Taking into account Inter-sentence Similarity for Update Summarization

Maâli Mnasri
CEA, LIST,
Univ. Paris-Sud,
Université Paris-Saclay.
maali.mnasri@cea.fr

Gaël de Chalendar
CEA, LIST,
Gif-sur-Yvette,
F-91191 France.
gael.de-chalendar@cea.fr

Olivier Ferret
CEA, LIST,
Gif-sur-Yvette,
F-91191 France.
olivier.ferret@cea.fr

Abstract

Following Gillick and Favre (2009), a lot of work about extractive summarization has modeled this task by associating two contrary constraints: one aims at maximizing the coverage of the summary with respect to its information content while the other represents its size limit. In this context, the notion of redundancy is only implicitly taken into account. In this article, we extend the framework defined by Gillick and Favre (2009) by examining how and to what extent integrating semantic sentence similarity into an update summarization system can improve its results. We show more precisely the impact of this strategy through evaluations performed on DUC 2007 and TAC 2008 and 2009 datasets.

1 Introduction

As recently shown by Hirao et al. (2017) from a theoretical viewpoint, there is still room for improvements in the extractive approach of Automatic Summarization (AS), which is the framework in which our work takes place. In this context, many methods have been developed for selecting sentences according to various features and aggregating the results of this selection for building summaries. All these methods aim at selecting the most informative sentences and minimizing their redundancy while not exceeding a maximum length.

In this article, we focus on Multi-Document Summarization (MDS) and more particularly on its update variant. Having two or more sets of documents ordered chronologically, the update summarization task consists in summarizing the newer documents under the assumption that a user has already read the older documents. Hence, the work done for tackling this task has mainly extended the work done for MDS by taking into account the notion of novelty through different means. Wan (2012) integrates this notion in the framework of graph-based methods for computing the salience of sentences while Delort and Alfonseca (2012) achieve this integration in the context of hierarchical topic models. Li et al. (2012) go one step further in hierarchical Bayesian models by applying the paradigm of Hierarchical Dirichlet Processes to the update task.

Another interesting way to consider the problem of AS is to formalize it as a constraint optimization problem based on Integer Linear Programming (ILP). ILP-based approaches are very flexible as they allow to jointly optimize several constraints and were found very successful for MDS, as illustrated by the ICSISumm system of Gillick and Favre (2009). They have also been developed for the update summarization task, with work such as (Li et al., 2015) about the weighting of concepts.

However, the most obvious weakness of such methods, particularly the one proposed by Gillick and Favre (2009), is their implicit way of modeling information redundancy. This prevents them from exploiting work about textual entailment or paraphrase detection, which could be especially relevant in the context of MDS. In this article, we aim at extending the update summarization frame-

Proceedings of the The 8th International Joint Conference on Natural Language Processing, pages 204–209,
Taipei, Taiwan, November 27 – December 1, 2017 ©2017 AFNLP

work of Gillick and Favre (2009) by integrating non-redundancy features in a more explicit way through sentences semantic similarities.

2 Summarization framework

For performing MDS with update, the approach we propose includes two main steps. First, we perform a semantic clustering over the input documents sentences, including the old and the new document sets. Then, we select a subset of sentences for the summary while considering the semantic information resulting from the clustering step. We detail each of these two steps hereafter.

2.1 Semantic clustering

As in basic MDS, update MDS aims to optimize the content relevance, the information redundancy within a document set and the final summary length. The additional constraint is to detect information novelty in the new documents in order to avoid repeating what has already been read.

Since emphasizing information novelty in the update summary is equivalent to penalizing old information in the new document set, we should start by identifying sentences from the new set that are equivalent to sentences from the old set. One way to achieve such identification is to perform semantic clustering over all the sentences, whatever their source. The aim of semantic clustering here is to group sentences conveying the same information, even when they are expressed in different ways. In addition to detecting redundancy over time, this filtering step allows to decrease the sentence selection cost by reducing the number of possible combinations of sentences. Furthermore, considering similarity at the sentence level rather than the sub-sentence level is more efficient since the number of sentences is much lower, which ensures less calculations of pairwise similarities.

Clustering method For performing the semantic clustering of sentences, we need a clustering algorithm that uses semantic similarity as a major feature with a low computing time. Partitioning algorithms like *k-means* require the number of clusters to be set in advance, which is inconsistent with our main need. Moreover, setting up a similarity threshold is less dependent on the size of the considered data than setting up the number of clusters. While the latter depends on the test data size and the content diversity, the former depends on the similarity measure itself and could be set up on large annotated corpora. Among the clustering methods relying on a similarity matrix as input, we chose the Markov Cluster Algorithm (MCL), a network-based clustering algorithm simulating the flow in graphs, known to be fast and scalable (van Dongen, 2000). It assumes that ”a random walk on a network that visits a dense cluster will likely not leave it until many of its vertices have been visited”. In our case, it turns the adjacency matrix of sentences into a transition matrix. Since our goal is to build small and tight clusters, we removed pairwise similarities under a given threshold. Finally, as MCL performs hard clustering, each sentence belongs to one cluster only.

Semantic similarity measure Sentence semantic similarity has gained a lot of interest recently, especially in the context of SemEval evaluations (Agirre et al., 2016). However, in practice, most proposed similarity measures for AS are subject to a time efficiency problem which tends to increase with the quality of the similarity measure. This is the case of the lexical word alignement based similarity that won the SemEval 2014 sentence similarity task (Sultan et al., 2014). We found it unusable in our set-up due to its computational complexity as we calculate about 5 million sentence pair similarities for some datasets while the SemEval 2014 corpus, for instance, gathers only 3,750 sentence pairs. Given this constraint, we chose a similarity measure relying on low dimensional word vectors from word embeddings. In fact, simply averaging word embeddings of all words in a sentence has proven to produce a sentence vector encoding its meaning and has shown a good performance in multiple tasks and particularly in text similarity tasks (Das et al., 2016; White et al., 2015; Gershman and Tenenbaum, 2015; Hill et al., 2016). We adopted this method to represent sentences and used only the embeddings of unigrams since bigrams and phrases are generally not well covered by the existing pre-trained embeddings[1]. Before building the sentence vectors, we did not perform any normalization of the words in documents (unigrams) as words in pre-trained embeddings are not normalized. Finally, we classically defined the similarity of two sentences as the cosine similarity of their vectors.

[1] Only 0.08% of TAC 2008 dataset bigrams are covered by the Glove840B embeddings.

2.2 Sentence selection

Extracting one sentence per cluster to generate summaries as in (Zopf et al., 2016) leads to poor results for update MDS. We have rather added the information brought by our semantic clustering to an ILP model for selecting sentences. This model is the ICSISumm model proposed by Gillick and Favre (2009). It is a maximum coverage model operating at the concept level where concepts are word bigrams. The score of a summary is the sum of its bigram weights. Each bigram is credited only once in the summary score to favor diversity at the lexical level. Thus, redundancy is globally and implicitly minimized. To address the update task, the value of concepts appearing in first sentences are up-weighted by a factor of 3 as in (Gillick and Favre, 2009). The ILP problem is formalized as follows:

$$\text{Maximize}: \sum_i w_i.c_i$$

$$\text{Subject to}: \sum_j s_j.l_j \leq L \tag{1}$$

$$s_j.Occ_{ij} \leq c_i, \forall i, j \tag{2}$$

$$\sum_j s_j.Occ_{ij} \geq c_i, \forall i \tag{3}$$

$$c_i \in \{0,1\} \, \forall i \text{ and } s_j \in \{0,1\} \, \forall j$$

where c_i is a variable indicating the presence of the concept i in the summary; w_i refers to the weight of the concept i, which is its document frequency; l_j represents the length of the sentence j while L is a constant representing the summary maximum length; s_j is the variable that indicates the presence of the sentence j in the summary and finally, Occ_{ij} is a constant parameter indicating the presence of concept i in sentence j. While the constraint (1) prevents the whole summary from exceeding the maximum length limit, constraints (2) and (3) ensure its consistency.

To take into account the semantic clustering of sentences in the ILP model, we focused on the weighting of bigrams since in such models, the concept weighting method is a key factor in the performance of the system (Li et al., 2015). As our aim is to reduce redundancy with the old information, we chose to penalize the weights of bigrams occurring in both old and new documents. If a bigram appears in a cluster including sentences from both old and new document sets, its weight is penalized by an α parameter as follows: $w'_i = \frac{w_i}{\alpha}$. The value of α is determined on a development set. As in (Gillick and Favre, 2009), bigrams whose weights are lower than a fixed threshold are pruned before solving the ILP problem for computational efficiency. However, since this pruning can have a negative impact results if it is too restrictive (Boudin et al., 2015), we carried out the penalization process after the bigram pruning step.

3 Experiments

3.1 Evaluation Setup

For our experiments, we used the DUC 2007 update corpus and TAC 2008 and 2009 update corpora. The 3 datasets are composed respectively of 10, 48 and 44 topics. They gather respectively about 25, 20 and 20 news articles per topic. The articles are ordered chronologically and partitioned into 3 sets, A to C, for DUC 2007 and two sets, A to B, for both TAC 2008 and 2009. We only considered sets A and B for all the datasets.

To evaluate our approach, we classically adopted the ROUGE[2] framework (Lin, 2004), which estimates a summary score by its n-gram overlap with several reference summaries (Rn). Although our method is unsupervised, we had to tune two parameters: the similarity threshold in the clustering step (for sparcifying the input similarity matrix) and the penalization factor α in the sentence selection. As training data, we used for each dataset the two other datasets. To set up these parameters, we followed a greedy sequential approach for optimizing ROUGE on each training set. We maximized the ROUGE-2 recall score (bigrams overlap) particularly since it has shown the best agreement with manual evaluations (Hong et al., 2014). Yet, we report in what follows three variants of ROUGE: ROUGE-1, which computes the overlap with reference summaries in terms of unigrams, ROUGE-2, described previously and ROUGE-SU4, which computes the overlap of skip-grams with a skip distance of 4 words at most. Again following (Hong et al., 2014), we only report the recall values of the ROUGE metrics because their precision and f-measure values are very close to them.

[2] ROUGE parameters: -n 2 -2 4 -m -l 100 -u -c 95 -p 0.5 -f A -r 1000

System/dataset	DUC 2007			TAC 2008			TAC 2009		
	R1	R2	RSU4	R1	R2	RSU4	R1	R2	R4
Baseline systems									
ICSISumm 2009	33.73	7.59	11.23	38.28	11.19	14.46	37.40	10.37	13.86
ICSISumm-BG-DOWN-1	34.46	7.91	11.74	36.99	10.15	13.66	37.39	10.25	13.87
ICSISumm-BG-DOWN-2	33.71	7.55	11.22	38.02	11.05	14.18	37.27	9.91	13.62
State-of-the-art systems									
Supervised ILP	-	-	-	-	9.99	13.61	-	9.61	13.77
Topic Modeling	-	-	-	36,73	10.41	13.79	36.42	9.58	13.53
CorrRank	-	-	-	36.71	9.70	13.19	36.87	9.73	13.59
Proposed systems									
MCL-W2V-ICSISumm	34.99	8.14	11.79	38.52	11.49	14.68	37.50	10.48	13.98
MCL-GLOVE-ICSISumm	**36.08**	**9.46**	**12.96**	**38.62**	**11.57**	**14.75**	**37.53**	**10.60**	**14.08**
MCL-ConceptNet-ICSISumm	35.23	8.30	11.98	38.28	11.21	14.49	**37.53**	10.38	13.91

Table 1: Average ROUGE recall scores on DUC 2007, TAC 2008 and TAC 2009 datasets

3.2 Results

We compare our system to three baselines and three high-level state-of-the-art references.

Baseline systems

- ICSISumm 2009. This is the system described in Section 2.2, on which we built our contribution. We report here the version with no sentence compression. It is worth noting that ICSISumm was still found the best performing system in (Hong et al., 2014).
- ICSISumm-BG-DOWN-1. This baseline is an adaptation of the ICSISumm 2009 system in which we down-weight the bigrams occurring in the chronologically first set of documents (A).
- ICSISumm-BG-DOWN-2. In this modified version of the ICSISumm 2009 system, we down-weight the bigrams whose frequency in the chronologically first set of documents (A) is greater than their frequency in the more recent document set (B).

The last two baselines, which do not include our semantic clustering of sentences, are tested to check how effective is this clustering and to what extent it is needed.

State-of-the-art systems

- Topic Modeling. This system uses topic probability distributions for salience determination and a dynamic modeling approach for redundancy control (Wang and Zhou, 2012).
- CorrRank. This algorithm selects sentences using a topic evolution pattern by filtering sentences from particular topic evolution categories (Lei and Yanxiang, 2010).
- Supervised ILP. This system predicts the bigrams weights by means of a supervised model using salience and novelty features at the sentence and bigram level. Sentence selection is done by an ILP model and a regression model for sentence reranking (Li et al., 2015).

Proposed systems We present the results of our system with different pre-trained word embeddings for evaluating sentence similarities. All the considered training sets showed that the optimal performance is reached when the penalization factor α is set to 3. No similarity threshold is set lower than 0.95, which guarantees a precision level for the similarity measure at least equal to the precisions reported in Table 2.

- MCL-W2V-ICSISumm. This version relies on 3 million vectors (300 dimensions) trained with the CBOW model of Word2Vec (Mikolov et al., 2013) on 100 billion words from a Google News dataset.
- MCL-GLOVE-ICSISumm. In this run, we used 2.2 million word vectors (300 dimensions) trained with GloVe (Pennington et al., 2014) on the 840 billion tokens from the Common Crawl repository.
- MCL-ConceptNet-ICSISumm. This version computes similarities with the ConceptNet

Dataset	Precision	Recall
MSRpara	91.44	17.69
SemEvaL STS 2014	88.00	14.17
SemEvaL STS 2015	90.60	11.46
SemEvaL STS 2016	88.28	25.98

Table 2: Results of our sentence semantic similarity with the minimum threshold value of 0.95

Vector Ensemble embeddings (Speer and Chin, 2016), which is a combination of GloVe and Word2Vec embeddings enriched with knowledge from the ConceptNet semantic network and PPDB.

Table 1 shows that for all the different word embeddings, our system outperforms all our references. The improvement is observed for the three ROUGE measures we used. The improvement over ICSISumm 2009, which has the same settings as our system, confirms the interest of handling redundancy explicitly in the update summarization task. However, the improvement over ICSISumm-BG-DOWN-1&2 also shows that basic methods for performing this handling are not efficient in ILP models, contrary to our sentence semantic clustering approach. Our second version using Glove pre-trained vectors reports higher results than those using Word2Vec or ConceptNet Ensemble word vectors. This could be explained by the size of the training sets for building the word vectors as the Common Crawl dataset is much larger than the Google News dataset. Moreover, the impact of the quality of the vectors on our results indirectly confirms the interest of our proposal.

Since the semantic similarity of sentences is central in our approach, we have tried to characterize *a posteriori* the similarity corresponding to the value of our similarity threshold as it was optimized on our development sets. We have applied our semantic similarity measure to reference evaluation datasets for sentence similarity[3]: the MSR Paraphrase Corpus (Dolan et al., 2004) and the SemEval STS datasets (Agirre et al., 2016). In order to calculate precision and recall scores on the SemEval datasets, we consider a result as a true positive if our similarity is higher than 0.95 and the gold standard similarity is higher than 3[4]. We present in Table 2, the evaluation of our similarity measure using the Google's pre-trained word vectors. On all datasets, our similarity shows a high precision but a weak recall. This trend is particularly noticeable on the MSR Paraphrase Corpus: when our system regroups two sentences, they are paraphrases in 91.44% of the cases, which fits our initial hypotheses and illustrates their validity.

4 Conclusion and Perspectives

For concluding, we showed that taking into account the semantic similarity of sentences for discarding redundancy in a maximal bigram coverage problem improves the update summarization performance and can be done by modifying the weights of bigrams in an ILP model according to the results of the semantic clustering of sentences.

The most direct perspective we will follow for extending this work is to improve the recall of the semantic similarity measure to increase the ability of our system to detect redundancy. In a more global extension, we will associate this criterion about redundancy with criteria more focused on information salience based on the discourse structure of documents.

References

Eneko Agirre, Carmen Banea, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Rada Mihalcea, German Rigau, and Janyce Wiebe. 2016. SemEval-2016 Task 1: Semantic Textual Similarity, Monolingual and Cross-Lingual Evaluation. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 497–511, San Diego, California.

Florian Boudin, Hugo Mougard, and Benot Favre. 2015. Concept-based Summarization using Integer Linear Programming: From Concept Pruning to Multiple Optimal Solutions. In *2015 Conference on Empirical Methods in Natural Language Processing (EMNLP 2015)*, pages 1914–1918, Lisbon, Portugal.

Arpita Das, Harish Yenala, Manoj Chinnakotla, and Manish Shrivastava. 2016. Together we stand: Siamese Networks for Similar Question Retrieval. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL 2016)*, pages 378–387, Berlin, Germany.

Jean-Yves Delort and Enrique Alfonseca. 2012. DualSum: a Topic-Model based approach for update

[3] To our knowledge, the only dataset specifically dedicated to the evaluation of sentence clustering in the context of MDS is described in (Geiss, 2009) but it is not publicly available.

[4] Starting from 3, two sentences are considered as equivalent in the SemEval gold standard.

summarization. In *13th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2012)*, pages 214–223, Avignon, France.

Bill Dolan, Chris Quirk, and Chris Brockett. 2004. Unsupervised construction of large paraphrase corpora: exploiting massively parallel news sources. In *Proceedings of the 20th international conference on Computational Linguistics (COLING 2004)*, pages 350–356.

Stijn van Dongen. 2000. *Graph Clustering by Flow Simulation*. Ph.D. thesis, University of Utrecht.

Johanna Geiss. 2009. Creating a Gold Standard for Sentence Clustering in Multi-Document Summarization. In *47th Annual Meeting of the Association for Computational Linguistics and 4th International Joint Conference on Natural Language Processing of the AFNLP (ACL-IJCNLP 2009), Student Research Workshop*, pages 96–104, Singapore.

S.J. Gershman and J.B. Tenenbaum. 2015. Phrase similarity in humans and machines. In *Proceedings of the 37th Annual Conference of the Cognitive Science Society*.

Dan Gillick and Benoit Favre. 2009. A Scalable Global Model for Summarization. In *Workshop on Integer Linear Programming for Natural Language Processing (ILP'09)*, pages 10–18, Boulder, Colorado.

Felix Hill, Kyunghyun Cho, and Anna Korhonen. 2016. Learning Distributed Representations of Sentences from Unlabelled Data. In *2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (HLT-NAACL 2016)*, pages 1367–1377, San Diego, California.

Tsutomu Hirao, Masaaki Nishino, Jun Suzuki, and Masaaki Nagata. 2017. Enumeration of Extractive Oracle Summaries. In *15th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2017)*, pages 386–396, Valencia, Spain.

Kai Hong, John Conroy, Benoit Favre, Alex Kulesza, Hui Lin, and Ani Nenkova. 2014. A Repository of State of the Art and Competitive Baseline Summaries for Generic News Summarization. In *Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1608–1616, Reykjavik, Iceland. ELRA.

Huang Lei and He Yanxiang. 2010. CorrRank: Update Summarization Based on Topic Correlation Analysis. In *Advanced Intelligent Computing Theories and Applications. With Aspects of Artificial Intelligence. 6th International Conference on Intelligent Computing (ICIC 2010)*, pages 641–648, Changsha, China.

Chen Li, Yang Liu, and Lin Zhao. 2015. Improving Update Summarization via Supervised ILP and Sentence Reranking. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1317–1322, Denver, Colorado.

Jiwei Li, Sujian Li, Xun Wang, Ye Tian, and Baobao Chang. 2012. Update Summarization using a Multilevel Hierarchical Dirichlet Process Model. In *COLING 2012*, pages 1603–1618, Mumbai, India.

Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *ACL-04 Workshop Text Summarization Branches Out*, pages 74–81, Barcelona, Spain.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *International Conference on Learning Representations 2013 (ICLR 2013), workshop track.*

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global Vectors for Word Representation. In *2014 Empirical Methods in Natural Language Processing (EMNLP 2014)*, pages 1532–1543.

Robert Speer and Joshua Chin. 2016. An Ensemble Method to Produce High-Quality Word Embeddings. *CoRR*, abs/1604.01692.

Md Arafat Sultan, Steven Bethard, and Tamara Sumner. 2014. DLS@CU: Sentence Similarity from Word Alignment. In *8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 241–246, Dublin, Ireland.

Xiaojun Wan. 2012. Update Summarization Based on Co-Ranking with Constraints. In *COLING 2012*, pages 1291–1300, Mumbai, India.

Hongling Wang and Guodong Zhou. 2012. Toward a Unified Framework for Standard and Update Multi-Document Summarization. *ACM Transactions on Asian Language Information Processing (TALIP)*, 11(2):1–18.

Lyndon White, Roberto Togneri, Wei Liu, and Mohammed Bennamoun. 2015. How Well Sentence Embeddings Capture Meaning. In *Proceedings of the 20th Australasian Document Computing Symposium (ADCS'15)*, pages 1–8, New York, NY, USA. ACM.

Markus Zopf, Eneldo Loza Mencía, and Johannes Fürnkranz. 2016. Sequential Clustering and Contextual Importance Measures for Incremental Update Summarization. In *26th International Conference on Computational Linguistics (COLING 2016)*, pages 1071–1082.