Modeling Literary Style for Semi-Automatic Generation of Poetry

Pablo Gervás

Departamento de Sistemas Informáticos y Programación, Universidad Complutense de Madrid, Ciudad Universitaria, 28040 Madrid, Spain. Tel.: (34) 91 3944293. Fax: (34) 91 3944602
pgervas@sip.ucm.es

Abstract. The generation of formal poetry involves both complex creativity - usually exercised by a human poet - and strict algorithmic restrictions regarding the metrical structure of the poem - determined by literary tradition. Starting from a generating system that enforces automatically the metrical restrictions, this paper presents a model for the literary style of a user based on four key features for user preferences - word selection, language structures, poem planning, and restrictions on realisation - governing the generation of poetry from input data provided by the user - a prose paraphrase of the intended message, a task specific vocabulary, and a corpus of construction patterns. The system exploits the CBR paradigm as a means to evolve a case base (a vocabulary / construction pattern grouping) that effectively models the style of a specific user as a result of multiple iterations through the CBR cycle.

Keywords: natural language generation, human-computer collaboration, task modeling

1 Introduction

Existing systems for the automatic generation of text have shown reasonable results in restricted domains [4,7]. The composition of poetry ranks among the most challenging problems of language generation, and is therefore a good test-bed for techniques designed to improve the quality of generated texts. There are currently at least two research efforts devoted to it, one in English [5,6] - dependent on having adequately rich lexicon, syntax and semantics for the language involved - and one in Spanish [2,3] - based on engineering solutions that achieve equivalent results without attempting to model the complexity of the human language.

These systems can automatically generate texts that conform to the rules of poetry, and they need to develop some way of modeling literary style, so that an automatic generation system can provide not only text that matches generic metrical rules for poetry, but that also presents a specified literary style.

To achieve this, an adequate set of parameters to characterise the elusive concept of literary style must be identified, and the process of poem generation must be adapted to take them into account.

2 Modeling Literary Style in Natural Language Generation

Statistic analysis of texts by computer is carried out in cases of disputed authorship [8,9], to identify the - unknown - author of a specific text by comparing it with texts known to have been written by specific authors. Frequency distribution of words of different lengths, sentence length, and combination of various mathematical analyses with word content analysis are generally regarded as having considerable validity in identifying differences between authors. This suggests such parameters may be acceptable as a trustworthy print of a specific style.

The process of generation of a natural language text can be divided in three different stages: gathering of the required data - identifying a specific message to be conveyed, the actual words that are going to be used, and the language structures that are going to be employed -, planning of the intended message over the text - splitting the message among the sentences in the text -, and realisation of the expression - building the final text from the data according to the planning.

A user model for the literary style of a particular poet, to be used in natural language generation, ought to include information about: *word selection*, *language structures*, *poem planning*, and *restrictions on the realisation* of the poem (including metric structure).

3 A CBR Approach to Modeling Literary Style

ASPERA [3] is a prose-to-poetry semiautomatic translator. The four aspects of a user model of literary style described in the previous section are covered by three kinds of input data provided to ASPERA by the user. The user is asked to provide a *prose paraphrase* of his intended message. The user must also provide a *task specific vocabulary*, a set of words to be used by the system. The system is provided with a *corpus of construction patterns* obtained from already validated verses (case-base). These last two kinds of data constitute a first approximation to the model of the literary style desired by the user. This is extended with explicit preferences represented in a user profile built beforehand by the user.

A Case Based Reasoning (CBR) approach [1] is applied to the input data together with the user profile to generate new verses. During a typical execution cycle, the ASPERA system performs the following sequence of operations: selects words and patterns useful for the poem from the task-specific vocabulary and corpus of verse patterns (CBR Retrieve step); generates each of the verses of the poem draft by mirroring the POS structure of the pattern of the retrieved verse pattern combined with the words in the selected vocabulary (CBR Reuse step); presents the draft to be validated or corrected by the user (CBR Revise step); and adds the corresponding information to its case-base for later use (CBR Retain step).

Word selection preferences are encoded in the user profile as a priority assignment to the three kinds of input data as preferred sources of vocabulary. Words to be added to the poem draft are initially looked for only among words with the highest priority, the search extending to words of lower priority only if none had been found earlier.

Language structure preferences are represented in the user profile in terms of restrictions that the elements in the corpus of construction patterns must satisfy. These

construction patterns are vectors of part-of-speech (POS) tags corresponding to the words being used in the case base.

Poem planning preferences appear in the user profile in terms of constraints on the appearance of sentence breaks within a line, and maximum and minimum number of lines that a sentence can span. Additionally, where patterns are used to represent complete stanzas, they also encode the distribution of words over lines (number and type of words per line, type of word at the end of a line...).

Metric preferences are stored in the user profile as initial parameters for the construction algorithm (chosen stanza, chosen verse length, rhyme pattern...).

Word selection and language structure preferences affect the CBR retrieve step. Poem planning and metric preferences affect the CBR reuse step.

The CBR revise and retain steps allow progressive refinement of the approximation to the desired literary style represented by the accumulated system vocabulary and the corpus of construction patterns. The system is continuously feeding back into the system the results that are being validated. From the moment the case base holds more user generated poems than poems in the original corpus, the data available to the system can be considered to embody the literary style of the user that has been interacting with it. In the same way as a CBR system employed consistently to solve a particular type of problem acquires with normal use knowledge that its designers were unaware of, such a poem generator would develop with continuous use into a model of the literary style of its users.

References

1. Aamodt, A. & Plaza, E. (1994). Case-Based Reasoning: Foundational Issues, Methodological Variations, and System Approaches. AI Communications, 7(i), pp 39-59.
2. Gervás, P., 'WASP: Evaluation of Different Strategies for the Automatic Generation of Spanish Verse', in: AISB-00 Symposium on Creative & Cultural Aspects and Applications of AI & Cognitive Science, 17th-18th April 2000, U. of Birmingham, England, pp 93-100.
3. Gervás, P., 'An Expert System for the Composition of Formal Spanish Poetry', in: Macintosh, A., Moulton, M., and Coenen, F. (eds.), Applications and Innovations in Intelligent Systems VIII, Springer Verlag, London Berlin Heidelberg, 2001, pp 19-34.
4. Horacek, H. and Busemann, S., 'Towards a Methodology for Developing Application-Oriented Report Generation', in: Günter, A. and Herzog, O. (eds.), 22nd German Conference on Artificial Intelligence (KI-98), Proceedings, Bremen, Germany, 1998.
5. Manurung, H.M., Ritchie, G., and Thompson, H., 'Towards a computational model of poetry generation', in: AISB-00 Symposium on Creative & Cultural Aspects and Applications of AI & Cognitive Science, 17th-18th April 2000, U. of Birmingham, England.
6. Manurung, H.M., Ritchie, G., and Thompson, H., 'A Flexible Integrated Architecture for Generating Poetic Texts', Informatics Research Report, EDI-INF-RR-0016, Division of Informatics, U. of Edinburgh, May 2000.
7. Nederhof, M.-J., 'Efficient generation of random sentences', Encyclopaedia of Computer Science and Technology, Vol.41, Marcel Dekker, 1999, pp 45-65.
8. Stratil, M., and Oakley, R.J., 'A Disputed Authorship Study of Two Plays Attributed to Tirso de Molina', Literary and Linguistic Computing, Vol. 2, No. 3, 1987, pp 153-160.
9. Tankard, J., 'The Literary Detective', Byte, February 1986, pp 231-238.