Issues for Corpus-Based Multimodal Generation

Mary Ellen Foster
Robotics and Embedded Systems Group
Faculty of Informatics, Technical University of Munich
Boltzmannstraße 3, 85748 Garching, Germany
foster@in.tum.de

Abstract

In recent years, data-driven methods have become increasingly popular in natural language generation. Multimodal generation can also benefit from using corpus data directly; however, there are several issues that arise when using corpora for multimodal generation that do not occur in the unimodal case, and that mean that existing multimodal corpora are often not suitable for being directly used in a generation system.

1 Introduction

In recent years, there has been an increasing amount of interest in the collection, annotation, and use of multimodal corpora—recorded collections of multimodal human behaviour, labelled and annotated for use in tasks such as analysis and summarisation. Another growing field of research is data-driven methods for natural language processing; this began with tasks such as parsing and machine translation, but more recently researchers in natural language generation have begun to take advantage of these data-driven methods as well.

Combining techniques and data from these two fast-growing fields to implement multimodal corpus-driven generation adds extra requirements that do not arise in each of the individual research fields: corpus-based techniques for text generation do not necessarily apply directly to the multimodal case, while general-purpose multimodal corpora are not always suitable for use in a generation system. This paper discusses several of the issues that must be taken into consideration if the two research areas are to be combined.

The paper is structured as follows. In Section 2, we first summarise the state of the art on multimodal corpora and the use of corpora in natural language generation. In Section 3, we next describe the generation task on which the most corpus-driven work has been done: generating non-verbal behaviour for an embodied conversational agent. After that, we summarise in Section 4 several issues that must be taken into consideration when designing a corpus-based generation system, using specific examples from the conversational-agent task. Finally, in Section 5, we give some conclusions and recommendations.

2 State of the Art

In order to fully appreciate the specific issues that arise in multimodal corpus-based generation, it is necessary to understand the related work in multimodal corpora and in corpus-based text generation. This section summarises the current state of the art in these two research areas.

2.1 Multimodal Corpora

A multimodal corpus is a recorded and annotated collection of communication modalities such as speech, gaze, hand gesture, body language, generally based on recorded human behaviour.[1]

[1] Although Chafai et al. (2006) used a corpus based on Tex Avery cartoons.

Recently, researchers in this area have increasingly been coming together to share raw and annotated data, as well as techniques and tools for annotation and analysis. At the most recent in a series of workshops on multimodal corpora (Martin et al., 2006), a number of papers were presented describing corpora and their applications in areas including meeting analysis, hand gestures, multimodality during conversation, and multimodal human-computer interaction.

The normal method for annotating a multimodal corpus is to annotate each of the individual communication modalities on its own layer, and to make explicit or implicit links between the layers. Standard tools for doing this type of annotation include NXT (Carletta et al., 2005), Anvil (Kipp, 2004), and ELAN (Hellweig and Van Uytvanck, 2006). The types of data that are annotated depend both on the corpus and the intended applications, and may range from low-level time-stamped motions to high-level discourse structures. For example, the raw data for the AMI meeting corpus (Carletta, 2006) consists of 100 hours of recorded multi-party meetings, including full video and audio recordings of all participants, with fully-transcribed and time-stamped speech. The data has been annotated on the following levels: dialogue acts, topic segmentation, abstractive and extractive summaries, named entities, individual actions and gestures, person location, focus of attention, emotional content, and argumentation structure. Many of these levels are linked directly to segments of the transcript, while others—such as gestures—are marked with starting and ending times.

At the moment, many multimodal corpora are built and used mainly for descriptive purposes such as analysis and summarisation. For example, the primary applications of the AMI meeting corpus include human-human communication modelling, multimedia indexing and retrieval, and meeting structure analysis and summarisation. Most papers in Martin et al. (2006) describe such applications; however, multimodal corpora have also been used for generating output, particularly for embodied conversational agents. For example, Kipp et al. (2006) use a corpus to generate gesturing behaviour. This work is discussed in more detail in Section 3.

2.2 Corpora in Natural Language Generation

In the then-current state of the art in natural language generation summarised by Reiter and Dale (2000), the primary purpose of a corpus was to serve as guidance for human developers of a generation system: the texts in a corpus were used as targets to help in specifying the rules or target outputs of the system, but were not themselves used directly in the process of creating or evaluating the output.

In recent years, the increasing availability of large textual corpora, both annotated and unannotated, has contributed to the explosive development of computational-linguistics techniques that make direct use of the data represented in a corpus. The areas where data-driven techniques have been successful include machine translation, part-of-speech tagging, parsing, chunking, and summarisation (Manning and Schütze, 1999).

Researchers in natural language generation have now also begun to adapt these data-driven techniques. Modern data-driven NLG systems make use of textual corpora in two ways. On the one hand, corpus data can act as a resource for decision-making at all levels of the generation process; on the other hand, the data can also be used to help evaluate the output of a generation system. The work presented at a recent workshop (Belz and Varges, 2005) includes generation systems that employ corpora in both of these roles.

Using corpus data directly in the generation process has several benefits. First, it provides a means for making decisions that are difficult to encode in rules, but that can easily be derived from data. The corpus can be used to control the entire generation process: Marciniak and Strube (2005), for example, used machine-learning classifiers trained on a corpus of route descriptions to make all of the decisions in generation. It is also possible to integrate corpus-based models into more traditional generation frameworks. Williams and Reiter (2005) used corpus data to create rules for content selection; at the other end of the generation pipeline, the OpenCCG surface realiser (White, 2005), for example, uses n-gram language models as a resource for making decisions such as adverb placement within a rule-based framework. Incorporating data-driven variation into the generation process can also produce output that is less repetitive and that is often preferred by human judges (e.g., Belz and Reiter, 2006; Foster and Oberlander, 2006).

In additionto being used in the generation process, corpus data can also be used to evaluate the output of a generation system, generally by measuring how close the generated output comes to the texts in the corpus. Note that there is a danger in using cross-validation alone to evaluate the output of a generation system. As pointed out above, human judges in several studies (Belz and Reiter, 2006; Foster and Oberlander, 2006) have been found to prefer output that includes data-driven variation; however, a pure cross-validation measure will penalise such outputs against those that do not diverge far, on average, from the contents of the corpus, giving a potentially false picture of the relative quality. However, cross-validation and other corpus-driven methods can still provide a useful and easily computed evaluation of output quality and system performance, and have been used to evaluate a number of systems. For example, White (2004) measured the accuracy and speed of the OpenCCG surface realiser through cross-validation against target texts; Marciniak and Strube (2005) also evaluated their realisation component through cross-validation; Wan et al. (2005) used cross-validation to measure the recall and precision of a stochastic summary-sentence generation system; while Karamanis and Mellish (2005) describe a number of corpus-based methods for evaluating information-ordering systems.

3 Generating Non-Verbal Behaviour for ECAs

For the rest of the paper, we will concentrate on the specific task of generating multimodal behaviour for embodied conversational agents, as that is the target for most current data-driven multimodal generation systems. To be sure, corpora have been used in other multimodal generation systems as a resource for developers, à la Reiter and Dale (2000)—Corio and Lapalme (1999) used a corpus of information graphics and their captions to help define the rules for their system, for example—but it does not appear that corpora have been used directly for any multimodal generation task other than embodied agents, so we will focus on that task here.

Embodied Conversational Agents (ECAs) are computer interfaces that are represented as human bodies, and that use their face and body in a human-like way in conversation with the user (Cassell et al., 2000). The main benefit of ECAs as a user-interface paradigm is that they allow users to interact with a computer in the most natural possible setting: face-to-face conversation. However, to take full advantage of this benefit, the conversational agent must produce high-quality output, both verbal and non-verbal. A number of existing systems have based the choice of non-verbal behaviours for an ECA on the behaviours of humans in conversational situations; the implementations vary as to how directly they use the human data.

In some systems, motion specications for the agent are created from scratch, using rules derived from studying human behaviour; this is similar to the classical Reiter and Dale view of the role of corpora in text generation. For the REA agent (Cassell et al., 2001a), for example, gesturing behaviour was selected to perform particular communicative functions, using rules based on studies of typical North American non-verbal displays. Similarly, the performative facial displays for the Greta agent (de Carolis et al., 2002) were selected using hand-crafted rules to map from affective states to facial motions.

In contrast, other ECA implementations have selected non-verbal behaviour based directly on motion-capture recordings of humans. Stone et al. (2004), for example, recorded an actor performing scripted output in the domain of the target system. They then segmented the recordings into coherent phrases and annotated them with the relevant semantic and pragmatic information, and combined the segments at run-time to produce complete performance specications that were then played back on the agent. Cunningham et al. (2005) and Shimodaira et al. (2005) used similar techniques to base the appearance and motions of their talking heads directly on recordings of human faces. This technique can produce extremely naturalistic and individual output; however, the technical requirements for doing the motion capture are high, and the procedure is quite invasive for the subject.

A middle ground between the above two implementation strategies is to use a purely synthetic agent—one whose behaviour is controlled by high-level instructions, rather than based directly on human motions—but to create the instructions for that agent using the data from an annotated corpus of human behaviour. Like a motion-capture implementation, this technique can also produce increased naturalism in the output over a purely rule-based system, and also allows choices to be based on the behaviour of a single individual if necessary. However, annotating a video corpus can be less technically demanding than capturing and directly re-using real motions, especially when the corpus and the number of features under consideration are small. This approach has been taken, for example, by Cassell et al. (2001b) to choose posture shifts, by Foster and Oberlander (2006) to select facial displays, and by Kipp et al. (2006) to select hand gestures.

4 Designing a Multimodal Corpus for Generation

As described in Section 2, both multimodal corpora and corpus-based generation are currently active and productive areas of research. However, bringing together these two areas for corpus-based multimodal generation raises several issues that do not arise, or that do not have the same impact, in the two individual research areas: corpus-based techniques for text generation do not necessarily apply directly to the multimodal case, while general-purpose multimodal corpora are not always suitable for use in a generation system. The considerations when designing a corpus-based multimodal generation system include the following:

1. The contextual information necessary for making generation decisions must be represented in the corpus.

2. The granularity of the annotation and of the cross-modal links must be appropriate to the generation task.

3. The generation system must be able to reproduce the corpus data in an appropriate way.

In the remainder of this section, we will discuss each of these issues in more detail.

4.1 Representing Contextual Information

In many cases, multimodal corpora are created based on naturally-occurring human behaviour; that is, the subjects being recorded are free to speak and act as they wish, and the annotators then analyse the behaviour based only on the recordings. The corpus resulting from such a recording cannot contain any more information than what is available from observing the behaviour, and—possibly—from annotators applying their own judgement to add extra information (such as the dialogue-act and topic-structure annotations on the AMI corpus).

For some generation contexts, this sort of surface-level annotation of context is sufficient; for example, for an ECA whose motion is selected entirely based on the features of the speech signal, such as that of Shimodaira et al. (2005), no deeper representation of the context is needed. However, in many cases, a generation system has available a much richer notion of context as it is planning its output. For example, Greta (de Carolis et al., 2002) represents the target information structure and affective content of its utterances, while the input to the talking head of Foster and Oberlander (2006) includes the intended prosodic, dialogue-history, and user-model contexts. All of this information can be useful in choosing the desired multimodal output behaviour; however, unless it is represented in the corpus, none of it can be used by the generation system.

The required contextual information can be included in the corpus in two ways. Either it can be manually added after the fact by annotators, or the corpus can be created in such a way that the required information is already present before the annotation. The latter can be achieved by using corpora based on scripted output in the domain of the eventual target system; if the human being recorded is following a known script, then all of the relevant contextual information can easily be added to the corpus at construction time. This approach was taken by Stone et al. (2004) and Foster and Oberlander (2006). It has the advantage that no additional manual effort is required; however, it also has the disadvantage that the corpus must be created specifically for the target application, which rules out using existing annotated corpora.

4.2 Representing Cross-Modal Links

In many multimodal corpora, each separate modality is represented on its own timeline, with the only links between modalities those that are implicitly represented by the timestamps. For example, in the AMI corpus, there are many levels of links corresponding to different aspects of the spoken signal; however, the gesturing behaviour is represented on its own timeline with its own start and stop times. This type of representation is adequate if the goal is to extract events or to analyse human behaviour. However, if the goal is to generate novel output based on the corpus, more explicit links between the modalities are useful, as the temporal structure may not coincide with the underlying generation process.[2]

One important decision is the level at which cross-modal links are represented—that is, the size of the segment on each channel that can be associated with segments on other channels. For example, when associating multimodal behaviours with speech, motions may be associated with phonemes or syllables, with single words, with syntactic constituents, or with arbitrary sequences of words. Which of these is chosen depends on the level at which the generation system will later be selecting these motions; if the assumptions are later changed, it may prove costly. For example, the original talking-head implementation described by Foster and Oberlander (2006) selected facial displays based on individual words in the output, and the corpus was annotated accordingly. However, that assumption proved to be unrealistic: the majority of displays did *not* in fact coincide with single words. In order to produce more realistic motions, the entire corpus had to be re-processed using a revised scheme that associated displays with word sequences, and the generation system was updated to use that updated corpus.

As well as the level of representation, the criteria for making a link must be established: is the choice based strictly on temporal or spatial coincidence, or is semantic information also used? The former is easier to annotate, and may even be automatically derived from an existing annotated corpus, but may not generalise as readily to new outputs; the latter requires a more involved annotation process that makes more demands on the annotators for careful judgement calls. For example, Kipp et al. (2006) chose to record temporal co-occurrence and lexical affiliation as separate attributes when annotating hand gestures for generation; temporal co-occurrence was derived largely automatically from the video, but annotating the lexical links relied on "gesture literature and sometimes intuition".

4.3 Reproducing Corpus Data

A unimodal corpus can be used for generation with a minimum of processing effort; in most cases, the data in the corpus can be simply be directly combined to produce the output. For example, when using a textual corpus to help make decisions in surface realisation, n-gram models can be built from the words in the corpus and used to guide the system towards high-scoring realisations, as was done by Langkilde-Geary (2002) and White (2005). Similarly, in speech synthesis, the technique of unit selection (Hunt and Black, 1996) involves segmenting recorded speech into diphones (phoneme-to-phoneme transitions) and then using a Viterbi-style algorithm to construct a sequence of diphones to synthesise a given string of words. The corpus data must be annotated with the contextual information necessary to select the right content in a given context; however, there is no need to do any processing on the actual data to use it for generation.

For multimodal generation, in contrast, it is generally not the case that corpus data can be directly combined to produce output in the way that diphones can be concatenated for speech synthesis, or words for text generation. In most cases, a multimodal generation system creates entirely synthetic output by specifying commands for each of the relevant output channels, rather than combining existing pieces of output directly. Even in cases where motion-capture data is used directly (e.g., Stone et al., 2004; Cunningham et al., 2005), the recorded motions must still be mapped to animation commands and synchronised with the speech. When the generated output is specified at a higher level, then more complex mappings must be made.

For example, Kipp et al. (2006) use an annotation scheme for hand gestures that makes the conscious decision not to represent every single feature of the motion, but rather to capture

[2] Kipp et al. (2006) "found that the claim that gesture stroke and lexical affiliate always co-occur is often wrong."

the essentials, in some cases using gesture "lexemes" to abstract over the data. To recreate a gesture schedule annotated using this scheme, the motion specifications are translated into specific commands for the animation engine. Foster and Oberlander (2006) use a similar mapping for their facial displays: based on the speech, a set of high-level displays are selected, and are then converted to motion specifications in the language of the talking head.

It is important that the final mapping between annotated events and output commands is sufficiently close that the corpus data is actually relevant to the generation task. If not, the resulting output may not be appreciated by the subjects. For example, Foster (2004) attempted to use an annotated corpus of humans making hand gestures to specify the motion of an on-screen pointer, with rather disappointing results on the human evaluation.

5 Conclusions

Both multimodal corpora and corpus-based generation are active areas of research at the moment. Large-scale multimodal corpus resources such as the AMI corpus are being created and made freely available, and it would be a positive development if the data-driven techniques being developed for text generation could be directly applied to multimodal generation and could make use of such available resources.

Unfortunately, in many cases, the additional requirements of a multimodal generation system mean that it makes more sense in practice to collect and annotate a special-purpose corpus for the specific generation task, instead of using existing corpora. An application-specific corpus has the advantage that it can be created entirely from in-domain recordings, possibly even based on scripts to ensure that the necessary contextual information is readily available. Also, care can be taken that the data in the corpus is represented at the correct level for use in the generation system and that the output generator is able to make coherent output by using the data in the corpus.

However, it seems a shame to disregard entirely the corpora that are now being created, and it may often be possible to adapt such a corpus for use in a particular generation task. If an existing corpus is to be used for generation, it will likely be necessary to do some additional annotation to incorporate the necessary contextual and cross-modal information, and to take care in the implementation that the corpus data can be easily—and sensibly—reproduced. However, in some cases, this extra effort may be justified if it allows the generation system to take advantage of the increasing range of multimodal data to improve the generation process or to produce higher-quality output.

References

Belz, A. and Reiter, E. (2006). Comparing automatic and human evaluation of NLG systems. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2006).*

Belz, A. and Varges, S., editors (2005). *Corpus Linguistics 2005 Workshop on Using Corpora for Natural Language Generation.*

Carletta, J. (2006). Announcing the AMI meeting corpus. *The ELRA Newsletter*, 11(1):3–5. Corpus available through `http://corpus.amiproject.org/`.

Carletta, J., Evert, S., Heid, U., and Kilgour, J. (2005). The NITE XML toolkit: Data model and query. *Language Resources and Evaluation Journal*, 39(4):313–334.

Cassell, J., Bickmore, T., Vilhjálmsson, H., and Yan, H. (2001a). More than just a pretty face: Conversational protocols and the affordances of embodiment. *Knowledge-Based Systems*, 14(1–2):55–64.

Cassell, J., Nakano, Y., Bickmore, T. W., Sidner, C. L., and Rich, C. (2001b). Non-verbal cues for discourse structure. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics (ACL 2001).*

Cassell, J., Sullivan, J., Prevost, S., and Churchill, E. (2000). *Embodied Conversational Agents.* MIT Press.

Chafai, N. E., Pelachaud, C., and Pelé, D. (2006). Analysis of gesture expressivity modulations from cartoons animations. In Martin et al. (2006).

Corio, M. and Lapalme, G. (1999). Generation of texts for information graphics. In *Proceedings of the 7th European Workshop on Natural Language Generation (EWNLG'99)*, pages 49–58.

Cunningham, D. W., Kleiner, M., Wallraven, C., and Bülthoff, H. H. (2005). Manipulating video sequences to determine the components of conversational facial expressions. *ACM Transactions on Applied Perception (TAP)*, 2(3):251–269.

de Carolis, B., Carofiglio, V., and Pelachaud, C. (2002). From discourse plans to believable behavior generation. In *Proceedings of the 2nd International Conference on Natural Language Generation (INLG 2002).*

Foster, M. E. (2004). Corpus-based planning of deictic gestures in COMIC. In *Proceedings of the INLG 2004 Student Session*, Brockenhurst, England.

Foster, M. E. and Oberlander, J. (2006). Data-driven generation of emphatic facial displays. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2006).*

Hellweig, B. and Van Uytvanck, D. (2006). EUDICO linguistic annotator (ELAN) version 2.6: Manual. `http://www.mpi.nl/tools/`.

Hunt, A. and Black, A. W. (1996). Unit selection in a concatenative speech synthesis system using a large speech database. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP96)*, volume 1.

Karamanis, N. and Mellish, C. (2005). A review of recent corpus-based methods for evaluating information ordering in text production. In Belz and Varges (2005).

Kipp, M. (2004). *Gesture Generation by Imitation - From Human Behavior to Computer Character Animation.* Dissertation.com.

Kipp, M., Neff, M., and Albrecht, I. (2006). An annotation scheme for conversational gestures: How to economically capture timing and form. In Martin et al. (2006).

Langkilde-Geary, I. (2002). An empirical verification of coverage and correctness for a general-purpose sentence generator. In *Proceedings of the 2nd International Language Generation Conference (INLG 2002).*

Manning, C. D. and Schütze, H. (1999). *Foundations of Statistical Natural Language Processing.* The MIT Press.

Marciniak, T. and Strube, M. (2005). Using an annotated corpus as a knowledge source for language generation. In Belz and Varges (2005).

Martin, J.-C., Kühnlein, P., Paggio, P., Stiefelhagen, R., and Pianesi, F., editors (2006). *LREC 2006 Workshop on Multimodal Corpora: From Multimodal Behaviour Theories to Usable Models.*

Reiter, E. and Dale, R. (2000). *Building Natural Language Generation Systems.* Cambridge University Press.

Shimodaira, H., Uematsu, K., Kawamoto, S., Hofer, G., and Nakai, M. (2005). Analysis and synthesis of head motion for lifelike conversational agents. In *Proceedings of the 2nd Joint Workshop on Multimodal Interaction and Related Machine Learning Algorithms (MLMI 2005).*

Stone, M., DeCarlo, D., Oh, I., Rodriguez, C., Lees, A., Stere, A., and Bregler, C. (2004). Speaking with hands: Creating animated conversational characters from recordings of human performance. *ACM Transactions on Graphics (TOG)*, 23(3):506–513.

Wan, S., Dale, R., Dras, M., and Paris, C. (2005). Statistically generated summary sentences: A preliminary evaluation using a dependency relation precision metric. In Belz and Varges (2005).

White, M. (2004). Reining in CCG chart realization. In *Proceedings of the 3rd International Conference on Natural Language Generation (INLG 2004)*.

White, M. (2005). Designing an extensible API for integrating language modeling and realization. In *Proceedings of the ACL 2005 Workshop on Software*.

Williams, S. and Reiter, E. (2005). Deriving content selection rules from a corpus of non-naturally occurring documents for a novel NLG application. In Belz and Varges (2005).