Creation of a New Domain and Evaluation of Comparison Generation in a Natural Language Generation System

Matthew Marge
Language Technologies Institute
School of Computer Science
Carnegie Mellon University
mrmarge@cs.cmu.edu

Amy Isard
ICCS/HCRC
School of Informatics
University of Edinburgh
Amy.Isard@ed.ac.uk

Johanna Moore
ICCS/HCRC
School of Informatics
University of Edinburgh
J.Moore@ed.ac.uk

Abstract

We describe the creation of a new domain for the Methodius Natural Language Generation System, and an evaluation of Methodius' parameterized comparison generation algorithm. The new domain was based around music and performers, and texts about the domain were generated using Methodius. Our evaluation showed that test subjects learned more from texts that contained comparisons than from those that did not. We also established that the comparison generation algorithm could generalize to the music domain.

1 Introduction

There has been research into tailoring natural language to a user's previous browsing history in a variety of domains such as medicine, museum collections, and animal descriptions (McKeown, 1985; Milosavljevic, 1997; Dale et al., 1998; O'Donnell et al., 2001). Another domain in which this could be applied is automated disc jockeys (DJs) that accompany a music stream such as Pandora[1] and discuss interesting trivia or facts about music tracks recently played to the user. User modeling could make these texts much more natural and less repetitive, and comparisons and contrasts between music artists or tracks could also provide users with a novel way to explore their music collection.

The Methodius system (Isard, 2007) continues in a line of research which began with ILEX (O'Donnell et al., 2001) and continued with M-PIRO (Isard et al., 2003) and now also NaturalOWL (Galanis and Androutsopoulos, 2007). Like these other systems, Methodius creates customizable descriptions of objects from an database, but it features a novel algorithm for generating comparisons between a new object and objects that have previously been encountered, which stands out from previous research in this area because it uses several explicit parameters to choose the most relevant and interesting comparisons given the context (Isard, 2007).

There have been previous evaluations of some of these systems, including (Cox et al., 1999; Karasimos and Isard, 2004). Karasimos and Isard conducted an evaluation of comparisons and aggregation in the M-PIRO system. The results showed that participants learned more and perceived that they learned more from texts that contained comparisons and aggregations than they did from texts that did not. In this study, we investigate whether these results generalize to our new domain, and we isolate the effect of comparisons from that of aggregation.

2 Knowledge Base Construction

2.1 Corpus Collection

We collected a small corpus to investigate the type of facts disc jockeys tend to say about music. We selected two genres where music descriptions between pieces were common, jazz and classical music. The programmes we used were broadcast on BBC Radio Three[2]. We transcribed sixty-four discussions; to maintain uniformity, we followed the Linguistic Data Consortium's transcription guidelines[3]. This

[1] http://www.pandora.com

[2] http://www.bbc.co.uk/radio3

[3] http://projects.ldc.upenn.edu/Transcription/quick-trans

was not a thorough corpus collection; the purpose of collecting examples was to gain a sense of what disc jockeys tend to discuss and compare.

2.2 Ontology Design

Based on the transcribed examples, we selected and hand-wrote twelve database entries for music tracks, using the authoring tool developed by the M-PIRO project (Androutsopoulos et al., 2007). We transformed the output of this tool into files suitable for Methodius using an ad-hoc collection of Perl and XSLT scripts, which also added the necessary information to the OpenCCG grammars (White, 2006) used by Methodius. We discuss future plans in this area in Section 5.

We created a single-inheritance ontology for a knowledge base of music pieces. First, we listed the high-level entity types in the music domain, such as "person", "instrument", "classical music period", and "jazz music period". We then added attributes commonly found in our disc jockey transcriptions. For each entity type, we defined a set of fields. For example, the classical-period field must contain an entity which expresses a classical music piece's time period. We also specified a microplanning expression for each field, which provides detail on how the field's information should be generated at the sentence level. We then added all the lexical items necessary for the music domain.

2.3 Ontology Population

We populated our domain with six classical music pieces and six jazz music pieces from the allmusic.com database[4]. The songs were selected to yield at least two interesting comparisons when placed in a specific order. We also added entities linked to the twelve songs, for example, each song's album, performer, and composer, and information about these entities. One challenge inherent in selecting these entities from a publicly available database was to eliminate as much common knowledge as possible about the music. In order to decrease background knowledge as a potential factor in our experiment, we selected songs that primarily did not contain popular performers, composers, and conductors. We were able to gauge the popularity of artists by their "popularity rank" in the allmusic.com database. However, we had to maintain a careful balance between obscure artists and the ability to generate interesting comparisons. Obscure artists had less detailed information in the allmusic.com database than popular music artists, so were forced to select a few popular music artists for our experiment, as their music pieces had multiple possible interesting comparisons.

[4]http://www.allmusic.com

```
"Avatar" was written by Gary Husband and it was
performed by Billy Cobham, who was influenced
by Miles Davis.  Billy Cobham originated from
Panama City, Panama and he played the drums; he
was active from the 1970s to the 1990s and he
participated in the Mahavishnu Orchestra.  He
was influenced by Miles Davis.  "Avatar" was
written during the Fusion period.
```

Figure 1: A generated description without comparisons.

```
Unlike "Fracture" and "A Mystery in Town",
which were written by Eddie "Lockjaw" Davis and
were performed by Fats Navarro, "Avatar" was
written by Gary Husband and it was performed by
Billy Cobham.  Cobham originated from Panama
City, Panama and he played the drums; he was
active from the 1970s to the 1990s and he
participated in the Mahavishnu Orchestra.  He
was influenced by Miles Davis.  "Avatar" was
written during the Fusion period.
```

Figure 2: A generated description with comparisons to previously described songs.

3 Experiment

We tried to maintain as many conditions from the previous, similar study (Karasimos and Isard, 2004) as possible to allow us to directly compare our results to theirs. The previous study established that people learned more and perceived that they learned more from text enriched with comparisons and aggregations of facts than from texts that contained neither. Our experimental design was similar to theirs but all conditions of our experiment contained text generated with aggregations of facts; our aim was to isolate the effects of comparisons from those of sentence aggregation.

For jazz texts, comparisons between songs involving performers, albums, composers, and time periods were possible. Classical texts could produce all four of these types of comparisons. In addition, classical texts could also include comparisons of conductors. Although the potential similarities

for classical and jazz texts were not equal, we decided to include the conductor as a potential comparison for classical music. This is because across both text types, we maintained the same number of generated comparisons for each text type by limiting Methodius to generating only one comparison or contrast per paragraph of text. We present examples of a paragraph of text generated by Methodius without (Figure 1) and with (Figure 2) comparisons. In both cases, we assume that the user has already seen texts about the songs "Fracture" and "A Mystery in Town", which expressed the facts about these previous songs which are used in the comparisons in Figure 2; the comparison text does not contain more new information.

3.1 Evaluation Design

For our user study, we created a web interface using WebExp2 Experiment Design software[5] that contained text generated by Methodius from our music fact knowledge base. Forty fluent English speakers were recruited and directed to a web page that gave detailed instructions. After providing some basic personal information including their name, age, gender, occupation and native languages, subjects started with a test page, where they read a sample paragraph and responded to one factual question, to make sure that they had understood the interface, and they then proceeded to the main experiment.

Participants read 6 paragraphs about either jazz or classical music, and answered 15 factual recall questions. They then read a further 6 paragraphs about the other type of music, followed by 15 factual recall questions on the second set of texts. Finally they completed a post-experimental survey of 12 Likert Scale questions (Likert, 1932). We used a within-subjects design, where each subject saw two sets of texts, one classical and one jazz, one with and one without comparisons, and the order in which text sets were presented was controlled. The multiple choice questions did not change given the condition; so every participant saw the same two sets of 15 multiple-choice questions in randomized orders. Seven multiple-choice questions of each fifteen-question set dealt with facts that may be reinforced by comparisons. The remaining eight questions in each section served as a control for this experiment.

[5] http://www.webexp.info

Group	Texts with comparisons	Texts without comparisons
A	**4.15** (1.814)	3.35 (1.872)
B	**4.45** (1.638)	3.10 (1.651)
All	**4.30** (1.713)	3.23 (1.747)

Table 1: Mean multiple choice scores with standard deviation in brackets.

On each page, the interface presented an image of a paragraph of text generated by Methodius. The users proceeded to the next paragraph when they were ready by pressing the "Next song" or "Next piece" button, depending on whether the music type was jazz or classical. The texts were presented as images for two reasons: so that the presentation of stimuli would remain consistent across the different computers and to prevent the text from being selected by the participant, thus discouraging them from copying the text and placing it into another window as a reference to answer the factual recall questions asked later.

4 Results

A summary of the participants' multiple choice scores are shown in Table 1. Group A read classical texts with comparisons and jazz texts without, and Group B read jazz texts with comparisons and classical texts without.

We performed a 2-way repeated measures ANOVA on our data and found that participants performed significantly better on questions about the texts which had comparisons (F (1, 36) = 11.131, $p < .01$). There were no ordering or grouping effects–the performance of participants did not depend on which type of texts they saw first, or on which type of texts contained comparisons.

In general, the Likert scores showed no significant differences between the texts which had comparisons and those which did not. Karasimos and Isard (2004) did find significant differences, but in their case, texts had either comparisons and sentence aggregations, or neither. In our study, all the texts had sentence aggregations, so it may be this factor which contributed to their higher Likert re-

sults on questions such as "I enjoyed reading about these songs" and the binary "Which text (quality, fluency) did you like more" question, for which we also found no significant difference. Details of results and statistics can be found in (Marge, 2007).

5 Conclusions and Future Work

We have shown that the Methodius comparison generation algorithm does generalize to new domains, and that it is possible to quickly author a new domain and generate fluent and readable text, using an appropriate authoring tool. We have also confirmed the findings of previous studies, and showed that the use of comparisons in texts does significantly improve participants' recall of the facts which they have read.

In future work, we would like to use the current text generation in an automatic DJ system with streaming music, and perform further user studies in order to make the texts as interesting and relevant as possible. We would also like to perform a study in which we compare the output of the comparison algorithm using different parameter settings, to see whether users express a preference.

Since this work was carried out, Methodius has been adapted to accept ontologies and sentence plans written in OWL/RDF. These can be created using the Protégé editor[6] with an NLG plugin developed at the Athens University of Economics and Business as part of the NaturalOWL generation system (Galanis and Androutsopoulos, 2007), which is available as an open source package[7]. A more principled method for the OpenCCG conversion process than the one described in Section 2.2 is in development, and we hope to publish a paper on this subject.

Acknowledgements

The authors would like to acknowledge the help and advice given by Colin Matheson, Ellen Bard, Keith Edwards, Ray Carrick, Frank Keller, and Neil Mayo and the comments of the anonymous reviewers. This work was funded in part by a grant from the Edinburgh-Stanford Link and by the Saint Andrew's Society of the State of New York. The music data in this study was used with the permission of the All Music Guide.

[6] http://www.protege.stanford.edu

[7] http://www.aueb.gr/users/ion/software/NaturalOWL.tar.gz

References

I. Androutsopoulos, J. Oberlander, and V. Karkaletsis. 2007. Source authoring for multilingual generation of personalised object descriptions. *Natural Language Engineering*, 13:191–233.

R. Cox, M. O'Donnell, and J. Oberlander. 1999. Dynamic versus static hypermedia in museum education: an evaluation of ILEX, the intelligent labelling explorer. In *Proceedings of the Artificial Intelligence in Education conference*, Le Mans.

R. Dale, J. Green, M. Milosavljevic, C. Paris, C. Verspoor, and S. Williams. 1998. The realities of generating natural language from databases. In *Proceedings of the 11th Australian Joint Conference on Artificial Intelligence*, Brisbane, Australia.

D. Galanis and I. Androutsopoulos. 2007. Generating multilingual descriptions from linguistically annotated OWL ontologies: the NaturalOWL system. In *Proceedings of ENLG 2007*.

A. Isard, J. Oberlander, I. Androutsopoulos, and C. Matheson. 2003. Speaking the users' languages. *IEEE Intelligent Systems*, 18(1):40–45. Special Issue on Advances in Natural Language Processing.

A. Isard. 2007. Choosing the best comparison under the circumstances. In *Proceedings of the International Workshop on Personalization Enhanced Access to Cultural Heritage*, Corfu, Greece.

A. Karasimos and A. Isard. 2004. Multi-lingual evaluation of a natural language generation system. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation*, Lisbon, Portugal.

R. Likert. 1932. A technique for the measurement of attitudes. *Archives of Psychology*, 22(140):1–55.

M. Marge. 2007. An evaluation of comparison generation in the methodius natural language generation system. Master's thesis, University of Edinburgh.

K. McKeown. 1985. *Text Generation: Using Discourse Strategies and Focus Constraints to Generate Natural Language Text*. Cambridge University Press, New York, NY, USA.

M. Milosavljevic. 1997. Content selection in comparison generation. In *6th European Workshop on Natural Language Generation)*, Duisburg, Germany.

M. O'Donnell, C. Mellish, J. Oberlander, and A. Knott. 2001. ILEX: An architecture for a dynamic hypertext generation system. *Natural Language Engineering*, 7:225–250.

M. White. 2006. Efficient realization of coordinate structures in combinatory categorial grammar. *Research on Language and Computation*, 4(1):39–75.