Representing Paraphrases Using Synchronous TAGs

Mark Dras

Microsoft Research Institute, Macquarie University

NSW Australia 2109

markd@mpce.mq.edu.au

Abstract

This paper looks at representing paraphrases using the formalism of Synchronous TAGs; it looks particularly at comparisons with machine translation and the modifications it is necessary to make to Synchronous TAGs for paraphrasing. A more detailed version is in Dras (1997a).

1 Introduction

The context of the paraphrasing in this work is that of Reluctant Paraphrase (Dras, 1997b). In this framework, a paraphrase is a tool for modifying a text to fit a set of constraints like length or lexical density. As such, generally applicable paraphrases are appropriate, so syntactic paraphrases—paraphrases that can be represented in terms of a mapping between syntax trees describing each of the paraphrase alternatives—have been chosen for their general applicability. Three examples are:

- (1)
 - a. The salesman made an attempt to wear Steven down.
 - b. The salesman attempted to wear Steven down.
- (2)
 - a. The compere who put the contestant to the lie detector gained the cheers of the audience.
 - b. The compere put the contestant to the lie detector test. He gained the cheers of the audience.
- (3)
 - a. The smile broke his composure.
 - b. His composure was broken by the smile.

A possible approach for representing paraphrases is that of Chandrasekar *et al* (1996) in the context of text simplification. This involves a fairly straightforward representation, as the focus is on paraphrases which simplify sentences by breaking them apart. However, for purposes other than sentence simplification, where paraphrases like (1) are used, a more complex representation is needed.

A paraphrase representation can be thought of as comprising two parts—a representation for each of the source and target texts, and a representation for mapping between them. Tree Adjoining Grammars (TAGs) cover the first part: as a formalism for describing the syntactic aspects of text, they have a number of desirable features. The properties of the formalism are well established (Joshi *et al*, 1975), and the research has also led to the development of a large standard grammar (XTAG Research Group, 1995), and a parser XTAG (Doran *et al*, 1994). Mapping between source and target texts is achieved by an extension to the TAG formalism known as Synchronous TAG, introduced by Shieber and Schabes (1990). Synchronous TAGs (STAGs) comprise a pair of trees plus links between nodes of the trees. The original paper of Shieber and Schabes proposed using STAGs to map from a syntactic to a semantic representation, while another paper by Abeillé (1990) proposed their use in machine translation. The use in machine translation is quite close to the use proposed here, hence the comparison in the following section; instead of mapping between possibly different trees in different languages, there is a mapping between trees in the same language with very different syntactic properties.

2 Paraphrasing with STAGs

Abeillé notes that the STAG formalism allows an explicit semantic representation to be avoided, mapping from syntax to syntax directly. This fits well with the syntactic paraphrases described in this paper; but it does not, as Abeillé also notes, preclude semantic-based mappings, with Shieber and Schabes constructing syntax-to-semantics mappings as the first demonstration of STAGs. Similarly, more semantically-based paraphrases are possible through an indirect application of STAGs to a semantic representation, and then back to the syntax.

One major difference between use in MT and paraphrase is in lexicalisation. The sorts of mappings that Abeillé deals with are lexically idiosyncratic: the English sentences *Kim likes Dale* and *Kim misses Dale*, while syntactically parallel and semantically fairly close, are translated to different

ken), which takes the past participle form. This sort of link is unnecessary when STAGs are used in MT, as the trees are lexicalised, and the information is shared in the transfer lexicon. Since, with paraphrasing, the transfer lexicon does not play such a role, the shared information is represented by this new type of link between the trees, where the links are labelled according to the information shared. Hence, node V_1 in the active tree has a TENSE link with node V_0 in the passive tree, where tense is the attribute in common; and a LEX link with node V_1 in the passive tree, where the lexeme is shared.³

3 Notation

In paraphrasing, the tree notation thus becomes fairly clumsy: as well as consuming a large amount of space (given the large derived trees), it fails to reflect the generality provided by the summary links. That is, it is not possible to define a mapping between two structures reflecting their common features if the structures are not, as is standard in STAG, entire elementary or derived trees. Therefore, a new and more compact notation is proposed to overcome these two disadvantages.

The new notation has three parts: the first part uniquely defines each tree of a synchronous tree pair; the second part describes, also uniquely, the nodes that will be part of the links; the third part links the trees via these nodes. So, let variables X and Y stand for any string of argument types acceptable in tree names; for example, X could be $nx1nx2$ and Y $n1$. Then, for example, the tree for (2a) can be defined as the adjunction of a $\beta N0nx0VX$ tree (generic relative clause tree, standing for, e.g., $\beta N0nx0Vnx1nx2$) into an $\alpha n0VY$ tree; the tree for (2b) can be defined as a conjoined S tree, having a parent S_m node and 2 child nodes $\alpha n0VX$ and $\alpha n0VY$.

Figure 3: Paraphrase with partial links

The second part of the notation requires picking out important nodes. The identification scheme

³The determination of a precise set of link labels is future work.

proposed here has a string comprising node labels with relations between them, signifying a relationship taken from the set {parent, child, left-sibling, right-sibling}, abbreviated {p, c, ls, rs}. The node NP_1 of the left-hand tree of Figure 2 can then be described by the string $NPpNPpS_r pNIL$; an associated mnemonic nickname might be $T_1 subjNP$.

The third part of the representation is then linking the nodes. Standard links are represented by an equal sign; other links are represented with the link type subscripted to the equal sign. Thus, for Figure 2, $T_1 subjNP = T_2 leftsubjNP$, where $T_2 leftsubjNP$ is $NPpS_r pS_m pNIL$ for the right-hand tree.

For a tabular representation using this notation, see Dras (1997a).

4 Conclusion

Synchronous TAGs are a useful representation for paraphrasing, the mapping between parallel texts of the same language which have different syntactic structure. A number of modifications need to be made, however, to properly capture the nature of paraphrases: the creation of a new type of summary link, to compensate for the increased importance of derived trees; the allowing of many-to-many links between trees; the creation of partial links, which allow some information to be shared; and a new notation which expresses the generality of paraphrasing.

References

- Abeillé, Anne, Y. Schabes and A. Joshi. 1990. Using Lexicalised Tags for Machine Translation. *Proc. of COLING90*, 1-6.
- Chandrasekar, R., C. Doran, B. Srinivas. 1996. Motivations and Methods for Text Simplification. *Proc. of COLING96*, 1041-1044.
- Doran, Christy, D. Egedi, B.A. Hockey, B. Srinivas and M. Zaidel. 1994. XTAG System - A Wide Coverage Grammar of English. *Proc. of COLING94*, 922-928.
- Dras, Mark. 1997a. Representing Paraphrases Using Synchronous Tree Adjoining Grammars. *1997 Australasian NLP Summer Workshop*, 17-24.
- Dras, Mark. 1997b. Reluctant Paraphrase: Textual Restructuring under an Optimisation Model. Submitted to *PACLING97*.
- Joshi, Aravind, L. Levy and M. Takahashi. 1975. Tree Adjunct Grammars. *J. of Computer and System Sciences*, 10(1).
- Shieber, Stuart and Y. Schabes. 1990. Synchronous Tree Adjoining Grammars. *Proc. of COLING90*, 253-258.
- XTAG Research Group. 1995. A Lexicalised Tree Adjoining Grammar for English. *Univ. of Pennsylvania Technical Report IRCS 95-03*.