Multi-document Summarisation and the PASCAL Textual Entailment Challenge

Nicola Stokes
NICTA Victoria Laboratory,
Department of Computer Science
and Software Engineering,
University of Melbourne.
nicola.stokes@nicta.com.au

Eamonn Newman
School of Computer Science
and Informatics,
University College Dublin,
Ireland.
eamonn.newman@ucd.ie

Abstract

A fundamental problem for systems that require natural language understanding capabilities is the identification of instances of semantic equivalence and paraphrase in text. The PASCAL Recognising Textual Entailment (RTE) challenge is a recently proposed research initiative that addressed this problem by providing an evaluation framework for the development of generic "semantic engines" that can be used to identify language variability in a variety of applications such as Information Retrieval, Machine Translation and Question Answering. This paper discusses the suitability of the RTE evaluation datasets as a framework for evaluating the problem of redundancy recognition in multi-document summarisation, i.e. the identification of repetitive information across documents. This paper also reports on the development of an additional dataset containing examples of informationally equivalent sentence pairs that are typically found in machine generated summaries. The performance of a competitive entailment recognition system on this dataset is also reported.

1 Introduction

The aim of multi–document summarisation (MDS) is to generate a concise and coherent summary given a cluster of related documents. Although this process is a natural extension of single–document summarisation, MDS poses a number of unique challenges such as, how to manage contradictory and repetitive information in the cluster, and how to order extracted information in the resultant summary. A popular approach to the MDS problem is to first identify and cluster repetitive information units across documents, then select representative sentences from the "dominant" clusters, and finally generate an extractive summary from these sentences. This approach assumes, like many others in text summarisation, that the repetition of information is an indication of information importance and consequently summary relevancy. The simplest method for determining commonality across documents is to group text units (e.g. sentences, paragraphs) that exhibit a high concentration of word overlap. However, this approximate method for recognising similar semantic content is often insufficient due to instances of language variability such as paraphrase and synonymy. Figure 1 shows two sentences (A and B) that are semantically equivalent but syntactically different.

Text A: Agassi's dream run is ended by world's number one player. Text B: Federer beats Agassi.

Figure 1: Paraphrases with minimal word overlap

In this paper we discuss the suitability of the recently proposed PASCAL Recognising Textual Entailment (RTE) challenge (Dagan et al., 2005a) as an evaluation methodology for determining the performance of redundant information identification techniques in the context of MDS. The aim of the RTE challenge is to aid the development of generic "semantic engines" that can be used in a number of applications such as Information Retrieval, Information Extraction, Text Summarisation and Machine Translation. Two types of language variability were investigated in this year's challenge: exact paraphrases and textual entailment or subsumption. The evaluation was defined as a binary classification problem where participating systems were required to identify entailment relationships between sentence pairs, i.e. a sentence A entails another sentence B if the meaning of B can be inferred from the meaning of A (Dagan et al., 2005b). During the data collection effort for the challenge, annotators were asked to limit the number of "difficult" cases of entailment

Proceedings of the Australasian Language Technology Workshop 2005, pages 215–223,
Sydney, Australia, December 2005.

that they included in the dataset. The entailment pairs shown in Figure 2 are representative of the level of difficulty of the subsumption relationships found in the data, where entailment, in the majority of cases, requires syntactic matching, and synonym/paraphrase recognition rather than complex logical inference. In this way techniques that recognise redundant information in MDS and entailment in the RTE challenge have a lot in common. This point will be discussed in more detail in Section 2 of the paper.

The RTE development and test sets are composed of entailment examples taken from seven distinct application settings. The "Comparable Documents" portion of the collection is intended to be representative of the types of entailment and semantic equivalence found in multi-document summarisation. In general, participating systems at the RTE workshop performed significantly better (achieving as high as 87% accuracy) on this portion of the corpus. This result suggests that the types of entailment and semantic equivalence found in MDS are significantly less challenging than entailment found in other application settings. In this paper we will show that this result is misleading, and that the difficulty of identifying language variability in MDS is comparable with the level of difficulty observed in the other application domains explored by the RTE initiative.

The following section motivates the need for evaluating sub–tasks in MDS such as redundant information identification, and provides a brief overview of the techniques that have been used to identify language variability in MDS and at the RTE challenge. Section 3 discusses the RTE framework in the context of MDS and argues for the inclusion of additional examples of language variablity that frequently require identification in MDS but are not represented in the current RTE evaluation dataset. Section 5 describes the University College Dublin (UCD) RTE system, which detects entailment between sentence pairs using linguistic and statistical language analysis techniques[1]. Section 6 discusses the performance of this system at the RTE workshop. In addition, the results of some initial experiments are provided that support the assertion that the performance of a competitive RTE system in an MDS application is comparable with its performance in other RTE applications settings. Finally Section 7, discusses some conclusions and future directions for this work.

Task=Comparable Documents; Paraphrase Example; Judgement=TRUE
Text A: Satomi Mitarai died of blood loss.
Text B: Satomi Mitarai bled to death.

Task=Comparable Documents; Textual Entailment Example; Judgement=TRUE
Text A: The Kota (Fort), or Old City, for example, sometimes called the downtown section, is the central business district and Indonesia's financial capital.
Text B: The Kota is the country's business center.

Task=Comparable Documents; Syntactic Variation Example; Judgement=TRUE
Text A: Jakarta floods easily during the rainy season.
Text B: Jakarta is easily flooded during the rainy season.

Figure 2: Examples of syntactic variation, paraphrase and information subsumption in the RTE dataset

2 MDS and RTE

So why is the RTE challenge an attractive sub–task evaluation methodology for MDS? Firstly, identifying semantic relationships and correctly clustering informationally equivalent sentences is a critical analysis component in many MDS systems for the following reasons: if sentences are incorrectly clustered then the commonality between the documents is harder to determine, and redundant (i.e. repetitive) information will be included in the summary – an outcome that summarisation systems want to avoid at all costs. Secondly, there are inherent limitations with the current summarisation evaluation standard provided by the Document Understanding Conference (DUC)[2], where both automatic and manual evaluation strategies are used to measure summary quality in terms of coverage, information redundancy, readability,

[1]The author was involved in the development of this system before moving to the NICTA Victoria Research Laboratory.

[2]DUC is an annual NIST sponsored workshop that provides participants with summarisation tasks and a corresponding evaluation framework, i.e. corpora, gold standard summaries and evaluation metrics.
`http://duc.nist.gov`

coherence and grammaticality. Since its creation in 2001, the DUC initiative has helped to ensure that real and transparent progress is being made in summarisation research; however, because the DUC evaluation methodology is determining the performance of many difficult natural language processing (NLP) components concurrently (i.e. semantic analysis, content selection, sentence ordering and natural language generation), it is often difficult to establish which techniques employed by a particular high performing summarisation system have contributed most to its overall success. As such summarisation researchers are recognising the need for distinct evaluation frameworks for each of these sub-components. For example, researchers at Columbia University have separately evaluated their sentence clustering algorithm, SimFinder, which is employed in their NewsBlaster summarisation system (McKeown et al., 2002). More recently Barzilay and Lapata (Barzilay and Lapata, 2005) describe an evaluation methodology for text coherence techniques, which are commonly used by summarisation systems to improve text readability. The following subsection provides a flavour of the Entailment and Semantic Equivalence techniques presented at the PASCAL RTE–2005 challenge, followed by a description of two important contributions made by Text Summarisation researchers in this area.

2.1 Language Variability Recognition Techniques

The 2005 PASCAL RTE challenge is described by the organisers as "an initial attempt to form a generic empirical task that captures major semantic inferences across applications" (Dagan et al., 2005b). Sixteen groups submitted their RTE system results to the workshop. The systems used a broad range of linguistic knowledge resources, statistical association metrics and logical inference mechanisms. As already stated, the simplest type of semantic equivalence measure that can be used to identify entailment is a measure of vocabulary overlap. Consequently, nearly all of the systems at the workshop considered uni-gram or n-gram overlap metrics when classifying entailment. A number of more sophisticated methods were also proposed. These measures either used statistical cooccurrence metrics (e.g. latent semantic indexing), lexical resources for detecting semantic relationships between verbs, nouns, and adjectives (e.g. WordNet (Millar, 1995)) or a combination of both. Syntactic–based overlap measures, which involves calculating the degree of match between parse tree representations of the sentence pair, were also popular. A few groups also incorporated a logical prover with some additional world knowledge resource such as a geospatial ontology or a semantic taxonomy. Many of the submitted systems, such as the UCD submission described in the following section, considered more than just one of these measures during the entailment recognition process. More specifically, these lexical, syntactic, semantic or logical–based inference measures were used as partial (rather than conclusive) evidence of the presence or absence of an entailment relationship between two sentences.

Overall the entailment recognition *accuracy* (see Section 6 for definition) of the participating systems at the workshop ranged from 50-60% where accuracy measures greater than 0.535 and 0.546 are better than chance at the 0.05 and 0.01 level, respectively (Dagan et al., 2005b). The general conclusion of the workshop was that relatively simple metrics used in combination performed better than more complex, "deeper" metrics such as logical inference or the incorporation of world knowledge into the classification computation. An obvious explanation for this outcome is that deep linguistic analysis methods are more prone to errors than simple term overlap metrics due to additional complexities such as word sense disambiguation.

So how do RTE techniques compare to the repetitive information detection methods used by the text summarisation community? Well as already stated, summarisation researchers have tended to favour simple similarity metrics based on the number of shared words. There are a couple of notable exceptions, however, which have been investigated by researchers at Columbia University.

Possibly the most well-known and successful approach to similarity detection in automatic summarisation is the SimFinder (Hatzivassiloglou et al., 2001) algorithm. This algorithm clusters sentences that share thematic content determined by a set of similarity features based on word, stem and Wordnet concept overlap as well as more complex features that capture match at a syntactic level such as subject-verb and verb-object relations. The subsequent clustering of sentences is then performed using a non-hierarchical clustering tech-

nique. Representative sentences from these clusters are then used to generate a summary.

(Barzilay and McKeown, 2005a) describe a revision strategy for improving the readability of the summary output of the SimFinder algorithm. Their revision system, MultiGen, searches for semantically equivalent textual units in the dependency tree graph representations of the summary sentences. Semantically similar words and phrases are identified using the WordNet taxonomy and a paraphrase dictionary, automatically constructed from parallel monolingual corpora. So once an overlapping paraphrase has been detected in the dependency trees this analysis then facilitates "information fusion", i.e. the generation of a single sentence that represents the information in the overlapping sentences. This text generation technique has been integrated into the Columbia NewsBlaster multi–document summarisation system (McKeown et al., 2002).

It is clear from this discussion that the Text Summarisation community has much to gain from, and contribute to, the advancement of Entailment and Semantic Equivalence recognition research.

3 RTE and language variability in MDS

In this section of the paper we comment on the coverage of the RTE evaluation corpora with respect to the type of real-world examples of semantic equivalence that require detection during multi-document summarisation. For the RTE 2005 challenge two development collections and one test collection where released to participants[3]. In each case, the datasets consisted of an even number of positive and negative examples of entailment between sentence pairs. During the development of these datasets annotators were asked to collect relevant examples that corresponded to typical success and failure settings in seven different applications, i.e. Information Retrieval (IR), Information Extraction (IE), Machine Translation (MT), Question Answering (QA), Paraphrase Acquisition (PP), Reading Comprehension (RC) and Comparable Documents–style tasks (CD) such as multi–document summarisation. A more detailed discussion of the annotation process can be found in (Dagan et al., 2005b).

[3]The RTE datasets can be downloaded from: `http://www.pascal-network.org/Challenges/RTE/Datasets`

As already stated, the motivation behind this paper is to establish whether or not these examples of language variablity are reflective of the types of information redundancy found in an MDS setting. Particularly in the case of the CD sentence pairs which are reportedly representive of the MDS task. To answer this question we considered Mani's analysis of this problem in his review of MDS methods, where he defines 4 distinct types of redundancy between text elements in MDS (Mani, 2001):

1. Two text elements are string identical when they are exact repetitions, i.e. the same sentence is repeated in multiple articles.
2. Two text elements are semantically equivalent when they are exact paraphrases of each other.
3. Two text elements are informationally equivalent if they are judged by humans to contain the same information.
4. A text element A informationally subsumes text element B if the information in element B is contained in A.

A manual examination of the RTE datasets shows that string identity and informational equivalence are not represented in these collections. Figure 2 provides examples of paraphrase and informational subsumption, i.e. textual entailment in the RTE data. The exclusion of string identical examples isn't considered critical as the detection of exact repetition is trivial. However, the lack of Mani's informational equivalence type examples is more troublesome. An example of informational equivalence is shown in Figure 3. What differentiates this example of language variablity from those in Figure 2, is that the common information unit is an embedded paraphrase surrounding in both sentences by additional information. More specifically, while Text A and B share the information unit: "American Airlines laid off flight attendants", they also contain additional non-overlapping information units, i.e. the federal judge turned aside a union bid to block the job losses; unions warned travellers to expect long delays due to protests. From our analysis we can conclude that examples of *exact* paraphrase and entailment are the exception rather than the rule in MDS and other CD–type applications. More often than not these systems will be required to deal with noisier instances of semantic equivalence where sentences repeat embed-

Task=MDS; Embedded Paraphrase Example; Judgement=TRUE
Text A: *American Airlines began laying off hundreds of flight attendants on Tuesday*, after a federal judge turned aside a union bid to block the job losses.
Text B: Unions have warned travellers that they can expect long delays this weekend as protests begin after *American Airlines let a large number of flight attendants go last week.*

Figure 3: An example of informational equivalence and embedded paraphrase

ded information units rather than exhibit complete semantic overlap (i.e. exact paraphrase) or subsumption.

In MDS, if the system can successfully detect these fuzzier examples of information redundancy it can make an informed decision on whether to: (a) substituted one sentence for another in the summary without any critical loss of information or (b) fuse these sentences together as proposed by (Barzilay and McKeown, 2005a). Sentence fusion would probably be the most appropriate option in the case of the embedded paraphrase example shown in Figure 3. With this type of natural language generation application in mind, it would be beneficial if the RTE classification task also required systems to explicitly identify and return the common information unit(s) between each sentence pair, i.e. the system must justify its classification decision.

4 An MDS-based Informational Equivalence Dataset

This section describes the development of a complementary RTE-style corpus of sentence–pairs that are more reflective of the types of information redundancy observed during multi-document summarisation.[4]. Annotators were asked to use Columbia's online NewsBlaster summarisation system[5] (a consistent top-performer at the annual DUC summarisation evaluation workshop) to aquire relevant sentence pairs. This curation strategy was employed to ensure that the MDS dataset was representive of the types of informational equivalence that are problematic in MDS. A subsequent analysis of the official DUC summary submissions to the multi-document summarisation task defined for the 2004 challenge (i.e. DUC task 2) indicates that these NewsBlaster examples are consistent with the types of repetitive information that were missed by sentence clustering strategies employed by other top performing summarisation systems at the workshop.

In line with the task-specific subsets in the RTE collection, the MDS dataset consists of 100 sentence pairs: 50 positive and 50 negative instances of informational equivalence. Figure 4 shows an example of each classification type. In the previous section it was explained that in order for a sentence pair to be tagged as a positive instance of informational equivalence it had to share an information unit; however, no formal definition of what constitutes such as unit was provided. The formulation of such a definition is a challenge in itself, and is currently receiving significant attention from the Text Summarisation community in the context of summarisation evaluation (Nenkova and Passonneau, 2004; Amigo, 2004). In the context of this task, an information unit is defined as a unit of text that contains at least one subject-verb relationship, (i.e. a noun phrase like "Air France Flight 358" is not a large enough information unit but "Air France Flight 358 crashed" is). In addition, when choosing these examples annotators were asked to be mindful of the underlying classification task in the context of a summarisation application, i.e. would the inclusion of both sentences result in unnecessary repetition in a summary. Any disagreement between annotator regarding the classification of certain pairs was discussed and resolved before experimentation on the corpus began.

From the MDS examples in Figure 4 it can also be seen that these sentences often make reference to vague temporal expressions such as "deadline...set for Monday" and "Monday deadline". In order to ground these temporal references to points in time the full text of the original source document would need to be analysed. However, temporal resolution is not necessary in this classification task since examples were carefully chosen to ensure that if an event (such as a "suicide bomb attack") is mentioned in both sentences, then the system can assume that this information unit is referring to the

[4]The MDS corpus can be downloaded from: `http://www.cs.mu.oz.au/~nstokes/TE/MDS_corpus_1.0.xml`

[5]The NewsBlaster summarisation system: `http://newsblaster.cs.columbia.edu`

same instance of the event in time.

Task=MDS; Pair Id=4; Judgement=TRUE;
Text A: The United States ratcheted up its pressure Saturday on Iraqi negotiators who are trying to meet a deadline for writing a draft constitution set for Monday.
Text B: With Iraq's parliament facing a Monday deadline to approve a new constitution, President Bush said Saturday that the document "is a critical step on the path to Iraqi self-reliance".

Task=MDS; Pair Id=62; Judgement=FALSE;
Text A: Discovery was loaded with nearly 7,000 pounds of garbage that had accumulated in the space station since it was last visited by a shuttle in December 2002.
Text B: The Discovery crew spent nine of their first 13 days in orbit transferring supplies to the space station.

Figure 4: Pair 4 and Pair 62 are examples of positive and negative informational equivalence in the MDS dataset.

With regard to the negative examples of information overlap in the MDS corpus, sentence pairs were picked from summaries that contained some word overlap, but which would still be considered unique information contributors to a summary. This helped to ensure that these negative sentence pairs were non–trivial.

During the creation of this corpus a number of examples of "contradiction" (i.e. conflicting news reports on the details of a specific event) between potential informationally equivalent sentence pairs were found. Although these examples represent another important problem in MDS, they were not included in the final version of the corpus because they frequently occur in the RTE challenge datasets in the form of negative entailment examples as shown in Figure 5.

In the following sections we describe the UCD RTE system, and compare its performance on the MDS dataset to its performance on the RTE test set. As already stated, this experiment is used to investigate our claim that the CD task data in the RTE challenge is unrepresentative of language variability in MDS.

Task=Comparable Documents; Judgement=False;
Text A: Jennifer Hawkins is the 21-year-old beauty queen from Australia.
Text B: Jennifer Hawkins is Australia's 20-year-old beauty queen.

Figure 5: An example of contradiction in the RTE data collection.

5 The UCD Textual Entailment Recognition System

In this section, we present an overview of the UCD Textual Entailment Recognition system, which was originally presented at the PASCAL RTE workshop (Newman et al., 2005). This system uses a decision tree classifier to detect an entailment relationship between pairs of sentences that are represented using a number of difference features such as lexical, semantic and grammatical attributes of nouns, verbs and adjectives. This entailment classifier was generated from the RTE training data using the C5.0 machine learning algorithm (Quinlan, 1993). The features used to train and test the classifier were calculated using the following similarity measures:

- The ROUGE (Recall–Oriented Understudy for Gisting Evaluation) (Lin and Hovy, 2004) n-gram overlap metrics, which have been used as a means of evaluating summary quality at the DUC summarisation workshop. The Rouge package provides measurement options such as uni-gram, bi-gram, tri-gram and 4-gram term overlap, and a weighted and unweighted longest common subsequence overlap measure.
- The Cosine Similarity metric calculates the cosine of the angle between the respective term vectors of the sentence pair.
- The Hirst–St-Onge WordNet–based measure (Millar, 1995), is an edge counting metric that estimates the semantic distance between words by counting the number of relational links between them in the WordNet taxonomy (Budanitsky and Hirst, 2001). This metric also defines constraints on the length of the path and the types of transitive relationships that are allowed between concepts (nodes) in the taxonomy. These constraints are important because unlike other WordNet–based

semantic relatedness measures (which only consider IS–A relationships) the Hirst–St Onge metric searches for paths that traverse the IS–A and HAS–A hierarchies in the noun taxonomy. Hence, this metric provides better coverage at an increased risk of detecting spurious relationships if unrestricted paths were allowed between concepts. This feature was implemented using the Perl Wordnet Similarity modules developed by (Patwardhan et al., 2003).

- A verb–specific semantic overlap metric, that uses the VerbOcean semantic network (Chklovski and Pantel, 2004b; Chklovski and Pantel, 2004a) to identify instances of antonymy and near-synonym between verbs. The relationships between verb–pairs in VerbOcean were gleaned from the web using lexico–syntactic patterns. Although WordNet provides a verb taxonomy, the VerbOcean data was used because it appears to provide better coverage of the types of relationships needed for detecting entailment.
- A Latent Semantic Indexing (Deerwester et al., 1990) measure, like the WordNet measure, attempts to calculated similarity beyond vocabulary overlap by identifying latent relationships between words though the analysis of cooccurrence statistics in an auxiliary news corpus.
- The final similarity measure is based on a more thorough examination of verb semantics. This measure finds the longest common subsequence in the sentence–pair, and then detects evidence of contradiction or entailment in the subsequence (such as verb negation, synonymy, near-synonymy, and antonymy) using the VerbOcean taxonomy. An example is shown in Figure 6.

A more detailed description of the UCD system can be found in (Newman et al., 2005).

6 Language Variability Recognition Experiments and Results

This section of the paper reports on the performance of the UCD RTE system on the RTE and MDS datasets. The RTE challenge defined two evaluation metrics:

- An *accuracy score* which is calculated as the number of correctly classified sentence pairs (positive and negative) returned by the system divided by the number of sentence pairs in the dataset.
- A *confidence-weighted score (CWS)* that ranges between 0 (no correct judgements at all) and 1 (perfect score), and rewards the system when it assigns a higher confidence score to correct judgements rather than to incorrect ones.

Task=Paraphrase Acquisition; Judgement=FALSE
Text A: *France on Saturday* flew *a planeload of United Nations aid into eastern Chad* where French soldiers prepared to deploy from their base in Abeche towards the border with Sudan's Darfur region.

Text B:*France on Saturday* crashed *a planeload of United Nations aid into eastern Chad*

Figure 6: The Longest Common Subsequence is highlighted in italics.

The UCD RTE and MDS results are shown in Table 1. The entailment classifier in the MDS and RTE experiments was trained using the RTE corpus training sets (`dev1` and `dev2`). The average accuracy and CWS scores (0.565 and 0.6 respectively), and the task results listed below this row in the table represent the official UCD results reported at the RTE 2005 workshop. A manual analysis of these results showed that many of the misclassified errors made by the UCD system could be attributed to the occurrence of equivalence phrasal and compositional paraphrases e.g. "X invented Y" = "Y was incubated in the mind of X". As explained in Section 5 the system can only identify word–level, atomic paraphrase units (e.g., child = kid; eat = devour) that are defined in the VerbOcean and WordNet lexical resources. A more detailed discussion of system misclassifications is provided in (Newman et al., 2005).

Out of 16 groups UCD's average accuracy and CWS scores were ranked 4th and 5th respectively, where system accuracy results ranged from 0.586 to 0.495 and CWS scores from 0.686 to 0.507. In general, systems performed significantly better on the CD entailment examples, and for many it was this score that added some respectability to their average accuracy score. The most plausible explanation for these high CD scores (as high as 87% accuracy), accord-

ing to (Dagan et al., 2005b), is that vocabulary overlap metrics performed very well on this task because sentence pairs containing common terms were more likely to have the same meanings than in the other tasks. This implies that MDS systems need nothing more than vocabulary overlap metrics, and that the negative effect of errors from this component of an MDS system is minimal. However, a comparison of the UCD system results on the CD and MDS language variablity examples suggests that redundant information detection is as difficult as the other tasks investigated, and that further research effort is also required in this area.

Task	**Accuracy**	**CWS**
MDS	0.5400	0.6006
RTE Average	*0.5650*	*0.6000*
CD	0.7400	0.7764
IE	0.4917	0.5260
IR	0.5444	0.6130
PP	0.5600	0.5006
MT	0.5083	0.5130
QA	0.5385	0.5006
RC	0.5286	0.5685

Table 1: RTE and MDS Accuracy and CWS results for the UCD entailment classifier.

7 Conclusions

This paper evaluates the RTE challenge as a potential evaluation framework for comparing the performance of redundant information recognition strategies used in multi–document summarisation (MDS) to detect informational equivalence across documents. Most MDS systems use simple word counts to identify repetitive information. The problem with this approach is that many sentences that convey the same information show little surface resemblance due to linguistic phenomenon such as paraphrase and synonymy. The RTE challenge provides an opportunity for summarisation researchers to evaluate more sophisticate redundancy identification techniques independent of the summarisation task. However, an analysis of the RTE development and test sets show that this data is not representative of the types of informational equivalence that require detection during the MDS process. More specifically, although subsumption relationships are a natural occurrence in applications such as Question Answering and Information Retrieval (where the answer/relevant document will always entails the question/query) this is not the case for Comparable Documents-style tasks. The results of an experiment on a complementary dataset of MDS informational equivalence examples using a competitive RTE system showed that identifying redundancy in MDS is more challenging than the results on the Comparable Documents portion of the RTE test set would suggest. Consequently, if the ultimate aim of the PASCAL RTE challenge is to build "generic semantic engines" then future evaluations will also have to consider the identification of embedded (semantic and syntactic) paraphrases across sentences.

An obvious extention of this work would be to incorporate the UCD RTE system into an MDS system, and compare its effect on summary performance against a baseline semantic equivalence measure such as cosine similarity. It would also be interesting to further investigate how well the RTE evaluation framework simulates the process of identifying repetitive information in MDS and other applications. In a paper by Barzilay and Elhadad (2003), on sentence alignment for monolingual comparable corpora, it was shown that the effectiveness of the alignment process increased when the context surrounding sentences was also considered. This conclusion suggests that future RTE evaluations should also consider evaluating the role of context in the entailment detection process, where additional context is provided by the document in which the sentence occurred.

8 Acknowledgements

The support of Enterprise Ireland and NICTA (National ICT Australia) is gratefully acknowledged.

References

E. Amigo. 2004. An empirical study of information synthesis tasks. In *Association for Computational Linguistics (ACL'04)*.

R. Barzilay and N. Elhadad. 2003. Sentence alignment for monolingual comparable corpora. In *Empirical Methods in Natural Language Processing (EMNLP'03)*.

R. Barzilay and M. Lapata. 2005. Modeling local coherence: An entity-based approach. In *Association for Computational Linguistics (ACL'05)*.

R. Barzilay and K. McKeown. 2005a. Sentence

fusion for multidocument news summarization. *Computational Linguistics*, 31(3).

A. Budanitsky and G. Hirst. 2001. Semantic distance in WordNet: An experimental, application-oriented evaluation of five measures. In *the Workshop on WordNet and Other Lexical Resources, NAACL'01.*

T. Chklovski and P. Pantel. 2004a. Global path-based refinement of noisy graphs applied to verb semantics. In *the International Joint Conference on NLP (IJCNLP-05)*, pages 11–13.

T. Chklovski and P. Pantel. 2004b. VerbOcean: Mining the web for fine–grained semantic verb relations. In *Empirical Methods in Natural Language Processing (EMNLP-04).*

I. Dagan, O. Glickman, and B. Magnini (eds). 2005a. In *the PASCAL Recognising Textual Entailment Challenge Workshop, April 11th-13th 2005, Southampton, UK.*

I. Dagan, O. Glickman, and B. Magnini. 2005b. The PASCAL recognising textual entailment challenge. In *the PASCAL Recognising Textual Entailment Challenge Workshop 2005*, pages 1–8.

S. Deerwester, S. Dumais, G. Furnas, T. Landauer, and R. Harshman. 1990. *Indexing by Latent Semantic Analysis.* Journal of the American Society for Information Science.

V. Hatzivassiloglou, J. Klavans, M. Holcombe, R. Barzilay, Min-Yen Kan, and K. McKeown. 2001. SimFinder: A flexible clustering tool for summarization. In *the Workshop on Automatic Summarization, NAACL-01.*

C.-Y. Lin and E. Hovy. 2004. Automatic evaluation of summaries using n-gram co–occurence statistics. In *the Document Understanding Conference (DUC'04), National Institute of Standards and Technology.*

I Mani. 2001. *Automatic Summarization.* John Benjamins (Natural language processing series, edited by Ruslan Mitkov, volume 3), Amsterdam.

K. McKeown, R. Barzilay, D. Evans, V. Hatzivassiloglou, J. Klavans, A. Nenkova, C. Sable, B. Schiffman, and S. Sigelman. 2002. Tracking and summarizing news on a daily basis with Columbia's Newsblaster. In *the Human Language Technology Conference (HLT'02).*

G. Millar. 1995. WordNet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

A. Nenkova and R. Passonneau. 2004. Evaluating content selection in summarization: The Pyramid Method. In *HLT–NAACL'04.*

E. Newman, N. Stokes, J. Dunnion, and J. Carthy. 2005. UCD IIRG approach to the Textual Entailment Challenge. In *the PASCAL Recognising Textual Entailment Challenge Workshop*, pages 53–56.

S. Patwardhan, J. Michelizzi, S. Banerjee, and T. Pedersen. 2003. *WordNet::Similarity Perl Module* `http://search.cpan.org/dist/wordnet-similarity/lib/wordnet/similarity.%pm`.

J.R. Quinlan. 1993. C5.0 machine learning algorithm. `http://www.rulequest.com`.