Unsupervised extractive summarization via coverage maximization with syntactic and semantic concepts

Natalie Schluter and **Anders Søgaard**
Center for Language Technology
University of Copenhagen
{natschluter,soegaard}@hum.ku.dk

Abstract

Coverage maximization with bigram concepts is a state-of-the-art approach to unsupervised extractive summarization. It has been argued that such concepts are adequate and, in contrast to more linguistic concepts such as named entities or syntactic dependencies, more robust, since they do not rely on automatic processing. In this paper, we show that while this seems to be the case for a commonly used newswire dataset, use of syntactic and semantic concepts leads to significant improvements in performance in other domains.

1 Introduction

State-of-the-art approaches to extractive summarization are based on the notion of coverage maximization (Berg-Kirkpatrick et al., 2011). The assumption is that a good summary is a selection of sentences from the document that contains as many of the important concepts as possible. The importance of concepts is implemented by assigning weights w_i to each concept i with binary variable c_i, yielding the following coverage maximization objective, subject to the appropriate constraints:

$$\sum_{i}^{N} w_i c_i \quad (1)$$

In proposing bigrams as concepts for their system, Gillick and Favre (2009) explain that:

> [c]oncepts could be words, named entities, syntactic subtrees or semantic relations, for example. While deeper semantics make more appealing concepts, their extraction and weighting are much more error-prone. Any error in concept extraction can result in a biased objective function, leading to poor sentence selection. (Gillick and Favre, 2009)

Several authors, e.g., Woodsend and Lapata (2012), and Li et al. (2013), have followed Gillick and Favre (2009) in assuming that bigrams would lead to better practical performance than more syntactic or semantic concepts, even though bigrams serve as only an approximation of these.

In this paper, we revisit this assumption and evaluate the maximum coverage objective for extractive text summarization with syntactic and semantic concepts. Specifically, we replace bigram concepts with new ones based on syntactic dependencies, semantic frames, as well as named entities. We show that using such concepts can lead to significant improvements in text summarization performance outside of the newswire domain. We evaluate coverage maximization incorporating syntactic and semantic concepts across three different domains: newswire, legal judgments, and Wikipedia articles.

2 Concept coverage maximization for extractive summarization

In extractive summarization, the unsupervised version of the task is sometimes set up as that of finding a subset of sentences in a document, within some relatively small budget, that covers as many of the important concepts in the document as possible. In the maximum coverage objective, concepts are considered as independent of each other. Concepts are weighted by the number of times they appear in a document. Moreover, due the NP-hardness of coverage maximization, for an exact solution to the concept coverage optimization problem, we resort to fast solvers for integer linear programming, under some appropriate constraints.

Bigrams. Gillick and Favre (2009) proposed to use bigrams as concepts, and to weight their contribution to the objective function in Equation (1)

Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Short Papers), pages 840–844,
Beijing, China, July 26-31, 2015.

by the frequency with which they occur in the document. Some pre-processing is first carried out to these bigrams: all bigrams consisting uniquely of stop-words are removed from consideration, and each word is stemmed. They also require bigrams to occur with a minimal frequency (cf. Section 3.2).

Named entities. We consider three new types of concepts, all suggested, but subsequently rejected by Gillick and Favre (2009). The first is simply to use named entities, e.g., *Court of Justice of the European Union*, as concepts. This reflects the intuition that persons, organizations, and locations are particularly important for extractive summarization. We use an NER maximum entropy tagger[1] to augment documents with named entities.

Syntactic dependencies. The second type of concept is dependency subtrees. In particular, we extract labeled and unlabeled syntactic dependencies, e.g., DEPENDENCY(walks,John) or SUBJECT(walks,John), from sentences and represent them by such syntactic concepts. We use the Stanford parser[2] to augment documents with syntactic dependencies. As was done for bigrams, each word in the dependency is stemmed. Syntactic dependency-based concepts are intuitively a closer approximation than bigrams to concepts in general.

Semantic frames. The intuition behind our use of frame semantics is that a summary should represent the most central semantic frames (Fillmore, 1982; Fillmore et al., 2003) present in the corresponding document—indeed, we consider these frames to be actual types of concepts. We extract frame names from sentences for a further type of concepts under consideration. We use SEMAFOR[3] to augment documents with semantic frames.

3 Experiments

3.1 Data

In order to investigate the importance of concept types across different domains, we evaluate our systems across three distinct domains, which we refer to as ECHR, TAC08, and WIKIPEDIA.

ECHR consists of judgment-summary pairs scraped from the European Court of Human Rights case-law website, HUDOC[4]. The document-summary pairs were split into training, development and test sets, consisting of 1018, 117, and 138 pairs, respectively. In the training set (pruning sentences of length less than 5), the average document length is 13,184 words or 455 sentences. The average summary length is 806 words or 28 sentences. For both documents and summaries, the average sentence length is 29 words.

TAC08 consists of 48 queries and 2 newswire document sets for each query, each set containing 10 documents. Document sets contain 235 input sentences on average, and the mean sentence length is 25 words. Summaries consist of 4 sentences or 100 words on average.

WIKIPEDIA consists of 992 Wikipedia articles (all labeled "good article"[5]) from a comprehensive dump of English language Wikipedia articles[6]. We use the Wikipedia abstracts (the leading paragraphs before the table of contents) as summaries. The (document,summary) pairs were split into training, development and test sets, consisting of 784, 97, and 111 pairs, respectively. In the training set (pruning sentences of length less than 5), the average document length is around 8918 words or 339 sentences. The average summary length is 335 words or 13 sentences. For both documents and summaries, the average sentence length is around 26 words.

In our main experiments, we use unsupervised summarization techniques, and we only use the training summaries (and not the documents) to determine output summary lengths.

3.2 Baseline and systems

Our baseline is the bigram-based extraction summarization system of Gillick and Favre (2009), `icsisumm`[7]. Their system was originally intended for multi-document update summarization, and summaries are extracted from document sentences that share more than k content words with some query. We follow this approach for the TAC08 data. For ECHR and WIKIPEDIA, the task is single document summarization, and the now irrelevant topic-document intersection preprocessing step is eliminated.

[1] `http://www.nltk.org/`

[2] `http://nlp.stanford.edu/software/lex-parser.shtml`

[3] `http://www.ark.cs.cmu.edu/SEMAFOR/`

[4] `http://hudoc.echr.coe.int/`

[5] `http://en.wikipedia.org/wiki/Wikipedia:Good_articles`

[6] `https://dumps.wikimedia.org/enwiki/latest/enwiki-latest-pages-articles-multistream.xml.bz2`

[7] `https://code.google.com/p/icsisumm/`

The original system uses the GNU linear programming kit[8] with a time limit of 100 seconds. For all experiments presented in this paper, we double this time limit; we experimented with longer time limits on the development set for the ECHR data, without any performance improvements. Once the summarizer reaches the time limit, a summary is output based on the current feasible solution, whether the solution is optimal or not. Moreover, the current `icsisumm` (v1) distribution prunes sentences shorter than 10 words. We note that we also tried replacing `glpk` by `gurobi`[9], for which no time limit was necessary, but found poorer results on the development set of the ECHR data.

The original system takes several important input parameters.

1. **Summary length,** for TAC08, is specified by the TAC 2008 conference guidelines as 100 words. For WIKIPEDIA and ECHR, we have access to training sets which gave an average summary length of around 335 and 805 words respectively, which we take as the standard output summary length.

2. **Concept count cut-off** is the minimum frequency of concepts from the document (set) that qualifies them for consideration in coverage maximization. For bigrams of the original system on TAC08, there are two types of document sets: 'A' and 'B'. For 'A' type documents, Gillick and Favre (2009) set this threshold to 3 and for 'B' type documents, they set this to 4. For WIKIPEDIA and ECHR, we take the bigram threshold to be 4. In our extension of the system to other concepts, we do not use any threshold.

3. **First concept weighting:** in multi-document summarization, there is the possibility for repeated sentences. Concepts from first-encountered sentences may be weighted higher: these concept counts from first-encountered sentences are doubled for 'B' documents and remain unchanged for 'A' documents in the original system on TAC08. For other concepts, we do not alter frequencies in this manner, which is justified by the task change to single-document summarization.

4. **Query-sentence intersection threshold,** is set to 1 for 'A' documents and 0 to 'B' documents in the original system on TAC08. This threshold is only for the update summarization task and therefore does not concern ECHR and WIKIPEDIA.

In addition to our baseline, we consider five single-concept systems using (a) named entities, (b) labeled dependencies, (c) unlabeled dependencies, (d) semantic frame names, and (e) semantic frame dependencies, as well as the five systems combining each of these new concept types with bigrams. For the combination of these new concepts with bigrams, we extend the objective function to maximise in, Equation (1), into two sums—one for bigram concepts and the other for the new concept type—with their relative importance controlled by a parameter α. N_1 and N_2 are the number of bigram and other concept types occurring with the permitted threshold frequency in the document, relatively. Given that we are carrying out unsupervised summarization, rather than tune α, we set $\alpha = 0.5$, so the concepts are considered in their totality (i.e., $N_1 + N_2$ concepts together) with no explicit favouring of one over the other that does not naturally fall out of concept frequency.

$$(1-\alpha)\sum_{i}^{N_1} w_i \texttt{bigram}_i + \alpha \sum_{j}^{N_2} w_j \texttt{new\_concept}_j$$

3.3 Results

We evaluate output summaries using ROUGE-1, ROUGE-2, and ROUGE-SU4 (Lin, 2004), with no stemming and retaining all stopwords. These measures have been shown to correlate best with human judgments in general, but among the automatic measures, ROUGE-1 and ROUGE-2 also correlate best with the Pyramid (Nenkova and Passonneau, 2004; Nenkova et al., 2007) and Responsiveness manual metrics (Louis and Nenkova, 2009). Moreover, ROUGE-1 has been shown to best reflect human-automatic summary comparisons (Owczarzak et al., 2012).

For single concept systems, the results are shown in Table 1, and concept combination system results are given in Table 2.

We first note that our runs of the current distribution of `icsisumm` yield significantly worse ROUGE-2 results than reported in (Gillick and Favre, 2009) (see Table 1, BIGRAMS): 0.081 compared to 0.110 respectively.

[8] `http://www.gnu.org/software/glpk/`
[9] `http://www.gurobi.com/`

On the TAC08 data, we observe no improvements over the baseline BIGRAM system for any ROUGE metric here. Hence, Gillick and Favre (2009) were right in their assumption that syntactic and semantic concepts would not lead to performance improvements, when restricting ourselves to this dataset. However, when we change domain to the legal judgments or Wikipedia articles, using syntactic and semantic concepts leads to significant gains across all the ROUGE metrics.

For ECHR, replacing bigrams by frame names (FRAME) results in an increase of +0.1 in ROUGE-1, +0.031 in ROUGE-2 and +0.046 in ROUGE-SU4. We note that FrameNet 1.5 covers the legal domain quite well, which may explain why these concepts are particularly useful for the ECHR dataset. However, labeled (LDEP) and unlabeled (UDEP) dependencies also significantly outperform the baseline.

For WIKIPEDIA, replacing bigrams by labeled or unlabeled syntactic dependencies results in significant improvements: an increase of +0.088 for ROUGE-1, +0.015 for ROUGE-2, and +0.03 for ROUGE-SU4. Interestingly, the NER system also yields significantly better performance over the baseline, which may reflect the nature of Wikipedia articles, often being about historical figures, famous places, organizations, etc.

We observe in Table 2, that for concept combination systems as well, ROUGE scores on TAC08 do not indicate any improvement in performance. However, best ROUGE-1 scores are produced for both ECHR and WIKIPEDIA data with systems that incorporate semantic frame names. For WIKIPEDIA, best ROUGE-2 and ROUGE-SU4 scores incorporate named-entity information.

ECHR			
concept	R-1 (95% conf.)	R-2 (95% conf.)	R-SU4 (95% conf.)
BIGRAMS	0.544 (0.528-0.562)	0.204 (0.195-0.215)	0.266 (0.257-0.277)
NER	0.549 (0.534-0.564)	0.184 (0.174-0.193)	0.254 (0.244-0.264)
LDEP	0.609 (0.597-0.621)	0.225 (0.217-0.235)	0.293 (0.285-0.302)
UDEP	0.612 (0.6-0.626)	0.227 (0.218-0.238)	0.295 (0.287-0.305)
FRAMES	**0.643** (0.63-0.657)	**0.235** (0.224-0.248)	**0.312** (0.302-0.323)
TAC08			
concept	R-1 (95% conf.)	R-2 (95% conf.)	R-SU4 (95% conf.)
BIGRAMS	**0.35** (0.34-0.36)	**0.081** (0.073-0.089)	**0.119** (0.113-0.126)
NER	0.307 (0.297-0.317)	0.054 (0.049-0.06)	0.093 (0.089-0.099)
LDEP	0.335 (0.325-0.346)	0.072 (0.065-0.08)	0.109 (0.103-0.116)
UDEP	**0.342** (0.331-0.353)	**0.075** (0.067-0.083)	**0.113** (0.106-0.12)
FRAMES	0.301 (0.292-0.31)	0.048 (0.042-0.053)	0.089 (0.085-0.094)
WIKIPEDIA			
concept	R-1 (95% conf.)	R-2 (95% conf.)	R-SU4 (95% conf.)
BIGRAMS	0.391 (0.364-0.415)	0.103 (0.094-0.113)	0.152 (0.134-0.163)
NER	**0.473** (0.46-0.487)	**0.114** (0.105-0.123)	**0.178** (0.169-0.186)
LDEP	**0.478** (0.461-0.495)	**0.116** (0.107-0.125)	**0.179** (0.169-0.188)
UDEP	**0.479** (0.462-0.497)	**0.118** (0.109-0.128)	**0.18** (0.17-0.189)
FRAMES	**0.476** (0.461-0.494)	**0.102** (0.094-0.112)	**0.172** (0.164-0.182)

Table 1: Single concept results on ECHR, TAC08, and WIKIPEDIA.

4 Related work

Most researchers have used bigrams as concepts in coverage maximization-based approaches to unsupervised extractive summarization. Filatova and Hatzivassiloglou (2004), however,use relations between named entities as concepts in extractive summarization. They use slightly different extraction algorithms, but their work is similar in spirit to ours. Nishikawa et al. (2010), also, use opinions – tuples of targets, aspects, and polarity – as concepts in opinion summarization. In early work on summarization, Silber and McCoy (2000) used WordNet synsets as concepts. Kitajima and Kobayashi (2011) replace words by syntactic dependencies in the Maximal Marginal Relevance Multidocument measure first proposed by Goldstein et al. (2000) for evaluating the importance of sentences in query-based extractive summarization, yielding improvements for their Japanese newswire dataset.

5 Conclusions

This paper challenges the assumption that bigrams make better concepts for unsupervised extractive summarization than syntactic and semantic concepts relying on automatic processing. We show that using concepts relying on syntactic dependencies or semantic frames instead of bigrams leads to significant performance improvements of coverage maximization summarization across domains.

References

Taylor Berg-Kirkpatrick, Dan Gillick, and Dan Klein. 2011. Jointly learning to extract and compress. In *Proc of ACL*, Portland, OR, USA.

ECHR			
concept	R-1 (95% conf.)	R-2 (95% conf.)	R-SU4 (95% conf.)
NER	0.605 (0.595-0.616)	0.228 (0.22- 0.237)	0.093 (0.288-0.303)
LDEP	0.614 (0.597-0.632)	0.235 (0.225-0.246)	0.301 (0.29-0.312)
UDEP	0.62 (0.605-0.634)	0.237 (0.227-0.247)	0.304 (0.294-0.313)
FRAMES	**0.65** (0.638-0.662)	**0.251** (0.24-0.262)	**0.322** (0.313-0.333)
TAC08			
concept	R-1 (95% conf.)	R-2 (95% conf.)	R-SU4 (95% conf.)
NER	0.35 (0.339-0.361)	0.082 (0.074-0.09)	0.119 (0.112-0.126)
LDEP	0.345 (0.334-0.355)	0.08 (0.072-0.088)	0.117 (0.11-0.124)
UDEP	0.347 (0.336-0.358)	0.08 (0.072-0.088)	0.12 (0.11-0.125)
FRAMES	0.344 (0.334-0.354)	0.078 (0.071-0.086)	0.115 (0.11-0.122)
WIKIPEDIA			
concept	R-1 (95% conf.)	R-2 (95% conf.)	R-SU4 (95% conf.)
NER	**0.496** (0.483-0.51)	**0.136** (0.127-0.146)	**0.195** (0.187-0.204)
LDEP	**0.495** (0.479-0.511)	**0.132** (0.122-0.141)	**0.192** (0.183-0.202)
UDEP	0.493 (0.478-0.511)	0.13 (0.121-0.14)	0.18 (0.181-0.199)
FRAMES	**0.497** (0.482-0.513)	0.124 (0.114-0.133)	0.187 (0.179-0.197)

Table 2: Results for systems combining bigrams with new concepts, on ECHR, TAC08 and WIKIPEDIA.

Elena Filatova and Vasileios Hatzivassiloglou. 2004. Event-based extractive summarization. In *ACL Workshop on Text Summarization Branches Out*.

Charles J. Fillmore, Christopher. R. Johnson, and Miriam R. L. Petruck. 2003. Background to framenet. *International Journal of Lexicography*, 16.

Charles J. Fillmore, 1982. *Linguistics in the Morning Calm*, chapter Frame Semantics, pages 111–137. Hanshin Publishing Co., Seoul, South Korea.

Dan Gillick and Benoit Favre. 2009. A scalable global model for summarization. In *Proc of ILP*, pages 10–18.

Jade Goldstein, Vibhu Mittal, Jaime Carbonell, and Mark Kantrowitz. 2000. Multi-document summarization by sentence extraction. In *Proc of the ANLP/NAACL Workshop on Automatic Summarization*, pages 40–48.

Risa Kitajima and Ichiro Kobayashi. 2011. A latent topic extracting method based on events in a document and its application. In *Proc of ACL-HLT Student Session*, pages 30–35, Portland, OR, USA.

Chen Li, Xian Qian, and Yang Liu. 2013. Using supervised bigram-based ilp for extractive summarization. In *Proc of ACL*, Sofia, Bulgaria.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Proc of WAS*, Barcelona, Spain.

Annie Louis and Ani Nenkova. 2009. Automatically evaluation content selection in summarization without human models. In *Proc of EMNLP*, pages 306–314, Singapore.

Ani Nenkova and Rebecca Passonneau. 2004. Evaluating content selection in summarization: The pyramid method. In *Proc. of HLT-NAACL*.

Ani Nenkova, Rebecca Passonneau, and Kathleen McKeown. 2007. The pyramid method:incorporating human content selection variation in summarization evaluation. *ACM Transactions on Speech and Language Processing*, 4.

Hitoshi Nishikawa, Takaaki Hasegawa, Yoshihiro Matsuo, and Genichiro Kikui. 2010. Opinion summarization with integer linear programming formulation for sentence extraction and ordering. In *COLING*.

Karolina Owczarzak, John M. Conroy, Hoa Trang Dang, and Ani Nenkova. 2012. An assessment of the accuracy of automatic evaluation in summarization. In *Proc of the Workshop on Evaluation Metrics and System Comparison for Automatic Summarization*, pages 1–9, Montréal, Canada.

H. Gregory Silber and Kathleen McCoy. 2000. An efficient text summarizer using lexical chains. In *INLG*.

Kristian Woodsend and Mirella Lapata. 2012. Multiple aspect summarization using integer linear programming. In *Proc of EMNLP/CoNLL*, pages 233–243, Jeju Island, Korea.