Centering Theory for Evaluation of Coherence in Computer-Aided Summaries

Laura Hasler

Research Institute in Information and Language Processing
University of Wolverhampton, Stafford St., Wolverhampton, WV1 1SB, UK
L.Hasler@wlv.ac.uk

Abstract

This paper investigates a new evaluation method for assessing the coherence of computer-aided summaries, justified by the inappropriacy of existing evaluation methods for this task. It develops a metric for Centering Theory (CT), a theory of local coherence and salience, to measure coherence in pairs of extracts and abstracts produced in a computer-aided summarisation environment. 100 news text summaries (50 pairs of extracts and their corresponding abstracts) are analysed using CT and the metric is applied to obtain a score for each summary; the summary with the higher score out of a pair is considered more coherent. Human judgement is also obtained to allow a comparison with the CT evaluation to assess the validity of the development of CT as a useful evaluation metric in computer-aided summarisation.

1. Introduction

Computer-aided summarisation (CAS) is an alternative to fully automatic summarisation which accounts for the fact that fully automatic summaries are not always of a high standard of quality (Orasan et al., 2003). CAS combines methods from the fields of both human and automatic summarisation, allowing users of a system to access the output and post-edit it to improve the summary. Guidelines, which aim to consistently improve the coherence and readability of extracts produced by CAS systems, have been developed to help users of such systems (Hasler, 2007). Summaries produced using these guidelines need to be evaluated to prove that such a resource is indeed useful. However, because the final summaries are produced via a mixture of extraction and human post-editing, existing evaluation methods used to assess quality are unsuitable.

This paper investigates a new evaluation method for assessing the coherence of computer-aided summaries. The aim is to take a step in the right direction to find a more objective evaluation method for coherence than those currently available for this task. A metric is developed for Centering Theory (CT) (Grosz et al., 1995), a discourse theory of local coherence and salience, to measure the coherence of pairs of extracts and the abstracts created from them using summary production guidelines. 50 pairs of news text summaries are analysed using CT and the metric is applied to obtain a score for each summary; the summary with the higher score out of a pair is considered more coherent. Human judgement is also obtained, which allows a comparison between human and CT evaluations of the same texts to assess the validity of the development of CT as a useful evaluation metric in computer-aided summarisation.

The remainder of this paper is structured as follows. Section 2. introduces current evaluation methods for informativeness and quality in summarisation. Centering Theory is described in Section 3., and more specifically for its application to evaluate coherence in computer-aided summaries in Section 4. The results of an evaluation of computer-aided summaries are discussed in Section 5. and a comparison with human judgement is given in Section 6.

2. Evaluation in Summarisation

Evaluation is a vital issue in the field of automatic summarisation (AS). If summaries produced by automatic systems are not evaluated, there is no way of knowing how well they perform and consequently how useful they are. To date, the majority of evaluation in automatic summarisation has focused on the information content of summaries. This reflects the current preference for automatic extraction as opposed to abstraction, with relatively little consideration for issues of coherence and readability. Evaluation in AS can be split into two main strands: informativeness and quality (Hirschman and Mani, 2003).[1] Sections 2.1. and 2.2. below briefly describe existing evaluation methods in the field of automatic summarisation and explain why they are not wholly suitable for assessing coherence in computer-aided summaries.

The human summarisation literature does not address evaluation in the same way as the automatic summarisation literature does. Instead, the focus is on the checking and editing of abstracts to ensure they adhere to the conventions and specifications of the organisation or publication for which they are produced. Hasler (2007) established that it is not desirable to assess abstracts solely in terms of guidance for professional summarisers. This is too restrictive and would result in abstracts being negatively evaluated when they are considered acceptable to a human judge and a human summariser working with guidelines. A corpus analysis proved that conventions regarding, for example, tense and voice, are not always strictly adhered to because texts have idiosyncracies in the ways they realise information.

2.1. Evaluation of Informativeness in AS

Although not directly relevant to the evaluation discussed in the rest of this paper, it is important to mention established methods for the evaluation of informativeness in AS, as these have recently received most attention from the AS research community. Evaluation of informativeness assesses the information content of a summary in comparison with a reference text, either the source text or an ideal summary

[1] The evaluation literature also distinguishes *intrinsic* and *extrinsic* and *on-line* and *off-line* evaluation.

created by a human. It can involve human judgements on how well the units in a summary cover the content of the source or an ideal summary. Relevance assessment, reading comprehension and the usefulness of the summary in completing other tasks are other common means of evaluation.[2] Automatic evaluation has recently become popular, with methods such as ROUGE (Lin, 2004) and the Pyramid method (Nenkova and Passonneau, 2004; Harnly et al., 2005) being incorporated in the Document Understanding Conferences (DUC: http://duc.nist.gov/). Systems can also be automatically compared against annotated corpora used as a gold standard, or against source texts using measures such as precision, recall and f-measure, and the cosine similarity (see Donaway et al. (2000) for a comparison of recall- and content-based evaluation measures). However, all of these evaluation methods are concerned with informativeness and therefore it is unfair to employ them in the assessment of the coherence of summaries.

2.2. Evaluation of Quality in AS

Quality evaluates how well a summary reads, by taking into consideration style via phenomena such as dangling anaphors and connectives, discourse ruptures and grammaticality (e.g. Minel et al. (1997)). Saggion and Lapalme (2000) use criteria such as good spelling and grammar, impersonal style, clear indication of the topic of the source document, and conciseness, as criteria for human judges to grade summaries. The SEE tool (Lin, 2001) allows humans to manually assess extracts for a variety of quality and informativeness phenomena, including coverage, completeness and grammatical fluency. Because the texts used in the evaluation experiment in this paper comprise abstracts as well as extracts (see Section 4.2.), it is assumed that the abstract which was produced by human post-editing of an extract will fulfil these quality criteria due to the effect of guidelines used in the process. Therefore, it is inappropriate to use evaluation methods such as these which will almost always score the abstracts very well because the criteria are very similar to the guidelines used to produce them.

Standard readability measures such as the Gunning-fog index (Gunning, 1988) and the Flesch-Kincaid index (Kincaid et al., 1975), which assess ease of reading based on average word and sentence length, can also be used to evaluate the quality of summaries. However, these have been criticised as extremely coarse methods due to their simplicity (Mani, 2001): word and sentence length do not determine a 'good' summary, and do not give many insights into how or why one summary is of a higher quality than another. Indeed, in the texts used for this evaluation, there were many instances of *merging*, where sentences in an abstract were made longer than their corresponding ones in the extract so that the text did not seem as 'choppy'. Sentences were also lengthened to shorten the abstract overall by incorporating one sentence into another.

Post-edit measures based on the number of corrections necessary to transform the output of a system into an acceptable state are another option. Whilst these measures may seem suitable for the evaluation of changes made when transforming an extract into an abstract, they are impractical due to the complexity of the operations applied during the process. Edit distance between sentences was tried, but because there were so many complex changes between extracts and abstracts, especially due to the merging and reordering of information, this proved unsuitable.

[2] See Mani (2001) for an overview.

3. Centering Theory

As Centering Theory attempts to explain local coherence and salience within a discourse, it is a prime candidate with which to evaluate summaries, where both coherence and salience are issues and the text is short and usually only about one (or two) main topic(s). CT is a parametric theory which deals with coherence by examining repetitions of entities across consecutive utterances, and the relationship between these repetitions. The main concepts and assumptions introduced in the earliest versions of Centering Theory (Brennan et al., 1987; Grosz et al., 1995) are presented in this section. Whilst the most popular application in the past has been anaphora resolution, the suitability of CT variations for other tasks has been shown in recent years by its application in natural language generation (e.g. Karamanis (2003)) and automatic summarisation (Orasan, 2006). Hasler (2004) and Lapata and Barzilay (2005) also prove CT's usefulness in evaluation.

3.1. Centers

As CT is a theory of *local* coherence, only two consecutive utterances are considered at any one time (U_n and U_{n+1}). Each utterance in a text introduces a number of *forward looking centers* (*Cf*s), which are noun phrases (NPs) referring to an entity. These *Cf*s must be realised explicitly in the utterance. In addition, each utterance except the first has precisely one *backward looking center* (the *Cb*), which is the link between one utterance and the previous utterance in the text. A weaker version of this is also offered, asserting that each utterance has at most one *Cb* (e.g. Walker et al. (1998)). The *Cb* of any current utterance (U_{n+1}) is the most highly ranked *Cf* of the previous utterance (U_n) which is realised in the current utterance (U_{n+1}).

The *Cf*s are ranked, usually according to grammatical function (see Section 3.4.). The more highly ranked a *Cf*, the more likely it is, in a 'coherent' text, to be the *Cb* of the next utterance. The most highly ranked *Cf* of an utterance is the *preferred center* (*Cp*), so the theory predicts that the *Cp* of U_n is most likely to be the *Cb* of U_{n+1}. If an entity within an utterance is pronominalised, it is most likely to be the *Cb*. Table 1 summarises these basic notions of Centering Theory.

3.2. Transitions

The relationships between *Cf*s and *Cb*s of utterances result in transitions between utterances, which have a definite order of preference; texts demonstrating certain transitions are considered to be more coherent than those demonstrating others. In the original formulation of the theory, three types of transition are described: CONTINUE, RETAIN, SHIFT. However, following Brennan et al. (1987) and Walker et al. (1998), amongst others, this discussion splits the SHIFT transition into two: SMOOTH SHIFT

Constraint 1	Each utterance has precisely 1 *Cb* (Weak version: each utterance has at most 1 *Cb*)
Constraint 2	Every element of $Cf(U_n)$ must be realised in U_n
Constraint 3	$Cb(U_{n+1})$ is the highest-ranked element of $Cf(U_n)$ which is realised in U_{n+1}
Rule 1	If some element of $Cf(U_n)$ is realised as a pronoun in U_n, then so is $Cb(U_{n+1})$ (Strong version: if $Cb(U_{n+1}) = Cb(U_n)$, a pronoun should be used)
Rule 2	In transitions, CONTINUE is preferred over RETAIN, which is preferred over SMOOTH SHIFT, which is preferred over ROUGH SHIFT

Table 1: Centering Theory Constraints and Rules

and ROUGH SHIFT. CONTINUE is preferred over RETAIN, which is preferred over SMOOTH SHIFT, which is in turn preferred over ROUGH SHIFT. The ordering of transitions reflects the idea that it is preferable for consecutive utterances to have the same *Cb*, i.e., for the same entity to provide the link between two utterances, and also for the most salient entity (the *Cp*) in one utterance to be the *Cb* of the next utterance. Table 2 presents CT's transitions in terms of the relationship between *Cb*s and *Cp*s.

3.3. Coherence

The interaction and positioning of entities (*Cb*s and *Cf*s) in consecutive utterances, which is encoded by the transitions in the table above, helps to create the impression that a text is about the same entity (in terms of summarisation, the 'main topic'). Consider the following examples taken from Grosz et al. (1995), who argue that the same information is present in both, but that (1) is more coherent because it suggests that the discourse is about the same thing (John). The changes in the subject (or ranking) of each sentence in (2) make it difficult for the reader to decide whether this particular example is about John or the store. Examples such as these help to validate arguments for the ranking of the *Cf* list and of transitions. In terms of a CT analysis adhering to the weak version of Constraint 1, (1) displays CONTINUE transitions, because the subject of the first utterance is kept as the *Cp* throughout and is also the *Cb* in the subsequent utterances. (2), on the other hand, displays a RETAIN followed by a CONTINUE, followed by a RETAIN, due to the fact that the *Cb* is the same throughout the utterances, but it is not the same entity as the *Cp*.

(1) **John**[*Cp*] went to his favorite music store to buy a piano. **He**[*Cp*], [*Cb*] had frequented the store for many years. **He**[*Cp*],[*Cb*] was excited that he could finally buy a piano. **He**[*Cp*], [*Cb*] arrived just as the store was closing for the day.

(2) **John**[*Cp*] went to his favorite music store to buy a piano. **It**[*Cp*] was a store **John**[*Cb*] had frequented for many years. **He**[*Cp*], [*Cb*] was excited that he could finally buy a piano. **It**[*Cp*] was closing just as **John**[*Cb*] arrived.

3.4. Parameters

Centering Theory is a notoriously underspecified theory, and this underspecification has prompted a substantial body of research into optimal parameters depending on text type, language and tasks (see Poesio et al. (2004) for a comprehensive overview). There are a wide variety of possible instantiations of CT, and parameters need to be specified before the theory can be used. In earlier work (Grosz et al., 1995), even the most basic notion of *utterance* is not defined, although an utterance is often considered to be a sentence because it is the simplest option. This view has been criticised by some researchers, and using different types of clauses as utterances has been proposed as an alternative (e.g., Kameyama (1998)), although this too has its critics.

Realisation is another parameter. It is possible to have *direct* or *indirect* realisations of *Cf*s, *Cp*s and *Cb*s. Direct realisations must be coreferential, whilst indirect realisations encompass other relationships between entities, such as part-whole and set-membership. Direct realisation is easier to employ due to the high number of possibilities of indirect realisations of an entity. *Ranking* of the *Cf* list is a third parameter to be specified. This is traditionally taken to be a grammatical/linear preference order: subject $>$ object $>$ others, which is sometimes further split to distinguish direct and indirect objects: subject $>$ direct object $>$ indirect object $>$ others. However, it has been argued that this is inappropriate because the 'topic' of an utterance is not always the subject. Strube and Hahn (1999) propose *functional centering* where information structure is taken into account and ranking is based on hearer-old and hearer-new information.

4. Centering Theory for Evaluation in CAS

This section discusses the instantiation of CT used for the evaluation of coherence in computer-aided summaries. A metric is formulated using CT to score summary coherence in 50 pairs of news text summaries.

4.1. Parameters for Summarisation

Section 3.4. discussed the parameters of CT that need to be specified before the theory can be applied. The specification of parameters adhered to in this paper is that which is most appropriate for summary evaluation. Due to the number of possible instantiations, it is best to start simply, addressing issues as and when they arise; there is no point in making the evaluation more complicated than necessary.[3]

The first parameter to be specified in the current instantiation is *utterance*. An utterance is considered to be a sentence, as that is the unit used to create the extracts analysed here so it would not be practical to equate an utterance with certain types of clause, for example. In addition, this enhances the generality of the evaluation method because it can be applied to summaries produced by a variety of methods and systems, which mainly operate at sentence level. It

[3]However, this is not to say that these parameters are the most appropriate for Centering Theory applications in general.

	$Cb(U_{n+1}) = Cb(U_n)$ or $Cb(U_n)$ undefined	$Cb(U_{n+1}) \neq Cb(U_n)$
$Cb(U_{n+1}) = Cp(U_{n+1})$	CONTINUE	SMOOTH SHIFT
$Cb(U_{n+1}) \neq Cp(U_{n+1})$	RETAIN	ROUGH SHIFT

Table 2: Centering Theory Transitions

is not feasible at this stage to use full indirect *realisation* between a *Cb* and a *Cf* as it is not easy to define what exactly indirect realisation could encompass. In addition, it is not desirable to branch too far from the 'main topic' of the text. However, the summaries used in this evaluation display a particular use of possessive pronouns. An entity such as a country, government or other 'organisation' is introduced in the first sentence, and then throughout the summary possessive pronouns are used to discuss slightly different aspects of this entity, but the entity itself could still be considered as the main topic, or very closely related to it. Therefore, direct realisation between entities, plus indirect realisation where possessive pronouns are involved, is used to identify the *Cb* of an utterance.

The grammatical/linear *ranking* of the *Cf* list for each utterance is employed: the subject of the sentence is most preferable, followed by the direct object, indirect object, and finally any other noun phrases in the sentence. Because sentences can be complex and do not always comprise only one clause, where there is a main and subordinate clause, the grammatical classes in the main clause appear higher in the *Cf* list, with those in the subordinate clause coming later. This reflects the assumption that the most important information in a sentence tends to be presented in its main clause. In a summary, the most important information from the source text is retained; a corpus analysis showed that there are fewer instances of main clause deletion than the deletion of subordinate clauses when post-editing extracts. This suggests that the most relevant information does indeed appear in main clauses. Where there is more than one clause of the same type in a sentence, the *Cf* list ranking reflects the linear organisation of the sentence.

The *weak version of Constraint 1* (each utterance has at most one *Cb*) is used, which allows two utterances to display a NO TRANSITION between them, in cases where there is no *Cb*. The instances of NO TRANSITION are split into two groups: NO TRANSITION (NO *Cb*), where there is no entity at all in common with a consecutive utterance, and NO TRANSITION (INDIRECT), which are due to the indirect realisation of an entity. Although NO CB is cited as a common transition, representing on average 20% of transitions found in various corpora (Karamanis et al., 2004), in terms of summaries it is the most damaging transition. In this type of short text, which should have one or perhaps two main topics over an average of 6 sentences, for even one pair of utterances to not have an entity in common has a negative effect in terms of coherence and readability.

4.2. Texts for CT Evaluation

The CT evaluation uses summaries produced both by humans and automatically. Twenty two human-produced extracts of news texts were taken from the CAST corpus (Hasler et al., 2003). A further 3 texts from New Scientist which had previously been annotated for summarisation in another project were added to make 25 in total. The source texts of the 25 human-produced extracts were fed into the CAST system (Orasan et al., 2003), which produced 30% automatic extracts of them using the *term weighting* method. Previous experiments showed that a professional summariser using CAST selected this method to produce extracts for post-editing (Orasan and Hasler, 2006). This means that the automatic extracts used in the evaluation are more likely to be similar to those which a user of a CAS system would work with. These extracts were then transformed into abstracts by a human summariser using a set of guidelines. All extracts adhere to a 30% compression rate, and all abstracts to 20%, based on experiments with a professional summariser (Orasan and Hasler, 2006). As a result, 100 summaries in total were analysed using Centering Theory: 25 human-produced extracts, 25 abstracts corresponding to these human-produced extracts, 25 automatically produced extracts, and 25 abstracts corresponding to these automatically produced extracts.

Using the parameters discussed in Section 4.1., these 100 summaries were analysed using Centering Theory. In order to perform a CT analysis, first of all, a summary was split into its constituent utterances and *Cf*s of the utterances identified. Based on grammatical ranking, the *Cp* of each utterance was marked and the *Cb* for each utterance (where present) was established. Transitions between consecutive utterances were assigned based on the relationship between the *Cb* and *Cp*. To illustrate the task, the following abstract from the evaluation texts is analysed using CT. Utterances are marked with {curly brackets}, *Cf*s with (normal brackets), and *Cp*s and *Cb*s are indicated by [square brackets] and are highlighted in **bold**. The transition holding between two utterances is indicated in between each pair.

U_1: {**(Everybody)**[*Cp*] should be ready for ((Monday)'s national championship game), despite (casualties in ((Saturday night)'s NCAA semifinal battles)).}

NO TRANSITION (INDIRECT)

U_2: {**(Jason Terry of (Arizona))**[*Cp*], [*Cb*] was injured.}

RETAIN

U_3: {"**(We)**[*Cp*] were going to put **(him)**[*Cb*] in late in (the game)," said (Arizona coach (Lute Olson)).}

ROUGH SHIFT

U_4: {"**(He)**[*Cp*] had played a lot before (that), of course, but when (we)'re protecting (a lead), **(we)**[*Cb*] like getting (four perimeter guys) in there and that gives (us) (another ball handler), gives (us) (another free throw shooter)."}

RETAIN

U_5: {**(Kentucky coach (Rick Pitino))**[*Cp*] predicted that ((Monday)'s championship game) would be also be physical, in view of ((Kentucky)'s all-out pressure defence) and ((**Arizona**)[*Cb*]'s blazing speed)).}

4.3. Evaluation Metric

For the evaluation to accurately reflect the relation between Centering Theory transitions and coherence and summarisation, a metric needs to be formulated which represents the positive and negative effects of the presence of certain transitions in summaries. This idea of formulating a metric for CT is not novel, but the development of a summary-motivated CT metric is. To reward those transitions which add to the coherence of a summary and to penalise those which negatively affect it, weights are assigned to each type of transition. The traditional order of preference for transitions is kept in this evaluation, although NO TRANSITIONs are treated differently (see Section 4.1.).

Although the presence of a NO CB in other text types can indicate a new discourse segment and therefore not be considered damaging for coherence, summaries are necessarily shorter than full texts. Due to the nature of a summary, summary sentences should generally be about the same topic. A sentence which does not contain even an indirect mention of an entity which has been repeated in other utterances throughout is detrimental to the flow of the text as the reader has to stop and work out why that particular sentence is included. In such a short text, the presence of even one or two utterances which do not have an entity in common is noticeable and affects coherence negatively. A preliminary investigation of the evaluation texts indicated that parts of extracts which are viewed intuitively as less coherent can be attributed to the presence of a NO TRANSITION (NO *Cb*). Therefore, in terms of summarisation, NO TRANSITION (NO *Cb*) is the most damaging to coherence and is weighted accordingly. NO TRANSITION (INDIRECT) does not damage a summary's coherence to the same extent because the reader can easily infer that there is some kind of relationship between two entities.

To obtain the average transition score per summary, the weights for each transition identified are added, and then divided by the number of transitions, which is the total number of utterances-1. Table 3 shows the scores assigned to each transition, based on their relation to coherence. It should be noted that the numbers themselves assigned to the transitions are subjective. The most important part of the metric is the difference between the scores for each transition because that represents how positive or negative the effect of a particular transition is on a summary.

Transition	Weight
CONTINUE	+3
RETAIN	+2
NO TRANSITION (INDIRECT)	+1
SMOOTH SHIFT	-1
ROUGH SHIFT	-2
NO TRANSITION (NO *Cb*)	-5

Table 3: Transition Weights for Summary Evaluation

According to this metric, the abstract used in the example above (Section 4.2.) would receive an average transition score of 0.8, based on the scores for its transitions (1 NO TRANSITION (INDIRECT), 2 RETAIN, 1 ROUGH SHIFT) and the number of transitions present (4). This score would then need to be compared to that of its corresponding extract to allow a meaningful interpretation.

5. CT Evaluation of Summary Coherence

This section discusses the results of the CT evaluation carried out using the texts, parameters and metric described above (Section 4.). The evaluation maintains the distinction between *human-based* summaries (*set 1*) and *automatically-based* summaries (*set 2*).

5.1. Results

In total, CT evaluated 78% of abstracts as more coherent than the extracts from which they were produced. 2% (1 instance) of pairs were considered to be of equal coherence, leaving 20% of extracts evaluated as more coherent than the human post-edited abstracts. If only set 1 is examined, the abstracts are judged as better in 84% of cases, and the extracts in 16%. For set 2, extracts are considered more coherent than their corresponding abstracts 24% of the time, and 4% are evaluated as demonstrating equal coherence, leaving 72% of abstracts classed as more coherent than their extracts. Table 4 displays the normalised transition scores for all summaries. The next section provides some explanations for this performance of CT in evaluating coherence in computer-aided summaries.

	Set 1		Set 2	
Text	**Extract**	**Abstract**	**Extract**	**Abstract**
475968	2.4	2.5	0.6	1
475997	0.1	-2	-0.1	0.8
476016	-0.2	0.8	0.1	1
476032	-1	1.7	0.6	1
476038	-0.3	1.7	1	2
476040	-0.9	2.3	1	2.3
476052	-1.5	-2	0.7	2
476056	-0.7	2	0	0.3
476057	-0.6	1.7	0.8	1.3
476058	0.3	1	0.2	1.3
476059	0	0.8	-1	1.3
476062	0.2	1.5	1.3	1
476074	1.3	2	0	2
476086	0.9	2.3	0.3	0.6
476093	-0.8	0.2	0.2	0
476097	-1.4	2.7	3	3
476143	2.8	2.5	1.8	0
476183	2.2	2	0.7	1.5
476316	-0.7	0.3	1	0.7
476501	-1.1	0.8	0.2	1
476208	-0.3	2.8	0.5	0.3
476520	-1.7	-0.3	-0.3	1.5
sci01	-0.1	0	-0.7	-0.3
sci03	-0.2	3	-0.2	0
sci37	0.6	1.7	1.5	1.3
Total score	0	1.3	0.5	1.1

Table 4: CT Transition Scores for All Summaries

5.2. Discussion

There are several reasons for the set 2 abstracts results in comparison with set 1. First of all, CT considers representations of the same information (an entity in consecutive utterances). The sentences extracted from source texts automatically are more likely to be repetitive, i.e., to focus on exactly the same aspects of the same topic, due to the way that term-based summarisation works. Such automatic methods are not concerned with the way sentences fit together, or the ways in which they repeat information. If repetitive sentences are presented to a human summariser, they are very likely to change them to make them more readable, which will involve some modification of the information or sentences to try to avoid repetition, as advised in their guidelines. However, the resulting abstract is considered less coherent by CT when such changes affect the ranking of elements within the sentences, or delete some of these altogether. If the same sentence is taken as the starting point in the extract and in the abstract, it will mean that the abstract is less coherent because the links in subsequent sentences to the initial ordering are disrupted. This illustrates the fact that information content cannot be completely divorced from issues of readability and coherence, which is emphasised by discussions with a human judge who evaluated the coherence and readability of the summaries using intuitive judgement (see Section 6.).

Related to these repetitive sentences is the presence of a 'headline' in automatically produced extracts. This headline can be repeated, once with a location and once without, for example:

RUSSIA: Threats get Russians to pay tax - but not much.
Threats get Russians to pay tax - but not much.

This would be analysed by CT as displaying the transition CONTINUE, although it is obviously not desirable to have such repetition in any kind of summary. This increased the number of CONTINUE and RETAIN transitions in automatic extracts (allowed because of the weak version of Constraint 1). If the same sentence is repeated immediately after its first instance, the utterances will be deemed optimally coherent due to the fact that the same entities are mentioned in exactly the same position. Because the transition weights in the evaluation metric reward the more coherent transitions, the average transition score is higher than it might have been had one of these headlines not been included. This should be taken into account in future uses of CT in evaluation of automatically produced summaries.

The third reason for these results is that the transformation of automatic extracts into abstracts was not as simple as that of human-produced extracts. Where human annotators have the option to indicate sentences which contain information vital to the full understanding of a sentence, such as the antecedent of a pronoun, this is not the case with extracts produced automatically. During the transformations of automatic extracts into abstracts, the source was accessed on a number of occasions in order to resolve a pronoun and replace it with the full NP to make the text understandable. However, improvements such as these are not reflected in the CT evaluation, because it does not consider that such operations may have to be applied, being developed for whole discourses rather than ones which are produced from parts of a whole. Issues such as whether a pronoun has an antecedent or not, and whether this has to be remedied to make the text coherent from the point of view of a human reader, are simply not addressed by the theory because it was not designed to deal with such issues. In these cases, the same transition will hold between utterances regardless of whether the entity was originally an unresolved pronoun or a full NP mention.

6. CT Evaluation vs Human Judgement

Because the evaluation using CT is a new method, not having been used in this form for this task before, human judgements about the summaries' coherence and readability are also obtained in an attempt to validate the evaluation method. Human readers are the ultimate users of computer-aided summaries which have been created by post-editing automatically produced extracts and so their judgement matters. In addition, being based on entity repetition, CT does not take into account aspects of extracts and abstracts such as the use of connectives to signal relations between units or the restructuring of NPs to avoid repetition. Other limitations of CT for the evaluation of summary coherence were discussed in Section 5.2. The human judge was asked to intuitively select the more readable and coherent summary out of an extract/abstract pair, but was not told which text was which.

6.1. Results of Human Judgement

In total, 82% of abstracts were judged to be more readable/coherent than the extracts from which they were produced. The human judge expressed complete uncertainty about one pair (2%), leaving 16% of extracts evaluated as more coherent than their corresponding abstracts. Similar to the CT evaluation of coherence, examining the two sets separately demonstrates a better evaluation of set 1 than set 2. Set 1 abstracts are judged as better in more cases: 92% of the time. Only 2 human-produced extracts (8%) were judged as more readable/coherent than the abstracts created by post-editing them. Automatic extracts are evaluated as better than their corresponding abstracts in 24% of cases, and there is one case of complete uncertainty (4%), meaning that 72% of the set 2 abstracts are considered to be more readable/coherent.

Automatic extracts do not always focus on the main topic or the same aspects of it, and in certain cases the human judge preferred summaries which gave information about various topics or aspects rather than focusing on the main topic. Where operations had been applied to delete units about different topics or different aspects of the same topic in the abstract, the extract contained different information. The judge commented that it was impossible to assess the summaries on readability and coherence alone, and that the information present in the summary was an important element in the evaluation. This is related to the discussion of the CT evaluation where it was pointed out that information content and coherence/readability cannot be completely separated.

6.2. Agreement between CT and Human Judge

In terms of agreement between the CT evaluation and the evaluation by the human judge, their total agreement was 70%, and total disagreement was 26%. Two cases (4%) were unable to be compared because in one case the human judge could not select the better text, and in the other case, CT evaluated the abstract and the extract as exactly the same in terms of coherence. As with the other results, the evaluation of set 1 was better on the whole than that of set 2. The human judge and CT agreed in 76% of cases in set 1, and disagreed in 24%. The summaries which could not be compared belong to the automatic group, constituting 8% of cases. However, in the instance of the CT evaluation of equal coherence, the human judge commented that it was extremely difficult to decide on a better summary because they were so similar, which is in keeping with the CT evaluation. For set 2, the human judge and CT agreed in 64% of cases and disagreed 28% of the time.

To establish whether the disagreement between the human judge and Centering Theory indicates a problem with the reliability of CT as an evaluation method, chi-square was calculated. Chi-square is normally calculated between a set of expected results and a set of observed results, to see whether there is a statistically significant difference. In the case of this particular evaluation, the expected result was that the human would agree with the CT evaluation on all pairs of extracts and abstracts. The number of agreements and disagreements on pairs between CT and the human judge is the observed result. Chi-square revealed that there is no statistically significant difference between the CT and human evaluations of pairs, with a confidence level of $p \leq 0.001$. This means that CT can be used as a reliable way of evaluating coherence because there is no statistical difference between its evaluation of the better summary out of a pair and human judgement on the same pair.

6.3. Discussion

The disagreement between the human judge and CT illustrates the fact that what a human considers to be a 'readable' and coherent text is not necessarily the same as something which is judged more coherent in terms of an 'objective' theory of local coherence. The comparison of the CT and human evaluations correlates with the findings of other researchers (Kibble, 2001; Poesio et al., 2004), who claim that CT alone is not always enough to account for the coherence of a text. Indeed, a reader does not assess a text solely on whether an entity is mentioned in consecutive utterances, although this can be an important part of their assessment. They also look at aspects such as rephrasing, reordering and conciseness. However, CT is still considered a useful tool in evaluating the local coherence of summaries (as chi-square proved), although at this stage in its development as an evaluation method it is wise to supplement it with other methods to take into account the wider variety of aspects of coherence and readability obvious to a human. Its appropriateness for the task was supported by discussions with the human judge. In most cases the judge found it very difficult to select the best summary out of a pair in terms of readability and coherence alone: the information contained in the summary nearly always affected their judgement. Therefore any assessment of coherence which is more objective than intuitive human judgement is useful for evaluation.

7. Conclusions

This paper presented an investigation into the use of Centering Theory for the evaluation of coherence in pairs of extracts and abstracts produced in a computer-aided summarisation environment. Current means of evaluating quality in automatic summaries, such as considering dangling anaphors, discourse ruptures, grammaticality, etc., are inappropriate for this task due to the way computer-aided summaries are produced. Users of CAS systems should have access to guidelines describing how best to post-edit the automatically produced extract to transform it into a readable and coherent abstract. These guidelines cover aspects which are often used in the evaluation of coherence/readability, meaning that using the same criteria for evaluation is unfair.

CT is considered an appropriate option for development as an evaluation method because it is a theory of local coherence considering consecutive utterances, which are important in short texts comprising parts of a whole, such as summaries. The CT parameters most suited to summaries were specified (utterance = sentence; realisation = direct + indirect for possessive pronouns; ranking = grammatical/linear; constraint 1 = weak), and a metric was formulated to represent the effect of CT transitions in summaries. The traditional preference order for transitions was adopted, with some alterations regarding NO TRANSITIONs, and the most damaging transition in a summary was found to be NO TRANSITION (NO *Cb*): CONTINUE > RETAIN > NO TRANSITION (INDIRECT) > SMOOTH SHIFT > ROUGH SHIFT > NO TRANSITION (NO *Cb*). 100 news text summaries (50 extract/abstract pairs) were subject to evaluation using CT and the metric developed.

Of the 50 pairs, CT evaluated 78% of abstracts as more coherent than the extracts from which they were produced. Human judgement was also obtained to validate the CT evaluation. The human judge considered 82% of abstracts as more coherent than the extracts from which they were produced. The CT evaluation and the human judge agreed in 70% of cases. Chi-square was calculated to assess whether their disagreement poses problems for the reliability of using CT in evaluation; it revealed that there is no statistically significant difference between the evaluations, with a confidence level of $p \leq 0.001$. This means that CT can be used as a reliable way of evaluating coherence in pairs of summaries.

The main reason for disagreements between CT and the human judge is the fact that CT only takes into account repetitions of entities across consecutive utterances in its assessment of coherence. Obviously, humans consider much more than this in their judgements of coherence and readability. There are also issues relating to the way automatic methods produce extracts, which can increase the number of very similar sentences in a summary. Despite these issues, the exploration of CT for evaluation proved useful in assessing coherence in pairs of computer-aided extracts and abstracts; in particular, it provides a more objective view

than human intuition alone. However, it is wise to supplement the CT evaluation with other methods, to take into account the wider variety of aspects of coherence and readability obvious to readers of summaries. In future, it would be interesting to experiment further with different instantiations of the theory, especially as other researchers have found that different instantiations result in different frequencies of occurrence of NO CB transitions, which were found to be the most damaging for summary coherence.

8. References

S. E. Brennan, M. A. Friedman, and C. J. Pollard. 1987. A Centering Approach to Pronouns. In *Proceedings of the 25th Annual Meeting of the Association for Computational Linguistics (ACL'87)*, pages 155–162.

R. Donaway, K. Drummey, and L. Mather. 2000. A Comparison of Rankings Produced by Summarization Evaluation Measures. In *Proceedings of the NAACL-ANLP2000 Workshop on Automatic Summarization*, pages 69–78.

B. J. Grosz, A. K. Joshi, and S. Weinstein. 1995. Centering: a Framework for Modelling the Local Coherence of Discourse. *Computational Linguistics*, 21(2):203–225.

R. Gunning. 1988. *The Technique of Clear Writing.* New York. McGraw-Hill.

A. Harnly, A. Nenkova, R. Passonneau, and O. Rambow. 2005. Automation of Summary Evaluation by the Pyramid Method. In *Proceedings of Recent Advances in Natural Language Processing 2005 (RANLP'05)*, pages 226–232.

L. Hasler, C. Orasan, and R. Mitkov. 2003. Building Better Corpora for Summarisation. In *Proceedings of Corpus Linguistics 2003*, pages 309–319.

L. Hasler. 2004. An Investigation into the Use of Centering Transitions for Summarisation. In *Proceedings of the 7th Annual Colloquium of the UK Special Interest Group in Computational Linguistics (CLUK'04)*, pages 100–107.

L. Hasler. 2007. *From Extracts to Abstracts: Human Summary Production Operations for Computer-Aided Summarisation.* Ph.D. thesis, University of Wolverhampton.

L. Hirschman and I. Mani. 2003. Evaluation. In R. Mitkov, editor, *The Oxford Handbook of Computational Linguistics*, pages 414–429. Oxford. Oxford University Press.

M. Kameyama. 1998. Intrasentential Centering: A Case Study. In M. A Walker, A. K. Joshi, and E. F. Prince, editors, *Centering Theory in Discourse*, pages 89–112. Oxford. Oxford University Press.

N. Karamanis, M. Poesio, C. Mellish, and J. Oberlander. 2004. Evaluating Centering-based Metrics of Coherence Using a Reliably Annotated Corpus. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL'04)*, pages 391–398.

N. Karamanis. 2003. *Entity Coherence for Descriptive Text Structuring.* Ph.D. thesis, University of Edinburgh.

R. Kibble. 2001. A Reformulation of Rule 2 of Centering Theory. *Computational Linguistics*, 27(4):579–587.

J. P. Kincaid, R. P. Fishburne, R. L. Rogers, and B. S. Chissom. 1975. Derivation of new readability formulas (automated readability index, Fog count and Flesch reading ease formula) for navy enlisted personnel. Research Branch Report 8-75, Naval Air Station, Memphis, TN, USA.

M. Lapata and R. Barzilay. 2005. Automatic Evaluation of Text Coherence: Models and Representations. In *Proceedings of the 19th International Conference on Artificial Intelligence (IJCAI'05)*, pages 1085–1090.

C.-Y. Lin. 2001. SEE - Summary Evaluation Environment. http://haydn.isi.edu/SEE/.

C.-Y. Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Proceedings of The ACL2004 Workshop Text Summarization Branches Out*, pages 74–81.

I. Mani. 2001. *Automatic Summarization.* Amsterdam/Philadelphia. John Benjamins.

J.-L. Minel, S. Nugier, and G. Piat. 1997. How to Appreciate the Quality of Automatic Text Summarization? Examples of FAN and MLUCE Protocols and their Results on SERAPHIN. In *Proceedings of the ACL/EACL'97 Workshop on Intelligent Scalable Text Summarization (ISTS'97)*, pages 25–31.

A. Nenkova and R. Passonneau. 2004. Evaluating Content Selection in Summarization: The Pyramid Method. In *Proceedings of the Human Language Technology Conference/North American Chapter of the Association for Computational Linguistics (HLT/NAACL2004)*, pages 145–152.

C. Orasan and L. Hasler. 2006. Computer-aided Summarisation: What the User Really Wants. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC2006)*, pages 1548–1551.

C. Orasan, R. Mitkov, and L. Hasler. 2003. CAST: a Computer-Aided Summarisation Tool. In *Proceedings of the 11th Conference of The European Chapter of the Association for Computational Linguistics (EACL'03)*, pages 135–138.

C. Orasan. 2006. *Comparative Evaluation of Modular Summarisation Systems Using CAST.* Ph.D. thesis, University of Wolverhampton.

M. Poesio, R. Stevenson, B. di Eugenio, and J. Hitzeman. 2004. Centering: A Parametric Theory and its Instantiations. NLE Technical Note TN-02-01/CS Technical Report CSM-369, University of Essex, UK.

H. Saggion and G. Lapalme. 2000. Concept Identification and Presentation in the Context of Technical Text Summarization. In *Proceedings of The NAACL-ANLP 2000 Workshop on Automatic Summarization*, pages 1–10.

M. Strube and U. Hahn. 1999. Functional Centering - Grounding Referential Coherence in Information Structure. *Computational Linguistics*, 25(3):309–344.

M. A. Walker, A. K. Joshi, and E. F. Prince. 1998. Centering in Naturally Occurring Discourse: An Overview. In M. A Walker, A. K. Joshi, and E. F. Prince, editors, *Centering Theory in Discourse*, pages 1–28. Oxford. Oxford University Press.