Seed and Grow: Augmenting Statistically Generated Summary Sentences using Schematic Word Patterns

Stephen Wan[†‡] **Robert Dale**[†] **Mark Dras**[†]
[†]Centre for Language Technology
Department of Computing
Macquarie University
Sydney, NSW 2113
swan,madras,rdale@ics.mq.edu.au

Cécile Paris[‡]
[‡]ICT Centre
CSIRO
Sydney, Australia
Cecile.Paris@csiro.au

Abstract

We examine the problem of content selection in statistical novel sentence generation. Our approach models the processes performed by professional editors when incorporating material from additional sentences to support some initially chosen key summary sentence, a process we refer to as *Sentence Augmentation*. We propose and evaluate a method called "Seed and Grow" for selecting such auxiliary information. Additionally, we argue that this can be performed using schemata, as represented by word-pair co-occurrences, and demonstrate its use in statistical summary sentence generation. Evaluation results are supportive, indicating that a schemata model significantly improves over the baseline.

1 Introduction

In the context of automatic text summarisation, we examine the problem of statistical novel sentence generation, with the aim of moving from the current state-of-the-art of sentence extraction to abstract-like summaries. In particular, we focus on the task of selecting content to include within a generated sentence.

Our approach to novel sentence generation is to model the processes underlying summarisation as performed by professional editors and abstractors. An example of the target output of this kind of generation is presented in Figure 1. In this example, the human authored summary sentence was taken verbatim from the executive summary of a United Nations proposal for the provision of aid addressing a particular humanitarian crisis. Such documents typically exceed a hundred pages.

Human-Authored Summary Sentence:
Repeated [poor seasonal rains]$_1$ [in 2004]$_2$, culminating in [food insecurity]$_3$, indicate [another year]$_4$ of crisis, the scale of which is larger than last year's and is further [exacerbated by diminishing coping assets]$_5$ [in both rural and urban areas]$_6$.

Key Source Sentence:
The consequences of [another year]$_4$ of [poor rains]$_1$ on [food security]$_3$ are severe.

Auxiliary Source Sentence(s):
However in addition to the needs of economic recovery activities for IDPs, [food insecurity]$_3$ [over the majority of 2004]$_2$ [has created great stress]$_5$ on the poorest families in the country, [both within the urban and rural settings]$_6$.

Figure 1: Alignment of a summary sentence to sentences in the full document. Phrases of similar meaning are co-indexed.

To write such summaries, we assume that the human abstractor begins by choosing key sentences from the full document. Then, for each key sentence, a set of auxiliary material is identified. The key sentence is revised incorporating these auxiliary sentences to produce the eventual summary sentence.

To study this phenomenon, a corpus of UN documents was collected and analysed.[1] Each document was divided into two parts comprising its executive summary, and the remainder, referred to here as the *source*. We manually aligned each executive summary sentence with one or more sentences from the source, by choosing a key sentence that provided

[1]This corpus is described in detail in Section 5.1.

Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing, pages 543–552,
Honolulu, October 2008.

evidence for the content of the summary sentence along with additional sentences that provided supporting material.

We refer to the resulting corpus as the UN Consolidated Appeals Process (UN CAP) corpus. It is a collection of sentence alignments, each referred to as an *aligned sentence tuple*, which consists of:

1. A human authored summary sentence from the executive summary;
2. A *key* sentence from the *source*;
3. Zero or more *auxiliary* sentences from the *source*.

The key and any auxiliary sentences are referred to collectively as the *aligned source sentences*.

We argue that some process that combines information from multiple sentences is required if we are to generate summary sentences similar to that portrayed in Figure 1. This is supported by our analysis of the UN CAP corpus. Of the 580 aligned sentence tuples, the majority, 61% of cases, appear to be examples of such a process.

Furthermore, the auxiliary sentences are clearly necessary. We found that only 30% of the open-class words in the summary are found in the key sentence. If one selects all the open-class words from aligned source sentences, recall increases to an upper limit of 45% without yet accounting for stemming. This upper bound is consistent with the upper limit of 50% found by Daumé III and Marcu (2005) which takes into account stemming differences.

This demonstrates that the auxiliary material is a valuable source of content which should be integrated into the summary sentence, allowing an improvement in recall of up to 15% prior to accounting for morphological, synonym and paraphrase differences. Of course, the trick is to improve recall without hurting precision. A naive addition of all words in the aligned source sentences incurs a drop in precision from 30% to 23%. The problem thus is one of selecting the relevant auxiliary content words without introducing unimportant content. We refer to this problem of incorporating material from auxiliary sentences to supplement a key sentence as *Sentence Augmentation*.

In this paper, sentence augmentation is modelled as a noisy channel process and has two facets: content selection and language modelling. This paper focuses on the former, in which the system must rank text segments—in this case, words—for inclusion in the generated sentence. Given a ranked selection of words, a language model would then order them appropriately, as described in work on sentence regeneration (for example, see Soricut and Marcu (2005); Wan et al. (2005)).

Provided with an aligned sentence tuple, the problem lies in effectively selecting words from the auxiliary sentences to bolster those taken from the key sentence. Given that there are on average 2.7 auxiliary sentences per aligned sentence tuple, this additional influx of words poses a considerable challenge.

We begin with the premise that, for documents of a homogeneous type (in this case, the genre is a funding proposal, and the domain is humanitarian aid), it may be possible to identify patterns in the organisation of information in summaries. For example, Figure 2 presents three summary sentences from our corpus that share the same patterned juxtaposition of two concepts *DisplacedPersons* and *HostingCommunities*. Documents may exhibit common patterns since they have a similar goal: namely, to convince donors to give financial support. In the above example, the juxtaposition highlights the fact that those in need are not just those people from the 'epicenter' of the crisis but also those that look after them.

We propose and evaluate a method called "Seed and Grow" for selecting content from auxiliary sentences. That is, we first select the core meaning of the summary, given here by the key sentence, and then we find those pieces of *additional* information that are conventionally juxtaposed with it.

Such patterns are reminiscent of *Schemata*, the organisations of propositional content introduced by McKeown (1985). Schemata typically involve a symbolic representation of each proposition's semantics. However, in our case, a text-to-text generation scenario, we are without such representations and so must find other means to encode these patterns.

To alleviate the situation, we turn to word-pair co-occurrences to approximate schematic patterns. Fig-

Sentence 1:
The increased number of [internally displaced persons]$_1$ and the continued presence of refugees have further strained the scarce natural resources of [host communities]$_2$, stretching their capacity to the limit.

Sentence 2:
100,000 people, a significant portion of the population, remain [displaced]$_1$, burdening the already precarious living conditions of [host families]$_2$ in Dili and the Districts.

Sentence 3:
The current humanitarian situation in Timor-Leste is characterised by: An estimated [100,000 displaced people]$_1$ (10% of the population) living in camps and with [host families]$_2$ in the districts; A total or partial destruction of over 3,000 homes in Dili affecting at least 14,000 IDPs

Figure 2: Examples of the pattern ⟨*DisplacedPersons*[1], *HostingCommunities*[2]⟩.

ure 2 showed that mentions of the plight of internationally displaced persons are often followed by descriptions of the impact on the host communities that look after them. In this particular example, this is realised lexically in the co-occurrences of the words *displaced* and *host*.

Corpus-based methods inspired by the notion of schemata have been explored in the past by Lapata (2003) and Barzilay and Lee (2004) for ordering sentences extracted in a multi-document summarisation application. However, to our knowledge, using word co-occurrences in this manner to represent schematic knowledge for the purposes of selecting content in a statistically-generated summary sentence has not previously been explored.

This paper seeks to determine whether or not such patterns exist in homogeneous data; and furthermore, whether such patterns can be used to better select words from auxiliary sentences. In particular, we propose the "Seed and Grow" approach for this task. The results show that even simple modelling approaches are able to model this schematic information.

In the remainder of this paper, we contrast our approach to related text-to-text research in Section 2. The Content Selection model is presented in Section 3. Section 4 describes how a binary classification model is used in a statistical text generation system. Section 5 describes our evaluation of the model for a summary generation task. We conclude, in Section 6, that domain-specific schematic patterns can be acquired and applied to content selection for statistical sentence generation.

2 Related Work

2.1 Content Selection in Text-to-Text Systems

Statistical text-to-text summarisation applications have borrowed much from the related field of statistical machine translation. In one of the first works to present summarisation as a noisy channel approach, Witbrock and Mittal (1999) presented a conditional model for learning the suitability of words from a news article for inclusion in headlines, or 'ultra-summaries'. Inspired by this approach, and with the intention of designing a robust statistical generation system, our work is also based on the noisy channel model. Into this, we incorporate our content selection model, which includes Witbrock and Mittal's model supplemented with schema-based information.

Roughly, text-to-text transformations fall into three categories: those in which information is *compressed*, *conserved*, and *augmented*. We use these distinctions to organise this overview of the literature.

In *Sentence Compression* work, a single sentence undergoes pruning to shorten its length. Previous approaches have focused on statistical syntactic transformations (Knight and Marcu, 2002). For content selection, discourse-level considerations were proposed by Daumé III and Marcu (2002), who explored the use of Rhetorical Structure Theory (Mann and Thompson, 1988). More recently, Clarke and Lapata (2007) use Centering Theory (Grosz et al., 1995) and Lexical Chains (Morris and Hirst, 1991) to identify which information to prune. Our work is similar in incorporating discourse-level phenomena for content selection. However, we look at schema-like information as opposed to chains of references and focus on the sentence augmentation task.

The work of Barzilay and McKeown (2005) on *Sentence Fusion* introduced the problem of converting multiple sentences into a single summary sen-

tence. Each sentence set ideally tightly clusters around a single news event. Thus, there is one general proposition to be realised in the summary sentence, identified by finding the common elements in the input sentences. We see this as an example of *conservation*. In our work, this general proposition is equivalent to the core information for the summary sentence *before* the incorporation of supplementary material.

In contrast to both *compression* and *conservation* work, we focus on *augmenting* the information in a key sentence. The closest work is that of Jing and McKeown (1999) and Daumé III and Marcu (2005), in which multiple sentences are processed, with fragments within them being recycled to generate the novel generated text.

In both works, recyclable fragments are identified by automatic means. Jing and McKeown (1999) use models that are based on "copy-and-paste" operations learnt from the behaviour of human abstractors as found in a corpus. Daumé III and Marcu (2005) propose a model that encodes how likely it is that different sized spans of text are skipped to reach words and phrases to recycle.

While similar in task, our models differ substantially in the nature of the phenomenon modelled. In this work, we focus on content-based considerations that model which words can be combined to build up a new sentence.

2.2 Schemata and Text Generation

There exists related work from Natural Language Generation (NLG) in finding material to build up sentences. As mentioned above, our content selection model is inspired by work on schemata from NLG (McKeown, 1985). Barzilay and Lee (2004) showed that it is possible to obtain schema-like knowledge automatically from a corpus for the purposes of *extracting* sentences and ordering them. However, their work represents patterns at the sentence level, and is thus not directly comparable to our work, given our focus on sentence *generation*.

In our system, what is required is a means to rank words for use in generation. Thus, we focus on commonly occurring word co-occurrences, with the aim of encoding conventions in the texts we are trying to generate. In this respect, this is similar to work by Lapata (2003), who builds a conditional model of words across adjacent sentences, focusing on words in particular semantic roles. Like Barzilay and Lee (2004), this model was used to order extracted sentences in summaries. In contrast, our work focuses on word patterns found within a summary sentence, not between sentences. Additionally, our tasks differ as we examine the statistical sentence generation instead of sentence ordering.

3 Linguistic Intuitions behind Word Selection

The "Seed and Grow" approach proposed in this paper divides the word-level content selection problem into two underlying subproblems. We address these with two separate models, called the *salience* and *schematic* models. The salience model chooses the key content for the summary sentence while the schematic model attempts to identify what else is typically mentioned given those salient pieces of information.

3.1 A Salience Model: Learning "Buzzwords"

There are a variety of methods for determining the salient information in a text, and these underpin most work in automatic text summarisation. As an example of a salience model trained on corpus data, Witbrock and Mittal (1999) introduced a method for scoring summary words for inclusion within news headlines. In their model, headlines were treated as 'ultra-summaries'. Their model learns which words are typically used in headlines and encodes, at least to some degree, which words are attention grabbing.

In the domain of funding proposals, key words that grab attention may amount to domain-specific buzzwords. Intuitively, a reader, perhaps someone in charge of allocating donations, tends to look for certain types of key information matching donation criteria, and so human abstract authors will target their summaries for this purpose.

We thus adapt the Witbrock and Mittal (1999) model to identify such domain specific buzzwords (BWM, for 'buzzword model'). For an aligned sentence tuple, the probability that a word is selected based on the salience of a word with respect to the domain is defined as:

$$\text{prob}_{bwm}(\text{select} = 1|w) = \frac{|\text{summary}_w|}{|\text{source}_w|} \quad (1)$$

where summary_w is the set of aligned sentence tuples that contain the word *w* in the summary sentence *and* in the source sentences. The denominator, source_w, is the set of aligned sentence tuples that have the word *w* in either the key or an auxiliary sentence.

As is implicit in this equation, we could just use this buzzword model to select content not only from the key sentence, but from the auxiliary sentences as well. While it is intended ultimately to find the key content of the summary, it can also serve as an alternative baseline for auxiliary content selection to compare against the "Seed and Grow" model.

3.2 A Schema Model: Approximation via Word co-Occurrences

To restate the problem at hand: the task is one of finding elements of secondary importance that schematically elaborate on the key information. We do this by examining sample summary sentences for conventional juxtapositions of concepts. As mentioned in Section 1, schemata are approximated here with patterns of word-pair co-occurrences. Using a corpus of human-authored summaries in the domain of our application, it is thus possible to learn what those common combinations of words are.

Roughly, the process is as follows. To begin with, a *seed set* of words is chosen. The purpose of the seed set is to represent the core proposition of the summary sentence.

In this work, this core proposition is given by the key sentence and so the non-stopwords belonging to it are used to populate the seed set. In the "Seed and Grow" approach, we check to see which words from auxiliary sentences pair well with words in the seed set.

3.2.1 Collecting Word-level Patterns

Each training case in the corpus contains a single human-authored summary sentence that can be used to learn which pairs of words conventionally occur in a summary. For each summary sentence, stopwords are removed. Then, each pairing of words in the sentence is used to update a pair-wise word co-occurrence frequency table. When looking up and storing a frequency, the order of words is ignored.

3.2.2 Scoring Word-Pair Co-occurrence Strength

For any two words, w_1 from the seed set and w_2 from an auxiliary sentence, the word-pair co-occurrence probability is defined as follows:

$$\text{prob}_{co\text{-}oc}(w_1, w_2) = \frac{\text{freq}(w_1, w_2)}{\text{freq}(w_1) + \text{freq}(w_2) - \text{freq}(w_1, w_2)} \quad (2)$$

where $freq(w_1, w_2)$ is a lookup in the word-pair co-occurrence frequency table. This table stores co-occurrence word pairs occurring in the summary sentence.

3.2.3 Combining a Set of Co-occurrence Scores

Each auxiliary word now has a series of scores, one for each comparison with a seed word. To rank each auxiliary word, these need to be combined into a single score for sorting.

When combining the set of co-occurrence scores, one might want to account for the fact that each pairing of a seed word with an auxiliary word might not contribute equally to the overall selection of that auxiliary word. Intuitively, a word in the seed set, derived from the key sentence, may only make a minor contribution to the core meaning of the summary sentence. For example, words that are part of an adjunct phrase in the key sentence might not be good candidates to elaborate upon. Thus, one might want to weight these seed words lower, to reduce their influence on triggering schematically associated words.

To allow for this, a seed weight vector is maintained, storing a weight per seed word. Different weighting schemes are possible. For example, a scheme might indicate the salience of a word. In addition to the buzzword model (BWM) described earlier, one might employ a standard vector space approach (Salton and McGill, 1983) from Information Retrieval, which uses term frequency scores weighted with an inverse document frequency factor, or *tf-idf*. We also implement the case in which all seed words are treated equally using binary weights, where 1 indicates the presence of a seed word, and 0 indicates its absence. In the evaluations described in Section 5, we refer to these three seed weighting schemes as *bwm* and *tf-idf*, and *binary* respectively.

To find the probability of selecting an auxiliary word using the schematic word-pair co-occurrence model (WCM), an averaged probability is found by normalising the sum of the weighted probabilities, where weights are provided by one of the three schemes above:

$$\text{prob}_{wcm}(w_i) = \frac{1}{Z} \times \sum_{k=0}^{|\text{seed}|} \text{weights}_k \times \text{prob}_{co\text{-}oc}(w_i, w_k) \quad (3)$$

where *seed* is the set of seed words and w_k is the k^{th} word in that set. The vector, *weights*, stores the seed weights. The normalisation factor for the weighted average, Z, is the number of auxiliary words.

Finally, since the WCM model only serves to select words from the *auxiliary* sentences, words from the key sentence must be given scores as well. For these words, the scoring is as follows:

$$\text{prob}_{wcm}(w) = \frac{1}{Z}\left(\frac{1}{|\text{seed}|} + \text{prob}_{wcm}(w)\right) \quad (4)$$

where Z is a normalisation across the set of seed words.

4 Combining Buzzwords and Word-Pair Co-Occurrence Models for Generation

As mentioned above, the noisy channel approach is used for producing the augmented sentence. Although the focus of this paper is on Content Selection, an overview of the end-to-end generation process is presented for completeness.

Sentence augmentation is essentially a text-to-text process: A key sentence and auxiliary material are transformed into a single summary sentence. Following Witbrock and Mittal (1999), the task is to search for the string of words that maximises the probability $prob(summary|source)$. Standardly reformulating this probability using Bayes' rule results in the following:

$$prob_{cm}(source|summary) \times prob_{lm}(summary) \quad (5)$$

In this paper, we are concerned with the first factor, $prob_{cm}(source|summary)$, referred to as the channel model (CM), which combines both the buzzword (BWM) and word-pair co-occurrence (WCM) models. An examination of differences between the two approaches revealed only a 20% word overlap on the Jaccard metric.

In order to combine multiple models, we intend to use machine learning approaches to combine the information in each model in a similar manner to Berger et al. (1996). We are currently exploring the use of logistic regression methods to learn a function that would treat, as features, the probabilities defined by the salience and schematic content selection models. Although generation is possible using each content selection model in isolation, evaluations of the combined model are on-going and are not presented in this paper.

5 Evaluation

In this evaluation, the task is to select n words from the aligned source sentences for inclusion in a summary. As a gold-standard for comparison, we simply examine what words were actually chosen in the summary sentence of the aligned sentence tuple. We are specifically interested in open-class words, and so a stopword list of closed-class words is used to filter the sentences in each test case.

We evaluate against the set of open-class words in the human-authored summary sentence using recall and precision metrics. Recall is the size of the intersection of the selected and gold-standard sets, normalised by the length of the gold-standard sentence (in words). This recall metric is similar to the ROUGE-1 metric, the unigram version of the ROUGE metric (Lin and Hovy, 2003) used in the Document Understanding Conferences[2] (DUC). Precision is the size of the intersection normalised by the number of words selected. We also report the F-measure, which is the harmonic mean of the recall and precision scores.

Recall, precision and F-measure are measured at various values of n ranging from 1 to the number of open-class words in the gold-standard summary sentence for a particular test case. For the purposes of evaluation, differences in tokens due to morphology were explored crudely via the use of Porter's stemming algorithm. However, the results from stemming are not that different from exact token matches when examining performance on the entire data set

[2] http://duc.nist.gov

Number of training cases	530
Average words in summary sentence	27.0
Average stopwords in summary sentence	10.3
Average number of auxiliary sentences	2.75
Word count: summary sentences	4630
Word count: source sentences	21356
Word type count in corpus	3800

Table 1: Statistics for the UN CAP training set

and so, for simplicity, these are omitted in this discussion.

5.1 The Data

The corpus is made up of a number of humanitarian aid proposals called Consolidated Appeals Process (UN CAP) documents, which are archived at the United Nations website.[3] 135 documents from the period 2002 to 2007 were downloaded by the authors. A preprocessing stage extracted text from the PDF files and segmented the documents into executive summary and source sections. These were then automatically segmented further into sentences.

Executive summary sentences were manually aligned by the authors to source key and auxiliary sentences, producing a corpus of 580 aligned sentence tuples referred to here as the UN CAP corpus. Of these, 230 tuples were paraphrase cases (i.e. without aligned auxiliary sentences). The remaining 550 cases were instances of sentence augmentation (with at least one auxiliary sentence).

Of the 580 cases, 50 cases were set aside for testing. The remaining 530 cases were used for training. Statistics for the training portion of the sentence augmentation set are provided in Table 1.

In this paper, aligned sentence tuples are obtained via manual annotation. Automatic construction of these sentence-level alignments is possible and has been explored by Jing and McKeown (1999). We also envisage using tools for scoring sentence similarity (for example, see Hatzivassiloglou et al. (2001)) for automatically constructing them; this is the focus of work by Wan and Paris (2008).

[3] http://ochaonline3.un.org/humanitarianappeal/index.htm

5.2 The Baselines

Three baselines were used in this work: the *random*, *tf-idf* and *position* baselines. A *random* word selector shows what performance might be achieved in the absence of any linguistic knowledge.

We also sorted all words in the aligned source sentences by their weighted *tf-idf* scores. This baseline selects words in order until the desired word limit is reached. This baseline is referred to as the *tf-idf* baseline.

Finally, we selected words based on their sentence order, choosing first those words from the key sentence. When these are exhausted, auxiliary sentences are sorted by their sentence positions in the original document. Words from the first auxiliary sentence are then chosen. This continues until either the desired number of words have been chosen, or no words remain. This baseline is known as the *position* baseline.

5.3 Content Selection Results

We compare the three baselines to the two models presented in Section 3. These are the buzzword salience model (BWM) and the schematic word-pair co-occurrence model (WCM).

We begin by presenting recall, precision and F-measure graphs when selecting from the aligned source sentences, comprising the key and auxiliary sentences. Figure 3 shows the results for the two models against the three baselines. The two models, the positional, and the *tf-idf* baselines perform better than the random baseline, as measured by a two-tailed Wilcoxon Matched Pairs Signed Ranks test ($\alpha = 0.05$).

The WCM consistently out-performs the BWM on all metrics, and the differences are statistically significant. In fact, the BWM also generally performs worse than the position and *tf-idf* baselines. WCM and the position baseline both significantly outperform the *tf-idf* baseline on all metrics for longer sentence lengths.

That the position baseline and WCM should perform similarly is not really surprising since, in effect, the position baseline first chooses words from the key sentence and then selects auxiliary words. The difference essentially lies in how the auxiliary words are chosen.

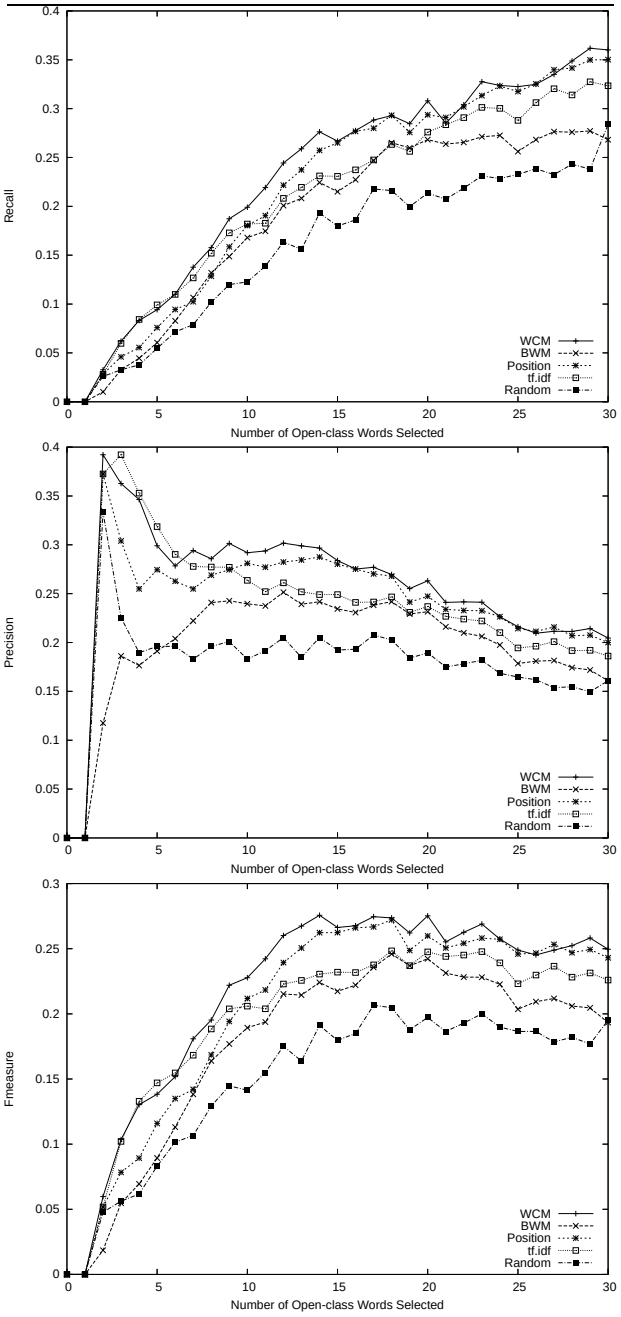

Figure 3: Recall, Precision and F-measure performance for open-class words from the entire input set (key and auxiliary). Models presented are the Buzzword Model (BWM), the Word-Pair Co-occurrence Model (WCM) and position, *tf-idf* and random baselines.

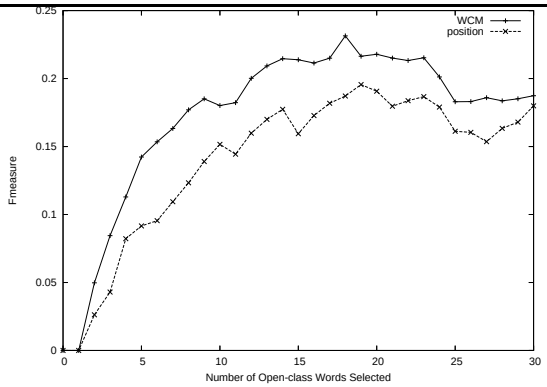

Figure 4: F-measure scores for content selection on just the auxiliary sentences. Models presented are the Word-Pair Co-occurrence model (WCM) and the position baseline.

The results of Figure 3 weakly support the hypothesis that using schematic word-pair co-occurrences helps improve performance over models without discourse-related features. The graphs show that WCM edges above the position baseline when the number of selected open-class words ranges from 10 to 15. Note that the average number of open-class words in a human authored summary sentence is 16. The only significant difference found was in the F-measure and precision scores for 19 selected open-class words. Nevertheless, a general trend can be observed in which WCM performs better than the position baseline.

Ultimately, however, what we want to do is select auxiliary content to supplement the key sentence. To examine the effect of two best performing approaches, WCM and the position baseline, on this task, were both modified so that the key sentence words were explicitly given a zero probability. Thus, the recall, precision and F-measure scores obtained are based solely on the ability of either to select auxiliary words. The F-measure scores are presented Figure 4. WCM consistently outperforms the position baseline for the selection of auxiliary words. Differences are significant for 6 or more selected open-class words.

The results show that even when considering only exact token matches, we can improve on the recall of open-class words, and do so without penalty in precision. Our working hypothesis is that such gains are possible because the corpus has a homo-

geneous quality and key patterns are sufficiently repeated even when the overall data set is of the order of hundreds of cases. The benefit of using a model encoding some schematic information is further shown by the performance of WCM over the position baseline when selecting words from auxiliary sentences.

This is an interesting finding given that domain independent methods are increasingly used on domain-specific corpora such as financial and biomedical texts, for which we may have access to only a limited amount of data. We anticipate that as we introduce methods to account for paraphrase and synonym differences, performance might rise further still.

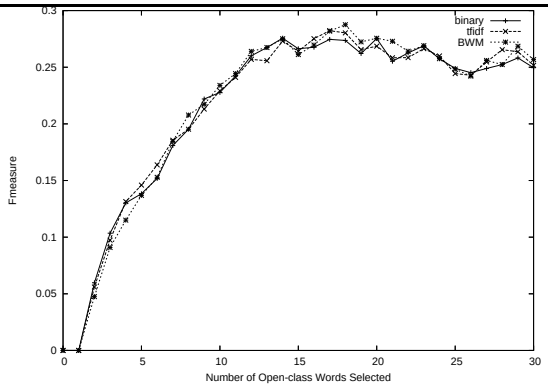

Figure 5: F-measure performance for open-class words from the entire input set (key and auxiliary). Models presented are variants of the Word-Pair Co-occurrence Model (WCM) that differ in the seed weighting schemes.

5.4 Testing Seed Weighting Schemes

We can also weight seed words in the "Seed and Grow" approach in a variety of ways. To test whether weighting schemes have any effect on content selection performance, we examined the use of three schemes. We were particularly interested in those schemes that indicate the contribution of a seed word to the core meaning of a sentence. These are the *binary*, *tf-idf* and *buzzword* weighting schemes described in Section 3. We present the F-measure graph for these three variants of the schematic word-pair co-occurrence model (WCM) in Figure 5.

The graphs show that there is no discernible difference between the seed weighting schemes. No scheme significantly outperforms another. Thus, we conclude that the choice of these particular seed weighting schemes has no effect on performance. In future work, we intend to examine whether weighting schemes encoding syntactic information might fare better, since such information might more accurately represent the contribution of a substring to the main clause of the sentence.

6 Conclusions and Future Work

In this paper, we argued a case for *sentence augmentation*, a component that facilitates abstract-like text summarisation. We showed that such a process can account for summary sentences as authored by professional editors. We proposed the use of schemata, as approximated with a word-pair co-occurrence model, and advocated a new schema-based "Seed and Grow" content selection model used for statistical sentence generation.

We also showed that domain-specific patterns, schematic word-pair co-occurrences in this case, can be acquired from a limited amount of data as indicated by modest performance gains for content selection using schemata information. We postulate that this is particularly true when dealing with homogeneous data.

In future work, we intend to explore other string matches corresponding to variations due to paraphrases and synonymy. We would also like to study the effects of corpus size when learning schematic patterns. Finally, we are currently investigating the use of machine learning methods to combine the best of the Salience and Schemata models in order to provide a single model for use in decoding.

7 Acknowledgments

We would like to thank the reviewers for their insightful comments. This work was funded by the CSIRO ICT Centre and Centre for Language Technology at Macquarie University.

References

Regina Barzilay and Lillian Lee. 2004. Catching the drift: Probabilistic content models, with applications to generation and summarization. In Daniel Marcu Susan Dumais and Salim Roukos, editors, *HLT-NAACL 2004: Main Proceedings*, pages 113–120, Boston,

Massachusetts, USA, May 2 - May 7. Association for Computational Linguistics.

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

Adam L. Berger, Stephen Della Pietra, and Vincent J. Della Pietra. 1996. A maximum entropy approach to natural language processing. *Computational Linguistics*, 22(1):39–71.

James Clarke and Mirella Lapata. 2007. Modelling compression with discourse constraints. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 1–11.

Hal Daumé III and Daniel Marcu. 2002. A noisy-channel model for document compression. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL – 2002)*, pages 449 – 456, Philadelphia, PA, July 6 – 12.

Hal Daumé III and Daniel Marcu. 2005. Induction of word and phrase alignments for automatic document summarization. *Computational Linguistics*, 31(4):505–530, December.

Barbara J. Grosz, Aravind K. Joshi, and Scott Weinstein. 1995. Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics*, 21(2):203–225.

V. Hatzivassiloglou, J. Klavans, M. Holcombe, R. Barzilay, M. Kan, and K. McKeown. 2001. Simfinder: A flexible clustering tool for summarization. pages 41–49. Association for Computational Linguistics.

Hongyan Jing and Kathleen McKeown. 1999. The decomposition of human-written summary sentences. In *Research and Development in Information Retrieval*, pages 129–136.

Kevin Knight and Daniel Marcu. 2002. Summarization beyond sentence extraction: a probabilistic approach to sentence compression. *Artificial Intelligence*, 139(1):91–107.

Mirella Lapata. 2003. Probabilistic text structuring: Experiments with sentence ordering. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 545–552, Sapporo, Japan.

Chin-Yew Lin and Eduard Hovy. 2003. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *NAACL '03: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, pages 71–78, Morristown, NJ, USA. Association for Computational Linguistics.

W. C. Mann and S. A. Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text*, 8(3):243–281.

Kathleen R McKeown. 1985. *Text Generation: Using Discourse Strategies and Focus Constraints to Generate Natural Language Text*. Cambridge University Press.

Jane Morris and Graeme Hirst. 1991. Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Computational Linguistics*, 17(1):21–48.

G. Salton and M. J. McGill. 1983. *Introduction to modern information retrieval*. McGraw-Hill, New York.

Radu Soricut and Daniel Marcu. 2005. Towards developing generation algorithms for text-to-text applications. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 66–74, Ann Arbor, Michigan, June. Association for Computational Linguistics.

Stephen Wan and Cécile Paris. 2008. In-browser summarisation: Generating elaborative summaries biased towards the reading context. In *Proceedings of ACL-08: HLT, Short Papers*, pages 129–132, Columbus, Ohio, June. Association for Computational Linguistics.

Stephen Wan, Robert Dale Mark Dras, and Cécile Paris. 2005. Towards statistical paraphrase generation: preliminary evaluations of grammaticality. In *Proceedings of The 3rd International Workshop on Paraphrasing (IWP2005)*, pages 88–95, Jeju Island, South Korea.

Michael J. Witbrock and Vibhu O. Mittal. 1999. Ultra-summarization (poster abstract): a statistical approach to generating highly condensed non-extractive summaries. In *SIGIR '99: Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 315–316, New York, NY, USA. ACM Press.