Improving Argument Overlap for Proposition-Based Summarisation

Yimai Fang and **Simone Teufel**
University of Cambridge Computer Laboratory
15 JJ Thomson Avenue
Cambridge CB3 0FD, United Kingdom
{yf261,sht25}@cam.ac.uk

Abstract

We present improvements to our incremental proposition-based summariser, which is inspired by Kintsch and van Dijk's (1978) text comprehension model. Argument overlap is a central concept in this summariser. Our new model replaces the old overlap method based on distributional similarity with one based on lexical chains. We evaluate on a new corpus of 124 summaries of educational texts, and show that our new system outperforms the old method and several state-of-the-art non-proposition-based summarisers. The experiment also verifies that the incremental nature of memory cycles is beneficial in itself, by comparing it to a non-incremental algorithm using the same underlying information.

1 Introduction

Automatic summarisation is one of the big artificial intelligence challenges in a world of information overload. Many summarisers, mostly extractive, have been developed in recent years (Radev et al., 2004; Mihalcea and Tarau, 2004; Wong et al., 2008; Celikyilmaz and Hakkani-Tür, 2011). Research is moving beyond extraction in various directions: One could perform text manipulation such as compression as a separate step after extraction (Knight and Marcu, 2000; Cohn and Lapata, 2008), or alternatively, one could base a summary on an internal semantic representation such as the proposition (Lehnert, 1981; McKeown and Radev, 1995).

One summarisation model that allows manipulation of semantic structures of texts was proposed by Kintsch and van Dijk (1978, henceforth KvD). It is a model of human text processing, where the text is turned into propositions and processed incrementally, sentence by sentence. The final summary is based on those propositions whose semantic participants (arguments) are well-connected to others in the text and hence likely to be remembered by a human reading the text, under the assumption of memory limitations.

Such a deep model is attractive because it provides the theoretical possibility of performing inference and generalisation over propositions, even if current NLP technology only supports shallow versions of such manipulations. This gives it a clear theoretical advantage over non-propositional extraction systems whose information units are individual words and their connections, e.g. centroids or random-walk models.

We present in this paper a new KvD-based summariser that is word sense-aware, unlike our earlier implementation (Fang and Teufel, 2014). §2 explains the KvD model with respect to summarisation. §3 and §4 explain why and how we use lexical chains to model argument overlap, a phenomenon which is central to KvD-style summarisation. §6 presents experimental evidence that our model of argument overlap is superior to the earlier one. Our summariser additionally beats several extractive state-of-the-art summarisers. We show that this advantage does not come from our use of lexical chains alone, but also from KvD's incremental processing.

Our second contribution concerns a new corpus of educational texts, presented in §5. Part of the reason why we prefer a genre other than news is the vexingly good performance of the lead baseline in the news genre. Traditionally, many summarisers struggled to beat this baseline (Lin and Hovy, 2003). We believe that the problem is partly due to the journalistic style, which calls for an abstract-like lead. If we want to measure the content selection ability of summarisers, alternat-

Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, pages 479–485,
Berlin, Germany, August 7-12, 2016.

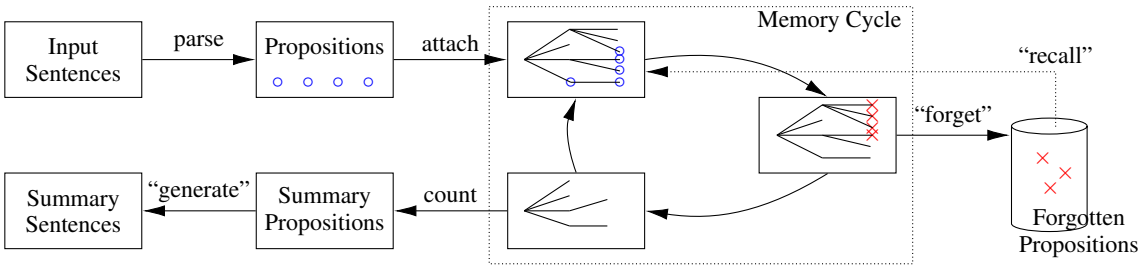

Figure 1: The KvD-inspired incremental summarisation model.

ive data sets are needed. Satisfyingly, we find that on our corpus the lead baseline is surpassable by intelligent summarisers.

2 The KvD Model

The KvD model is a cognitive account of human text comprehension. In our KvD-inspired model (Figure 1), the summariser constructs a list of propositions as a meaning representation from a syntactic parse of the input text. A batch of new propositions (○ in the figure) are processed for each sentence. At the beginning of a memory cycle, these new propositions are added to a *coherence tree*, which represents the working memory. They attach to the existing propositions on the tree with which they have the strongest overlap in arguments. At the end of a cycle, as a simulation of limited memory, only a few important propositions are carried over to the next cycle, while the others are "forgotten" (represented by ×). This selection is based on the location of propositions in the tree, using the so-called *leading edge strategy*; propositions that are on more recent edges, or that are attached higher, are more likely to be retained. The model attempts all future attachments using only the propositions in working memory, and allows to reuse forgotten ones only if this strategy runs into problems (when a new proposition could not otherwise be attached).

KvD suggest that the decision whether a proposition should be included in the final summary depends on three factors: a) the number of cycles where it was retained in working memory, b) whether it is a generalisation, and c) whether it is a meta-statement (or macro-proposition).

For its explanatory power and simplicity, the model has been well-received not only in the fields of cognitive psychology (Paivio, 1990; Lave, 1988) and education (Gay et al., 1976), but also in the summarisation community (Moens et al., 2003; Uyttendaele et al., 1998; Hahn and Reimer, 1984).

We presented the first computational prototype of the model that follows the proposition-centric processing closely (Fang and Teufel, 2014). Of the factors mentioned above, only the first is modelled in this summariser (called FT14). That is, we use the frequency of a proposition being retained in memory as the only indicator of its summary-worthiness. This is a simplification due to the fact that robust inference is beyond current NLP capability. Additionally, macro-propositions depend on domain-specific schema, whereas our system aims to be domain-independent.

Zhang et al. (2016) presented a summariser based on a later cognitive model by Kintsch (1998). Instead of modelling importance of propositions directly, their summariser computes the importance of words by spreading activation cyclically, but extracts at proposition level.

Although the summariser presented in the current paper, a newer version of FT14, is capable of sub-sentential content selection, we present its output in the form of extracted sentences that contain the most summary-worthy propositions. This is different from FT14, where we used a token-based extraction method. A better output would of course be an abstract based on the selected propositions, but we currently do not have a language generation module and can therefore evaluate only the content selection ability of our summariser.

3 Argument Overlap

The central mechanism of the KvD model is *argument overlap* of propositions, and it is key to successful content selection. This is because there are often multiple propositions on the tree where a new proposition could attach, of varying attractiveness. The task therefore boils down to ranking attachments, for instance by the strength of overlap, and the position in the tree.

Figure 2 is an example of competing attachment

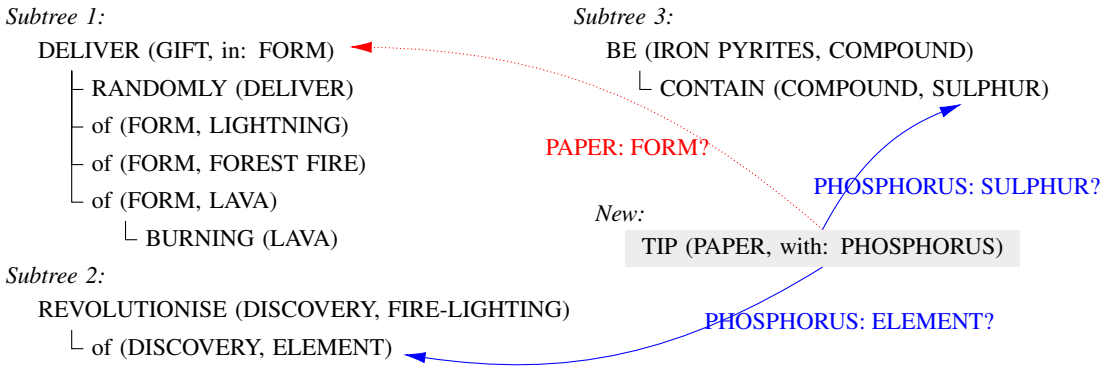

Figure 2: Possible attachments of a new proposition.

sites. Three subtrees in the working memory are shown, containing propositions that correspond to the text pieces 1) *[fire was] a gift randomly delivered in the form of lightning, forest fire or burning lava*, 2) *fire-lighting was revolutionised by the discovery of the element*, and 3) *iron pyrites, a compound that contains sulphur*, respectively. The new proposition corresponds to the text *paper tipped with phosphorus*. It can attach in subtree 2, because *phosphorus* is a kind of *element*; it can also attach in subtree 3, because both *phosphorus* and *sulphur* are chemicals.

The definition of argument overlap is conceptually simple, namely reference of the arguments to the same concept, which can be an entity, an event, or a class of things. In KvD's manual demonstration of the algorithm, the resolution of textual expressions to concepts relies on human intelligence. A "perfect" coreference resolver is arguably all we need, but coreference as currently defined excludes generics, abstract concepts, paraphrases, bridging connections (Weischedel et al., 2007) and several other relevant linguistic phenomena. This means an insufficient number of possible overlaps are found by current coreference systems, if no further information is used. How exactly to model argument overlap for a KvD summariser is therefore open to exploration.

We use other sources of information that addresses topicality and semantic relatedness, in combination with coreference resolution. In FT14, that source was the distributional similarity of words, normalised with respect to their distractors in context to achieve numerically comparable overlap scores. In this paper, we argue that using the shared membership in lexical chains as the other source provides a better basis for ranking argument overlap.

FT14's overlap detection runs into problems in the situation above (Figure 2). Under FT14's definition of argument overlap as distributional semantic distance, the link between *paper* and *form* is as strong as the other possibilities, which leads to the attachment of the new proposition as a child node of the root proposition of subtree 1 due to higher tree level. This attachment uses the wrong sense of the polysemous word *form* ("`form/8` – a printed document with spaces in which to write"). In our new ranking of attachment sites, lexical chains enable us to reject the spurious attachment, as we will now explain.

4 Our Lexical Chain-Based System

In our new model, argument overlap is computed using lexical chains (Barzilay and Elhadad, 1997), a construct that combines the ideas of topicality and word sense clusters. A lexical chain is an equivalence class of expressions found in the text whose presumed senses in context are related to the same concept or topic. For the example in the last section, in our system *form* is correctly resolved to sense 2, not sense 8, and as `form/2` and `paper/1` are not members of the same lexical chain, the wrong attachment is prevented.

Lexical chain algorithms typically use WordNet (Miller, 1995) to provide the lexical relations needed, whereby each synset (synonym set) represents a concept. Hypernyms and hyponyms are related to the same topic, and they may be in a coreference relationship with the concept. To a lesser extent, the potential for coreference also holds for siblings of a concept. WordNet relations therefore give information about concept identity and topical relatedness, both of which are aspects of argument overlap.

We implemented Galley and McKeown's (2003, henceforth GM03) chaining algorithm, which im-

proves over Barzilay and Elhadad's and Silber and McCoy's (2002) chain definition by introducing the limitation of "one sense per discourse", i.e. by enforcing that all occurrences of the same word take the same sense in one document. Initially designed to improve word sense disambiguation accuracy, GM03's method has been shown to improve summarisation quality as well (Ercan and Cicekli, 2008).

In GM03, the edge weight between possible word senses of two word occurrences depends on the lexical relation and the textual distance between them. Each word is disambiguated by choosing the sense that maximises the sum of weights of the edges leaving all its occurrences. Edges that are based on non-selected senses are then discarded. Once the entire text has been processed, each connected component of the graph represents a lexical chain.

As far as nouns[1] are concerned, we follow GM03's edge weights, but unlike GM03, we also allow verbs to enter into chains. We do this in order to model nominalised event references, and to provide a sufficient number of possible connections. Table 1 provides the distance of relations; weights of verb and derivation relations equal to the weights of noun relations on the same row. Instead of assigning an overlap value of 1 to all pairs of words in the same chain, the extent of overlap is given as $a^{\sum_{e \in E} d_e}$, where E is the set of edges in the shortest path between the two words in the graph of lexical relations, d_e the distance of the lexical relation of e, and a an attenuation factor we set at 0.7. This models the transition from concept sameness to broader relatedness. We found empirically that the introduction of verbs and the graded overlap value using relation distance improves the performance of our KvD summariser.

Lexical coverage of this algorithm is good: WordNet covers 98.3% of all word occurrences allowed into our lexical chains in the experiment in §6, excluding those POS-tagged as proper nouns. For unknown words, the system's backoff strategy is to form overlap only if the surface strings match.

The structuring of information in a memory tree and the incremental addition of information, including the concept of "forgetting", are key claims of the KvD model. But do these manipulations actually add any value beyond the information contained in a global network representing *all* connections between all propositions in the text? In such a network without forgetting or discourse structure, standard graph algorithms could be used to determine central propositions. This hypothesis is tested in §6.

Distance	Noun	Verb	Derivation
0	synonymy		
1	hypernymy	synonymy	noun-to-verb
2	sibling	hypernymy	

Table 1: Distance of lexical relations.

5 New Corpus of Texts and Summaries

We introduce new evaluation materials, created from the reading sections of Academic Tests of the *Official IELTS Practice Materials* (British Council et al., 2012).

The IELTS is a standardised test of English proficiency for non-native speakers. The texts cover various general topics, and resemble popular science or educational articles. They are carefully chosen to be of the same difficulty level, and understandable by people of any cultural background. Unlike news text, they also presuppose less external knowledge, such as US politics, which makes it easier to demonstrate the essence of proposition-based summarisation.

Out of all 108 texts of volumes 1–9, we randomly sampled 31. We then elicited 4 summaries summary for each, written by 14 members of our university, i.e., a total of 124 summaries.[2] We asked the summarisers to create natural-sounding text, keeping the length strictly to 100 ± 2 words. They were allowed but not encouraged to paraphrase text.

6 Experiment

6.1 Systems and Baselines

We test 7 automatic summarisers against each other on this evaluation corpus. Our summariser (O) runs the KvD memory cycles and uses lexical chains to determine argument overlap. It is not directly comparable to FT14 due to the difference in generation method, described in §2. In order to be able to compare to FT14 nevertheless, we created a version that uses our new sentence extraction module together with an argument over-

[1] Following Silber and McCoy (2002), we create an additional chain for each named entity, in addition to those chains defined by WordNet synsets.

[2] Max number of summaries per person 31, min number 2. The summaries are available for download at `http://www.cl.cam.ac.uk/~sht25`.

	O	D	C	M	LR	TR	L
1	**.376**	.349	.351	.343	.341	.343	.341
2	**.122**	.094	.088	.092	.100	.094	.100
L	**.345**	.320	.318	.308	.314	.309	.314
SU4	**.154**	.131	.129	.128	.132	.130	.132

Table 2: ROUGE F-scores by four metrics.

lap module very similar to FT14 but with an even stronger model for semantic similarity, the cosine similarity of word embeddings pre-trained using word2vec (Mikolov et al., 2013) on part of the Google News dataset ($\sim$ 100 billion words), and we call this system D.

Another variant, C, tests the hypothesis that the recurrent KvD processing is not superior than simpler network analysis. Summariser C constructs only one graph, where every two propositions are connected by an edge whose length is the reciprocal of their argument overlap, and uses betweenness centrality to determine proposition importance. We choose betweenness centrality because we found it to outperform other graph algorithms, including closeness centrality and eigenvector centrality.

We also test against the lead baseline (L) and three well-known lexical similarity-based single document summarisers: MEAD (Radev et al., 2004, M), TextRank (Mihalcea and Tarau, 2004, TR), and LexRank (Erkan and Radev, 2004, LR).

Because the evaluation tool we use is sensitive to text length, fair evaluation demands equal length of all summaries tested. We obtain output of exactly 100 ± 2 words from each summariser by iteratively requesting longer summaries, and unless this results in a sentence break within 2 tokens of the 100-word limit, we cut the immediately longer output to exactly 100 words.

6.2 Results

For automated evaluation, we use ROUGE (Lin, 2004), which evaluates a summary by comparing it against several gold standard summaries. Table 2 shows our results in terms of ROUGE-1, 2, L and SU4.[3] The metrics are based on the co-occurrence of unigrams, bigrams, longest common subsequences, and skip-bigrams (within distance of 4 and including unigrams), respectively. Our summariser outperforms all other summarisers,[4] and is the only summariser that beats the lead baseline.

The fact that our summariser beats D, our KvD summariser using FT14-style distributional semantics for argument overlap, is clear evidence that our method of lexical chaining provides a superior model of argument overlap. On this genre, D performs indistinguishably from the other summarisers. This is in line with our earlier findings for FT14 on DUC (Over and Liggett, 2002) news texts, where the token extraction-based summariser was comparable to extractive summarisers but was outperformed by MEAD. In a qualitative analysis, we found that a main source of error in FT14's system was that it favoured related but semantically and pragmatically incompatible terms over compatible paraphrases. This is a side-effect of the use of co-occurrence, which relies on syntagmatic rather than paradigmatic similarities, and which is insensitive to word senses. As a result, context-unaware distributional semantics allows too many spurious overlaps.

The fact that summariser C is significantly worse than our summariser shows that the idea of incrementally maintaining a KvD-style structured memory is effective for summarisation, despite the simplifications we had to make. This naturally points to the direction of modelling incremental memory updates for summarisation, which also makes modelling with a recurrent neural network plausible in the future.

The current experiment can be seen as a demonstration of the superiority of KvD proposition-based *content selection* on a genre of commonsense, naturally occurring texts. This was the case even with a inferior "generation" method, namely sentence extraction. Reading through the propositions, we had the impression that they manage to capture relevant information about the text in a much shorter and more modular form than extracted sentences, although this cannot be demonstrated with a surface-based methodology such as ROUGE. Content selection is of course only the first step of summarisation; we are currently working on a grammar-based re-generation from the selected propositions.

Acknowledgments

The CSC Cambridge International Scholarship for the first author is gratefully acknowledged.

[3] The scores of L and LR are very close, but not identical.

[4] We use the paired Wilcoxon test (two-tailed). Differences between O and each other summariser at $p < 0.01$. All differences between all summarisers other than O are insignificant ($p > 0.05$).

References

Regina Barzilay and Michael Elhadad. 1997. Using lexical chains for text summarization. *Proceedings of the ACL Workshop on Intelligent Scalable Text Summarization*.

British Council, IDP Education Australia, and University of Cambridge Local Examinations Syndicate. 2012. *Official IELTS Practice Materials Volume 1. Paperback with CD*. Klett Ernst /Schulbuch.

Asli Celikyilmaz and Dilek Hakkani-Tür. 2011. Discovery of topically coherent sentences for extractive summarization. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 491–499. Association for Computational Linguistics.

Trevor Cohn and Mirella Lapata. 2008. Sentence compression beyond word deletion. In *Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1*, pages 137–144. Association for Computational Linguistics.

Gonenc Ercan and Ilyas Cicekli. 2008. Lexical cohesion based topic modeling for summarization. In *Computational Linguistics and Intelligent Text Processing*, pages 582–592. Springer.

Günes Erkan and Dragomir R Radev. 2004. Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research*, pages 457–479.

Yimai Fang and Simone Teufel. 2014. A summariser based on human memory limitations and lexical competition. *EACL 2014*, page 732.

Michel Galley and Kathleen McKeown. 2003. Improving Word Sense Disambiguation in Lexical Chaining. In *IJCAI*, pages 1486–1488.

Lorraine R Gay, Geoffrey E Mills, and Peter W Airasian. 1976. *Educational research: Competencies for analysis and application*. Merrill Columbus, OH.

Udo Hahn and Ulrich Reimer. 1984. Computing text constituency: An algorithmic approach to the generation of text graphs. In *Proceedings of the 7th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '84, pages 343–368, Swinton, UK. British Computer Society.

Walter Kintsch and Teun A. van Dijk. 1978. Toward a model of text comprehension and production. *Psychological review*, 85(5):363–394.

Walter Kintsch. 1998. *Comprehension: A paradigm for cognition*. Cambridge university press.

Kevin Knight and Daniel Marcu. 2000. Statistics-based summarization-step one: Sentence compression. In *AAAI/IAAI*, pages 703–710.

Jean Lave. 1988. *Cognition in practice: Mind, mathematics and culture in everyday life*. Cambridge University Press.

Wendy G Lehnert. 1981. Plot units and narrative summarization. *Cognitive Science*, 5(4):293–331.

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-Volume 5*, pages 73–80. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, pages 74–81.

Kathleen McKeown and Dragomir R Radev. 1995. Generating summaries of multiple news articles. In *Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 74–82. ACM.

Rada Mihalcea and Paul Tarau. 2004. Textrank: Bringing order into texts. In *EMNLP 2004*. Association for Computational Linguistics.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

George A Miller. 1995. WordNet: a lexical database for English. *Communications of the ACM*, 38(11):39–41.

Marie-Francine Moens, Roxana Angheluta, and Rik De Busser. 2003. Summarization of texts found on the world wide web. In *Knowledge-Based Information Retrieval and Filtering from the Web*, pages 101–120. Springer.

Paul Over and W Liggett. 2002. Introduction to duc: an intrinsic evaluation of generic news text summarization systems. *Proc. DUC. http://wwwnlpir. nist. gov/projects/duc/guidelines/2002. html*.

Allan Paivio. 1990. *Mental representations*. Oxford University Press.

Dragomir Radev, Timothy Allison, Sasha Blair-Goldensohn, John Blitzer, Arda Celebi, Stanko Dimitrov, Elliott Drabek, Ali Hakim, Wai Lam, Danyu Liu, et al. 2004. MEAD – a platform for multidocument multilingual text summarization. In *Proceedings of LREC*.

H. Gregory Silber and Kathleen F. McCoy. 2002. Efficiently Computed Lexical Chains as an Intermediate Representation for Automatic Text Summarization. *Computational Linguistics*, 28(4):487–496, December.

Caroline Uyttendaele, Marie-Francine Moens, and Jos Dumortier. 1998. Salomon: automatic abstracting of legal cases for effective access to court decisions. *Artificial Intelligence and Law*, 6(1):59–79.

Ralph Weischedel, Sameer Pradhan, Lance Ramshaw, Michelle Franchini, Mohammed El-bachouti, Martha Palmer, Mitchell Marcus, Ann Taylor, Craig Greenberg, Eduard Hovy, Robert Belvin, and Ann Houston. 2007. Co-reference Guidelines for English OntoNotes. Technical report, Linguistic Data Consortium.

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-Volume 1*, pages 985–992. Association for Computational Linguistics.

Renxian Zhang, Wenjie Li, Naishi Liu, and Dehong Gao. 2016. Coherent narrative summarization with a cognitive model. *Computer Speech & Language*, 35:134–160.