Models for Computer Generated Parody

Abstract

This paper outlines two approaches to the construction of computer systems that generate prose in the style of a given author. The first involves using intuitive notions of stylistic trademarks to construct a grammar that characterizes a particular author?in this case, Ernest Hemingway. The second uses statistical methods for inferring a grammar from samples of an author's work?in this instance, Thomas Hardy.

The relative merits of each approach are discussed, and text generated from the resulting grammars is assessed in terms of its parodic quality. Further to its esoteric interest, a discussion of parody generation as a useful technique for measuring the success of grammatical inferencing systems is included, along with suggestions for its practical application in areas of language modeling and text compression.

1 Introduction

Computer programs that generate prose and poetry may represent the last bastions of Artificial Intelligence research still seeking to pass the notorious Turing Test. That is, these programs attempt to generate text with surface forms that do not reveal their computational source.

The objective of parody generating systems need not be expressed quite so strongly. They seek simply to generate surface forms reminiscent of the style of a particular author. As a consequence, the success of a parody generating system tends to be measured in terms of the level of mimicry achieved in its output.

A spectrum of design methodologies has emerged implicitly from research in parody generation. At one end of this spectrum are programs with underlying models designed to offer psychologically plausible accounts of an author's effective grammar. At the other end are those that employ any means possible to generate surface forms of sufficient parodic quality.

This paper outlines two computer programs that generate parody?one from either end of the design spectrum. The first is based on a manually constructed grammar built from an intuitive assessment of Ernest Hemingway's stylistic trademarks. The second infers a grammar for Thomas Hardy through statistical analyses of a large sample text?in this case Far From The Madding Crowd.

Though the two approaches adopt highly disparate techniques to arrive at a final grammar, many of the intermediate objectives and internal components of each system are the same. For example, each must

1. derive an appropriate vocabulary,

2. construct a suitable set of phrase production rules, and

3. develop probability distributions for the selection of vocabulary items and phrase constructs in the generation stage.

This paper describes how these components are achieved within the constraints of the two methodologies. For continuity, we describe each system in its entirety before attempting to evaluate their respective merits. The first system relies on basic intuitions about authorship style, whereas the second is based on principled pattern inference. Some background material on grammar induction is included before discussion of the second system.

Finally, the quality of parody demonstrated in the resulting generated texts is analyzed, along with some suggestions for improvement. Some practical applications for parody generation in other areas of computer science are touched upon.

model

Few authors have been parodied to the same extent as Ernest Hemingway. His pervasive run-on sentence, unlikely juxtaposition, and imaginative exposition of the insipid have made him a ready target for mimicry.

In 1978, an advertising man named Paul Keye noted Hemingway's attachment to Harry's Bar in Venice and recognized the potential promotional value of running an Imitation Hemingway Competition for the Century City Harry's Bar & American Grill in California. Of the thousands of submissions received by the competition every year, forty-five were published in the 1989 publication ?The Best of Bad

Hemingway? [Pli89]. Even a modest familiarity with Hemingway's work allows one to appreciate within these parodies the notable characteristics of his infamous style.

Prompted by the entries to this competition, an attempt was made to characterize Hemingway's style for the purpose of generating parodies by a computer program. Much of the task of ascertaining Hemingwayesque attributes had already been accomplished by the authors of the entries to the Imitation Hemingway Competition. A model of the author's style was gleaned from a quick survey of characteristic similarities present in these so-called ?bad Hemingway? parodies.

2.1 The vocabulary

Barring the most frequent word sequence ?Harry's Bar & American Grill? present in every entry to the Imitation Hemingway contest as part of the submission requirements, most of the commonly adopted words in the parodies seemed to play on Hemingway's distinctive topics. Specific proper nouns like Kilimanjaro and Madrid were frequently included, perhaps as a reflection of Hemingway's cosmopolitan vernacular. Other nouns, like fisherman, cafe, matador and aficionado also appeared with notable frequency. The pervasive, often sublime, pathos of

Hemingway may explain the repeated use of beggar, martini and death, along with the conspicuous presence of verbs that express argument, deception or conspiracy. The third-person, omniscient narrative used by Hemingway offers a possible explanation for the frequent use of introspective verbs like thought, believed, knew, and assumed.

These, and similar terms, were gleaned from the sample parodies to construct a final vocabulary that consists of 96 words, which can be roughly tagged to 31 nouns, 19 verbs, 14 prepositions, 10 adjectives, 8 determiners, 4 adverbs, 4 auxiliary verbs, 3 pronouns, and 3 conjunctions.

2.2 The production rules

Much of the vocabulary arose incidentally from the selection of certain key phrases found within the Hemingway parodies. A few explicit expressions, like the snows of Kilimanjaroand in a well lighted room, were lifted from the sample texts and used as the basis for assembling the entire grammar. Subsequent additions to the vocabulary arose from efforts to tie these isolated phrases together into wellformed expressions. Complete noun phrases were constructed to express caricature persona, and general formulations were derived for restrictive and relative clauses to support complex sentence structures.

The grammar, shown in its entirety in Figure 1, was thus constructed piecemeal from instantiated Generalized Phrase Structures [GKPS85, Jac84]. That is, prepositional phrases are expressed in fixed surface forms, as are substructures of noun and verb phrase constructs. However, the form and content of every generatable sentence is determined by random instantiation of a general propositionalform. Each sentence begins with a subject phrase (SubP) under an optional restrictive clause (ResC). The subject is followed by a past participle auxiliary connective and verb phrase (AuxC), or by a past tense verb and relative clause (RelC), or by some compound phrase combination of AuxC and RelC (CompP). A ResC is comprised of the word who and either RelC or a short form of AuxC (SAuxC).

The number preceding each token in the production rules indicates the probability with which that construct is to be selected for generation purposes. The symbol + indicates string concatenation while || indicates disjunction. Braces are used to clarify precedence relations.

2.3 The parody

The grammar uses a pseudo-random number generator to produce a piece of text between 4 and 13 sentences in length. The limitations placed on text length were made arbitrarily and do not reflect a limitation of the grammar. Random numbers are further used to select phrase substructures in accordance with their associated probability distribution. These probabilities were arrived at primarily through trial and error with the objective of giving a sufficient illusion of free form composition in the resulting text.

Figure 2 shows a sample text generated from the grammar in Figure 1. Some capitalization and minor formatting has been added for readability.

Without any semantic component to guide the generation process the output generally lacks any sort of consistent focus of attention. Even so, the wellformed expressions and limited selection for subject and object allow the reader to endow the text with a marginal amount of coherence.

3 Hardy?the descriptive model

O'Donnell [O'D70] and others [Bri63, Ell62] have often used statistical methods in efforts to identify authors of unknown texts, or to challenge traditional authorship claims. Vocabulary size, sentence length, distribution of sentence complexity, and other ?stylistic variables? are adopted as a metric to test, say, whether the works attributed to Thackeray and Dickens were penned by the same author or, similarly, whether Shakespeare and Marlowe are one and the same.

These techniques can be adapted to construct a probability based grammar for a particular author. Vocabulary and phrase structures are gathered from a statistical analysis on a large number of sample expressions. This ?authentic? grammar and associated probability distribution can, at least in principle, be used to generate parody text.

This section describes an attempt at such a system for Thomas Hardy using Far From The Madding Crowd as the source of sample expressions. A small amount of background material on grammar induction is useful before the details of the approach are presented.

3.1 Grammatical inference

The construction of a grammar as a generalization of regularities present in sample expressions is known as grammatical inference, and has been an important ongoing area of research in computer science [BP87, CL82, PCF77, PC78]. Though natural languages allow many levels of abstraction (e.g. syntax, morphology, semantics, phonetics, etc.), grammatical inference by automata focuses on the discovery of an acceptable description for the syntax of a language based on a finite set of sample strings constructed from a finite vocabulary [Fel70, Ang80]. Despite their inherent limitations, it is generally accepted that context-free grammars are adequate enough models for generalizations or hypotheses about the syntax of natural languages [Mac82].

Syntactic wellformedness is (minimally) the property of a linear arrangement of lexical categories (like nouns and verbs) instantiated in conformance with established linguistic features (like tense and number agreement) such that the expression is deemed grammatical by speakers of the language in question [OD87]. More plainly, given the sentence ?The cowboy saddled his horse?, the words cowboy and horse may exchange places without affecting the wellformedness of the sentence. This is true because the words cowboy and horse are of the same grammatical category (i.e. they are both nouns) and the underlying structure of the sentence is expressed (at least) in terms of such categories. From this we conclude that it is inherently more useful (in English) to study occurrences of the sequence ?determiner-adjectivenoun? than it is to study occurrences of the sequence ?the-spotted-dog?.

We refine our notion of the goal of the induction process to be the discovery of an acceptable contextfree grammar based on a finite set of sample strings constructed from a finite vocabulary of lexical categories.

3.2 The vocabulary

Cultivating a vocabulary and associated lexical frequencies from any machine readable text is an essentially trivial task. However, pattern analysis on plain text is generally unfruitful andgrammars derived from such tend to degenerate into an unprincipled description of random word sequences. More useful patterns can be brought to the surface by assigning each word of the sample text to an appropriate grammatical category (i.e. tagging) and thereafter forming generalizations about sequences of category symbols.

Many techniques exist for tagging lexical items [Kup89, Mer91, GR71]. The approach adopted here is the closed-class vocabulary method described by Smith and Witten [SW93], which infers lexical categories for all novel words as proximity relations to a small set of functional elements. The adoption of inferential tagging moves the text generation system closer to its limit as an induction based design.

When this lexical inference process is applied to Thomas Hardy's novel Far From The Madding Crowd?a work of 140,632 words in length?it tags each of the 11,589 distinct vocabulary items to one of 62 categories.

3.3 The production rules

Once lexical categories have been established, the sentences of the sample text are rewritten as corresponding strings of category symbols. These sequences are generalized using Smith's [Smi93] lattice overlay technique. Each unique sequence of symbols is a partial ordering on the set of category symbols and can be expressed as a nonbranching nondirected graph called a chain. Chains with matching initial and terminal nodes are overlaid to produce more complex lattices which are subsequently rewritten as production rules such that each sublattice is expressed as a disjunction on the righthand side.

When this induction process is applied to the 7282 unique sentences in Far From The Madding Crowd, which have an average length of 19.31 words, it yields a final grammar of 6288 production rules with an average length of 6.97 nonterminal symbols. Despite a reduction of nearly 70%, the final grammar is too large to reproduce in this paper.

3.4 The parody

Precise counts for each lexical category and each production string are garnered during inferencing and

used as the probability distribution in the generation process.

Figure 3 shows a segment of text generated from the grammar inferred from Far From The Madding Crowd. Once again, some capitalization and minimal formatting has been added to assist the reader.

As with the Hemingway parody, it is still possible to endow the generated text with a certain amount of semantic coherency, though admittedly this requires a bit more effort. Whether or not any Hardyesque style is discernible is left for the reader to judge.

4 Analysis

The texts produced from both modeling paradigms exhibit two general forms of ?complexity failure??a breakdown in sentence coherence, and a disintegration of discourse topic.

4.1 Sentence degeneration

Failures arising from increased sentence complexity are common for many types of natural language processor. Parsers, for example, that experience little difficulty performing hierarchical decomposition on short, simple sentences tend to see their rejection rates grow proportionately with an increase in input sentence lengths.

The Hemingway model attempts to preserve sentence coherency by placing strong restrictions on both its sentence forms and the vocabulary. The model is constructed from thematically uniform nominal and prepositional phrase segments [Spe90] that can only be used in fixed juxtaposition to the main predicate.

All main predicates are primitive equivalents [Sch80], and restrictive clauses can only be assigned to an agentive noun phrase.

Sentence coherence fails more conspicuously in the Hardy parody than it does in the Hemingway output. This reflects a propagation of generalization side effects from the inductive processes. That is, as aspects of the grammar are weakened to accommodate a greater variety of sample expressions and a larger vocabulary, these weaknesses amplify one another when the language stream is reversed for generation. Though the increased vocabulary and relaxed syntactic constraints allow for a more flamboyant, free form output, it entails a corresponding decrease in the epiphenomenal coherence.

4.2 Discourse incoherency

Generation from the two grammars is probability driven and thus is effectively semantics free. The intention of each system was to examine static stylistic attributes and thus the output should not be judged too harshly for lacking continuity. Even so, it is interesting to explore possible methods for controlling the semantic thread of generated text.

Inasmuch as the Hemingway parody is more semantically cohesive, it achieves this presumably because of the way its limited vocabulary imposes restrictions on possible meaning projections for the reader. Other potential ?focus-of-attention? mechanisms that could be added might include fixing the subject, or object, for a particular number of sentences, or imposinga constraint that the direct object of one sentence must be the subject of the following sentence.

Many other more semantically oriented techniques have been explored [SA77, Bru75, McK85]. Smith and Witten [SW90, SW91] outline a method for planning coherent story plots by likening storytelling to game playing. Characters are given objectives and strategies to achieve them within a story environment. Simple sentence text is generated as a report on character activities. Adapting such a planning mechanism to incorporate more elaborate sentence forms, like those from the Hardy grammar, might yield more engaging parody.

5 Conclusions

The July 31st, 1993 issue of The [Christchurch] Press carried an article (reprinted from The Los Angeles Times) about a new Jaqueline Susann novel posthumously produced (at least in part) by a computer program. The news-worthiness of the article reflects the esoteric interest computer generated text has for the general public.

Inasmuch as parody based grammars capture stylistic attributes of authors, similar techniques could be employed to construct music grammars from which novel melodies might be generated in the style of, say, Haydn or Charlie Parker.

Parody generation also has useful applications in a variety of computer science research areas. The veracity of inferred and theoretically compiled grammars can to some degree be evaluated by the quality of parody they are able to generate. And grammatical models can be adapted to lossy text compression techniques by reducing text transmissions to a grammar and pseudo-random number generator seed [WBM+92].

References

[Ang80] D. Angluin. Inductive inference of formal languages from positive data. Information
Control, 45:117?135, 1980.

[BP87] R. C. Berwick and S. Pilato. Learning syntax by automata induction. Machine
Learning, 2(1):9?38, 1987.

[Bri63] Claude S. Brinegar. Mark Twain and the Quintus Curtius Snodgrass letters: A
statistical test of authorship. American
Statistical Association Journal, March
1963.

[Bru75] Bertram C. Bruce. Generation as a social action. In Theoretical Issues in Natural
Language Processing-1, pages 64?67.
Association for Computational Linguistics,

1975. Urbana-Champaign.

[CL82] J. Case and C. Lynes. Inductive inference and language identification. In
Proceedings of the International Colloquium
on Algorithms, Languages, and

Programming (ICALP) 82, pages 107?
115, New York, June 1982. SpringerVerlag.
Barcelona, Spain.

[Ell62] Alvar A. Ellegard. A Statistical Method for Determining Authorship. Goteborg,
Holland, 1962.

[Fel70] J. A. Feldman. Some decidability results on grammatical inference and complexity.
AI Memo 93.1, Computer Science
Dept., Stanford University, Stanford,
California, 1970.

[GKPS85] Gerald Gazdar, Ewan Klein, Geoffrey Pullam, and Ivan Sag. Generalized
Phrase Structure Grammar. Basil
Blackwell, Oxford, UK, 1985.

[GR71] B. B. Greene and G. M. Rubin. Automatic grammatical tagging of English.
Technical report, Brown University,
Providence, Rhode Island, 1971.

[Jac84] Pauline Jacobson. Connectivity in generalized phrase structure grammar. Natural
Language and Linguistic Theory,
1:535?81, 1984.

[Kup89] J. M. Kupiec. Augmenting a Hidden Markov Model for phrase-dependent
word tagging. In Proceedings of
the 1989 DARPA Speech and Natural
Language Workshop, pages 92?98,
Philadelphia, 1989. Morgan Kaufmann.

[Mac82] B. MacWhinney. Basic processes in syntactic acquisition. In S.

A. Kuczaj,
editor, Language Development: Vol. 1,
Syntax and Semantics. Lawrence Erlbaum,
Hillsdale, New Jersey, 1982.

[McK85] K. McKeown. Discourse strategies for generating natural
language-text. Artificial
Intelligence, 27:1?42, 1985.

[Mer91] B. Merialdo. Tagging text with a probabilistic model. In
Proceedings of
ICASSP-91, pages 809?812, Toronto,
Canada, 1991.

[O'D70] Bernard O'Donnell. An Analysis of Prose Style to Determine
Authorship.
Mouton & Company, The Netherlands,
1970.

[OD87] William O'Grady and Michael Dobrovolsky, editors. Contemporary
Linguistic
Analysis. Copp Clark Pittman Ltd.,
Toronto, 1987.

[PC78] T. W. Pao and J. W. Carr. A solution of the syntactical
induction-inference
problem for regular languages. Computer
Languages, 3:53?64, 1978.

[PCF77] R. C. Parkison, K. M. Colby, and W. S. Faught. Conversational
language
comprehension using integrated patternmatching
and parsing. Artificial Intelligence,
9:111?134, 1977.

[Pli89] George Plimpton, editor. The Best of Bad Hemingway. Harcourt
Brace Jovanovich
Publishers, New York, 1989.
Choice entries from the Harry's Bar &
American Grill Imitation Hemingway
Competition.

[SA77] R. Schank and R. Abelson. Scripts, Plans, Goals and
Understanding.

Lawrence Erlbaum Assoc., Hillsdale,
NJ, 1977.

[Sch80] R. C. Schank. Language and memory. Cognitive Science,
4(3):243?284, 1980.
Ablex Publishing.

[Smi93] Tony C. Smith. Language inference from a closed-class vocabulary. Master's
thesis, University of Calgary,
Canada, March 1993.

[Spe90] Margaret Speas. Phrase structure in natural language. Kluwer Academic
Publishers, Dordrecht, The Netherlands,
1990.

[SW90] Tony C. Smith and Ian H. Witten. A planning mechanism for generating
story text. In H. Schanze, editor, ALLC-ACH 90 Proceedings, pages 201?204,
Siegen, Germany, June 1990. University
of Siegen.

[SW91] Tony C. Smith and Ian H. Witten. A planning mechanism for text generation.
Literary & Linguistic Computing,
6(2):119?126, 1991.

[SW93] Tony C. Smith and Ian H. Witten. Language inference from function words.
Working Paper Series 1170-487X-1993-3, Department of Computer Science,
University of Waikato, Hamilton, New
Zealand, August 1993.

[WBM+92] I. H. Witten, T. C. Bell, A. Moffat, C. G. Nevill-Manning, and T. C. Smith. Semantic
and generative models for lossy
text compression. Working Paper Series
1992/8, Department of Computer Science,
University of Waikato, Hamilton,
New Zealand, August 1992. submitted
to The Computer Journal.