Towards Robust Abstractive Multi-Document Summarization: A Caseframe Analysis of Centrality and Domain

Jackie Chi Kit Cheung
University of Toronto
10 King's College Rd., Room 3302
Toronto, ON, Canada M5S 3G4
jcheung@cs.toronto.edu

Gerald Penn
University of Toronto
10 King's College Rd., Room 3302
Toronto, ON, Canada M5S 3G4
gpenn@cs.toronto.edu

Abstract

In automatic summarization, *centrality* is the notion that a summary should contain the core parts of the source text. Current systems use centrality, along with redundancy avoidance and some sentence compression, to produce mostly extractive summaries. In this paper, we investigate how summarization can advance past this paradigm towards robust abstraction by making greater use of the domain of the source text. We conduct a series of studies comparing human-written model summaries to system summaries at the semantic level of *caseframes*. We show that model summaries (1) are more abstractive and make use of more sentence aggregation, (2) do not contain as many topical caseframes as system summaries, and (3) cannot be reconstructed solely from the source text, but can be if texts from in-domain documents are added. These results suggest that substantial improvements are unlikely to result from better optimizing centrality-based criteria, but rather more domain knowledge is needed.

1 Introduction

In automatic summarization, *centrality* has been one of the guiding principles for content selection in extractive systems. We define centrality to be the idea that a summary should contain the parts of the source text that are most similar or representative of the source text. This is most transparently illustrated by the Maximal Marginal Relevance (MMR) system of Carbonell and Goldstein (1998), which defines the summarization objective to be a linear combination of a centrality term and a non-redundancy term.

Since MMR, much progress has been made on more sophisticated methods of measuring centrality and integrating it with non-redundancy (See Nenkova and McKeown (2011) for a recent survey). For example, term weighting methods such as the signature term method of Lin and Hovy (2000) pick out salient terms that occur more often than would be expected in the source text based on frequencies in a background corpus. This method is a core component of the most successful summarization methods (Conroy et al., 2006).

While extractive methods based on centrality have thus achieved success, there has long been recognition that abstractive methods are ultimately more desirable. One line of work is in text simplification and sentence fusion, which focus on the ability of abstraction to achieve a higher compression ratio (Knight and Marcu, 2000; Barzilay and McKeown, 2005). A less examined issue is that of aggregation and information synthesis. A key part of the usefulness of summaries is that they provide some synthesis or analysis of the source text and make a more general statement that is of direct relevance to the user. For example, a series of related events can be aggregated and expressed as a trend.

The position of this paper is that centrality is not enough to make substantial progress towards abstractive summarization that is capable of this type of semantic inference. Instead, summarization systems need to make more use of domain knowledge. We provide evidence for this in a series of studies on the TAC 2010 guided summarization data set that examines how the behaviour of automatic summarizers can or cannot be distinguished from human summarizers. First, we confirm that abstraction is a desirable goal, and

provide a quantitative measure of the degree of sentence aggregation in a summarization system. Second, we show that centrality-based measures are unlikely to lead to substantial progress towards abstractive summarization, because current top-performing systems already produce summaries that are more "central" than humans do. Third, we consider how domain knowledge may be useful as a resource for an abstractive system, by showing that key parts of model summaries can be reconstructed from the source plus related in-domain documents.

Our contributions are novel in the following respects. First, our analyses are performed at the level of *caseframes*, rather at the level of words or syntactic dependencies as in previous work. Caseframes are shallow approximations of semantic roles which are well suited to characterizing a domain by its slots. Furthermore, we take a *developmental* rather than *evaluative* perspective—our goal is not to develop a new evaluation measure as defined by correlation with human responsiveness judgments. Instead, our studies reveal useful criteria with which to distinguish human-written and system summaries, helping to guide the development of future summarization systems.

2 Related Work

Domain-dependent template-based summarization systems have been an alternative to extractive systems which make use of rich knowledge about a domain and information extraction techniques to generate a summary, possibly using a natural language generation system (Radev and McKeown, 1998; White et al., 2001; McKeown et al., 2002). This paper can be seen as a first step towards reconciling the advantages of domain knowledge with the resource-lean extraction approaches popular today.

As noted above, Lin and Hovy's (2000) signature terms have been successful in discovering terms that are specific to the source text. These terms are identified by a log-likelihood ratio test based on their relative frequencies in relevant and irrelevant documents. They were originally proposed in the context of single-document summarization, where they were calculated using in-domain (relevant) vs. out-of-domain (irrelevant) text. In multi-document summarization, the in-domain text has been replaced by the source text cluster (Conroy et al., 2006), thus they are now used as a form of centrality-based features. In this paper, we use guided summarization data as an opportunity to reopen the investigation into the effect of domain, because multiple document clusters from the same domain are available.

Summarization evaluation is typically done by comparing system output to human-written model summaries, and are validated by their correlation with user responsiveness judgments. The comparison can be done at the word level, as in ROUGE (Lin, 2004), at the syntactic level, as in Basic Elements (Hovy et al., 2006), or at the level of summary content units, as in the Pyramid method (Nenkova and Passonneau, 2004). There are also automatic measures which do not require model summaries, but compare against the source text instead (Louis and Nenkova, 2009; Saggion et al., 2010).

Several studies complement this paper by examining the best possible extractive system using current evaluation measures, such as ROUGE (Lin and Hovy, 2003; Conroy et al., 2006). They find that the best possible extractive systems score higher or as highly than human summarizers, but it is unclear whether this means the oracle summaries are actually as useful as human ones in an extrinsic setting. Genest et al. (2009) ask humans to create extractive summaries, and find that they score in between current automatic systems and human-written abstracts on responsiveness, linguistic quality, and Pyramid score. In the lecture domain, He et al. (1999; 2000) find that lecture transcripts that have been manually highlighted with key points improve students' quiz scores more than when using automated summarization techniques or when providing only the lecture transcript or slides.

Jing and McKeown (2000) manually analyzed 30 human-written summaries, and find that 19% of sentences cannot be explained by *cut-and-paste* operations from the source text. Saggion and Lapalme (2002) similarly define a list of transformations necessary to convert source text to summary text, and manually analyzed their frequencies. Copeck and Szpakowicz (2004) find that at most 55% of vocabulary items found in model summaries occur in the source text, but they do not investigate where the other vocabulary items might be found.

Sentence:	
At one point, two bomb squad trucks sped to the school after a backpack scare.	
Dependencies:	
$num(point, one)$	$prep\_at(sped, point)$
$num(trucks, two)$	$nn(trucks, bomb)$
$nn(trucks, squad)$	$nsubj(sped, trucks)$
$root(ROOT, sped)$	$det(school, the)$
$prep\_to(sped, school)$	$det(scare, a)$
$nn(scare, backpack)$	$prep\_after(sped, scare)$
Caseframes:	
$(speed, prep\_at)$	$(speed, nsubj)$
$(speed, prep\_to)$	$(speed, prep\_after)$

Table 1: A sentence decomposed into its dependency edges, and the caseframes derived from those edges that we consider (in black).

3 Theoretical basis of our analysis

Many existing summarization evaluation methods rely on word or N-gram overlap measures, but these measures are not appropriate for our analysis. Word overlap can occur due to shared proper nouns or entity mentions. Good summaries should certainly contain the salient entities in the source text, but when assessing the effect of the domain, different domain instances (i.e., different document clusters in the same domain) would be expected to contain different salient entities. Also, the realization of entities as noun phrases depends strongly on context, which would confound our analysis if we do not also correctly resolve coreference, a difficult problem in its own right. We leave such issues to other work (Nenkova and McKeown, 2003, e.g.).

Domains would rather be expected to share *slots* (a.k.a. *aspects*), which require a more semantic level of analysis that can account for the various ways in which a particular slot can be expressed. Another consideration is that the structures to be analyzed should be extracted automatically. Based on these criteria, we selected *caseframes* to be the appropriate unit of analysis. A caseframe is a shallow approximation of the semantic role structure of a proposition-bearing unit like a verb, and are derived from the dependency parse of a sentence[1].

[1]Note that caseframes are distinct from (though directly inspired by) the similarly named *case frames* of Case Grammar (Fillmore, 1968).

Relation	**Caseframe Pair**	**Sim.**
Degree	$(kill, dobj)$ $(wound, dobj)$	0.82
Causal	$(kill, dobj)$ $(die, nsubj)$	0.80
Type	$(rise, dobj)$ $(drop, prep\_to)$	0.81

Figure 1: Sample pairs of similar caseframes by relation type, and the similarity score assigned to them by our distributional model.

In particular, they are $(gov, role)$ pairs, where gov is a proposition-bearing element, and $role$ is an approximation of a semantic role with gov as its head (See Figure 1 for examples). Caseframes do not consider the dependents of the semantic role approximations.

The use of caseframes is well grounded in a variety of NLP tasks relevant to summarization such as coreference resolution (Bean and Riloff, 2004), and information extraction (Chambers and Jurafsky, 2011), where they serve the central unit of semantic analysis. Related semantic representations are popular in Case Grammar and its derivative formalisms such as frame semantics (Fillmore, 1982).

We use the following algorithm to extract caseframes from dependency parses. First, we extract those dependency edges with a relation type of subject, direct object, indirect object, or prepositional object (with the preposition indicated), along with their governors. The governor must be a verb, event noun (as defined by the hyponyms of the WordNet EVENT synset), or nominal or adjectival predicate. Then, a series of deterministic transformations are applied to the syntactic relations to account for voicing alternations, control, raising, and copular constructions.

3.1 Caseframe Similarity

Direct caseframe matches account for some variation in the expression of slots, such as voicing alternations, but there are other reasons different caseframes may indicate the same slot (Figure 1). For example, $(kill, dobj)$ and $(wound, dobj)$ both indicate the victim of an attack, but differ by the degree of injury to the victim. $(kill, dobj)$ and $(die, nsubj)$ also refer to a victim, but are linked by a causal relation. $(rise, dobj)$ and

$(drop, prep\_to)$ on the other hand simply share a named entity type (in this case, numbers). To account for these issues, we measure caseframe similarity based on their distributional similarity in a large training corpus.

First, we construct vector representations of each caseframe, where the dimensions of the vector correspond to the lemma of the head word that fills the caseframe in the training corpus. For example, *kicked the ball* would result in a count of 1 added to the caseframe $(kick, dobj)$ for the context word $ball$. Then, we rescale the counts into pointwise mutual information values, which has been shown to be more effective than raw counts at detecting semantic relatedness (Turney, 2001). Similarity between caseframes can then be compared by cosine similarity between the their vector representations.

For training, we use the AFP portion of the Gigaword corpus (Graff et al., 2005), which we parsed using the Stanford parser's typed dependency tree representation with collapsed conjunctions (de Marneffe et al., 2006). For reasons of sparsity, we only considered caseframes that appear at least five times in the guided summarization corpus, and only the 3000 most common lemmata in Gigaword as context words.

3.2 An Example

To illustrate how caseframes indicate the slots in a summary, we provide the following fragment of a model summary from TAC about the *Unabomber trial*:

(1) *In Sacramento, Theodore Kaczynski faces a 10-count federal indictment for 4 of the 16 mail bomb attacks attributed to the Unabomber in which two people were killed. If found guilty, he faces a death penalty. ... He has pleaded innocent to all charges. U.S. District Judge Garland Burrell Jr. presides in Sacramento.*

All of the slots provided by TAC for the *Investigations and Trials* domain can be identified by one or more caseframes. The DEFENDANT can be identified by $(face, nsubj)$, and $(plead, nsubj)$; the CHARGES by $(face, dobj)$; the REASON by $(indictment, prep\_for)$; the SENTENCE by $(face, dobj)$; the PLEAD by $(plead, dobj)$; and the INVESTIGATOR by $(preside, nsubj)$.

4 Experiments

We conducted our experiments on the data and results of the TAC 2010 summarization workshop. This data set contains 920 newspaper articles in 46 topics of 20 documents each. Ten are used in an initial guided summarization task, and ten are used in an update summarization task, in which a summary must be produced assuming that the original ten documents had already been read. All summaries have a word length limit of 100 words. We analyzed the results of the two summarization tasks separately in our experiments.

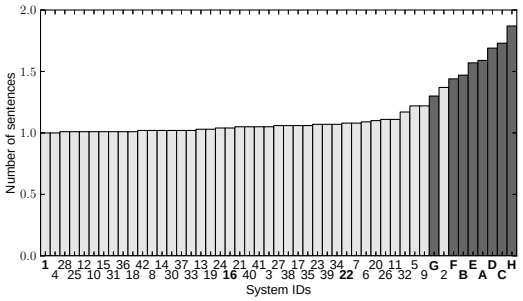
The 46 topics belong to five different categories or domains: *Accidents and natural disasters*, *Criminal or terrorist attacks*, *Health and safety*, *Endangered resources*, and *Investigations and trials*. Each domain is associated with a template specifying the type of information that is expected in the domain, such as the participants in the event or the time that the event occurred.

In our study, we compared the characteristics of summaries generated by the eight human summarizers with those generated by the peer summaries, which are basically extractive systems. There are 43 peer summarization systems, including two baselines defined by NIST. We refer to systems by their ID given by NIST, which are alphabetical for the human summarizers (A to H), and numeric for the peer summarizers (1 to 43). We removed two peer systems (systems 29 and 43) which did not generate any summary text in the workshop, presumably due to software problems. For each measure that we consider, we compare the average among the human-written summaries to the three individual peer systems, which we chose in order to provide a representative sample of the average and best performance of the automatic systems according to current evaluation methods. These systems are all primarily extractive, like most of the systems in the workshop:

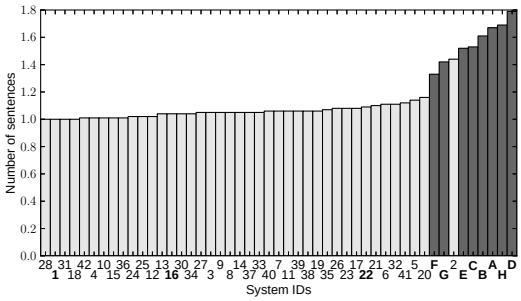
Peer average The average of the measure among the 41 peer summarizers.

Peer 16 This system scored the highest in responsiveness scores on the original summarization task and in ROUGE-2, responsiveness, and Pyramid score in the update task.

Peer 22 This system scored the highest in ROUGE-2 and Pyramid score in the original summarization task.



(a) Initial guided summarization task



(b) Update summarization task

Figure 2: Average sentence cover size: the average number of sentences needed to generate the caseframes in a summary sentence (Study 1). Model summaries are shown in darker bars. Peer system numbers that we focus on are in bold.

Condition	Initial	Update
Model average	1.58	1.57
Peer average	1.06	1.06
Peer 1	1.00	1.00
Peer 16	1.04	1.04
Peer 22	1.08	1.09

Table 2: The average number of source text sentences needed to cover a summary sentence. The model average is statistically significantly different from all the other conditions $p < 10^{-7}$ (Study 1).

Peer 1 The NIST-defined baseline, which is the leading sentence baseline from the most recent document in the source text cluster. This system scored the highest on linguistic quality in both tasks.

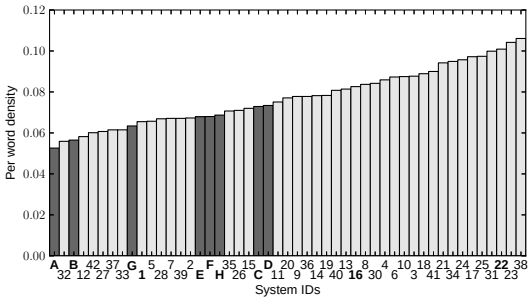
4.1 Study 1: Sentence aggregation

We first confirm that human summarizers are more prone to sentence aggregation than system summarizers, showing that abstraction is indeed a desirable goal. To do so, we propose a measure to quantify the degree of sentence aggregation exhibited by a summarizer, which we call **average sentence cover size**. This is defined to be the minimum number of sentences from the source text needed to cover all of the caseframes found in a summary sentence (for those caseframes that can be found in the source text at all), averaged over all of the summary sentences. Purely extractive systems would thus be expected to score 1.0, as would systems that perform text compression by removing constituents of a source text sentence. Human summarizers would be expected to score higher, if they actually aggregate information from multiple points in the source text.

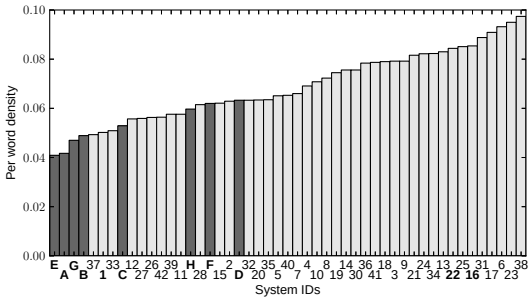
To illustrate, suppose we assign arbitrary indices to caseframes, a summary sentence contains caseframes $\{1, 2, 3, 4, 5\}$, and the source text contains three sentences with caseframes, which can be represented as a nested set $\{\{1, 3, 4\}, \{2, 5, 6\}, \{1, 4, 7\}\}$. Then, the summary sentence can be covered by two sentences from the source text, namely $\{\{1, 3, 4\}, \{2, 5, 6\}\}$.

This problem is actually an instance of the minimum set cover problem, in which sentences are sets, and caseframes are set elements. Minimum set cover is NP-hard in general, but the standard integer programming formulation of set cover sufficed for our data set; we used ILOG CPLEX 12.4's mixed integer programming mode to solve all the set cover problems optimally.

Results Figure 2 shows the ranking of the summarizers by this measure. Most peer systems have a low average sentence cover size of close to 1, which reflects the fact that they are purely or almost purely extractive. Human model summarizers show a higher degree of aggregation in their summaries. The averages of the tested conditions are shown in Table 2, and are statistically significant. Peer 2 shows a relatively high level of aggregation despite being an extractive system. Upon inspection of its summaries, it appears that Peer 2 tends to select many datelines, and without punctuation to separate them from the rest of the summary, our automatic analysis tools incorrectly merged many sentences together, resulting in incorrect parses and novel caseframes not found in



(a) Initial guided summarization task



(b) Update summarization task

Figure 3: Density of signature caseframes (Study 2).

Topic: Unabomber trial
$(charge, dobj)$, $(kill, dobj)$,
$(trial, prep\_of)$, $(bombing, prep\_in)$
Topic: Mangrove forests
$(beach, prep\_of)$, $(save, dobj)$
$(development, prep\_of)$, $(recover, nsubj)$
Topic: Bird Flu
$(infect, prep\_with)$, $(die, nsubj)$
$(contact, dobj)$, $(import, prep\_from)$

Figure 4: Examples of signature caseframes found in Study 2.

Condition	**Initial**	**Update**
Model average	0.065	0.052
Peer average	0.080*	0.072*
Peer 1	0.066	0.050
Peer 16	0.083*	0.085*
Peer 22	0.101*	0.084*

Table 3: Signature caseframe densities for different sets of summarizers, for the initial and update guided summarization tasks (Study 2). *: $p < 0.005$.

the source text.

4.2 Study 2: Signature caseframe density

Study 1 shows that human summarizers are more abstractive in that they aggregate information from multiple sentences in the source text, but how is this aggregation performed? One possibility is that human summary writers are able to pack a greater number of salient caseframes into their summaries. That is, humans are fundamentally relying on centrality just as automatic summarizers do, and are simply able to achieve higher compression ratios by being more succinct. If this is true, then sentence fusion methods over the source text alone might be able to solve the problem. Unfortunately, we show that this is false and that system summaries are actually more central than model ones.

To extract topical caseframes, we use Lin and Hovy's (2000) method of calculating signature terms, but extend the method to apply it at the caseframe rather than the word level. We follow Lin and Hovy (2000) in using a significance threshold of 0.001 to determine signature caseframes[2]. Figure 4 shows examples of signature caseframes for several topics. Then, we calculate the **signature caseframe density** of each of the summarization systems. This is defined to be the number of signature caseframes in the set of summaries divided by the number of words in that set of summaries.

Results Figure 3 shows the density for all of the summarizers, in ascending order of density. As can be seen, the human abstractors actually tend to use fewer signature caseframes in their summaries than automatic systems. Only the leading baseline is indistinguishable from the model average. Table 3 shows the densities for the conditions that we described earlier. The differences in density between the human average and the non-baseline conditions are highly statistically significant, according to paired two-tailed Wilcoxon signed-rank tests for the statistic calculated for each topic cluster.

These results show that human abstractors do

[2]We tried various other thresholds, but the results were much the same.

Threshold	0.9		0.8	
Condition	**Init.**	**Up.**	**Init.**	**Up.**
Model average	0.066	0.052	0.062	0.047
Peer average	0.080	0.071	0.071	0.063
Peer 1	0.068	0.050	0.060	0.044
Peer 16	0.083	0.086	0.072	0.077
Peer 22	0.100	0.086	0.084	0.075

Table 4: Density of signature caseframes after merging to various threshold for the initial (**Init.**) and update (**Up.**) summarization tasks (Study 2).

not merely repeat the caseframes that are indicative of a topic cluster or use minor grammatical alternations in their summaries. Rather, a genuine sort of abstraction or distillation has taken place, either through paraphrasing or semantic inference, to transform the source text into the final informative summary.

Merging Caseframes We next investigate whether simple paraphrasing could account for the above results; it may be the case that human summarizers simply replace words in the source text with synonyms, which can be detected with distributional similarity. Thus, we merged similar caseframes into clusters according to the distributional semantic similarity defined in Section 3.1, and then repeated the previous experiment. We chose two relatively high levels of similarity (0.8 and 0.9), and used complete-link agglomerative (i.e., bottom-up) clustering to merge similar caseframes. That is, each caseframe begins as a separate cluster, and the two most similar clusters are merged at each step until the desired similarity threshold is reached. Cluster similarity is defined to be the minimum similarity (or equivalently, maximum distance) between elements in the two clusters; that is, $\max_{c \in C_1, c' \in C_2} -sim(c, c')$. Complete-link agglomerative clustering tends to form coherent clusters where the similarity between any pair within a cluster is high (Manning et al., 2008).

Cluster Results Table 4 shows the results after the clustering step, with similarity thresholds of 0.9 and 0.8. Once again, model summaries contain a lower density of signature caseframes. The statistical significance results are unchanged. This indicates that simple paraphrasing alone cannot account for the difference in the signature caseframe densities, and that some deeper abstraction or semantic inference has occurred.

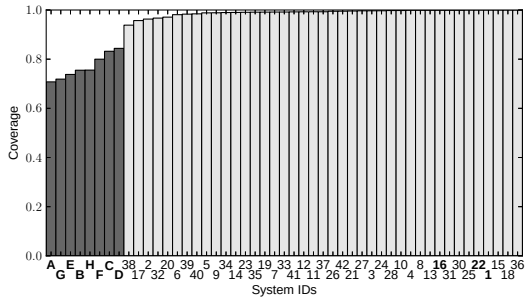
Note that we are not claiming that a lower density of signature caseframes necessarily correlates with a more informative summary. For example, some automatic summarizers are comparable to the human abstractors in their relatively low density of signature caseframes, but these turn out to be the lowest performing summarization systems by all measures in the workshop, and they are unlikely to rival human abstractors in any reasonable evaluation of summary informativeness. It does, however, appear that further optimizing centrality-based measures alone is unlikely to produce better informative summaries, even if we analyze the summary at a syntactic/semantic rather than lexical level.

4.3 Study 3: Summary Reconstruction

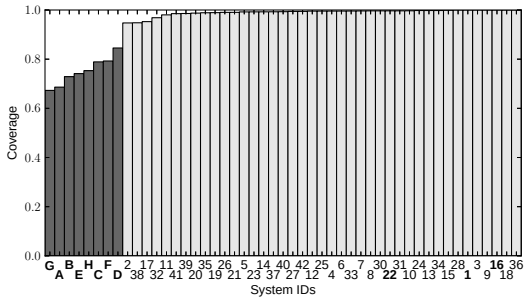
The above studies show that the higher degree of abstraction in model summaries cannot be explained by better compression of topically salient caseframes alone. We now switch perspectives to ask how model summaries might be automatically generated at all. We will show that they cannot be reconstructed solely from the source text, extending Copeck and Szpakowicz (2004)'s result to caseframes. However, we also show that if articles from the same domain are added, reconstruction then becomes possible. Our measure of whether a model summary can be reconstructed is **caseframe coverage**. We define this to be the proportion of caseframes in a summary that is contained by some reference set. This is thus a score between 0 and 1. Unlike in the previous study, we use the full set of caseframes, not just signature caseframes, because our goal now to create a hypothesis space from which it is in principle possible to generate the model summaries.

Results We first calculated caseframe coverage with respect to the source text alone (Figure 5). As expected, automatic systems show close to perfect coverage, because of their basically extractive nature, while model summaries show much lower coverage. These statistics are summarized by Table 5. These results present a fundamental limit to extractive systems, and also text simplification and sentence fusion methods based solely on the source text.

The Impact of Domain Knowledge How might automatic summarizers be able to acquire these



(a) Initial guided summarization task



(b) Update summarization task

Figure 5: Coverage of summary text caseframes in source text (Study 3).

Condition	Initial	Update
Model average	0.77	0.75
Peer average	0.99	0.99
Peer 1	1.00	1.00
Peer 16	1.00	1.00
Peer 22	1.00	1.00

Table 5: Coverage of caseframes in summaries with respect to the source text. The model average is statistically significantly different from all the other conditions $p < 10^{-8}$ (Study 3).

Reference corpus	Initial	Update
Source text only	0.77	0.75
+out-of-domain	0.91	0.91
+in-domain	0.98	0.97

Table 6: The effect on caseframe coverage of adding in-domain and out-of-domain documents. The difference between adding in-domain and out-of-domain text is significant $p < 10^{-3}$ (Study 3).

caseframes from other sources? Traditional systems that perform semantic inference do so from a set of known facts about the domain in the form of a knowledge base, but as we have seen, most extractive summarization systems do not make much use of in-domain corpora. We examine adding in-domain text to the source text to see how this would affect coverage.

Recall that the 46 topics in TAC 2010 are categorized into five domains. To calculate the impact of domain knowledge, we add all the documents that belong in the same domain to the source text to calculate coverage. To ensure that coverage does not increase simply due to increasing the size of the reference set, we compare to the baseline of adding the same number of documents that belong to another domain. As shown in Table 6, the effect of adding more in-domain text on caseframe coverage is substantial, and noticeably more than using out-of-domain text. In fact, nearly all caseframes can be found in the expanded set of articles. The implication of this result is that it may be possible to generate better summaries by mining in-domain text for relevant caseframes.

5 Conclusion

We have presented a series of studies to distinguish human-written informative summaries from the summaries produced by current systems. Our studies are performed at the level of caseframes, which are able to characterize a domain in terms of its slots. First, we confirm that model summaries are more abstractive and aggregate information from multiple source text sentences. Then, we show that this is not simply due to summary writers packing together source text sentences containing topical caseframes to achieve a higher compression ratio, even if paraphrasing is taken into account. Indeed, model summaries cannot be reconstructed from the source text alone. However, our results are also positive in that we find that nearly all model summary caseframes can be found in the source text together with some in-domain documents.

Current summarization systems have been heavily optimized towards centrality and lexical-semantical reasoning, but we are nearing the bottom of the barrel. Domain inference, on the other hand, and a greater use of in-domain documents as a knowledge source for domain inference, are very promising indeed. Mining useful caseframes

for a sentence fusion-based approach has the potential, as our experiments have shown, to deliver results in just the areas where current approaches are weakest.

Acknowledgements

This work is supported by the Natural Sciences and Engineering Research Council of Canada.

References

Regina Barzilay and Kathleen R. McKeown. 2005. Sentence fusion for multidocument news summarization. *Computational Linguistics*, 31(3):297–328.

David Bean and Ellen Riloff. 2004. Unsupervised learning of contextual role knowledge for coreference resolution. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*.

Jaime Carbonell and Jade Goldstein. 1998. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 335–336. ACM.

Nathanael Chambers and Dan Jurafsky. 2011. Template-based information extraction without the templates. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 976–986, Portland, Oregon, USA, June. Association for Computational Linguistics.

John M. Conroy, Judith D. Schlesinger, and Dianne P. O'Leary. 2006. Topic-focused multi-document summarization using an approximate oracle score. In *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, pages 152–159, Sydney, Australia, July. Association for Computational Linguistics.

Terry Copeck and Stan Szpakowicz. 2004. Vocabulary agreement among model summaries and source documents. In *Proceedings of the 2004 Document Understanding Conference (DUC)*.

Marie-Catherine de Marneffe, Bill MacCartney, and Christopher D. Manning. 2006. Generating typed dependency parses from phrase structure parses. In *In LREC 2006*.

Charles Fillmore. 1968. The case for case. In E. Bach and R. T. Harms, editors, *Universals in Linguistic Theory*, pages 1–88. Holt, Reinhart, and Winston, New York.

Charles J. Fillmore. 1982. Frame semantics. *Linguistics in the Morning Calm*, pages 111–137.

Pierre-Etienne Genest, Guy Lapalme, and Mehdi Yousfi-Monod. 2009. Hextac: the creation of a manual extractive run. In *Proceedings of the Second Text Analysis Conference, Gaithersburg, Maryland, USA. National Institute of Standards and Technology*.

David Graff, Junbo Kong, Ke Chen, and Kazuaki Maeda. 2005. English gigaword second edition. *Linguistic Data Consortium, Philadelphia*.

Liwei He, Elizabeth Sanocki, Anoop Gupta, and Jonathan Grudin. 1999. Auto-summarization of audio-video presentations. In *Proceedings of the Seventh ACM International Conference on Multimedia*. ACM.

Liwei He, Elizabeth Sanocki, Anoop Gupta, and Jonathan Grudin. 2000. Comparing presentation summaries: slides vs. reading vs. listening. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '00, pages 177–184, New York, NY, USA. ACM.

Eduard Hovy, Chin-Yew Lin, Liang Zhou, and Junichi Fukumoto. 2006. Automated summarization evaluation with Basic Elements. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC)*, pages 899–902.

IBM. *IBM ILOG CPLEX Optimization Studio V12.4*.

Hongyan Jing and Kathleen R. McKeown. 2000. Cut and paste based text summarization. In *Proceedings of the 1st North American Chapter of the Association for Computational Linguistics Conference*, pages 178–185.

Kevin Knight and Daniel Marcu. 2000. Statistics-based summarization-step one: Sentence compression. In *Proceedings of the National Conference on Artificial Intelligence*.

Chin-Yew Lin and Eduard Hovy. 2000. The automated acquisition of topic signatures for text summarization. In *Proceedings of the 18th Conference on Computational Linguistics - Volume 1*, COLING '00, pages 495–501, Stroudsburg, PA, USA. Association for Computational Linguistics.

Chin-Yew Lin and Eduard Hovy. 2003. The potential and limitations of automatic sentence extraction for summarization. In *Proceedings of the HLT-NAACL 03 on Text Summarization Workshop*. Association for Computational Linguistics.

Chin Y. Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In Stan Szpakowicz and Marie-Francine Moens, editors, *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, pages 74–81, Barcelona, Spain, July. Association for Computational Linguistics.

Annie Louis and Ani Nenkova. 2009. Automatically evaluating content selection in summarization without human models. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language*

Processing. Association for Computational Linguistics.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze, 2008. *Introduction to Information Retrieval*, chapter 17. Cambridge University Press.

Kathleen R. McKeown, Regina Barzilay, David Evans, Vasileios Hatzivassiloglou, Judith L. Klavans, Ani Nenkova, Carl Sable, Barry Schiffman, and Sergey Sigelman. 2002. Tracking and summarizing news on a daily basis with Columbia's Newsblaster. In *Proceedings of the Second International Conference on Human Language Technology Research*, pages 280–285. Morgan Kaufmann Publishers Inc.

Ani Nenkova and Kathleen McKeown. 2003. References to named entities: a corpus study. In *Companion Volume of the Proceedings of HLT-NAACL 2003 - Short Papers*. Association for Computational Linguistics.

Ani Nenkova and Kathleen McKeown. 2011. Automatic summarization. *Foundations and Trends in Information Retrieval*, 5(2):103–233.

Ani Nenkova and Rebecca Passonneau. 2004. Evaluating content selection in summarization: The pyramid method. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, volume 2004, pages 145–152.

Dragomir R. Radev and Kathleen R. McKeown. 1998. Generating natural language summaries from multiple on-line sources. *Computational Linguistics*, 24(3):470–500.

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

Horacio Saggion, Juan-Manuel Torres-Moreno, Iria Cunha, and Eric SanJuan. 2010. Multilingual summarization evaluation without human models. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pages 1059–1067. Association for Computational Linguistics.

Peter Turney. 2001. Mining the web for synonyms: PMI-IR versus LSA on TOEFL. In *Proceedings of the Twelth European Conference on Machine Learning (ECML-2001)*, pages 491–502.

Michael White, Tanya Korelsky, Claire Cardie, Vincent Ng, David Pierce, and Kiri Wagstaff. 2001. Multidocument summarization via information extraction. In *Proceedings of the First International Conference on Human Language Technology Research*. Association for Computational Linguistics.