSEM Workshop on Prepositions*, pages 45–50, Prague, Czech Republic.

De Felice, Rachele and Stephen G. Pulman. 2008. A classifier-based approach to preposition and determiner error correction in L2 english. In *Proceedings of the 22nd International Conference on Computational Linguistics*, pages 169–176, Manchester, UK.

De Felice, Rachele and Stephen G. Pulman. 2009. Automatic detection of preposition errors in learner writing. *CALICO Journal*, 26(3):512–528.

Donahue, Steven. 2001. Formal errors: Mainstream and ESL students. Presented at the 2001 Conference of the Two-Year College Association (TYCA); cited by Leacock et al. 2010.

Douglas, Shona and Robert Dale. 1991. Towards a taxonomy of errors in technical texts. Technical report, Human Communication Research Centre, University of Edinburgh.

Douglas, Shona and Robert Dale. 1992. Towards robust PATR. In *Proceedings of the 15th International Conference on Computational Linguistics*, pages 468–474, Nantes, France.

Eeg-Olofsson, Jens and Ola Knuttson. 2003. Automatic grammar checking for second language learners: the use of prepositions. In *Proceedings of the 14th Nordic Conference in Computational Linguistics*, Reykjavik, Iceland.

Faigley, Lester and Stephen Witte. 1981. Analyzing revision. *College Composition and Communication*, 32(4):400–414, Dec.

Flower, Linda and John R. Hayes. 1981. A cognitive process theory of writing. *College Composition and Communication*, 32(4):365–387.

Gamon, Michael. 2010. Using mostly native data to correct errors in learners writing. In *Proceedings of the Eleventh Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 163–171, Los Angeles.

Gamon, Michael, Jianfeng Gao, Chris Brockett, Alex Klementiev, William B. Dolan, Dmitriy Belenko, and Lucy Vanderwende. 2008. Using contextual speller techniques and language modeling for ESL error correction. In *Proceedings of the International Joint Conference on Natural Language Processing*, pages 449–456, Hyderabad, India.

Golding, Andrew. 1995. A bayesian hybrid method for context sensitive spelling correction. In *Proceedings of the Third Workshop on Very Large Corpora (WVLC-3)*, page 3953.

Golding, Andrew R. and Yves Schabes. 1996. Combining trigram-based and feature-based methods for context-sensitive spelling correction. In *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics*, pages 71–78, Santa Cruz, CA.

Han, Na-Rae, Martin Chodorow, and Claudia Leacock. 2006. Detecting errors in English article usage by non-native speakers. *Natural Language Engineering*, 12(2):115–129.

Han, Na-Rae, Joel Tetreault, Soo-Hwa Lee, and Jin-Young Ha. 2010. Using error-annotated ESL data to develop an ESL error correction system. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation*, Malta.

Heidorn, George. 2000. Intelligent writing assistance. In Robert Dale, Herman Moisl, and Harold Somers, editors, *Handbook of Natural Language Processing: Techniques and Applications for the Processing of Language as Text*. Marcel Dekker, pages 181–207.

Heidorn, George E., Karen Jensen, Lance A. Miller, Roy J. Byrd, and Martin Chodorow. 1982. The EPISTLE text-critiquing system. *IBM Systems Journal*, 21:305–326.

Hermet, Matthieu, Alain Dsilets, and Stan Szpakowicz. 2008. Using the web as a linguistic resource to automatically correct lexico-syntactic errors. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation*, pages 390–396, Marrekech, Morocco.

Hirst, Graeme and Alexander Budanitsky. 2005. Correcting real-word spelling errors by restoring lexical cohesion. *Natural Language Engineering*, 11(1):87–111, March.

Jensen, K., G. Heidorn, and S. Richardson, editors. 1993. *Natural Language Processing: The PNLP Approach*. Kluwer Academic Publishers.

Jensen, Karen, George E. Heidorn, Lance A. Miller, and Yael Ravin. 1983. Parse fitting and prose fixing: Getting a hold on ill-formedness. *American Journal of Computational Linguistics*, 9(34):147–160.

Kernighan, M. D., K.W. Church, and W.A. Gale. 1990. A spelling correction program based on a noisy channel model. In *Proceedings of the Thirteenth International Conference on Computational Linguistics*, pages 205–210.

Knight, Kevin and Ishwar Chander. 1994. Automated postediting of documents. In *Proceedings of the Twelfth National Conference on Artificial Intelligence*, pages 779–784, Seattle, WA.

Kohut, Gary F. and Kevin J. Gorman. 1995. The effectiveness of leading grammar/style software packages in analyzing business students' writing. *Journal of Business and Technical Communication*, 9:341–361.

Leacock, C., M. Chodorow, M. Gamon, and J. Tetreault. 2010. *Automated Grammatical Error Detection for Language Learners*. Synthesis Lectures on Human Language Technologies. Morgan and Claypool.

Lee, John and Stephanie Seneff. 2008. Correcting misuse of verb forms. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics:Human LanguageTechnology*, pages 174–182, Columbus, OH.

Lee, Kong Joo, Cheol Jung Kweon, Jungyun Seo, and Gil Chang Kim. 1995. A robust parser based on syntactic information. In *Proceedings of the Seventh Conference of the European Association for Computational Linguistics*, pages 223–228.

Lin, Dekang. 1999. Automatic identification of non-compositional phrases. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 317–324.

Macdonald, N. H. 1983. The UNIX writer's workbench software: Rationale and design. *Bell System Technical Journal*, 62:1891–1908.

Mangu, Lidia and Eric Brill. 1997. Automatic rule acquisition for spelling correction. In *Proceedings of the 14th International Conference on Machine Learning*, pages 734–741, Nashville, Tennessee.

Mann, William C. and Sandra A. Thompson. 1988. Rhetorical Structure Theory: Toward a functional theory of text organization. *Text*, 8(3):243–281.

Mays, Eric, Fred J. Damerau, and Robert L. Mercer. 1991. Context based spelling correction. *Information Processing and Management*, 27(5):517–522.

Mellish, Chris S. 1989. Some chart-based techniques for parsing ill-formed input. In *Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 102–109.

Pelletier, F. Jeffry. 1975. Non-singular reference: some preliminaries. *Philosophia*, 5:451–465.

Peterson, James L. 1980. Computer programs for detecting and correcting spelling errors. *Communications of the ACM*, 23(12):676–687, December.

Reddy, M. J. 1979. The conduit metaphor – a case of frame conflict in our language about language. In A. Ortony, editor, *Metaphor and Thought*. Cambridge University Press, pages 284–297.

Reiter, Ehud and Robert Dale. 2000. *Building Natural Language Generation Systems*. Cambridge University Press.

Rohman, Gordon. 1965. Pre-writing: the stage of discovery in the writing process. *College Composition and Communication*, 16(2):106–112.

Schneider, David and Kathleen McCoy. 1998. Recognizing syntactic errors in the writing of second language learners. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*, pages 1198–1204, Montreal, Canada.

Tetreault, Joel and Martin Chodorow. 2008a. Native judgments of non-native usage: Experiments in preposition error detection. In *Proceedings of the Workshop on Human Judgments in Computational Linguistics at the 22nd International Conference on Computational Linguistics (COLING)* ., pages 24–32.

Tetreault, Joel and Martin Chodorow. 2008b. The ups and downs of preposition error detection in ESL writing. In *Proceedings of the 22nd International Conference on Computational Linguistics*, pages 865–872, Manchester, UK.

Tetreault, Joel and Martin Chodorow. 2009. Examining the use of region web counts for ESL error detection. In *Proceedings of the Web as CorpusWorkshop (WAC-5)*, San Sebastian, Spain.

Toutanova, Kristina and Robert Moore. 2002. Pronunciation modeling for improved spelling correction. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*, pages 144–151, Philadelphia, PA.

van Berkel, Brigitte and Koenraad de Smedt. 1988. Triphone analysis: a combined method for the correction of orthographical and typographical errors. In *Proceedings of the Second Conference on Applied Natural Language Processing*, pages 77–83. Association for Computational Linguistics.

Weischedel, R. M. and J. Black. 1980. Responding-to potentially unparseable sentences. *American Journal of Computational Linguistics*, 6:97–109.

Whitelaw, Casey, Ben Hutchinson, Grace Y Chung, and Ged Ellis. 2009. Using the Web for language independent spellchecking and autocorrection. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 890–899, Singapore.

Yannakoudakis, E. J and D. Fawthrop. 1983. The rules of spelling errors. *Information Processing and Management*, 19(2):87–99.