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

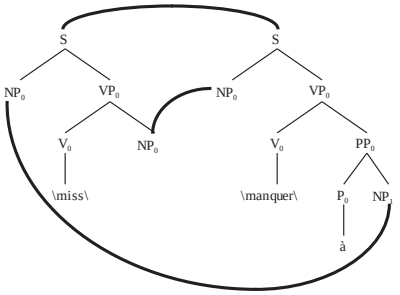

Figure 1: **STAGs:** *miss-manquer à*

syntactic structures in French; see Figure 1. The actual mappings depend on the properties of words, so any TAGs used in this synchronous manner will necessarily be lexicalised. Here, however, the sorts of paraphrases which are used are lexically general: splitting off a relative clause, as in (2), is not dependent on any lexical attribute of the sentence.

Related to this is that, at least between English and French, extensive syntactic mismatch is unusual, much of the difficulty in translation coming from lexical idiosyncrasies. A consequence for machine translation is that much of the synchronising of TAGs is between elementary trees. So, even with a more complex syntactic structure than the translation examples above, the changes can be described by composing mappings between elementary trees, or just in the transfer lexicon. Abeillé notes that there are occasions where it is necessary to replace an elementary tree by a derived tree; for example, in *Hopefully, John will work* becomes *On espère que Jean travaillera*, *hopefully* (an elementary tree) matches *on espère que* (derived).

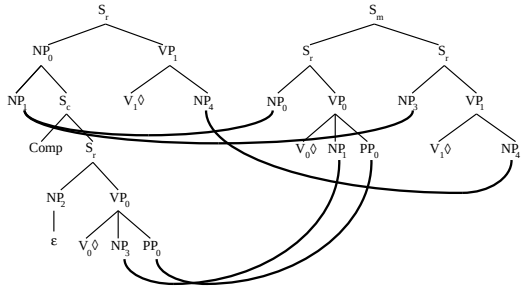

Figure 2: **Relative clause paraphrase**

The situation is more complex in paraphrasing: by definition, the mappings are between units of text with differing syntactic properties. For example, the mapping of examples (2a) and (2b) involves the pairing of two derived trees, as in Figure 2. In this case, both trees are derived ones. A problem with the STAG formalism in this situation is that it doesn't capture the generality of the mapping between (2a) and (2b); separate tree pairings will have to be made for verbs in the matrix clause which have complementation patterns different from that of the above examples; the same is true for verbs in the subordinate clause. For more complex matchings, the making and pairing of derived trees becomes combinatorially large.

A more compact definition is to have links, of a kind different from the standard STAG links, between nodes higher in the tree. In STAG, a link between two nodes specifies that any substitution or adjunction occurring at one node must be replicated at the other. This new proposed link would be a summary link indicating the synchronisation of an entire subtree: more precisely, each subnode of the node with the summary link is mapped to the corresponding node in the paired tree in a synchronous depth-first traversal of the subtree. Naturally, this can only be defined for pairs of nodes which have the same structure[1]; that is, in the context of paraphrasing, it is effectively a statement that the paired subtrees are identical. So, for example, a mapping between the nodes labelled VP_1 in each of the trees of the example described above would be an appropriate place to have such a summary link: by establishing a mapping between each subnode of VP_1, this covers different types of matrix clauses.

Another feature of using STAGs for paraphrasing is that the links are not necessarily one-to-one. In the right-hand tree of the Figure 2 pairing, the subject NPs of both sentences are linked to NP_1 of the left-hand tree; this is a statement that both resulting sentences have the same subject. This does not, however, change the properties in any significant way.[2]

It is also useful to add another type of link which is non-standard, in that it is not just a link between nodes at which adjunction and substitution occur, but which represents shared attributes. It connects nodes such as the main verb of each tree, and indicates that particular attributes are held in common. For example, mapping between active and passive voice versions of a sentence is represented by the tree in Figure 3. The verb in the active version of (3) (*broke*) shares the attribute of tense with the auxiliary verb $\backslash be\backslash$, and the lexical component is shared with the main verb of the passive tree (*bro-*

[1] More precisely, they need only have the same number and type of argument slots.

[2] This is equivalent to there being m dummy child nodes of the node at the multiple end of an m:1 link, each child node being exactly the same as the parent with fully re-entrant feature structures, with one link being systematically allocated to each child.

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 βN0nx0VX tree (generic relative clause tree, standing for, e.g., βN0nx0Vnx1nx2) into an αn0VY tree; the tree for (2b) can be defined as a conjoined S tree, having a parent S_m node and 2 child nodes αn0VX and αn0VY.

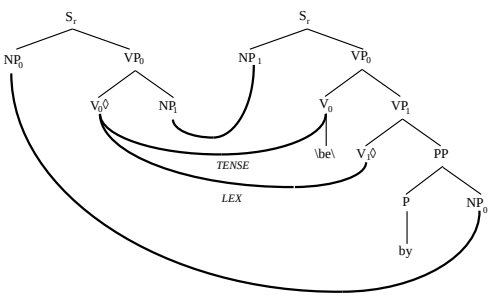

Figure 3: **Paraphrase with partial links**

The second part of the notation requires picking out important nodes. The identification scheme 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 $NPpNPpS_rpNIL$; an associated mnemonic nickname might be $T_1subjNP$.

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, $T_1subjNP$=$T_2leftsubjNP$, where $T_2leftsubjNP$ is $NPpS_rpS_mpNIL$ 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.*

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