Reluctant Paraphrase: Textual Restructuring under an Optimisation Model

Mark Dras
Language Technology Group, Microsoft Research Institute
School of MPCE, Macquarie University
Sydney NSW Australia 2109
markd@mpce.mq.edu.au

Abstract

This paper develops a computational model of paraphrase under which text modification is carried out reluctantly; that is, there are external constraints, such as length or readability, on an otherwise ideal text, and modifications to the text are necessary to ensure conformance to these constraints. This problem is analogous to a mathematical optimisation problem: the textual constraints can be described as a set of constraint equations, and the requirement for minimal change to the text can be expressed as a function to be minimised; so techniques from this domain can be used to solve the problem.

The work is done as part of a computational paraphrase system using the XTAG system [5] as a base. The paper will present a theoretical computational framework for working within the Reluctant Paraphrase paradigm: three types of textual constraints are specified, effects of paraphrase on text are described, and a model incorporating mathematical optimisation techniques is outlined.

1 Framework

The work this paper describes is done as part of a computational paraphrase system using the XTAG system [5] as a base. Although the goal of the system is to modify text to achieve some objective, it is fundamentally unlike existing systems which paraphrase text, such as style checkers [13], in that the context of paraphrasing is different; this context, Reluctant Paraphrasing, is described below, with a theoretical framework for the paraphrasing presented in the rest of the paper.

Reluctant Paraphrase (RP) can best be defined by contrasting it with the remedial sort of paraphrases suggested by style checkers, or in style guides such as Strunk and White [17], and so on. The starting point under this remedial style of paraphrase is an imperfect text which has to be corrected, the corrections being determined by some prescriptive advice such as "make the text more active". The text is run through a style checker, or past an editor, and flaws of vocabulary or grammar or style are corrected. In contrast, imagine the completion of an ideal document: it says exactly what the author intends, and no more; every word captures all the nuances the author wants to convey. However, it has to be changed because of external constraints. These constraints might be the need to cut down an academic paper by one page for conference publication; or the need to make a technical document conform to house style readability requirements; or some combination of these or other sorts of external constraints. Thus, the text has to be paraphrased, albeit reluctantly, in order to meet these externally imposed constraints.

Dealing with this reluctant sort of paraphrase, rather than the remedial sort, has a number of advantages. Firstly, it avoids representational problems that are otherwise inherent in paraphrasing. In remedial paraphrasing, paraphrase requirements can be of arbitrary complexity, ranging from "change sentence voice" to "fix incoherent theme". This arbitrariness of complexity makes developing a consistent representation near impossible. However, under RP the paraphrases don't embody the correction in the same way that remedial paraphrases do; instead, they are just tools which are used to alter the text so that it conforms to the imposed constraints. Given that the paraphrases are just tools, it is possible to pick a limited set of them and still attempt to cover all of them with a consistent representation.

Secondly, it avoids the debate about making text 'better'. There are longstanding arguments in the literature about particular techniques and their efficacy in improving text: examples are the passive to active voice paraphrase, relative pronoun deletion and the avoidance of nominalisation. In RP, by contrast, taking the standpoint that the original text is ideal means that any change will be undesirable, so only the minimal level of change to the text in keeping with the constraint satisfaction should be made.

The computational paraphrase system within RP that this paper discusses thus has three components: a set of paraphrase techniques which is used to achieve the text modification; a set of constraints to which the text must adhere after the modification; and an effect—that of the change to the text caused by the paraphrases applied—which is to be minimised. This parallels closely a mathematical optimisation model, with, respectively, a set of decision variables, a set of constraint equations and an optimisation function. The rest of this paper presents a formulation of RP which draws on ideas from the field of mathematical optimisation: Section 2 discusses numeric constraints on text; Section 3 looks at quantifying text effects of paraphrases; and Section 4 describes the actual model.

2 Textual Constraints

This section describes three measures of text, those of length, readability and lexical density. These measures are often used in the production of text; their numeric quality is what makes them particularly amenable to the optimisation model of this paper.

Length is the simplest measure, and is frequently used in practice as a constraint. For example, restricting the length of a text is standard for academic conferences—like this conference with its 3000 word limit on abstracts—and meeting this constraint often involves cutting down a longer draft version. It is also typical in other areas such as the editing of newspaper text [2]. Constraining text length is also a feature of computational language generation systems, either as a general directive implementing the Gricean maxim of conciseness, as in the Epicure system [4], or as an explicit limit on the length of an individual text unit, as in the STREAK system [16].

Another common measure comes from readability formulae, such as the Flesch Reading Ease Score or the Dale-Chall formula [15]. Standard readability formulae are basically equations which attempt to *predict*, rather than evaluate, the readability of text; in form they are generally linear combinations of factors which correlate with text complexity. These factors are of fairly simple types: a measure of sentence complexity, usually average sentence length; and a measure of word complexity, such as average word length in syllables, or proportion of infrequent words. The weightings for these terms are assigned by calculating a correlation with tests of readers' comprehension.

The most accurate way of determining readability would be by testing readers' comprehension directly. However, this would be expensive in terms of time and other resources; readability formulae were constructed as an attempt to predict the readability that would be measured by these tests. This, together with the numerical phrasing of the readability, is the reason for using readability formulae here. Moreover, the faults of readability formulae—documented in, for example, [7]—are not significant in the context of RP, for a number of reasons.

Firstly, use of readability formulae can be defended on practical grounds: readability formulae are used as criteria for writing public documents in the US, such as insurance policies, tax forms, contracts and jury instructions [3], for producing military documents [14], and so on. In these situations the use of readability formulae is mandatory; so for a system which models realistic constraints on text, using the formulae as a constraint is reasonable.

Secondly, most objections are based on the use of readability formulae in the strong sense—when actual readability levels are predicted—rather than when used in their weak sense—when readability formulae are used to rank texts relative to each other in order of reading complexity [12]; and under Reluctant Paraphrase, this is not a problem, as the texts, one of which is a paraphrase of the other, are just ranked relative to each other.

Lexical density is a textual measure discussed by Halliday [10]; it attempts to capture the 'condensedness' of text by measuring the proportion of non-content (or function) words to total text. Halliday uses this idea of condensedness to distinguish between written and spoken forms of language: written language tends to be more condensed than spoken, with constructions of type (1a) more prevalent in writing and those of type (1b) more prevalent in speech.

(1) a. Sex determination varies in different organisms.

b. The way sex is determined varies in different organisms.

The concept is also useful in the context of this paper's optimisation model, as a constraint counterbalancing the readability one. Under a typical readability formula, the readability value is generally correlated with average sentence length, so the formula value can be improved by the sort of paraphrases which compress text, such as the mapping of (1b) to

(1a). Compression to too great an extent can lead to text that is difficult to understand; the use of lexical density as a constraint can act as a counterweight to the readability constraint, to prevent excessive text compression.

3 Paraphrases

As noted in Section 1, paraphrases can be of arbitrary complexity. In keeping with their use in RP as broad-coverage tools, the most appropriate paraphrases, and hence the ones that are used in this work, are ones that are syntactic in nature. An example of this type, modelled on work by Jordan [11], is the splitting off of a noun post-modifier to form a separate sentence:

(2) a. Sarah warily eyed the page filled with topicalisations and other linguistic phenomena.

b. Sarah warily eyed the page. It was filled with topicalisations and other linguistic phenomena.

The paraphrases used here are taken from three different types of sources: popular (style guides such as [17]); academic (work on textual analysis involving paraphrasing, such as [11] and [16]); and practical (the actual practices of people involved in paraphrasing text, such as editors and journalists [2]).

These paraphrases will cause some change to the text, and, under RP, any change effected by a paraphrase is taken to be a negative one. Developing an optimisation model thus requires a quantification of the effects that imposing a paraphrase on a text will have on that text. The rest of this section sketches methods for assigning a quantification to a paraphrase, which will lead to a minimisation function for the model. There are two types of effects analysed in this work, effects on meaning and effects on discourse structure. These two types are then combined to give the minimisation function.

3.1 Meaning Effects

One way in which a paraphrase can affect a text is in terms of its truth-conditional meaning; or, in Hallidayan terms, its ideational metafunction. A unit of text, such as a sentence, can be viewed as a statement about the world, which is either true or false[1]; an alternative, but related, view is that the truth of the statement is represented by a set of possible worlds in which the statement is true[2]. A paraphrase is consequently defined more precisely as consisting of two sentences where the set of possible worlds in which one sentence is true is a (not necessarily proper) subset of the possible worlds in which the other is true. Take the following examples:

(3) a. Onlookers scrambled to avoid the car which was flashing its headlights.

b. Onlookers scrambled to avoid the car flashing its headlights.

(4) a. The salesman made an attempt to wear Steven down.

b. The salesman attempted to wear Steven down.

(5) a. There was a girl standing in the corner.

b. There was a girl in the corner.

(6) a. Tempeste approached Blade, a midnight dark and powerful figure, and gave him a resounding slap.

b. Tempeste approached Blade and gave him a resounding slap.

These examples give a range of different magnitudes in the size of the sets representing the possible worlds in which each of the paraphrase alternatives is true. Example (3) represents a fairly minimal difference: (3b) can be a paraphrase either of (3a) or of *Onlookers scrambled to avoid the car which* is *flashing its headlights*. The possible worlds in which (3b) is true is a proper superset of the possible worlds in which (3a) is true; but intuition suggests the sets are relatively close in size, (3b) only covering two different cases with respect to the altered constituents. Example (4) represents a slightly bigger paraphrase: (4b) can paraphrase statements asserting one attempt—equivalent to (4a)—two attempts, seven attempts, or many attempts. The size of the set difference here is consequently relatively larger than in (3). In (5), the difference is larger still, in that (5b) can describe situations where the girl is sitting, lying, dancing, and so on. The largest difference is in (6), where (6b) includes in its set of possible worlds, over and above the possible worlds in which (6a) is true, worlds in which Blade is described by any other appositive.

A way of approximating the intuition about the difference in the relative sizes of possible world sets is by using parts of speech. An alteration in less significant parts of speech corresponds to a small relative difference in set size, and so on. So in (3), the changed parts of speech are a relative pronoun, which causes no difference in truth-conditional meaning, and the auxiliary verb *be*, which leads to the relatively small

[1] Only declarative sentences are dealt with in this paper.

[2] This is a much simplified summary of work on truth-conditional meaning presented in, for example, [1].

difference. In comparison, the deletion of the open-class constituent in (5), the present participle *standing*, leads to a much greater set difference; and deleting multiple open-class words in (6) has a still larger effect.

A possible refinement of this approximation involves considering lexical factors. For example, the paraphrase in (5) is less significant than if (5a) had been *the girl coruscating in the corner*; the latter option is much more unexpected, and so it can be argued that its removal alters the text to a much greater extent. As they are related to frequency, these lexical factors could be estimated through collocational analysis within a corpus, although this has not been done as yet.

3.2 Discourse Effects

As well as affecting the truth-conditional meaning of the text, a paraphrase can alter the discourse features of the text; or, in Hallidayan terms again, the textual metafunction. Because of the assumption behind RP that the author has deliberately chosen a particular way of packaging the information in a sentence, any paraphrase which alters the packaging structure is altering the author's intention and hence should be included in the measurement of change and the consequent minimisation function. Work in the area of information packaging includes [8], [9] and [18]; although approaches differ, all have some concept of syntactic structures reflecting packaging of information—which part is known to the reader, and which is new. An example is an *it*-cleft sentence and its standard declarative paraphrase:

(7) a. It was the balcony and its scholarly discourse which irresistibly drew Ryan.

b. The balcony and its scholarly discourse irresistibly drew Ryan.

In (7a), the fact that Ryan has been irresistibly drawn is indicated as a given or topic, and the balcony-as-drawer as the new piece of information. In (7b) there is no such marking.

A rough numerical measure of this can be gained by counting the difference in the questions to which the sentence can be an answer. So (7a) can only be an answer to the narrow-focus *What irresistibly drew Ryan?*, while (7b) can answer not only this question but also *What did the balcony and its scholarly discourse do to Ryan?*, *What did the balcony and its scholarly discourse do?*, or the wide-focus *What happened?*.

4 An Optimisation Approach

The optimisation model for the computational paraphrase system requires a formal specification of the paraphrases and their attributes—their effect on the text in terms of the parameters, such as number of words or sentences, affected by each constraint; and their effect on the text's meaning and information structure. The paraphrases are formally specified using the representation formalism as proposed in [6]; however, an informal description of the paraphrase is adequate for discussion of the paraphrase effects and their inclusion into the optimisation model.

This section presents a mathematical optimisation model of paraphrasing. The basic techniques are those of integer programming (see, for example, [19]), which describes the constraints and function to be minimised in terms of linear combinations of integer variables. The integer programming approach is useful because it provides a set of techniques for guaranteeing an optimal solution, heuristics for cutting the search space, and methods for model analysis[3]. After a formal presentation of the model, an example is given for clarification.

4.1 The Model

In developing an optimisation model, it is first necessary to identify the DECISION VARIABLES: that is, those factors about which a decision is to be made. In this case, it is the paraphrase mappings: for each paraphrase, the decision is whether this paraphrase should be applied to the text to move it towards satisfying the constraints while minimally perturbing the text. In this situation, the choice is binary, whether or not to apply the paraphrase. Given this, the decision variables are

p_{ij} = a 0/1 valued variable representing the jth potential paraphrase for sentence i

The OBJECTIVE FUNCTION, the function to be optimised, is, for RP, a measure of the change to the text, as described in Section 3. With c_{ij} being the effect (or cost) of each paraphrase, if applied, this function has the form

$$z = \sum c_{ij}.p_{ij}$$

The constraints take the form "total length must be decreased by at least some constant value", or "readability value must be no greater than some constant value". Expressed mathematically, the length constraint is

[3]This last feature is not discussed in this paper.

$$\sum w_{ij}.p_{ij} \leq k_1$$

where

w_{ij} = change to length of sentence i caused by paraphrase ij
k_1 = required change to the length of text in words; $k_1 \leq 0$

A simplified readability constraint[4], using only the average sentence length component, is

$$\frac{W + \sum w_{ij}.p_{ij}}{S + \sum s_{ij}.p_{ij}} \leq k_2$$

that is,

$$\sum (w_{ij} - k_2.s_{ij})p_{ij} \leq k_2 S - W$$

where

s_{ij} = change to number of sentences in the text by paraphrase ij
W = total number of words in original text
S = total number of sentences in original text
k_2 = required average sentence length[5]; $k_2 \geq 0$

The lexical density constraint requires the proportion of function words, taken here to be all closed class words, to total words to be greater than some constant value. It has the form

$$\frac{F + \sum f_{ij}.p_{ij}}{W + \sum w_{ij}.p_{ij}} \geq k_3$$

that is,

$$\sum (f_{ij} - k_3.w_{ij})p_{ij} \geq k_3 W - F$$

where

f_{ij} = change to number of function words caused by paraphrase ij
F = total number of function words in original text
k_3 = required proportion of function words to total words; $0 \leq k_3 \leq 1$

Given that there are j paraphrases for each sentence (with j varying for each sentence), there is a potential conflict for the paraphrases. To simplify the application of the paraphrases, an extra constraint is added, stating that there can be at most one paraphrase for each sentence:

$$\sum_j p_{ij} \leq 1$$

Although it is possible in particular cases for paraphrases to overlap and produce satisfactory text, there is no easy way in advance to decide this; so for an automated system the above constraint is necessary, at least until a much more detailed analysis of paraphrase interaction has been carried out.

An example is presented in the next section, to illustrate the model. The small size of this example does not allow a real demonstration of the usefulness of the approach, since the problem can be solved almost by inspection. However, in larger problems this method of modelling allows the use of techniques such as branch-and-bound [19] which make the solution of the problem feasible, where the solution would otherwise be impractical because of the problem's exponential complexity.

4.2 An Example

As an example, take the short text:

(8) a. The cat sat on the mat which was by the door.
b. It ate the cream ladled out by its owner.
c. The owner, an eminent engineer, had a convertible used in a bank robbery.

The values of F, W and S are 17, 33 and 3 respectively.

Possible paraphrases of individual sentences, using just relative pronoun deletion, post-modifier split, and parenthetical deletion, are:

(9) p_{11}. The cat sat on the mat by the door.
p_{21}. It ate the cream. It had been ladled out by its owner.
p_{31}. The owner, an eminent engineer, had a convertible. It had been used in a bank robbery.
p_{32}. The owner had a convertible used in a bank robbery.

[4]This simplification means that non-linear, quadratic programming techniques do not have to be introduced at this stage.

[5]While the choice of a particular k_1 is straightforward, choosing a reasonable value for k_2 requires more effort: for example, analysing average sentence length in a corpus which satisfies typical readability targets (such as "senior high school level" in the Flesch Reading Ease score). The constant k_3 can be ascertained similarly.

	number of words	avg sent. length
original text	1791	24.88
num. words minimised	1531	23.92
avg sent. minimised	1784	17.66

Table 2: **Maximal text flexibility**

paraphrase ij	f_{ij}	w_{ij}	s_{ij}
11	-2	-2	0
21	+3	+3	+1
31	+3	+3	+1
32	-1	-3	0

Table 1: **Variable coefficients**

This gives decision variables p_{11}, p_{21}, p_{31}, and p_{32}, with associated coefficients in Table 1.

For the example, the constraint values are (arbitrarily) chosen as $k_1 = 0$ (at worst no compression of text length), $k_2 = 10$ (average sentence length no greater than 10), and $k_3 = 0.525$ (function words no less than 52.5% of the text).

Through the process of integer programming, there are two alternatives which are feasible solutions:

$$p_{11} = p_{21} = p_{31} = 0, p_{32} = 1$$

$$p_{31} = 0, p_{11} = p_{21} = p_{32} = 1$$

This gives two values for the objective function, $z = c_{32}$ and $z = c_{11} + c_{21} + c_{32}$. Since $\forall(ij)c_{ij} > 0$—under the Reluctant Paraphrase assumption all changes involve a positive cost—the best alternative is the first, with only the second paraphrase for sentence number three being applied. The resulting text is then:

(10) a. The cat sat on the mat which was by the door.

b. It ate the cream ladled out by its owner.

c. The owner had a convertible used in a bank robbery.

4.3 Actual Text

Current work involves applying this technique to actual text, taken from the periodical *The Atlantic Monthly*. This source was chosen as it has reasonably complex text on which a large range of paraphrases can be applied. The text consists of 72 sentences and totals 1791 words; there are 84 possible paraphrases, over 45 of the sentences.

In order to determine possible constraint values for real text, it is first necessary to evaluate the flexibility of the text: to what extent can the length be altered, say, or the readability changed? Choosing sets of paraphrases which maximise the relevant constraint, regardless of the value of the cost function or the effect on other constraints, the results given in Table 2 were obtained.

So at best it is possible, for this text, to reduce word length by about 15%, and the average sentence length by about 30%. This information is then used to set reasonable constraint limits.

One way in which the task of applying the model to actual text is more complicated than the example is in the need to set numeric values for the objective function coefficients. In the example, because of the small number of objective function coefficients, it is generally possible to just compare the result of the function algebraically. Taking as a first attempt at a numeric objection function the assignment of constant differences between the classes of textual change described in Section 3, the approach was applied to the first 19 sentences of the *Atlantic Monthly* text. Modelling the problem as a optimisation one, combined with branch-and-bound techniques, reduced the search space by 41.5% from 2^{19} possible solutions to 306828 candidates.

5 Conclusion

The paper has drawn on diverse areas of linguistics and mathematics to present a nonetheless fairly natural view of paraphrase as a mathematical optimisation problem. This phrasing of paraphrase as an optimisation problem has three main components. Firstly, three appropriate constraints have been chosen and modelled as constraint equations. Secondly, a method for quantifying the effects of paraphrase on text, and their expression as an optimisation objective function, has been discussed. Thirdly, the model has been described with an application to a small example text given. Application to actual text has shown the extent to which the technique can be applied: for example, the length constraint is not meant to mimic summarisation, but rather to enable the massaging of a text that is not too far from what

is required.

Current work involves a deeper application of the model to actual text: a larger number of constraints, more paraphrases, and an objective function which can be numerically evaluated. This then enables an analysis of text using the sensitivity analysis which is a corollary of linear programming, answering questions such as:

- What are the characteristics of ELASTIC text, that is, one which responds a lot to small changes?
- What is the sensitivity of text to changes in model assumptions, and would the same paraphrases be chosen given these changes?
- What are the equivalence classes for the paraphrases used, that is, which paraphrases are in effect interchangeable?

References

[1] Allwood, Jens, Lars-Gunnar Andersson and Östen Dahl. 1977. *Logic in Linguistics*. Cambridge University Press. Cambridge, UK.

[2] Bell, Allan. 1991. *The Language of News Media*. Blackwell. Oxford, UK.

[3] Bruce, Bertram, Andee Rubin and Kathleen Starr. 1981. Why Readability Formulas Fail. *IEEE Transactions on Professional Communication*, 24(1), 50-52.

[4] Dale, Robert. 1992. *Generating Referring Expressions*. MIT Press, Cambridge, MA.

[5] Doran, Christy, D. Egedi, B.A. Hockey, B. Srinivas and M. Zaidel. 1994. XTAG System - A Wide Coverage Grammar of English. *Proceedings of COLING94*, 922-928.

[6] Dras, Mark. 1997. Representing Paraphrases Using Synchronous Tree Adjoining Grammars. *Australasian Natural Language Processing Summer Workshop*, 17-24.

[7] Duffy, Thomas 1985. Readability Formulas: What's the Use? In Duffy, T. and R. Waller (eds), *Designing Usable Texts*. Academic Press. Orlando, FL.

[8] Givón, Talmy. 1979. *On understanding grammar*. Academic Press. New York, NY.

[9] Halliday, Michael. 1985. *An Introduction to Functional Grammar*. Edward Arnold. London, UK.

[10] Halliday, Michael. 1985. *Spoken and Written Language*. Oxford University Press. Oxford, UK.

[11] Jordan, Michael. 1994. Toward Plain Language: A Guide to Paraphrasing Complex Noun Phrases. *The Journal of Technical Writing and Communication*, 24(1), 77-96.

[12] Kern, R. 1979. *Usefulness of readability formulas for achieving Army readability objectives: research and state-of-the-art applied to the Army's problem*. Technical Advisory Service, US Army Research Institute. Fort Benjamin Harrison, IN.

[13] Kieras, David. 1990. *The Computerized Comprehensibility System Maintainer's Guide*. University of Michigan Technical Report no. 33.

[14] Kincaid, J. Peter, James Aagard, John O'Hara and Larry Cottrell. 1981. Computer Readability Editing System. *IEEE Transactions on Professional Communication*, 24(1), 38-41.

[15] Klare, George. 1974-75. Assessing Readability. *Reading Research Quarterly, Number 1, 1974-1975*, 62-102.

[16] Robin, Jacques. 1994. *Revision-Based Generation of Natural Language Summaries Providing Historical Background*. Columbia University Technical Report CUCS-034-94.

[17] Strunk, William and E. B. White. 1979. *The Elements of Style, 3rd edition*. MacMillan Publishing Co. New York, NY.

[18] Vallduví, Enric. 1993. *Information Packaging: A Survey*. University of Edinburgh Technical Report HCRC/RP-44.

[19] Winston, Wayne. 1987. *Operations Research Applications and Algorithms*. PWS-Kent Publishing Company. Boston, MA.