TUNA: Towards a UNified Algorithm for the Generation of Referring Expressions

Final Report

1 Background and Aims

TUNA arose from two observations concerning the state of the art in the Generation of Referring Expressions (GRE). Firstly, we observed that the area contained a number of separate algorithms, each of which addresses a separate aspect of reference (e.g., reference to sets, or the treatment of salience) in isolation, without trying to address several aspects in one algorithm. Secondly, we observed that algorithms in this area were lacking an empirical basis; even though some algorithms were partially motivated by psycholinguistic concepts, it was unclear to what extent these algorithms reflected actual language use by human speakers or writers, or to what extent the descriptions generated by them were useful to hearers or readers. In response to these shortcomings, TUNA promised empirical results, and progress towards a unified algorithm.

2 Objectives

TUNA promised algorithms that are able to deal in an empirically justifiable way with

1. references to sets as well as individuals, also involving negation and disjunction
2. references involving context-dependence and vagueness
3. references involving under- and overspecification
4. references involving pointing

In order to give an empirical basis to these algorithms, we expected to do extensive experimental work. As a basis for building a unified algorithm, we saw the labelled di-graphs of Krahmer and colleagues as a plausible option. We knew that these graphs could be extended to cover sets and vagueness and wondered whether it might be possible to do all these things at the same time, giving us a basis for the unified algorithm that TUNA had promised to work towards.

3 Project plan review

In broad outline, we have adhered to the TUNA project plan. There is, however, one whole exception, and one partial exception. The "whole" exception is the treatment of context-dependence and salience. In this respect, our project plan was too ambitious. Contextual aspects of reference are clearly of great importance, but the issues involving "one-shot" descriptions (where the linguistic context of the description is assumed to be empty) are so difficult that it seemed better to postpone these issues largely to a follow-up project, a proposal for which is now being drafted. The "partial" exception is the issue of pointing. Only a limited amount of time was devoted to pointing in the project itself, except by stimulating the second postdoc on the project (Dr van der Sluis) to continue the research that she had started during her PhD work at Tilburg, (under supervision of Dr Krahmer, her supervisor at the time, and Visiting Fellow to the TUNA project).

Our adherence to the project plan did not mean that the project was carried out entirely as planned. The main discontinuity was caused by the departure to Stanford (later: Trento) of the first postdoc on the project, Dr Sebastian Varges. His departure, which happened around the

halfway point, was a considerable loss to the project. Before entering TUNA, Dr Varges had designed an "instance-based" approach to Linguistic Realisation, the idea of which is to generate expressions whose linguistic form resembles at least one or the expressions in a corpus. Progress was made, during the first year of TUNA, towards extending this instance-based method in such a way that it becomes applicable to GRE, where the selection of semantic content (rather than linguistic form) is a large part of the problem (Varges 2004, 2005a, 2005b, Varges and van Deemter 2005). When Dr Varges was replaced by Dr van der Sluis, it soon became apparent that continuation of this research without its main instigator was not the best option. We therefore changed course, by focussing on an experimental (i.e., essentially psycholinguistic) approach instead of a corpus-based one. As we shall make clear below, this new direction has proven to be very fruitful.

4 Main contributions

We start with some strictly computational contributions, followed by a discussion of empirical and other contributions.

4.1 Extensions and limitations of the graph-based model

Early on, we discovered that the graph-based model loses some of its appeal when the knowledge base (KB) from which GRE takes its starting point becomes very complex. In Van Deemter and Krahmer (2007), we show that relations, sets, Booleans, and many other phenomena can each be captured by the graph-based approach. The idea is as follows. Suppose you want to describe a set of objects as "the things that are either trumpets or trombones", based on a KB that only contains atomic information. This can be achieved by *expanding* the initial KB in such a way that every object initially marked as a trumpet is now also marked with the disjunctive label $trumpet \vee trombone$ (and similarly for trombones). This new disjunctive property can be used by the graph-based algorithm in the normal way.

However, disjunctions might involve complex rather than atomic properties. Suppose, for example, that the intended referents of a description are those people who *either* blow a trombone *or* smash a guitar (as was once fashionable). To make the graph expansion method work, one would have to introduce complex disjunctive labels saying "blowing a trombone or smashing a guitar", "blowing a new trombone and smashing an expensive guitar", and so on. Even the simplest initial graph would have to be expanded into a huge one, in which each object is associated with a huge number of disjunctive properties. In the face of these difficulties, it is no longer clear that graphs are preferable over other representational methods.

Considerations of this kind have meant that graphs have played a less central role in the project than foreseen. This does not mean that graphical representations do not have potential for GRE. Developments in the final stages of the project are worth mentioning in this connection, in which Krahmer et al's ideas about graphs were linked with the representational and inferential power of Conceptual Graphs (CGs). Dr Croitoru, then a postdoc at Aberdeen (not employed on the TUNA grant), suggested replacing Krahmer's di-graphs with full CGs. Croitoru and Van Deemter (2007a and 2007b) showed that this CG-based approach can do everything that can be done with di-graphs, and much more besides. One advantage is that CGs are a well-understood representational framework, in which facts are systematically linked with ontological ("T-box") information. Moreover, CGs make it possible in principle to perform GRE on KBs that contain non-atomic information. For example, one could have a KB saying "every person has a mother" and "John is a person", automatically enabling the reference "John's mother", even if no such person is explicitly represented in the KB. Such exciting generalisations of GRE will be explored in future work.

4.2 A unified algorithm for GRE

In the early phases of the project, we were still focussing on extending the semantic coverage of GRE algorithms. A turning point was Varges and Van Deemter (2004). In this paper (whose representational style was still graph-oriented), we showed how referring expressions can be generated that contain quantifiers, as in "the man who plays three instruments". We also argued, however, that there is no end to this kind of extension. Any sentence can be turned into a relative clause, and thereby become a part of a referring expression. The GRE problem as a whole can therefore be argued to be "NLG complete" (i.e., a solution to all of GRE implies a solution for all of NLG). From this time onwards, we re-focussed on generating fairly *simple* descriptions *well*.

For a long time after Varges' departure, it seemed as if no actual 'unified algorithm' was going to be built, since we focussed more on gathering empirical insights that would be useful on the road "towards" a unified algorithm (cf. the acronym TUNA). The developments of section 4.7, however, made it desirable to implement a number of key algorithms within a unified framework, addressing most the referential phenomena that were listed as Objectives (section 2). The resulting implementation was an API, built by Mr Gatt, which was broadly inspired by Bohnet and Dale's (2005) search-based perspective on GRE. The API facilitates the implemention of new GRE algorithms, but also contains implementations of existing algorithms, all of which are generalised to deal with gradable properties, disjunction and plurality, and negation. This generalisation is made possible through a separation of (a) representational issues, to do with how KB information is represented and manipulated (e.g., by disjoining literals or performing inference on vague properties), and (b) the search procedure which selects content from the available properties. Although this unified algorithm does not combine all the phenomena of which the TUNA proposal speaks (e.g. salience is lacking at the moment), it has proven to be a very useful piece of software which seems likely to be used and extended by others.

4.3 Pointing

After joining TUNA in the winter of 2005/6, Dr Van der Sluis continued working on the topic of her PhD thesis (under supervision of Dr Krahmer, Visiting Fellow to TUNA) for some time. The main outcome is a journal paper (van der Sluis and Krahmer 2007, also Van der Sluis and Krahmer 2004a, 2004b, 2005), which presents a computational model for the generation of *multimodal* referring expressions, based on observations in human communication. The algorithm makes use of a so-called Flashlight Model for pointing. The Flashlight Model accounts for various types of pointing gestures of different precisions. Guided by a notion of expressive effort (e.g., pointing to something small takes more effort than to something large), the algorithm produces referring expressions combining language and pointing gestures. The algorithm was evaluated using two controlled production experiments, which showed that its output coincided to a large extent with the utterances of the participants. However, the participants tended to produce far more overspecified referring expressions than the algorithm. Results of this kind raise the question whether anything can be said about the reasons why and the circumstances in which speakers overspecify.

4.4 Overspecification

Any systematic empirical study into referring expressions will automatically answer some questions about under- and overspecification. (This is true, for example, of the evaluation studies of section 4.7, even though these were not set up explicitly to address these issues.) Around the sidelines of TUNA, however, the PI conducted an investigation (Paraboni and Van Deemter 2006, Paraboni et al. 2006a and 2006b, Paraboni et al. 2007) that focusses specifically on the question how much overspecification is desirable, paying specific attention to situations in which the KB is hierarchically structured, in the way in which a city may be structured into

neighbourhoods, streets, houses, etc. When referring to a house, we often deliberately add information beyond what is logically necessary, saying 'Belvidere Street in Rosemount, Aberdeen', rather than just 'Belvidere Street in Aberdeen'. In collaboration with Dr Paraboni (University of Sao Paulo, a previous PhD student of the PI) and Dr Masthoff (University of Aberdeen, who contributed to the design and implementation of the experiment), we carried out experiments in which we measured the *effort* that readers needed to find a referent, as a function of the amount of overspecification. These experiments provided strong support for the models proposed earlier by the same authors, which prescribe overspecification in a few well-specified situations (called Dead End and Lack of Orientation). We believe this to be the first study into the effects of generated referring expressions on readers/hearers, and we are excited by the prospects for our experimental method, which is based on counting the number of clicks (following the hyperlinks on a web page) performed by readers while searching the referent.

4.5 Reference to sets

Reference to sets has turned out to be one of the main aspects of reference studied in TUNA, partly because it became the focus of Mr Gatt's PhD project. His PhD thesis details a systematic attempt to discover what is the best way to refer to a set. Far from only proposing and evaluating algorithms, his approach is underpinned by new psycholinguistic experiments that aim to find out why one description may be better than another, and by an attempt to understand how plurals are different from singulars. Perhaps the main factor identified by Mr Gatt's work is that of lexical/conceptual coherence. Suppose you can refer to a pair of people as either 'the Italian and the Spaniard' or (equivalently) 'the Italian and the cook'. We propose an account of such phenomena that rests on lexical similarity (and ultimately on lexical priming), and implements this account in an algorithm that honours lexical coherence (e.g., between 'Italian' and 'Spaniard') without loosing the property of logical completeness (Gatt and Van Deemter 2006, 2007, Gatt and Van Deemter to appear). The model differs from models proposed in the formal semantics literature by allowing less coherent descriptions if no fully coherent one is possible (for example because there are two Spaniards in the domain, so that 'the Spaniard' is referentially unclear). Evaluation of the algorithm revealed an unexpected degree of success, suggesting that coherence is more important than brevity. It turns out, for example, that subjects prefer 'The Italian, the Frenchman and the Englishman (...)' over 'The Italian and the bachelors (...)' (in a situation where the Frenchman and the Englishman are the only bachelors). Space limitations prevent more elaborate discussion of these and related findings.

4.6 Vagueness and location

The starting year of the TUNA project was the year in which the PI finalised his proposal for the use of vague/gradable concepts in referring expressions (Van Deemter 2006). The TUNA project was able to build on this work in connection with *geographical* references. Examples abound in the experimental settings discussed in section 4.7, where subjects are exposed to pictures of furniture on a computer screen. Suppose the screen shows two sofas, both solidly on the righthand half of the screen, but one to the left of the other. In such situations, many subjects speak of 'the sofa on the left', even though it is on the righthand side of the screen. Such phenomena can be captured by Van Deemter's algorithm, if 'left' is modelled as a gradable adjective ({sofa, leftmost}). An extension, inspired by Thorisson (1994), was proposed in Gatt (2006a), where an algorithm is proposed that performs a similar task while ensuring that any group (containing any numer of elements) that is referred to is perceptually *grounded* in the sense that its elements are sufficiently close to each other to actually be perceived as a group. This was later incorporated in an algorithm for the generation of collective spatial descriptions ('the group of objects in the top left corner', Gatt 2006b).

4.7 Evaluating GRE algorithms

Towards the end of the project, we started to feel that there was something perverse in pursuing ever more complex variants of reference, when so little is known about the success of GRE algorithms in simple cases. For although algorithms like Dale and Reiter's Incremental Algorithm were broadly motivated by psycholinguistic research (particularly the idea that descriptions are not always minimal, and that certain attributes are more popular than others), existing algorithms were never tested and compared against each other in a controlled setting that focusses on identification of the referent. (For a discussion of the virtues and shortcoming of studies by Gupta and Stent, Jordan and Walker, and Dale and Viethen, see Van Deemter et al. 2006, Gatt et al. 2007a, 2007b.) A large elicitation experiment was designed and performed, which led to a corpus of some 2300 descriptions, each of which identifies either a single object or set of two. Descriptions in two different domains were annotated with semantic information, which made the corpus *semantically transparent*, in that every description was paired with complete information regarding its properties and the properties of all distractor objects. It is also *pragmatically transparent* because the task performed by the subjects was always such that *identification of the referent* was their only goal.

We used this transparent corpus to automatically evaluate a number of existing algorithms in this area, based on calculation of the average similarity between the descriptions generated by the algorithm in question, compared with all descriptions generated by human subjects. (Here we made crucial use of the API described in section 4.2, which allowed us to implement each algorithm quickly.) Our two main findings hinge on the fact that the Incremental Algorithm makes use of a *preference order* between attributes, which determines the order in which attributes are inspected by the algorithm (to see whether they make a useful contribution to the description generated so far). We found that, on the one hand, there was always at least one preference order which caused the IA to outperform all other algorithms (including a greedy algorithm and one resulting in optimally brief descriptions). But we also found that there was always at least one preference order which caused the IA to perform much *worse* than many other algorithms. Moreover, it appeared to be extremely difficult to predict in advance which preference orders might lead to a good algorithm, especially in the more complex of the two domains, where the number of attributes to choose from was large. We expect to spend more time analysing the TUNA corpus before publishing a journal article to address the experiment as a whole.

5 Research impact

We believe the impact of the TUNA project to be considerable. As one indication of the importance of GRE within the larger area of Natural Language Generation, six of the 24 papers in the main programme of the Fourth International Natural Language Generation Conference (INLG-06) focussed on referring expressions, with three contributions from Aberdeen; for the 11th European Workshop on NLG (ENLG-07) this figure was four out of 18, with two from Aberdeen, until one of our papers was withdrawn to be presented at EMNLP instead; for the Workshop on Shared Tasks and Comparative Evaluation in Natural Language Generation, organised in Arlington by the Ohio State University, three out of 15 papers focussed on GRE, with one from Aberdeen. Perhaps most significantly, the workshop at Arlington led to a proposal for a first Shared Task Evaluation Campaign (STEC) in NLG. This STEC will focus on GRE *"since this area has been the focus of intensive research over the past decade, leading to greater consensus over basic problem definition, inputs and outputs than in most NLG subfields"* (Call for Participation, 1 May 2007). The first installment of this STEC will be built around the TUNA corpus (see section 4.7) and the automatic evaluation measures proposed in TUNA. Although the success of this STEC (also in terms of its wider effects on the NLG community) cannot be predicted with certainty, we are pleased that TUNA was chosen as its basis.

6 Explanation of expenditure

Van Deemter and Gatt's move to Aberdeen, and the subsequent departure of Varges, has led to an extension by 5 months (because of the time when there was no RF in place). This change has essentially been financially neutral. Contributions from the Visiting Fellow, Krahmer, have concentrated on the issue of pointing (section 4.3), on experiments involving vague descriptions (section 4.6), and on the organisation of the MOG workshop (see below). Lower than expected expenditure on consumables was offset by slightly higher expenditure on salaries.

7 Dissemination, exploitation, and further work

The project has led to 5 journal papers, 1 book chapter, and 22 papers for conferences and workshops. Mr Gatt's PhD thesis was submitted in March 2007. Project members co-organised a number of international events, including (1) the workshop Coherence in Generation and Dialogue (http://www.doc.gold.ac.uk/∼mas01rk/esslli2006.html) at ESSLLI-2006 (which is also leading to a Special Issue of the journal JoLLI), (2) the workshop Using Corpora for Natural Language Generation (UCNLG, http://www.itri.brighton.ac.uk/ucnlg/ucnlg05/) and (3) the successful workshop Multimodal Output Generation (MOG, http://www.csd.abdn.ac.uk/mog2007/).

The TUNA corpus (including the annotation manual and API) and the related STEC (mentioned above) is arguably TUNA's main exploitation result, see http://www.csd.abdn.ac.uk/research/evaluation/ . Additionally, we believe the TUNA bibliography on GRE (see web page) to be a valuable resource for other researchers.

Further work includes a survey paper on GRE that project members are planning to submit to *Computational Linguistics*, and plans for a new EPSRC project on GRE in a dialogue (focussing on referential collaboration and alignment). A wide-ranging journal paper on the TUNA corpus and its analysis has also been planned for the near future.

TUNA *publications* are listed at http://www.csd.abdn.ac.uk/research/tuna/. Five of these publications were selected on the Je-S page as representative of TUNA. One of these (Gatt et al. 2007b), although only a workshop paper, was selected because it is representative of our corpus-related work, to which we wish to draw attention even though it is still in progress.
Below we list only *journal* publications relating to the TUNA project. Masthoff and Gatt (2006) is not listed on the TUNA web page, because it is not directly related to the generation of referring expressions. Mr Gatt's PhD thesis is not listed until after the Viva (June 2007).

Journal publications:

– Gatt, A., and Van Deemter, K. (to appear). Lexical choice and conceptual perspective in the generation of plural referring expressions. To appear in *Journal of Logic Language and Information* (JoLLI).
– Masthoff, J., and Gatt, A. (2006). In pursuit of satisfaction and the prevention of embarrassment: Affective state in group recommender systems. *Journal of User Modelling and User-Adaptive Interaction* 16: 281-319.
– Paraboni, I., Van Deemter, K., and Masthoff, J. (2007). Generating Referring Expressions: Making Referents Easy to Identity. To appear in *Computational Linguistics* 33(2).
– van Deemter, K. (2006). Generating Referring Expressions that involve gradable properties. *Computational Linguistics* 32(2).
– van der Sluis, I., and Krahmer, E. (2007). Generating Multimodal References. To appear in *Discourse Processes*.

Kees van Deemter, Aberdeen, 25 May 2007.