Concept-based Summarization using Integer Linear Programming: From Concept Pruning to Multiple Optimal Solutions

Florian Boudin, Hugo Mougard, Benoit Favre

Concept-based Summarization using Integer Linear Programming: From Concept Pruning to Multiple Optimal Solutions

Florian Boudin[1] **Hugo Mougard**[1] **Benoit Favre**[2]

[1] LINA - UMR CNRS 6241, Université de Nantes, France
{florian.boudin,hugo.mougard}@univ-nantes.fr

[2] LIF - UMR CNRS 7279, Université Aix-Marseille, France
benoit.favre@lif.univ-mrs.fr

Abstract

In concept-based summarization, sentence selection is modelled as a budgeted maximum coverage problem. As this problem is NP-hard, pruning low-weight concepts is required for the solver to find optimal solutions efficiently. This work shows that reducing the number of concepts in the model leads to lower ROUGE scores, and more importantly to the presence of multiple optimal solutions. We address these issues by extending the model to provide a single optimal solution, and eliminate the need for concept pruning using an approximation algorithm that achieves comparable performance to exact inference.

1 Introduction

Recent years have witnessed increased interest in global inference methods for extractive summarization. These methods formulate summarization as a combinatorial optimization problem, i.e. selecting a subset of sentences that maximizes an objective function under a length constraint, and use Integer Linear Programming (ILP) to solve it exactly (McDonald, 2007).

In this work, we focus on the concept-based ILP model for summarization introduced by (Gillick and Favre, 2009). In their model, a summary is generated by assembling the subset of sentences that maximizes a function of the unique concepts it covers. Selecting the optimal subset of sentences is then cast as an instance of the budgeted maximum coverage problem[1].

As this problem is NP-hard, pruning low-weight concepts is required for the ILP solver to find optimal solutions efficiently (Gillick and Favre, 2009; Qian and Liu, 2013; Li et al., 2013). However, reducing the number of concepts in the model has two undesirable consequences. First, it forces the model to only use a limited number of concepts to rank summaries, resulting in lower ROUGE scores. Second, by reducing the number of items from which sentence scores are derived, it allows different sentences to have the same score, and ultimately leads to multiple optimal summaries.

To our knowledge, no previous work has mentioned these problems, and only results corresponding to the first optimal solution found by the solver are reported. However, as we will show through experiments, these multiple optimal solutions cause a substantial amount of variation in ROUGE scores, which, if not accounted for, could lead to incorrect conclusions. More specifically, the contributions of this work are as follows:

- We evaluate (Gillick and Favre, 2009)'s summarization model at various concept pruning levels. In doing so, we quantify the impact of pruning on running time, ROUGE scores and the number of optimal solutions.
- We extend the model to address the problem of multiple optimal solutions, and we sidestep the need for concept pruning by developing a fast approximation algorithm that achieves near-optimal performance.

2 Concept-based ILP Summarization

2.1 Model definition

Gillick and Favre (2009) introduce a concept-based ILP model for summarization that casts sentence selection as a maximum coverage problem. The key assumption of their model is that the value of a summary is defined as the sum of the weights of the unique concepts it contains. That way, redundancy within the summary is addressed implicitly at a sub-sentence level: a summary only benefits from including each concept once.

[1] Given a collection S of sets with associated costs and a budget L, find a subset $S' \subseteq S$ such that the total cost of sets in S' does not exceed L, and the total weight of elements covered by S' is maximized (Khuller et al., 1999).

Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 1914–1918,
Lisbon, Portugal, 17-21 September 2015.

Formally, let w_i be the weight of concept i, c_i and s_j two binary variables indicating the presence of concept i and sentence j in the summary, Occ_{ij} an indicator of the occurrence of concept i in sentence j, l_j the length of sentence j and L the length limit for the summary, the concept-based ILP model is described as:

$$\max \quad \sum_i w_i c_i \tag{1}$$

$$s.t. \quad \sum_j l_j s_j \leq L \tag{2}$$

$$s_j Occ_{ij} \leq c_i, \quad \forall i, j \tag{3}$$

$$\sum_j s_j Occ_{ij} \geq c_i, \quad \forall i \tag{4}$$

$$c_i \in \{0, 1\} \, \forall i$$

$$s_j \in \{0, 1\} \, \forall j$$

The constraints formalized in equations 3 and 4 ensure the consistency of the solution: selecting a sentence leads to the selection of all the concepts it contains, and selecting a concept is only possible if it is present in at least one selected sentence.

Choosing a suitable definition for concepts and a method to estimate their weights are the two key factors that affect the performance of this model. Bigrams of words are usually used as a proxy for concepts (Gillick and Favre, 2009; Berg-Kirkpatrick et al., 2011). Concept weights are either estimated by heuristic counting, e.g. document frequency in (Gillick and Favre, 2009), or obtained by supervised learning (Li et al., 2013).

2.2 Pruning to reduce complexity

The concept-level formulation of (Gillick and Favre, 2009) is an instance of the budgeted maximum coverage problem, and solving such a problem is NP-hard (Khuller et al., 1999). Keeping the number of variables and constraints small is then critical to reduce the model complexity.

In previous work, efficient summarization was achieved by pruning concepts. One way to reduce the number of concepts in the model is to remove those concepts that have a weight below a given threshold (Gillick and Favre, 2009). Another way is to consider only the top-n highest weighted concepts (Li et al., 2013). Once low-weight concepts are pruned, sentences that do not contain any remaining concepts are removed, further reducing the number of variables and constraints in the model. As such, this can be regarded as a way to approximate the problem.

Pruning concepts to reduce complexity also cuts down the number of items from which summary scores are derived. As we will see in Section 3.2, this results in a lower ROUGE scores and leads to the production of multiple optimal summaries.

The concept weighting function also plays an important role in the presence of multiple optimal solutions. Limited-range functions, such as frequency-based ones, yield many ties and increase the likelihood that different sentences have the same score. Redundancy within the set of input sentences exacerbate this problem, since highly similar sentences are likely to contain the same concepts.

2.3 Summarization parameters

For comparison purposes, we use the same system pipeline as in (Gillick et al., 2009), which is described below.

Step 1: clean input documents; a set of rules is used to remove bylines and format markup.

Step 2: split the text into sentences; we use splitta[2] (Gillick, 2009) and re-attach multi-sentence quotations.

Step 3: compute parameters needed by the model; we extract and weight the concepts.

Step 4: prune sentences shorter than 10 words, duplicate sentences and those that begin and end with a quotation mark.

Step 5: map to ILP format and solve; we use an off-the-shelf ILP solver[3].

Step 6: order selected sentences for inclusion in the summary, first by source and then by position.

Similar to previous work, we use bigrams of words as concepts. Although bigrams are rough approximations of concepts, they are simple to extract and match, and have been shown to perform well at this task. Bigrams of words consisting of two stop words[4] or containing a punctuation mark are discarded. Stemming[5] is then applied to allow more robust matching.

Concepts are weighted using document frequency, i.e. the number of source documents

[2]We use splitta v1.03, `https://code.google.com/p/splitta/`

[3]We use glpk v4.52, `https://www.gnu.org/software/glpk/`

[4]We use the stoplist in nltk, `http://www.nltk.org/`

[5]We use the Porter stemmer in nltk.

	DUC'04				TAC'08			
DF	1	2	3	4	1	2	3	4
# solutions	1.3	1.3	1.5	1.5	1.2	1.3	1.8	4.8
# concepts	2 955	676	247	107	2 909	393	127	56
# sentences	184	175	159	139	174	167	149	129
Avg. time (sec)	22.3	1.7	0.5	0.3	21.5	0.8	0.3	0.2

Table 1: Average number of optimal solutions, concepts and sentences for different minimum document frequencies. The average time in seconds for finding the first optimal solution is also reported.

where the concept was seen. Document frequency is a simple, yet effective approach to concept weighting (Gillick and Favre, 2009; Woodsend and Lapata, 2012; Qian and Liu, 2013). Reducing the number of concepts in the ILP model is then performed by pruning those concepts that occur in fewer than a given number of documents.

ILP solvers usually provide only one solution. To generate alternate optimal solutions, we iteratively add new constraints to the problem that eliminate already found optimal solutions and rerun the solver. We stop the iterations when the value of the objective function returned by the solver changes.

3 Experiments

3.1 Datasets and evaluation measures

Experiments are conducted on the DUC'04 and TAC'08 datasets. For DUC'04, we use the 50 topics from the generic multi-document summarization task (Task 2). For TAC'08, we focus only on the 48 topics from the non-update summarization task. Each topic contains 10 newswire articles for which the task is to generate a summary no longer than 100 words (whitespace-delimited tokens).

Summaries are evaluated against reference summaries using the ROUGE automatic evaluation measures (Lin, 2004). We set the ROUGE parameters to those[6] that lead to highest agreement with manual evaluation (Owczarzak et al., 2012), that is, with stemming and stopwords not removed.

3.2 Results

Table 1 presents the average number of optimal solutions at different levels of concept pruning. Overall, the average number of optimal solutions increases along with the minimum document frequency, reaching 4.8 for TAC'08 at DF $= 4$. Pruning concepts also greatly reduces the number of variables in the ILP formulation, and consequently improves the run-time for solving the problem.

Interestingly, we note that, even without any pruning, the model produces multiple optimal solutions. The choice of document frequency for weighting concepts is responsible for this as it generates many ties. Finer-grained concept weighting functions such as frequency estimation (Li et al., 2013) should therefore be preferred to limit the number of multiple optimal solutions.

The mean ROUGE recall scores of the multiple optimal solutions for different minimal document frequencies are presented in Table 2. Here, the higher the concept pruning threshold, the higher the variability of the generated summaries as indicated by the standard deviation. Best ROUGE scores are achieved without concept pruning while the best compromise between effectiveness and run-time is given when DF ≥ 3, confirming the findings of (Gillick and Favre, 2009).

To show in a realistic scenario how multiple optimal solutions could lead to different conclusions, we compare in Table 3 the ROUGE-1 scores of the summaries generated from the first optimal solution found by three off-the-shelf ILP solvers against that of the systems[7] that participated at TAC'08. We set the minimum document frequency to 3, which is often used in previous work (Gillick and Favre, 2009; Li et al., 2013), and use a two-sided Wilcoxon signed-rank to compute the number of systems that obtain significantly lower and higher ROUGE-1 recall scores[8].

Despite being comparable (p-value > 0.4), the solutions found by the three solvers support different conclusions. The solution found using GLPK

[6] We use ROUGE-1.5.5 with the parameters: `n 4 -m -a -l 100 -x -c 95 -r 1000 -f A -p 0.5 -t 0`

[7] 71 systems participated at TAC'08 but we removed ICSI1 and ICSI2 systems which are based on the concept-based ILP model.

[8] ROUGE-1 recall is most accurate metric to identify the better summary in a pair (Owczarzak et al., 2012).

	DUC'04			TAC'08		
DF	ROUGE-1	ROUGE-2	ROUGE-4	ROUGE-1	ROUGE-2	ROUGE-4
1	37.74 ±0.07	**9.48** ±0.05	1.45 ±0.02	**37.65** ±0.10	**10.63** ±0.08	**2.23** ±0.04
2	37.25 ±0.08	9.14 ±0.02	1.37 ±0.01	37.16 ±0.11	9.96 ±0.07	2.05 ±0.03
3	37.37 ±0.11	9.16 ±0.06	1.41 ±0.02	37.39 ±0.15	10.62 ±0.07	2.13 ±0.03
4	**37.96** ±0.10	9.38 ±0.05	**1.57** ±0.02	36.73 ±0.12	10.10 ±0.08	1.78 ±0.07

Table 2: Mean ROUGE recall and standard deviation for different minimum document frequencies.

Solver	ROUGE-1	↓ / ↑
GLPK	37.33	54 / 0
Gurobi	37.20	52 / 1
CPLEX	37.17	51 / 1

Table 3: ROUGE-1 recall scores for the first optimal solution found by different solvers along with the number of systems that obtain significantly lower (↓) or higher (↑) scores (p-value < 0.05).

indicates that the concept-based model achieves state-of-the-art performance whereas the solutions provided by Gurobi and CPLEX do not do so. The reason for these differences is the use of different solving strategies, involving heuristics for finding feasible solutions more quickly. This example demonstrates that multiple optimal solutions should be considered during evaluation.

3.3 Solving the multiple solution problem

Multiple optimal solutions occur when concepts alone are not sufficient to distinguish between two competing summary candidates. Extending the model so that it provides a single solution can therefore not be done without introducing a second term in the objective function. Following the observation that the frequency of a non-stop word in a document set is a good predictor of a word appearing in a human summary (Nenkova and Vanderwende, 2005), we extend equation 1 as follows:

$$\max \quad \sum_i w_i c_i + \mu \sum_k f_k t_k \qquad (5)$$

where f_k is the frequency of non-stop word k in the document set, and t_k is a binary variable indicating the presence of k in the summary. Here, we want to induce a single solution among the multiple optimal solutions given by concept weighting, and thus set μ to a small value (10^{-6}). We add further constraints, similar to equations 3 and 4, to ensure the consistency of the solution.

This extended model succeeds in giving a single solution that is at least comparable to the mean score of the multiple optimal solutions. However, it requires about twice as much time to solve which makes it impractical for large documents.

3.4 Fast approximation

Instead of pruning concepts to reduce complexity, one may consider using an approximation if results are found satisfactory. Here, similarly to (Takamura and Okumura, 2009; Lin and Bilmes, 2010) we implement the greedy heuristic proposed in (Khuller et al., 1999) that solve the budgeted maximum coverage problem with a performance guarantee $1/2 \cdot (1 - 1/e)$. Table 4 compares the performance of the model that achieves the best trade off between effectiveness and runtime, that is when $DF \geq 3$, with that of the greedy approximation without pruning.

Overall, the approximate solution is over 96% as good as the average optimal solution. Although the ILP solution marks an upper bound on performance, its solving time is exponential in the number of input sentences. The approximate method is then relevant as it marks an upper bound on speed (less than 0.01 seconds to compute) while having performance comparable to the ILP model with concept pruning (p-value > 0.3).

Dataset	ROUGE-1	ROUGE-2
DUC'04	37.14 (−0.7%)	9.37 (+2.3%)
TAC'08	36.90 (−1.3%)	10.27 (−3.3%)

Table 4: ROUGE recall scores of the approximation. The relative difference from the mean score of the multiple optimal solutions is also reported.

4 Conclusion

Multiple optimal solutions are not an issue as long as alternate solutions are equivalent. Unfortunately, summaries generated from different sets of

sentences are likely to differ. We showed through experiments that concept pruning leads to the presence of multiple optimal solutions, and that the latter cause a substantial amount of variation in ROUGE scores. We proposed an extension of the ILP that obtains unique solutions. If speed is a concern, we showed that a near-optimal approximation can be computed without pruning. The implementation of the concept-based summarization model that we use in this study is available at `https://github.com/boudinfl/sume`.

In future work, we intend to extend our study to compressive summarization. We expect that the number of optimal solutions will increase as multiple compression candidates, which are likely to be similar in content, are added to the set of input sentences.

Acknowledgments

This work was partially supported by the GOLEM project (grant of CNRS PEPS FaSciDo 2015, `http://boudinfl.github.io/GOLEM/`). We thank the anonymous reviewers and Rémi Bois for their insightful comments.

References

Taylor Berg-Kirkpatrick, Dan Gillick, and Dan Klein. 2011. Jointly learning to extract and compress. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 481–490, Portland, Oregon, USA, June. Association for Computational Linguistics.

Dan Gillick and Benoit Favre. 2009. A scalable global model for summarization. In *Proceedings of the Workshop on Integer Linear Programming for Natural Language Processing*, pages 10–18, Boulder, Colorado, June. Association for Computational Linguistics.

Dan Gillick, Benoit Favre, Dilek Hakkani-Tur, Berndt Bohnet, Yang Liu, and Shasha Xie. 2009. The icsi/utd summarization system at tac 2009. In *Proceedings of the Second Text Analysis Conference*, Gaithersburg, Maryland, USA. National Institute of Standards and Technology.

Dan Gillick. 2009. Sentence boundary detection and the problem with the u.s. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*, pages 241–244, Boulder, Colorado, June. Association for Computational Linguistics.

Samir Khuller, Anna Moss, and Joseph (Seffi) Naor. 1999. The budgeted maximum coverage problem. *Information Processing Letters*, 70(1):39 – 45.

Chen Li, Xian Qian, and Yang Liu. 2013. Using supervised bigram-based ilp for extractive summarization. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1004–1013, Sofia, Bulgaria, August. Association for Computational Linguistics.

Hui Lin and Jeff Bilmes. 2010. Multi-document summarization via budgeted maximization of submodular functions. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 912–920, Los Angeles, California, June. Association for Computational Linguistics.

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

Ryan McDonald. 2007. A study of global inference algorithms in multi-document summarization. In *Proceedings of the 29th European Conference on IR Research*, ECIR'07, pages 557–564, Berlin, Heidelberg. Springer-Verlag.

Ani Nenkova and Lucy Vanderwende. 2005. The impact of frequency on summarization. *Microsoft Research, Redmond, Washington, Tech. Rep. MSR-TR-2005-101*.

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

Xian Qian and Yang Liu. 2013. Fast joint compression and summarization via graph cuts. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1492–1502, Seattle, Washington, USA, October. Association for Computational Linguistics.

Hiroya Takamura and Manabu Okumura. 2009. Text summarization model based on maximum coverage problem and its variant. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 781–789, Athens, Greece, March. Association for Computational Linguistics.

Kristian Woodsend and Mirella Lapata. 2012. Multiple aspect summarization using integer linear programming. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 233–243, Jeju Island, Korea, July. Association for Computational Linguistics.