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

Stephen Wan[1,2] **Robert Dale**[1] **Mark Dras**[1]

[1]Centre for Language Technology
Macquarie University
swan,rdale,madras@ics.mq.edu.au

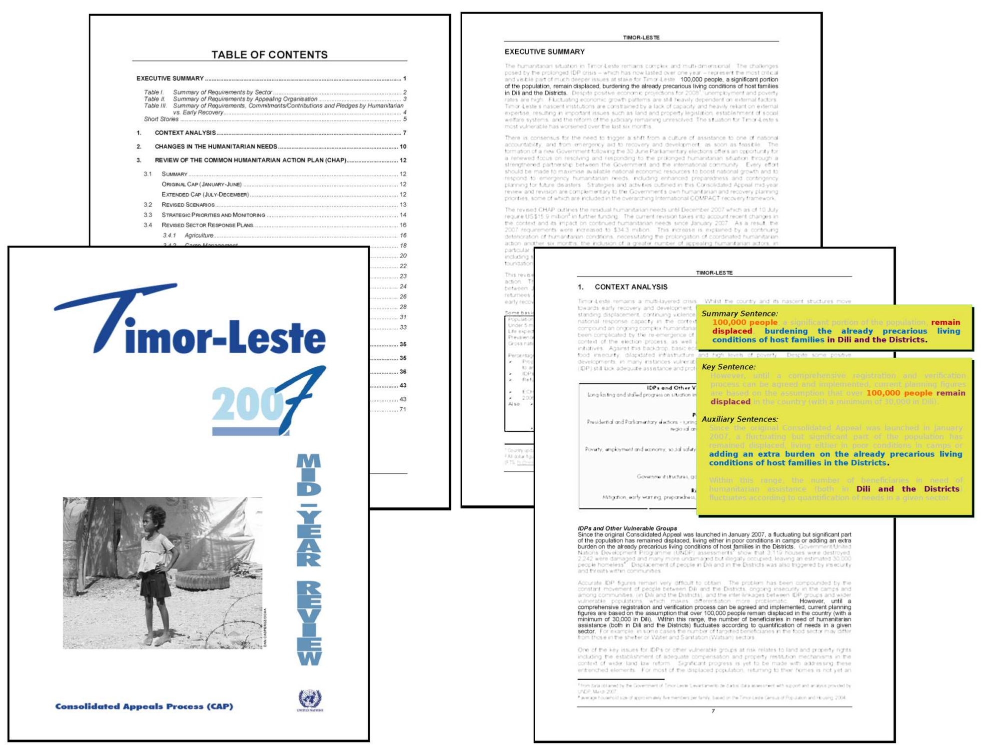
Cécile Paris[2]

[2]Information and Communication Technologies
CSIRO
Cecile.Paris@csiro.au

1 Sentence Augmentation

Sentence Augmentation is the process of supplementing a sentence with additional information to produce a novel (summary) sentence.

1.1 Application Scenario



Summarisation Procedure

1. Choose key sentences from the input document
   (a) For each key sentence, choose auxiliary sentences.
   (b) Revise key sentence incorporating auxiliary information

1.2 The UN CAP Corpus

- The UN CAP corpus is based on a set of funding proposals for meeting humanitarian crises.
- Sentences in the *executive summary* are aligned with one or more sentences from the rest of the document, or the *source*.
- The result, an *Aligned Sentence Tuple*, contains:
  1. A summary sentence from the executive summary;
  2. A *key* sentence from the *source*;
  3. Zero or more *auxiliary* sentences from the *source*.
- The corpus is a collection of these aligned sentence tuples.

1.3 The Problem: Auxiliary Content Selection

Given the key and auxiliary sentences, determine which words from the auxiliary sentence bests supplements the key sentence content.

Auxiliary Information is Important

- Of the 580 aligned sentence tuples in our corpus, the majority, 61% of cases, align to multiple sentences.
- Only 30% of the open-class words in the summary sentence are found in the key sentence.
- Selecting all open-class words from both key and auxiliary sentences increases recall to 45% (without stemming).
- **The challenge: Improve recall without hurting precision**

2 Our Approach

An Observation: Data is Homogeneous

- Genre: a funding proposal
- Domain: humanitarian aid; world events
- Style: conforms to an editorial style guide

"Seed and Grow" Approach

- Homogeneous documents may exhibit common patterns since they have a similar goal: in this case, to convince donors to give financial support.
- If so, look for schematic patterns [9] that reveal the organisation of information in summaries.
- We approximate schemata as word juxtapositions patterns.
- For related work on content selection using discourse features, see [4] and [3]; For related work in corpus-based approaches to learning schemata, see [8] and [1].

3 Word-Pair Co-Occurrences as Schematic Word Patterns

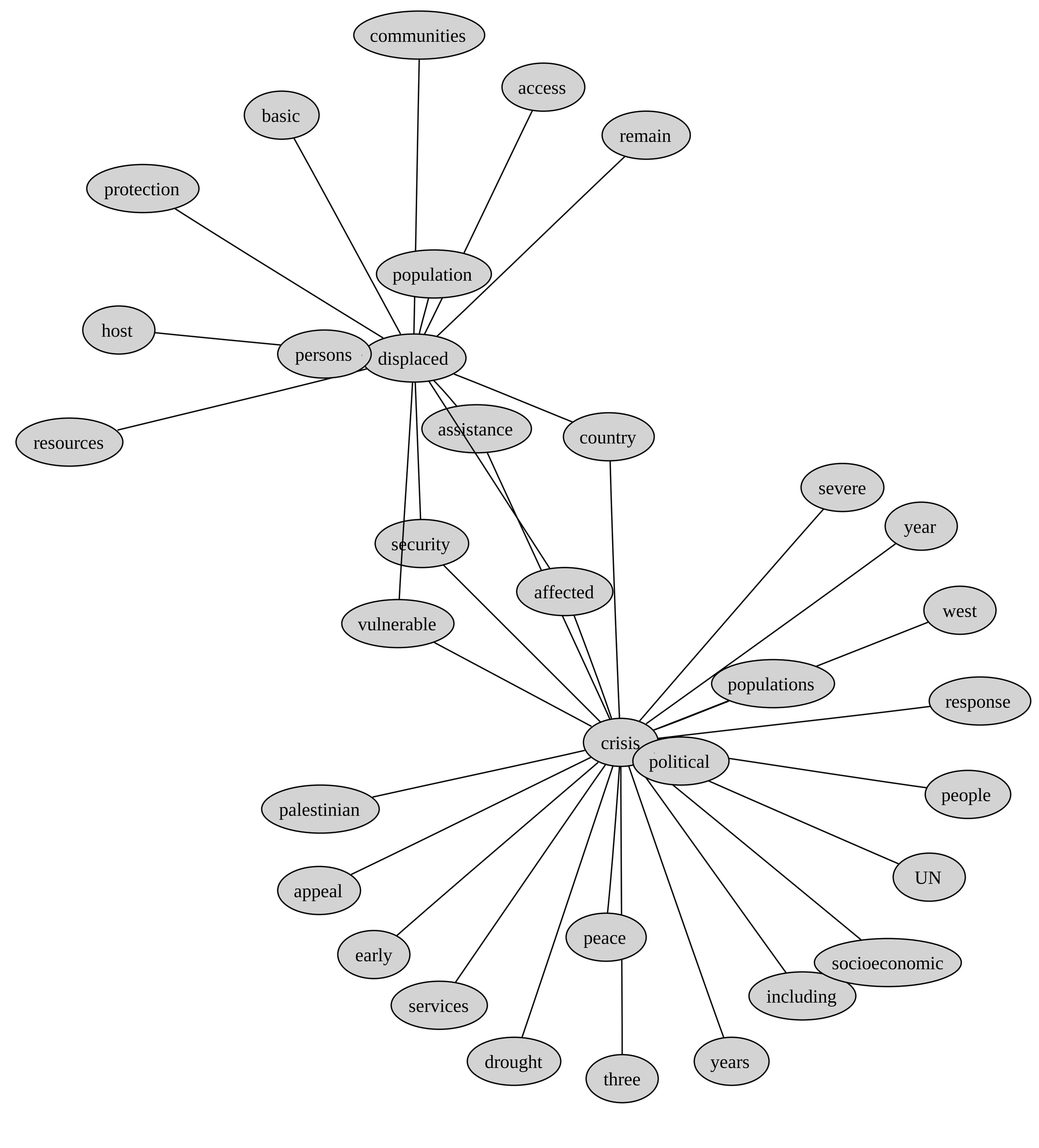
Example Pattern in Summary Sentences

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

- Training:
  - Count frequency of each *word pair* in a *summary sentence*
- Runtime:
  - Given the key sentence, for each auxiliary word
    - Rank candidate auxiliary words based on probability of the juxtaposition: ⟨ key word, auxiliary word ⟩
- Model:



4 Evaluation: Selecting Words

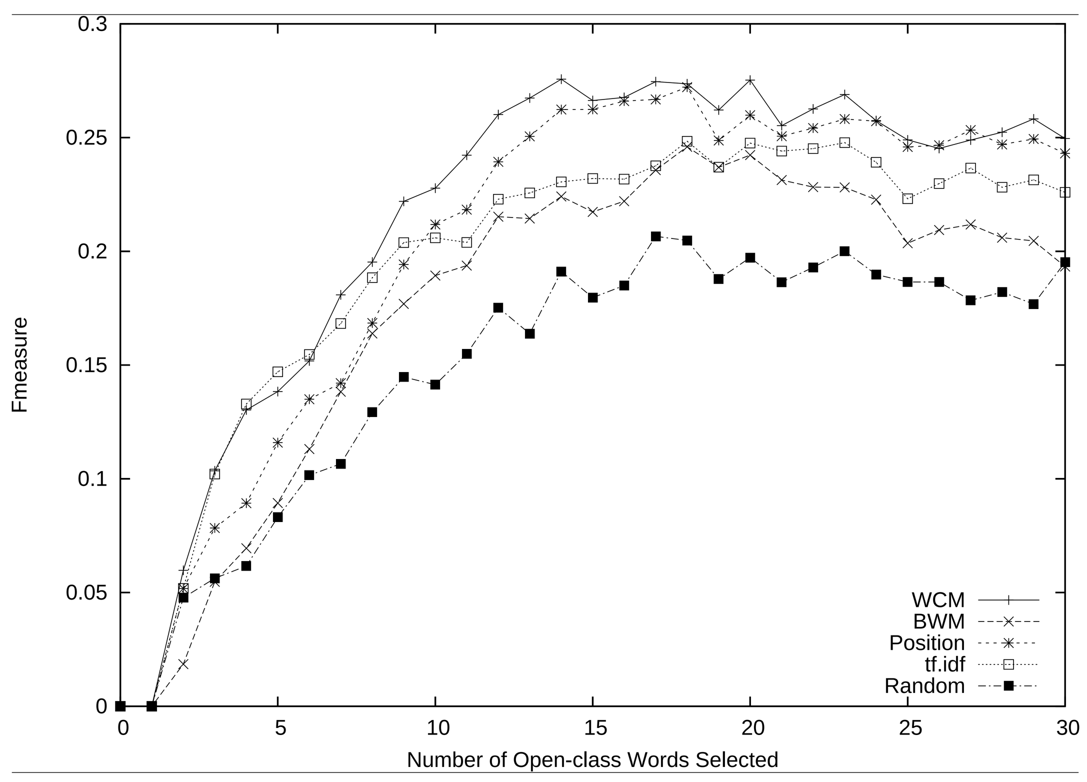
Test Evaluation

- Data: 50 unseen aligned sentence tuple test cases
- Task: Predict word selection in the summary sentence given the key and auxiliary sentences (c.f. [2], [6], [5])
- Evaluate: Measured via Recall, Precision and F Measure (Significance tested using two-tailed Wilcoxon)

Systems and Baselines

WCM: Word Co-occurrence Model: Schematic Word Patterns)
BWM: Buzzword Model based on [10])
position: Baseline based on the sentence position
tf.idf: Baseline based on tf.idf scores for words
random: Random word selection

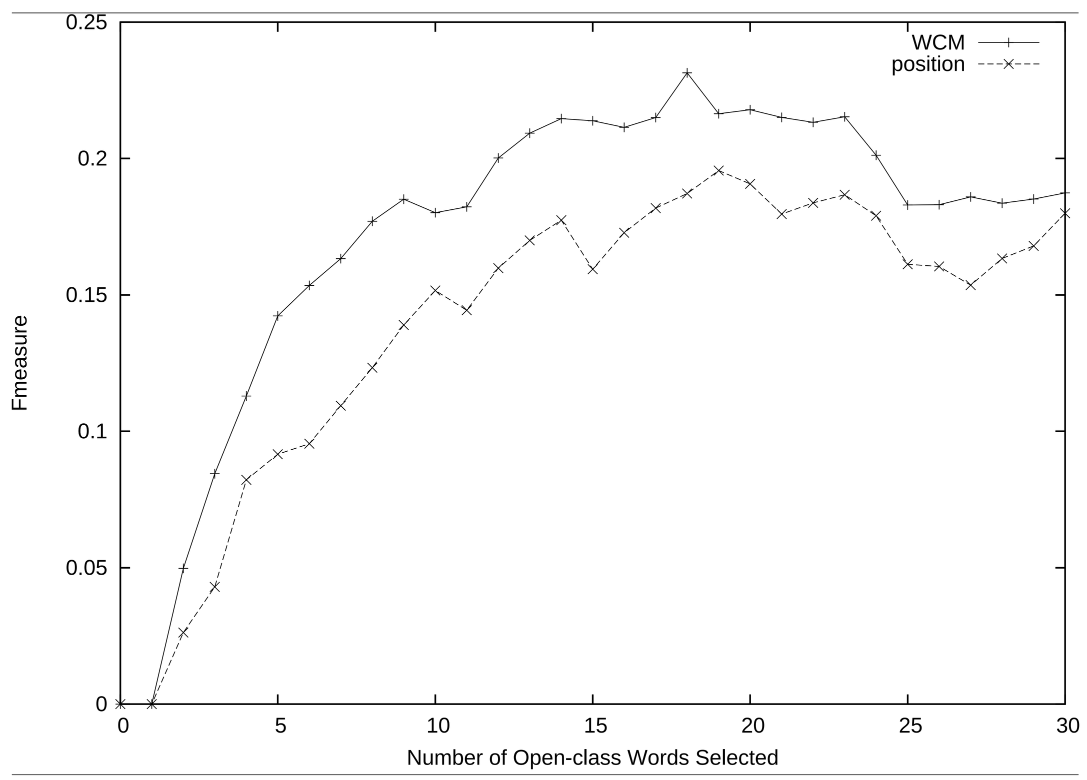
Do Schematic Word Patterns help Word Selection Overall?



Results:

- Weak trend suggests schematic word patterns help (see WCM curve).
- Conclude: On overall task, no loss of performance.

Do Schematic Word Patterns Select Better Auxiliary Words?



Results:

- Improvement of the WCM over the position baseline from 6-10 ($p < 0.01$) and 11-20 ($p < 0.05$) selected words.
- Conclude: schematic word patterns help in selecting auxiliary words.

Conclusions

1. We argued a case for *sentence augmentation*, a component that facilitates abstract-like text summarisation.
2. We proposed the use of schemata for selecting auxiliary content, as approximated with a word-pair co-occurrence model in an approach called "Seed and Grow".
3. Domain-specific patterns, specifically schematic word-pair co-occurrences in this case, can be acquired from homogenous data, as demonstrated by the observed improvement in F Measure for selecting words.

References

[1] Regina Barzilay and Lillian Lee. 2004. Catching the drift: Probabilistic content models, with applications to generation and summarization. In *Proceedings of HLT-NAACL 2004*, pages 113–120.

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

[3] 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.

[4] 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*, pages 449 – 456.

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

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

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

[8] 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.

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

[10] Michael J. Witbrock and Vibhu O. Mittal. 1999. Ultra-summarization (poster abstract): a statistical approach to generating highly condensed non-extractive summaries. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 315–316.