Multi-Document Summarisation Using Generic Relation Extraction

Ben Hachey
Centre for Languate Tecnology
Macquarie University
NSW 2109 Australia

Capital Markets CRC Limited
GPO Box 970
Sydney NSW 2001

bhachey@cmcrc.com

Abstract

Experiments are reported that investigate the effect of various source document representations on the accuracy of the sentence extraction phase of a multi-document summarisation task. A novel representation is introduced based on generic relation extraction (GRE), which aims to build systems for relation identification and characterisation that can be transferred across domains and tasks without modification of model parameters. Results demonstrate performance that is significantly higher than a non-trivial baseline that uses *tf*idf*-weighted words and at least as good as a comparable but less general approach from the literature. Analysis shows that the representations compared are complementary, suggesting that extraction performance could be further improved through system combination.

1 Introduction

The goal of summarisation is to take an information source, extract content from it, and present the most important content in a condensed form (Mani, 2001). The field of automatic summarisation (Mani, 2001; Spärck Jones, 2007) aims to create tools that address various summarisation tasks with minimal human intervention. Extractive approaches to automatic summarisation create representations of the source document that are generally based on an easily identified text sub-unit such as sentences or paragraphs. These representations are then used to identify representative or otherwise important snippets of text to place in the summary.

Following Spärck Jones (2007), summarisation systems can be characterised with respect to their approach to three main sub-tasks: 1) interpretation, 2) transformation and 3) generation. The input consists of the source document (or a collection of source documents in the case of multi-document summarisation). The first step (interpretation) creates a representation of the source document by performing some level of interpretation. A simple approach here represents sentences by their tokens (i.e., as an unordered bag-of-words). The next step (transformation) is the compaction step where the source representation is converted into the summary representation, e.g. by identifying sentences whose words are most representative of the full text. Finally, in the last step (generation), the output summary is created. In the case of sentence extraction, this includes various operations to maximise coherence such as ensuring that entity references are comprehensible and arranging the sentences in a sensible order.

The current work investigates several representations of source documents. In particular, an approach from the literature based on atomic events (Filatova and Hatzivassiloglou, 2004) is compared to a novel approach based on generic relation extraction (GRE), which aims to build systems for relation identification and characterisation that can be transferred across domains and tasks without modification of model parameters (Hachey, 2009). The various representations are substituted in the interpretation phase of a multi-document summarisation task and used as the basis for extracting sentences to be placed in the summary. System summaries are compared by calculating term overlap with reference summaries created by human analysts.

2 Motivation

In seminal work on automatic summarisation, Luhn (1958) introduces a representation based on content words. These are defined as non-function words from the source document that are neither too frequent nor too infrequent. Luhn uses frequency to weight content words and ex-

Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, pages 420–429,
Singapore, 6-7 August 2009.

tracts sentences with the highest combined content scores to form the summary. Subsequent work adapted the *tf*idf* weighting scheme, where term frequency (*tf*) is combined with inverse document frequency (*idf*), an inverse measure of term occurrence across documents that serves to downweight common words (Spärck Jones, 1972). In modern work, *tf*idf* representations are often used as simple but non-trivial baselines. The problem is that these shallow features often break down where underlying linguistic content needs to be compared rather than just surface structure.

The use of representations based on information extraction (IE) has been suggested as one approach to capturing deeper semantic information. This is based on the notion that IE definitions of types for entities, relations and events provide a level of abstraction that is appropriate for automatic summarisation. Several approaches in the literature have explored the use of IE-based representations for extractive summarisation: McKeown et al. (1998) incorporate patient characteristic templates for matching potential treatments to specific patients in a medical summarisation system; White and Cardie (2002) incorporate a bootstrapped IE system based on Autoslog (Riloff, 1996) for filling event templates; and Harabagiu and Maiorano (2002) incorporate a hybrid approach that uses conventional supervised IE techniques for known topics and a more general approach based on WordNet for unknown topics.[1] The problem with these systems is that they all use supervised approaches to IE that require that the IE templates be known in advance and additionally require significant investment in writing extraction rules or in annotating data for training. Where more general techniques are used, they still require domain-specific resources, e.g. White and Cardie (2002) bootstrapping approach still requires that the extraction templates be known in advance and Harabagiu and Maiorano (2002) approach depends on the WordNet lexical database, for which coverage is not guaranteed for arbitrary domains.

Filatova and Hatzivassiloglou (2004) introduce methods using more general IE representations that are not based on supervised learning. Given a named entity recogniser, the representation is automatically derived and consists of $<Ent, Connector, Ent>$ event triples, where connectors are verbs or action nouns that occur in between the two NEs. Thus, the approach aims to perform a generic IE task that the authors refer to as atomic event extraction. This representation is shown to outperform a *tf*idf* baseline on a multi-document summarisation task. As we will see in Section 4.3 below, Filatova and Hatzivassiloglou's approach has three main shortcomings. First, it focuses exclusively on simple atomic events (i.e., entity mention pairs with an intervening verbal connector), meaning that it will not be able to address tasks where relations are at least as important as events (e.g., biographical summarisation). Second, it relies on exact matching between connectors, which is not capable of capturing latent semantic similarities (e.g., between 'work for' and 'employed by'). Third, its performance is subject to the coverage of WordNet, which is used to identify action nouns.

Generic relation extraction (GRE) aims to build systems that can be transferred across domains and tasks without modification of model parameters (Hachey, 2009). For relation identification (i.e., extraction of relation forming entity mention pairs), this is achieved by using general rule-based approaches and, for relation characterisation (i.e., assignment of types to relation mentions), this is achieved by using unsupervised machine learning. Hachey (2009) introduces a GRE approach that addresses the shortcomings of the atomic event approach mentioned above. First, it models a type of IE that includes relations. Second, it uses a connector model based on latent Dirichlet allocation (Blei et al., 2003), which provides a mechanism for capturing latent semantic similarities between connectors. Third, it does not rely on domain-specific resource like WordNet. The GRE models used here do rely on dependency parsing. However, they still generalise across formal domains as the relation identification and characterisation systems, developed on news data, achieve comparable performance when applied directly to a relation extraction task in the biomedical domain (see Hachey (2009) for details). Furthermore, grammatical relations obtained from dependency parsing provide a means for constraining relation identification and supplying more linguistically meaningful features for relation characterisation.

[1] Comparable approaches using IE in the context of abstractive–as opposed to extractive–summarisation include work by DeJong (1982), Hahn and Reimer (1999), White et al. (2001) and Saggion and Lapalme (2002).

	c_1	c_2	c_3	c_4	c_5
t_1	1	1	0	1	1
t_2	1	0	0	1	0
t_3	0	1	0	0	1
t_4	1	0	1	1	1

Table 1: Text × concept matrix for set cover approach to automatic summarisation (Filatova and Hatzivassiloglou, 2004).

3 Algorithm for Set Cover Extraction

For the sake of comparison, the current evaluation adopts the Filatova and Hatzivassiloglou (2004) summarisation framework. This defines an extraction approach based on a mapping between *textual* units and *concepts*. To illustrate, consider the matrix in Table 1 where rows represent textual units (e.g., sentences, paragraphs) and columns represent concepts (e.g., words, events, relations) in the input text. Each concept is either absent or present in a given textual unit. Additionally, each concept has a weight associated with it. Looking at the problem in this way makes it natural to formulate it as follows: *the summary should select textual units such that there is maximal coverage of the salient conceptual units*.[2] This is essentially the maximum coverage problem, which has been shown to be reducible to the set covering problem, for which there are approximation algorithms in the literature that run in polynomial time or better (Hochbaum, 1997; Bienstock and Iyengar, 2004).

Filatova and Hatzivassiloglou define several greedy algorithms that can be parametrised in terms of the general SUMMARISE function in Figure 1, which takes the text × concept matrix D and the maximum summary length k as input. The SUMMARISE function first initialises the summary $\mathcal{S}$ to the empty set. Then it enters a loop that continues until the summary reaches the desired length. Within the loop, a text unit is extracted and added to the summary after which the text × concept matrix is updated The output of the algorithm is a set $\mathcal{S}$ comprising the text units that make up the summary. For the experiments reported here, the text units t are sentences and LENGTH(t_i) returns the count of word tokens in sentence t_i.

```
SUMMARISE : D, k
1  S ← {}
2  while Σ_{t_i∈S} LENGTH(t_i) < k
3        t_j ← EXTRACT(D)
4        S ← S ∪ t_j
5        D ← UPDATE(D, t_j)
6  return S
```

Figure 1: Generalised function for Filatova and Hatzivassiloglou (2004) approach to extractive summarisation.

```
EXTRACT: D
1  c_j ← arg max_{c_j∈cols(D)} Σ_{t_i∈rows(D)} D[t_i, c_j]
2  t_k ← arg max_{t_k∈rows(D)&D[t_k,c_j]>0} SCORE(D, t_k)
3  return t_k
```

```
UPDATE: D, t_i
1  for each c_j ∈ cols(D)
2        if D[t_i, c_j] > 0
3              for each t_k ∈ rows(D)
4                    D[t_k, c_j] ← 0
5  D ← DELETE(D, t_i)
6  return D
```

Figure 2: Extraction and update functions for Filatova and Hatzivassiloglou (2004) modified adaptive algorithm.

Figure 2 contains the EXTRACT and UPDATE functions used here.[3] The EXTRACT function first identifies the concept c_j not yet covered in the summary that has the highest overall weight in the text × concept matrix D. Then it selects the text unit t_k with the highest score from among the text units that contain concept c_j. The SCORE function is the sum of concept weights for the given text unit, i.e.:

$$\text{SCORE} : D, t_i \mapsto \textbf{return} \sum_{c_j \in cols(D)} D[t_i, c_j] \quad (1)$$

The UPDATE function in Figure 2 aims to minimise redundancy in the summary by globally maximising the number of conceptual units covered in the output. In addition to removing the row representing the extracted text unit from the text × concept matrix D, it iterates through the remaining text units and assigns zero weights to all concepts that are covered by the extracted text unit.

[2] While not considered in the current experiments, a more discourse-oriented approach could be derived within the set cover framework by down-weighting conceptual units that occur e.g. in portions of the source documents that describe background information, where text segments containing background information could be identified using a sentence-level rhetorical status classifier like that developed by Teufel and Moens (2002).

[3] The EXTRACT and UPDATE functions in Figure 2 correspond to Filatova and Hatzivassiloglou (2004) modified adaptive algorithm and were found in preliminary experiments to be the better than the simple greedy and adaptive greedy algorithms (see Hachey (2009) for details).

Bush worked as an oil lease negotiator for Amoco in Denver and later started his own oil company, JNB.
*tf*idf* (*TF*) jnb:3.55, amoco:3.13, oil:3.05, negotiator:3.04, lease:2.58, denver:2.45, bush:2.44, worked:2.28, started:2.21, later:2.13, own:1.96, company:1.94, ...
event (*EV*) <PER_bush,worked,XFN_oil>:0.00023, <PER_bush,worked,ORG_amoco>:0.00011, <PER_bush,worked,LOC_denver>:0.00011, <XFN_oil,started,ORG_jnb>:0.00011, ...
relation (*RL*) <ORG_amoco,rd94,LOC_denver>:0.00039, <ORG_amoco,rd505,LOC_denver>:0.00039, <XFN_oil,rd92,ORG_jnb>:0.00002, <XFN_oil,rd712,ORG_jnb>:0.00002, ...
entity pair$_{ev}$ (*EE*) <PER_bush,XFN_oil>:0.00244, <PER_bush,LOC_denver>:0.00122, <PER_bush,ORG_jnb>:0.00044, <LOC_denver,XFN_oil>:0.00033, ...
entity pair$_{rl}$ (*ER*) <ORG_amoco,LOC_denver>:0.00311, <ORG_jnb,XFN_oil>:0.00155

Figure 3: Example sentence and various representations of sentence content.

4 Models

Figure 3 contains an example sentence and its representations corresponding to the various models of sentence content explored here.[4] These are described in detail in the rest of this section.

4.1 Baseline *tf*idf* Representation (TF)

The baseline model represents sentences as *tf*idf*-weighted bags-of-words (*TF*). Document frequencies for terms are derived from the same resource used by Filatova and Hatzivassiloglou (2004)–a frequency list compiled from a large sample of web pages. Term weighting is calculated using *tf*idf* as:

$$w(i,j) = \sqrt{(1 + \log(tf_{i,j})) * \log\left(\frac{N}{df_i}\right)} \quad (2)$$

where $tf_{i,j}$ is the number of times term i occurs in sentence j and df_i is the number of documents in which term i occurs. An example sentence and its *tf*idf* representation can be seen in Figure 3.

[4]The sentence was selected from document set d47 (from the data set described in Section 5.1 below), which contains articles about Neil Bush and his role in the collapse of Silverado Savings and Loan during the U.S. Savings and Loan crisis of the 1980s and 1990s.

4.2 Event Representation (EV)

We also compare to Filatova and Hatzivassiloglou (2004) atomic events (*EV*). This consists of $<Ent_i, Connector_j, Ent_k>$ event triples, where *connectors* are verbs or action nouns (i.e., nouns that are hyponyms of event or activity in WordNet) that occur in between the two entity mentions. Given a named entity recogniser and a lexical resource (WordNet), these are derived automatically from the text as follows. In the first step, all pairs of entity mentions that occur together in a sentence are identified. Next, the algorithm characterises the entity mention pairs using the connector words from the intervening context and discards pairs without an intervening connector word.

Event triple weighting is calculated by combining entity pair and connector weights as:

$$w_{ev}(i,j,k) = w_{ne}(i,k) * w_{cn}(j,i,k) \quad (3)$$

where $w_{ne}(i,k)$ is the weight of the entity pair $<i,k>$ consisting of entities i and k and $w_{cn}(j,i,k)$ is the weight of connector j in the context of entity pair $<i,j>$. $w_{ne}(i,k)$ is calculated as the normalised entity pair count, i.e.:

$$w_{ne}(i,k) = \frac{C_{ne}(<i,k>)}{C_{ne}(<*,*>)} \quad (4)$$

where $C_{ne}(<i,k>)$ is the count of mentions of entity pair $<i,k>$[5] and $C_{ne}(<*,*>)$ is the total count of entity mention pairs. And, $w_{cn}(j,i,k)$ is calculated as the normalised count of connector j in the context of the entity pair, i.e.:

$$w_{cn}(j,i,k) = \frac{C_{cn}^{<i,k>}(j)}{C_{cn}^{<i,k>}(*)} \quad (5)$$

where $C_{cn}^{<i,k>}(j)$ is the count of occurrences of connector j in the context of entity pair $<i,k>$ and $C_{cn}^{<i,k>}(*)$ is the total count of connectors in the context of entity pair $<i,k>$. An example sentence and its *event* representation can be seen in Figure 3. Event triples generated include <PER_bush,worked,ORG_amoco> and <PER_bush,started,ORG_jnb>.

Some erroneous event triples are also generated. The first error has to do with the fact that entities

[5]Coreference between entity mentions is computed by exact string match after removing punctuation, converting to all lower case, and prefixing the entity type. For example, the entity mention string "JNB" with type ORGANISATION is normalised to ORG_jnb.

include named entities identified in the preprocessing as well as the ten most frequent nouns in the document set. In the example sentence from Figure 3, the most frequent nouns include 'oil' but not 'negotiator' or 'company'. Therefore, 'oil' is labelled as an entity and extracted in a number of triples such as <PER_bush,worked, XFN_oil> (as opposed to <PER_bush, worked,XFN_negotiator>). Another problem illustrated by the example sentence has to do with the noisy nature of the surface-level approach to identifying entity mention pairs and connectors which tends to generate many false positive events, e.g. <ORG_amoco,started, ORG_jnb>. If the algorithm was constrained based on the underlying grammatical structure, it should be able to identity that the arguments of 'worked' are 'Bush' and 'Amoco' (i.e., <PER_bush,worked,ORG_amoco>) and that 'worked' does not describe an event involving 'Amoco' and 'JNB'.

4.3 Relation Representation (RL)

The focus of the current evaluation is a novel representation based on generic relation extraction (GRE). As mentioned above, GRE is a minimally supervised approach to the relation extraction task that aims to build systems for relation identification and characterisation that can be transferred across domains and tasks without modification of model parameters. Relation mentions are identified by taking pairs of entity mentions that have either 1) no more than two intervening words in the surface order of the sentence or 2) no more than one edge intervening on the shortest path through a dependency parse (see Hachey (2009) for details and experiments comparing different window configurations). This is stricter than the Filatova and Hatzivassiloglou approach in that entity mentions have to occur much closer or be connected by a single dependency relation. At the same time, it is less strict in the sense that an action- or event-denoting word is not required in the context, which makes it a more general model of IE.

Relation *connectors* are derived from a model of relation types based on latent Dirichlet allocation (Blei et al., 2003) that incorporates word, entity and dependency path features from the context of a relation-forming entity mention pair (see Hachey (2009) for details). This outputs a topic distribution for each entity mention pair that corresponds to the type of relation that is described. This representation 1) models a type of generic IE that includes relations, 2) uses a connector model that abstracts away from surface-level event descriptors used by Filatova and Hatzivassiloglou (2004) and 3) does not rely on domain-specific resources like WordNet.[6] For the purpose of comparison, relation triples are weighted in the same way as event triples using Equations 3 and 4 above. However, the connector pair weighting is modified to use the distribution over topics given by the LDA output.[7]

Relation triples generated for the example sentence in Figure 3 include <ORG_amoco, rd94,LOC_denver> and <ORG_amoco, rd505,LOC_denver>, where the connectors (i.e., rd94 and rd505) are identifiers that index particular topics from the LDA output. Here, rd94 and rd505 index topics that correspond to *located-in* relations so the respective triples both describe *located-in* relations between Amoco and Denver. Relation triples generated for the example sentence also include <XFN_oil, rd92,ORG_jnb> and <XFN_oil,rd712, ORG_jnb>. These are erroneous for the same reason as some of the event triples above (i.e., due to the noise inherent in the approach to identifying nominal entity mentions by identifying the ten most frequent nouns in the document set).

4.4 Entity Pair Representations (EE, ER)

Finally, we investigate the performance of representations that do not model event or relation type information. These are identical to the EV and RL representations above, except they are $<Ent, Ent>$ 2-tuples instead of $<Ent, Connector, Ent>$ 3-tuples. That is, entity pairs are included here provided that they meet the relation mention identification constraints. They are weighted using the normalised entity pair count (Equation 4 above). Relation-based entity pairs generated for the example sentence in Figure 3 include <LOC_denver,ORG_amoco> and <ORG_jnb,XFN_oil>.

[6] The GRE representation here does rely on dependency parsing, however, Hachey (2009) shows that it is still directly portable between the news and biomedical domains without modification of model parameters.

[7] Distributions for entity mention pairs tend to have a long uniform tail and only a few topics with higher probability. In converting to a weighting scheme, topic representations here are converted to a sparse representation where all topics in the uniform tail are removed.

5 Experimental Setup

5.1 Data

The experiments here use the multi-document summarisation data from the 2001 Document Understanding Conference (DUC),[8] which is the same data used by Filatova and Hatzivassiloglou (2004). This comprises 30 test document sets, each of which include approximately 10 news stories. Each document set is collected by a human and focuses on a particular topic. Example topics include the nomination of Clarence Thomas to the American Supreme Court, Neil Bush's role in the collapse of Silverado Savings and Loan and the Exxon Valdez oil spill. Gold standard summaries are provided for each document set for summary lengths of 50, 100, 200 and 400 words. This helps to ensure that the systems are not over-tuned to specific summary lengths. For each summary task (i.e., all 120 document set $\times$ summary length combinations), there are three distinct gold standard summaries created by different human analysts.

Pre-processing includes sentence boundary identification, segmentation of words (tokenisation), labelling words with part-of-speech tags, identification of noninflected base word forms (lemmatisation) from the LT-TTT tools (Grover et al., 2000). It also includes dependency parsing using Minipar (Lin, 1998) and automatic named entity recognition using the C&C tagger (Curran and Clark, 2003) trained on the data from the MUC-7 shared task (Chinchor, 1998). Weights for the various IE-based representations are calculated over each input document set.

5.2 Evaluation

The evaluation uses Rouge[9] to determine which representation selects content that overlaps most with human summaries. Rouge estimates the coverage of appropriate concepts (Lin, 2004) in a summary by comparing it to several human-created reference summaries. Rouge-1 does so by computing recall based on macro-averaged unigram overlap. Rouge-SU4 does so by calculating skip-bigram overlap where bigrams are allowed to be composed of non-contiguous words (with as many as four words intervening). Rouge-SU4 also includes unigrams to decrease the chances of zero scores where there is no skip-bigram overlap.

The configuration is based on comparisons between Rouge and human judgements of content coverage (Lin, 2004), which suggest that Rouge-1 and Rouge-SU4 with stemming and removal of stop words are good measures for evaluating multi-document summarisation tasks, consistently achieving Pearson's correlation scores above 0.72 and as high as 0.9 for longer summaries. Paired Wilcoxon signed ranks tests across document sets are used to check for significant differences between systems. The paired Wilcoxon signed ranks test is a non-parametric analogue of the paired t test. The null hypothesis is that the two populations from which the scores are sampled are identical.

1	50	100	200	400
TF	0.0797	0.1113	0.1742	0.2467
EV	**0.1360**	**0.1776**	0.2315	**0.3019**
RL	**0.1360**	0.1766	**0.2412**	0.3014

SU4	50	100	200	400
TF	0.0173	0.0259	0.0442	0.0693
EV	**0.0376**	**0.0494**	0.0692	**0.0950**
RL	0.0356	0.0491	**0.0701**	0.0939

Table 2: Comparison of Rouge scores for the *tf*idf* (*TF*), *event* (*EV*) and *relation* (*RL*).

6 Results

Can extractive summarisation be improved using representations based on generic information extraction? Table 2 contains results for *tf*idf* (*TF*), *event* (*EV*) and *relation* (*RL*) representations. Columns contain results for different lengths of summary (50, 100, 200 and 400 words). The best representation for each summary length is in bold and representations that are statistically distinguishable from the best (i.e., $p \leq 0.05$) are underlined. The results demonstrate unambiguously that the *event* and *relation* representations outperform the *tf*idf* representation, with strongly significant p-values less than 0.001 for both Rouge measures and all summary lengths. The *event* and *relation* representations are indistinguishable for both Rouge measures and all summary lengths.

[8] http://www-nlpir.nist.gov/projects/duc/index.html

[9] Rouge stands for recall-oriented understudy for gisting evaluations. While current versions also compute precision and f-score of system summaries, the evaluation here uses recall alone, which is sufficient when the length of the summaries being compared is the same. Rouge can be obtained from http://haydn.isi.edu/ROUGE/.

1	50	100	200	400
ER	**0.1497**	**0.1929**	**0.2527**	**0.3123**
EE	0.1442	0.1705	0.2288	0.3061

SU4	50	100	200	400
ER	**0.0419**	**0.0537**	**0.0786**	**0.1008**
EE	0.0364	0.0447	0.0643	0.0963

Table 3: Comparison of Rouge scores for entity pairs based on relations (*ER*) and events (*EE*).

How does entity pair identification for generic relations compare to entity pair identification for atomic events? Table 3 contains results for the representations described in Section 4.4. Rows correspond to entity pair identification for relations (*ER*) and events (*EE*).[10] Results suggest that the entity pair model based on GRE data outperforms the entity pair model based on atomic events, at least for medium sized summaries of 100 and 200 words where *ER* is significantly better than *EE* for both Rouge measures.

How do the event and relation representations perform with respect to corresponding entity pair representations? The scores for the entity pair representations reported in Table 3 are statistically indistinguishable from those for corresponding *relation* and *event* representations in Table 2 above. This appears to be a mixed result for both the *relation* representation introduced here and the Filatova and Hatzivassiloglou *event* representation. And, while GRE is shown to have a positive effect on Rouge scores when compared to atomic events, the same cannot be said of approaches to characterising relation and event types. However, as the correlation analysis (Section 7.1 below) demonstrates, RL and ER do not necessarily perform well on the same document sets. This suggests that they are actually complementary to some degree, meaning that a combined system based on both representations would outperform RL and ER on their own.

[10] In contrast to the results for the *tf*idf*, *relation* and *event* representations which use the modified adaptive algorithm described above, results for entity pair representations use a simplified version of the EXTRACT function that picks the text unit that has the highest score. This performed significantly better than the modified adaptive algorithm ($p \leq 0.01$) for all summary lengths for *ER* and was indistinguishable for *EE*. See Hachey (2009) for details.

7 Analysis and Comparison

7.1 Complementarity

Figure 4 contains results for a correlation analysis comparing the various representations. This also includes a comparison to the human upper bound (*HU*), computed by leave-one-out cross validation. Cells in the matrix contain the correlations values measured across document set Rouge-SU4 scores[11] using Spearman's ρ rank correlation coefficient (r_S). Here, high values mean that two representations tend to perform well on the same document sets such that an ordering of document sets by Rouge scores is similar for the representations being compared. In the figure, correlation strength is represented by shading where light-toned squares indicate strong correlation (and the darkest squares indicate weak negative correlation). For example, the upper left cell contains r_S between the TF and EV representations. The four squares correspond to r_S values of -0.085, 0.199, 0.245 and 0.267 respectively for summaries of 50, 100, 200 and 400 words.

The analysis illustrates a number of interesting points. First, it demonstrates that none of the representations correlate highly with the human upper bound, meaning that the automatic systems do not necessarily do well on the document sets that may be considered easier as measured by human agreement using Rouge. This suggests that task difficulty does not need to be considered as a possible underlying cause of correlation between the automatic systems. The analysis also illustrates that there is no clear and consistent relationship between summary length and correlation values. Some cells suggest that correlation may have a monotonic linear relationship increasing with length (e.g., TF*EV) while others seem to suggest inverse linear (e.g., TF*RL), quadratic (e.g., EV*HU) and invariant (e.g., EV*EE) relationships with length.

Looking at correlation between automatic systems (i.e., TF, EV, RL, EE and ER), correlation values closer to zero suggest that the systems do well on different document sets and that a combined system might therefore be better. By this reasoning, the largest gains would come from combining TF with any other representation. Among the other automatic systems, the *relation*

[11] Correlation across document set Rouge-1 scores shows similar trends.

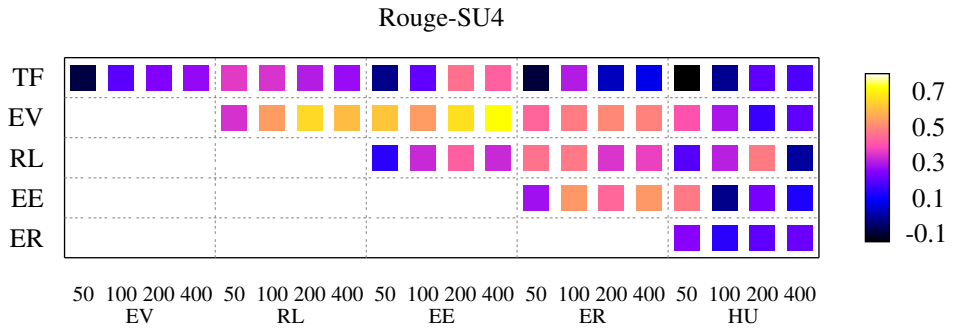

Figure 4: Comparison of representations using Spearman's r_s. Row and column labels correspond to *tf*idf* (*TF*), *event* (*EV*), *relation* (*RL*), *event entity pair* (*EE*), *relation entity pair* (*ER*) and human (*HU*) representations. Lighter toned squares indicate stronger correlation.

representation (RL) shows moderately high potential for combination with its corresponding entity pair representation (ER) with Spearman's r_S values in the range from 0.348 to 0.476. This suggests that ER should not necessarily be considered a simpler representation of the same information captured by RL when comparing results. The *event* representation (EV), by contrast, shows the strongest correlation of any comparison with its corresponding entity pair representation (EE) with r_S values in the range from 0.541 to 0.725.

7.2 Error Analysis

Four document sets were considered for error analysis. These were selected to cover different relative rankings of representations. Rows in Table 4 give the document set ID and list the representations in order of their Rouge-SU4 scores. Inspection of the corresponding document sets suggests that the different approaches compared here are appropriate for different types of summary tasks. Specifically, it suggests that *relation* and *event* representations perform poorly on summarisation tasks that are oriented towards sentiment, description or analysis. However, they do well on document sets that are oriented towards factual information typical of information extraction tasks (though current representations do not capture date, time or numeric information). This supports the notion from the previous section that the different representations evaluated here are complementary.

The document set (d06) for the summaries in Figure 5 illustrates a case where the *relation* and *event* representations perform well with respect

Set	Rank 1	Rank 2	Rank 3
d15	TF (0.046)	RL (0.035)	EV (0.023)
d39	TF (0.033)	EV (0.024)	RL (0.014)
d06	RL (0.094)	EV (0.060)	TF (0.016)
d53	RL (0.078)	TF (0.035)	EV (0.020)

Table 4: DUC 2001 document sets chosen for error analysis and corresponding Rouge-SU4 scores.

to *tf*idf*. The gold standard summary describes a beating event, addressing the basic facts of the Rodney King beating by Los Angeles police as well as the political aftermath which consists primarily of an investigation and a summary of related police brutality events. The difference in performance seems to be due to the fact that relations and events are central to all aspects of this summary and the *relation* and *event* representations clearly do better than *tf*idf* at capturing this information. This summary also illustrates an unintended side-effect of the *relation* representation where the generic relation identification algorithm finds relations between components of lexical compounds or multi-word phrases. The representation for the third sentence in the *RL* summary, for example, includes a relation between `ORG_police` and `XFN_chief` in addition to true positive relations e.g. between `ORG_police` and `PER_darylgates` and false positive relations e.g. between `PER_tombradley` and `ORG_police`.

7.3 Comparison to Supervised Extraction

Related work by Wong et al. (2008) also compares representations for sentence extraction on

TF (0.016) (20/29)	**[S1]** Mr. Williams likened the report to the Knapp Commission, a 1970s blue-ribbon study that exposed widespread corruption in the New York Police Department and led to significant improvements there. **[S2]** "There's no doubt in our mind that the only reason they stopped Joe Morgan was because he is black and he was the first black who happened to come by," said William Barnes, one of the attorneys representing the former ballplayer. **[S3]** Joseph McNamara, retired chief of San Jose's department and now a fellow at Stanford University's Hoover Institution, said he has been getting calls all summer from **[END]** cities around the country about racism and brutality in their departments.
EV (0.060) (9/29)	**[S1]** A high-ranking commission appointed after the beating, under the chairmanship of Mr Warren Christopher, a lawyer and former deputy secretary of state, concluded that the Los Angeles police department got results, in terms of arrests, but had developed a 'siege mentality that alienates the officer from the community'. **[S2]** The images of Los Angeles police swinging nightsticks at King as he lay on the ground, played repeatedly on national news programs, were burned into the national conscience and led to widespread calls for investigation of police brutality. **[S3]** Besides recommending that Mr Gates should go, the Christopher commission urged a policy **[END]** of community policing with more foot patrols, as well as measures to discipline racist police officers and to improve the investigation of complaints about police brutality.
RL (0.094) (3/29)	**[S1]** Mr. Gates opposed the Police Corps because its members would not be professionals. **[S2]** Shortly after Rodney King's beating, a news program on ABC illustrating police brutality showed a still photo of police using a martial-arts weapon against a person being arrested, but there was no mention that the episode involved Operation Rescue. **[S3]** The report was issued yesterday by a commission appointed by Mayor Tom Bradley and Police Chief Daryl Gates in the wake of the videotaped beating March 3 of a black motorist, Rodney King, by Los Angeles police. **[S4]** Investigations have been launched by the FBI, the Los **[END]** Angeles County district attorney's office and the Long Beach Police Department.
HU (0.400) (15/29)	The most important of the many cases of police brutality reported in southern California 1989-1992, was the beating of Rodney King by four Los Angeles officers on March 3, 1991. An investigating commission outlined steps for improvement of the police department and called for the resignation of Chief Gates. Gates did not resign until the following year after the acquittal of the four officers caused massive rioting. Other cases of police brutality arose in Minneapolis, Chicago and Kansas City. Operation Rescue claimed that its non-violent anti-abortion demonstrators were seriously injured by excessive police tactics in more than **[END]** 50 cities.

Figure 5: Example system and *human* (*HU*) summaries where *relation* (*RL*) and *event* (*EV*) representations perform well with respect to the *tf*idf* (*TF*) representation: Police Brutality Document Set (d06).

the DUC 2001 data. However, it uses supervised machine learning (probabilistic support vector machines) to derive a salience function while we focus on unsupervised approaches that can be ported to new domains and tasks without annotation or training. Interestingly, Wong et al.'s results suggest that adding events to a word-based feature set increases the precision of supervised sentence extraction but reduces the recall. By contrast, the current results and analysis provide evidence that word and generic IE-based representations are complementary when using unsupervised salience functions for sentence extraction.

The Wong et al. (2008) paper also provides useful results for comparison to state-of-the art. On the 200 word summarisation task, Wong et al. report Rouge-1 scores of 0.352 and 0.344 respectively for word-based and event-based representations. On the same task, our unsupervised approach achieves Rouge-1 scores of 0.174, 0.232, 0.229, 0.241 and 0.253 respectively for the *tf*idf*, *event*, *event entity pair*, *relation* and *relation entity pair* representations. Wong et al.'s best overall score is 0.396 using a representation that combines surface, content and relevance features.

8 Conclusion

Experiments were presented that compare the effect of various source document representations on the accuracy of automatic summarisation. This serves as an extrinsic evaluation of generic relation extraction, a domain-neutral and fully portable approach to relation identification and characterisation. Results demonstrate that GRE is an effective representation for sentence extraction for multi-document summarisation. Performance for the *relation* representation is significantly better than a non-trivial *tf*idf* baseline across the range of summary lengths explored. Performance is also at least as good as a comparable but less general representation based on event extraction. Correlation analysis suggests that different representations are complementary due to the fact that they perform well on different document sets. Error analysis supports this conclusion, suggesting that the *relation* and *event* representations perform poorly on summarisation tasks that are oriented towards e.g. sentiment, description or analysis while they perform well on tasks that focus on fact-oriented information.

Acknowledgments

This work was supported by Scottish Enterprise Edinburgh-Stanford Link grant R37588 as part of the EASIE project at the University of Edinburgh. It would not have been possible without the guidance of Claire Grover and Mirella Lapata.

References

Daniel Bienstock and Garud Iyengar. 2004. Faster approximation algorithms for packing and covering problems. Technical Report TR-2004-09, Columbia University.

David Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

Nancy Chinchor. 1998. Overview of MUC-7. In *Proceedings of the 7th Message Understanding Conference*, Fairfax, VA, USA.

James R. Curran and Stephen Clark. 2003. Language independent NER using a maximum entropy tagger. In *Proceedings of the 7th Conference on Natural Language Learning*, Edmonton, Alberta, Canada.

Gerald DeJong. 1982. An overview of the FRUMP system. In Wendy G. Lehnert and Martin H. Ringle, editors, *Strategies for Natural Language Processing*, pages 149–176. Lawrence Erlbaum Associates, Hillsdale, NJ.

Elena Filatova and Vasileios Hatzivassiloglou. 2004. Event-based extractive summarization. In *Proceedings of the ACL Text Summarization Branches Out Workshop*, Barcelona, Spain.

Claire Grover, Colin Matheson, Andrei Mikheev, and Marc Moens. 2000. LT TTT—a flexible tokenisation tool. In *Proceedings of the 2nd International Conference on Language Resources and Evaluation*, Athens, Greece.

Ben Hachey. 2009. *Towards Generic Relation Extraction*. Ph.D. thesis, University of Edinburgh.

Udo Hahn and Ulrich Reimer. 1999. Knowledge-based text summarization: Salience and generalization operators for knowledge base abstraction. In Inderjeet Mani and Mark T. Maybury, editors, *Advances in Automatic Text Summarization*, pages 215–232. MIT Press, Cambridge, MA.

Sanda M. Harabagiu and Steven J. Maiorano. 2002. Multi-document summarization with GISTexter. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation*, Las Palmas, Spain.

Dorit S. Hochbaum. 1997. Approximating covering and packing problems: set cover, vertex cover, independent set and related problems. In Dorit S. Hochbaum, editor, *Approximation Algorithms for NP-Hard Problems*, pages 94–143. PWS Publishing Company, Boston, MA.

Dekang Lin. 1998. Dependency-based evaluation of MINIPAR. In *Proceedings of the LREC Workshop Evaluation of Parsing Systems*, Granada, Spain.

Chin-Yew Lin. 2004. ROUGE: a package for automatic evaluation of summaries. In *Proceedings of the ACL Text Summarization Branches Out Workshop*, Barcelona, Spain.

Hans P. Luhn. 1958. The automatic creation of literature abstracts. *IBM Journal of Research and Development*, 2(2).

Inderjeet Mani. 2001. *Automatic Summarization*. John Benjamins, Amsterdam/Philadelphia.

Kathleen R. McKeown, Desmond A. Jordan, and Vasileios Hatzivassiloglou. 1998. Generating patient-specific summaries of online literature. In *Proceedings of the AAAI Spring Symposium on Intelligent Text Summarization*, Stanford, CA, USA.

Ellen Riloff. 1996. Automatically generating extraction patterns from untagged text. In *Proceedings of the 14th National Conference on Artificial Intelligence*, Portland, OR, USA.

Horacio Saggion and Guy Lapalme. 2002. Generating indicative-informative summaries with SumUM. *Computational Linguistics*, 28(4):497–526.

Karen Spärck Jones. 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28(1):11–21.

Karen Spärck Jones. 2007. Automatic summarising: The state of the art. *Information Processing and Management*, 43:1449–1481.

Simone Teufel and Marc Moens. 2002. Summarizing scientific articles – experiments with relevance and rhetorical status. *Computational Linguistics*, 28(4):409–445.

Michael White and Claire Cardie. 2002. Selecting sentences for multidocument summaries using randomized local search. In *Proceedings of the ACL Workshop on Automatic Summarization*, Philadelphia, PA, USA.

Michael White, Tanya Korelsky, Claire Cardie, Vincent Ng, David Pierce, and Kiri Wagstaff. 2001. Multidocument summarization via information extraction. In *Proceedings of the 1st International Conference on Human Language Technology Research*, San Diego, CA, USA.

Kam-Fai Wong, Mingli Wu, and Wenjie Li. 2008. Extractive summarization using supervised and semi-supervised learning. In *Proceedings of the 22nd International Conference on Computational Linguistics*, Manchester, UK.