Automatic slides generation in the absence of training data

Automatic slides generation in the absence of training data

Luca Cagliero
Dipartimento di Automatica e Informatica
Politecnico di Torino
Corso Duca degli Abruzzi, 24
10129 Turin, Italy
0000-0002-7185-5247

Moreno La Quatra
Dipartimento di Automatica e Informatica
Politecnico di Torino
Corso Duca degli Abruzzi, 24
10129 Turin, Italy
0000-0001-8838-064X

***Abstract*—Disseminating the main research findings is one of the main requirements to become a successful researcher. Presentation slides are the most common way to present paper content. To support researchers in slide preparation, the NLP research community has explored the use of summarization techniques to automatically generate a draft of the slides consisting of the most salient sentences or phrases. State-of-the-art methods adopt a supervised approach, which first estimates global content relevance using a set of training papers and slides, then performs content selection by optimizing also section-level coverage. However, in several domains and contexts there is a lack of training data, which hinders the use of supervised models. This paper focuses on addressing the above issue by applying unsupervised summarization methods. They are exploited to generate sentence-level summaries of the paper sections, which are then refined by applying an optimization step. Furthermore, it evaluates the quality of the output slides by taking into account the original paper structure as well. The results, achieved on a benchmark collection of papers and slides, show that unsupervised models performed better than supervised ones on specific paper facets, whereas they were competitive in terms of overall quality score.**

***Index Terms*—Automatic slides generation, unsupervised text summarization, sentence extraction**

I. Introduction

Dissemination is crucial for sharing theoretical and applied research findings [1]. The most common way to disseminate knowledge to the research community is to attend conferences, workshops, and seminars and present the latest ideas and achievements. However, generating presentation slides of academic papers is known to be a time-consuming task activity. Hence, to present their work at conferences or workshops, authors commonly pick, on first approximation, slide content from the original paper. The first draft of the slides is commonly aligned to the original sequence of paper sections (e.g., introduction, method, experimental results, and conclusions). Next, authors refine the presentation to make it more effective and attractive by adding enumerated lists, images, tables, animations, and videos. This paper addresses the automatic generation of a first draft of the presentation slides including textual content only. Unlike traditional methods (e.g., [2], [3]) input documents are neither manually annotated nor aligned with the slide content.

The aforesaid problem can be reformulated as an extractive text summarization task. Extractive summarization entails selecting part of the existing content of a document corpus to compose its summarized version [4]. In recent years, it has been largely investigated by the Natural Language Processing community since it applies to a wide range of application contexts, among which timeline extraction [5], papers highlight extraction [6], and learning analytics [7]. Most of the existing summarization methods aim at generating a summary that consist of a shortlist of the most significant sentences or phrases. Sentence/phrase significance is usually evaluated in terms of *relevance* and *conciseness*. Relevance guarantees that the summarization process preserves the most informative content, whereas conciseness quantifies the redundancy of the extracted content.

A draft of presentation slides may consist of a summary of the corresponding paper content. State-of-the-art approaches to automatic slides generation (e.g., [8], [9]) adopt a supervised two-step process: firstly, to each paper sentence a global relevance score is assigned. The aforesaid task is accomplished by training a regression model on a set of training papers and slides. The model predicts sentence relevance scores based on the similarity between the considered sentence and those occurring in the training set. Then, an optimization step is applied to select the most appropriate slides content. The evaluation process analyze sentence relevance and conciseness at both paper- and section-levels. The latter task is addressed using Integer Linear Programming (ILP) solvers.

State-of-the-art slides generation methods have the following drawbacks: (A) Since they mostly rely on supervised techniques (e.g., [8], [9]), they require a sufficiently large set of papers annotated with presentation slides. (B) The outcomes of the slides generation process are evaluated using standard summary evaluation metrics, based on n-gram comparison [10]. However, since they are commonly applied to the whole paper, the evaluation process disregards the underlying paper structure.

In the absence of training data (issue A), existing approaches are not applicable, unless trusting papers and slides in related domains or similar contexts. This hinders the use of state-of-the-art solutions in many real scenarios and prompts the need

for new, unsupervised methods.

When the output slides need to be focused on specific discourse facets such as *Introduction, Method, Results, or Discussion* [11] (issue B), researchers would be more interested in getting high-quality outcomes on specific slides. In the latter scenario, it would be advisable to assess slides generator performance not only *globally* but also *locally* within facet.

To tackle the above issues, this paper proposes to integrate *unsupervised* summarization methods into the automatic slides generation pipeline and to evaluate the corresponding outcomes both globally and locally within facet, according to the IMRaD classification of the scientific paper structure [11]. Since many unsupervised summarization methods inherently provide a ranked shortlist of sentences/phrases, we explore two pipeline variants: (i) the standard one (hereafter denoted as *ILP-based*), which relies on a cascade of summarization and optimization steps. (ii) A simplified version (hereafter denoted as *summarization-only*), which neglects the optimization step and exclusively relies on the content selection and ranking produced by the summarizer. The idea behind is to also investigate the impact of the optimization phase, since most summarizers inherently provide a sentence rank.

We tested the proposed approach on a benchmark collection of scientific papers and presentation slides. The achieved results, evaluated in terms of global quality of the slide content, reflect the expectation: supervised methods are averagely superior to unsupervised ones, even if the performance differences were not always statistically significant. Furthermore, the best performing unsupervised ILP-based strategy outperformed all the corresponding summarization-only versions.

Conversely, while focusing on specific paper facets things did not always go as we expected. Specifically, for specific sections (e.g., *Introduction*) some unsupervised methods performed significantly better then all the supervised ones. This evidences that the training phase was not able to capture the most salient trends in that particular part of the paper structure. Therefore, to generate slides tailored to specific discourse facets the use of unsupervised methods is particularly appealing.

The paper is organized as follows. Section II describes the original slides generation pipeline. Section III details the pipeline variants presented in this work. Section IV reports the standard evaluation metrics and presents the newly proposed faceted version. Finally, Sections V and VI summarize the main results and draw conclusions of this work, respectively.

II. The slides generation pipeline

The slide generation task entails automatically generating presentation slides for scientific papers. Presentation slides are useful, for instance, for a presenter who has to give a talk on her/his most recent research findings. The idea behind is to propose an automated, data-driven strategy to generate a draft version of the slides. The generated slides will include textual content solely, i.e., handling multimedia content is out of the scope of the current work.

A. Preliminaries

The draft version of the presentation consists of a set of slides SL whose content is aligned to that of the main paper sections SE. More specifically, the content of each section $se_x \in SE$ in the paper is assumed to be aligned to k presentation slides $sl^x_y, sl^x_{y+1}, \ldots, sl^x_{y+k-1} \in SL$ ($k \geq 1$), where k may vary from one section to another.

Each slide sl^x_y consists of (i) a title, (ii) a list of key phrases that summarize the most relevant topics covered by the section, and (iii) a separate list of bullet points per key phrase, which provide more insights into the corresponding topic.

B. Pipeline description

The standard pipeline for automatic slides generation [8], [9] is depicted in Figure 1. It consists of six key steps, which are briefly summarized below.

a) Preprocessing: This step focuses on preparing the training set of scientific papers and presentation slides on top of which the machine learning-based slide generation process is executed. The training set consists of a set of paper-slides pairs whose content ranges over the same domain of the *target paper* (i.e., the paper whose presentation slides are currently missing and need to be created).

Since the textual content of papers and slides is usually not promptly usable, a preliminary text extraction phase is performed. It requires properly handling documents of various document formats while preserving text sectioning. Similarly, the textual content of the slides is extracted by separating the titles from the remaining content.

b) Feature extraction: Based on the hypothesis that sentences occurring in the author-written slides are likely to be representative of the main paper content, we label sentences in each paper of the training corpus with a numerical score that quantifies the similarity between the considered sentence and the author-written slides at the sentence level. Specifically, to enable supervised learning from the prepared data we label each sentence s of the paper with the following relevance score $score(s)$:

$$score(s) = max_{s^*_i \in S_{SL}} \Big(sim(s, s^*_i) \Big) \quad (1)$$

where S_{SL} is the set of sentences occurring in the paper slides.

A sentence s is labeled with the maximal similarity score between s and an arbitrary sentence in S_{SL}. Next, each sentence in a paper of the training corpus is described by a set of features described in [8]. The feature set can be also extended by including features extracted by Deep NLP models (see, for example, [9]).

Sentence-level paper descriptions are collected into a labeled dataset, which stores for each sentence the corresponding feature values and label. Then, a supervised regression model is trained on the labeled dataset.

c) Sentence importance evaluation: This step aims at predicting for each sentence s_p its relative importance in the target paper importance based on the information available in the training corpus. Sentence importance is quantified by the score $score_{pr}(s_p)$ returned by the regression model.

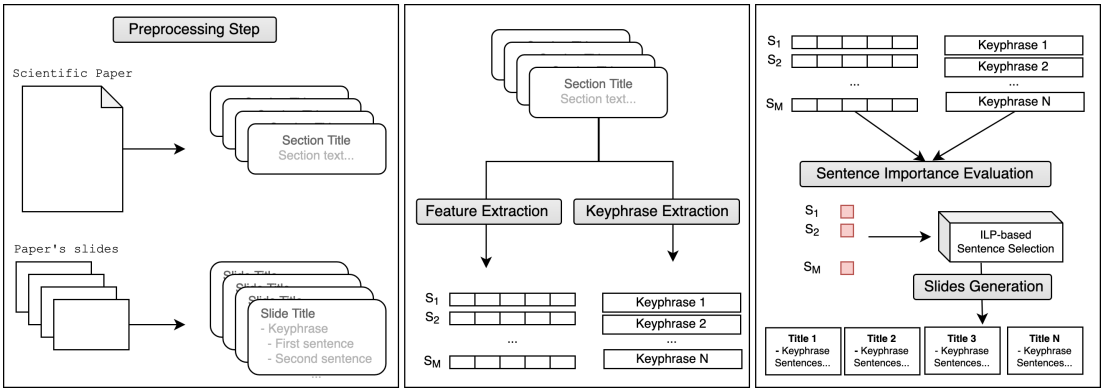

Fig. 1. The standard slide generation pipeline [8].

d) Keyphrase extraction: Key phrases are extracted from the sentences in the target paper using a NLP parsing tool. The goal here is to automate the extraction of the level-1 key phrases from the same sentences that will be used as level-2 bullet points.

e) Sentence selection: This step focuses on shortlisting the sentences and key phrases of the target paper that are worth appearing the presentation slides. The selection criteria can be summarized as follows: (i) maximize the *overall sentence importance* (i.e., the sum of the sentence scores). (ii) maximize *sentence diversity*, which is estimated as the syntactic overlap between the selected paper sentences and the presentation slides. (iii) maximize the relevance of the selected key phrases and their coverage in the level-2 bullet points. In [8] the problem stated above is modelled and solved by using an Integer Linear Programming (ILP) optimizer.

f) Slides generation: The presentation slides are an ensemble of the previously shortlisted sentences. The slides generation procedure entails the following steps:

- Each slide contains at most 4 sentences.
- Sentences that refer to the same keyphrase are grouped together and placed below the same bullet point.
- The bullet point is titled with the reference keyphrase.
- Each slide contains up to 2 bullet points, together with the corresponding sentences.
- The slide title is set by using the titles of the section to which the first sentence in the slide belongs to (truncated to the first 5 token, whenever necessary).

III. Proposed solution

We propose a variant of the standard slide generation pipeline, which can be applied in the absence of a training set of papers and slides. The idea behind is to rely on unsupervised text summarization algorithms [4], which, by construction, do not require a training phase.

Extractive sentence-based summarization entails generating a summary of a document corpus that consists of the most relevant sentences. A huge body of work has been devoted to applying unsupervised methods such as clustering algorithms (e.g., graph-based techniques (e.g., CoreRank [12], LexRank [13], TextRank [14]), Latent Semantic Analysis (e.g., LSARank [15]) and word embedding techniques (e.g., [16]). A recent survey of summarization methods can be found in [4].

The goal here is to explore the applicability of various unsupervised summarization methods for generating presentation slides for academic papers. To this aim, Figure 2 depicts a sketch of the proposed pipeline variants relying on unsupervised text summarization. They will be hereafter denoted as variants (b) and (c), respectively. The main differences with the standard pipeline, denoted by (a), are enumerated below.

1) **Feature extraction**: since most text summarization methods take as input raw text, the feature extraction and labeling phase are no longer required.
2) **Split paper content into section**: In the proposed variants, section-level content splits are instrumental in building separate per-section summaries, which incorporate the most salient information about a specific paper section. Conversely, in the standard pipeline the paper structure is neglected by supervised learning whereas it is re-considered for the subsequent optimization-based content selection step.
3) **Sentence importance evaluation**: In variant (b) the ILP-based sentence evaluation is omitted thus sentence ranking exclusively relies on the sentences' shortlists produced by the summarization process. In variant (c) ILP-based optimization is applied as in the standard pipeline, but the sentence scores used to infer sentence importance and diversity are no longer inferred by a supervised learner, but rather produced by the unsupervised summarization process.
4) **Slide generation**: In variant (b) a greedy strategy similar to those presented in [9] is used in place of the ILP-based selection process. Specifically, summary sentences are picked in order of decreasing importance until a maximum length is reached (10% in our experiments). Variant (c) applies the same strategy used in the standard pipeline.

Point (3) deserves a more detailed explanation. Let S_p be the set of sentences in the target paper. A text summarizer takes S_p as input and produces a summary Sum consisting of a ranked

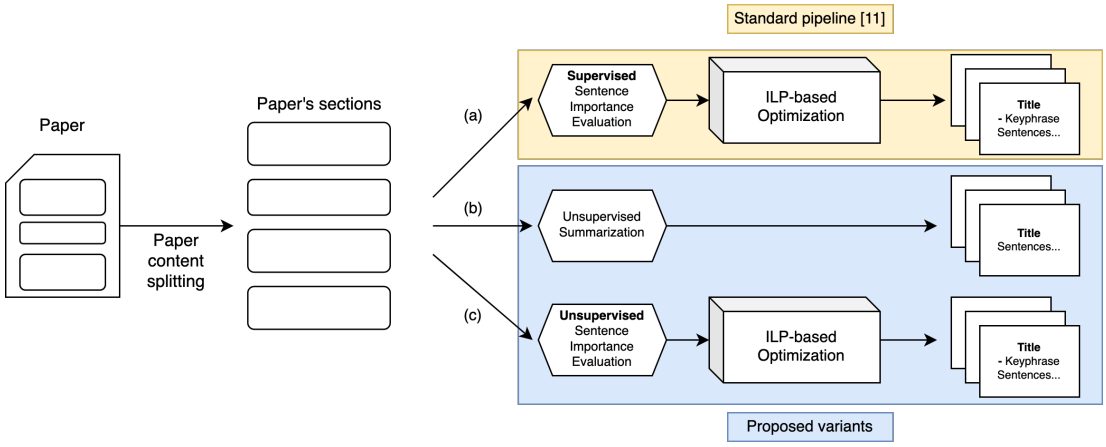

Fig. 2. Proposed variants of the slide generation pipeline.

list of sentences s_p^{1st}, s_p^{2nd}, ..., s_p^{q-th} in S_p. We focus on the subset of text summarizers that inherently provide as output a sentence score $score(\cdot)$ such that $score(s_p^{1st}) \geq score(s_p^{2nd}) \geq \ldots \geq score(s_p^{q-th})$.

The proposed pipeline variants aim at replacing the sentence score in Equation 1 with the score produced by the text summarizers. In principle, every summarization method that is able to produce a ranked list of sentence can be applied to the target paper in place of the supervised learner. Notably, the designed variants do not require a training set.

IV. Evaluation metrics and proposed variants

Previous studies (e.g., [8], [9]) have assessed the quality of the automatically generated slides by using the standard Rouge toolkit, which is established for summary evaluation [10]. It counts the unit overlaps between the two snippets of text. Depending on unit type (e.g., unigrams, bigrams), different Rouge metrics can be selected (e.g., Rouge-1, Rouge-2). the Rouge scores indicate the precision, recall, and F-measure values [17] achieved by the summarization system according to a specific metric.

In the slide generation context, the empirical comparisons carried out in previous studies were focused on evaluating the overlap between the whole content of the automatically generated slides and that of the hand-written ones. However, doing them this way implies neglecting the separate contribution of each paper section. For instance, the quality of the slide content relative to the introductory and experimental sections can be rather different. Currently, such a difference cannot be quantified.

To address the aforesaid issue, we propose to tailor slide evaluation to specific paper facets according to the official IMRAD classification of the scientific paper structure [11]. More specifically, slide content is first classified as *Introduction*, *Methods*, *Results*, or *Discussion* according to the particular section it refers to. Next, separate Rouge scores, hereafter denoted by *Facet-specific* scores, are produced for each IMRAD class. This allows us to differentiate between the slide generation techniques that achieved variable performance on different paper sections due to the peculiar properties that characterize the text of each section.

V. Experimental results

We report here an empirical comparison between the performance of the proposed variants and that of the standard pipeline version. Throughout the evaluation process, we considered as reference metrics the Rouge-1 (R1), Rouge-2 (R2), and Rouge-SU4 (RSU4) F-measure scores, as they are standard for evaluating slides generation performance [8].

A. Paper and slide data collection

For evaluation purposes we exploited a real data collection[1] consisting of 195 academic papers enriched with the corresponding presentation slides. Benchmark data, provided by the project owner upon request, include also the ground-truth slides. Slides were manually created by the respective papers' authors to disseminate their research findings.

B. Tested methods

For the standard pipeline we evaluated the performance of the slide generators based on several (supervised) regression methods. Specifically, we tested Support Vector Machines (SVR), MultiLayer Perceptron (MLP), Gradient Boosting (GB), Decision Tree (DT), and Random Forest (RF). We exploited the algorithm implementations and settings provided by the Scikit-Learn library [18].

To assess the newly proposed pipeline variants, we tested several unsupervised summarization algorithms, which provide as output a sentence ranking. Specifically, Graph-based (TextRank [14], LexRank [13]), LSA-based summarization (LSARank [15], [19]), Embedding-based methods (Centroid-with-BERT [16], using the sentence embedding approach proposed in [20]).

To study the impact of the ILP-based optimization on the performance of variants (b) vs. (c), we tested a larger set of text summarizers, including also the graph-based strategies

[1]The data collection is relative to a Open GitHub project available at https://github.com/hairav/SlideSpawn (latest access: January 2021)

that produce an output summary without explicitly assigning a importance score to each sentence, i.e., CoreRank [12] and TextRankBM25 [21].

C. Overall comparisons between supervised and unsupervised methods

Table I compares the results of the supervised and unsupervised pipelines embedding the ILP-based sentence selection, i.e., variant (a) vs. variant (c). The results were computed in terms of the standard Rouge scores. We recall that they are *independent* of the paper structure.

The performance of the supervised pipeline variant (a) is superior to that of the corresponding unsupervised version. The performance improvements were statistically relevant for four out of five regressors. The best performing approach relies on a Random Forest Regressor and the overall quality scores achieved by the majority of the tested models are roughly comparable with each other.

To study the effect of the optimization step, Table I compares the two variants of the unsupervised pipeline, i.e., variant (b) vs. variant (c). The results show that the ILP-based methods performed better than summarization-only ones. The performance gap is particularly evident for LSA- and embedding-based strategies.

D. Section-level comparisons between supervised and unsupervised methods

We deepen the comparative analysis between supervised and unsupervised strategies separately on each paper facet specified by the IMRaD classification [11].

Table II reports the results of the comparison of the two pipeline variants (a) and (c) (i.e., the ILP-based supervised and unsupervised strategies) by integrating multiple regressors/summarizers.

The results show significant changes from one paper section to another. Specifically, the results on *Method* reflect the overall results discussed earlier, whereas the outcomes achieved on *Introduction* and *Discussion* reverse the situation. In the latter cases, the unsupervised LSA-based approach [19] performed significantly better than most of the supervised models. A possible reason is the potential inability of the NLP features used in the standard pipeline [8] to capture the text correlations within each separate discourse facet.

VI. Conclusions and future works

The paper explores the integration of unsupervised text summarization methods deeply into the pipeline of automatic slides generation for scientific papers. The main goal is to avoid the supervised learning phase, which requires the availability of a sufficiently large corpus of training data, including both papers and slides. It investigates also the influence of paper structure on the quality of the generated slides. To the best of our knowledge, the aforesaid contributions have never been investigated in previous studies.

The results achieved by using the standard evaluators confirmed the expectation: supervised methods have shown to be on average superior to unsupervised techniques. Conversely, while deepening the analysis on specific paper sections (e.g., the introductory and discussion parts) the results were opposite: the unsupervised pipeline that integrates a state-of-the-art text summarizer outperformed all the supervised methods.

The future research agenda will address (1) the study and development of new unsupervised methods based on contextualized embeddings, (2) the training of separate supervised models each one tailored to a different IMRaD class, and (3) The study of innovative multi-modal approaches that would be able to produce slides containing not only text but also images, videos, and audio.

Acknowledgment

The research leading to these results has been funded by the SmartData@PoliTO center for Big Data and Machine Learning technologies.

References

[1] E. Rowley-Jolivet and S. Carter-Thomas, "The rhetoric of conference presentation introductions: context, argument and interaction," *International Journal of Applied Linguistics*, vol. 15, no. 1, pp. 45–70, 2005.

[2] M. Utiyama and K. Hasida, "Automatic slide presentation from semantically annotated documents," in *Coreference and Its Applications@ACL 1999, College Park, Maryland, USA, June 22, 1999*. Association for Computational Linguistics, 1999.

[3] T. Hayama, H. Nanba, and S. Kunifuji, "Alignment between a technical paper and presentation sheets using a hidden markov model," in *Proceedings of the 2005 International Conference on Active Media Technology, AMT 2005, Kagawa International Conference Hall, Takamatsu, Kagawa, Japan, May 19-21, 2005*, H. Tarumi, Y. Li, and T. Yoshida, Eds. IEEE, 2005, pp. 102–106. [Online]. Available: https://doi.org/10.1109/AMT.2005.1505278

[4] W. S. El-Kassas, C. R. Salama, A. A. Rafea, and H. K. Mohamed, "Automatic text summarization: A comprehensive survey," *Expert Systems with Applications*, vol. 165, p. 113679, 2021.

[5] Y. Duan, A. Jatowt, and K. Tanaka, "Discovering typical histories of entities by multi-timeline summarization," in *Proceedings of the 28th ACM Conference on Hypertext and Social Media*, ser. HT '17. New York, NY, USA: Association for Computing Machinery, 2017, p. 105–114. [Online]. Available: https://doi.org/10.1145/3078714.3078725

[6] L. Cagliero and M. L. Quatra, "Extracting highlights of scientific articles: A supervised summarization approach," *Expert Syst. Appl.*, vol. 160, p. 113659, 2020. [Online]. Available: https://doi.org/10.1016/j.eswa.2020.113659

[7] L. Cagliero, L. Farinetti, and E. Baralis, "Recommending personalized summaries of teaching materials," *IEEE Access*, vol. 7, pp. 22 729–22 739, 2019.

[8] Y. Hu and X. Wan, "Ppsgen: Learning-based presentation slides generation for academic papers," *IEEE Transactions on Knowledge and Data Engineering*, vol. 27, no. 4, pp. 1085–1097, 2015.

[9] A. Sefid and J. Wu, "Automatic slide generation for scientific papers," in *Third International Workshop on Capturing Scientific Knowledge co-located with the 10th International Conference on Knowledge Capture (K-CAP 2019), SciKnow@ K-CAP 2019*, 2019.

[10] C.-Y. Lin and E. Hovy, "Automatic evaluation of summaries using n-gram co-occurrence statistics," in *Proceedings of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, 2003, pp. 71–78.

[11] L. B. Sollaci and M. G. Pereira, "The introduction, methods, results, and discussion (imrad) structure: a fifty-year survey," *Journal of the medical library association*, vol. 92, no. 3, p. 364, 2004.

[12] A. Tixier, P. Meladianos, and M. Vazirgiannis, "Combining graph degeneracy and submodularity for unsupervised extractive summarization," in *Proceedings of the workshop on new frontiers in summarization*, 2017, pp. 48–58.

[13] G. Erkan and D. R. Radev, "Lexrank: Graph-based lexical centrality as salience in text summarization," *Journal of Artificial Intelligence Research*, vol. 22, pp. 457–479, 2004.

TABLE I

ROUGE-BASED COMPARISON BETWEEN SUPERVISED / UNSUPERVISED SENTENCE SCORING AND STANDARD SUMMARIZATION METHODOLOGIES. STATISTICALLY SIGNIFICANT VARIATIONS WITH RESPECT TO THE BEST PERFORMING ALGORITHM (P-VALUE < 0.05) ARE STARRED.

Method	**R1-P**	**R1-R**	**R1-F**	**R2-P**	**R2-R**	**R2-F**	**RSU4-P**	**RSU4-R**	**RSU4-F**
Supervised ILP-based approach									
MLP	0.281	0.239	0.256*	0.116	0.099	0.106	0.117	0.100	0.107
SVR	0.282	0.240	0.257*	0.115	0.098	0.105	0.118	0.100	0.107
GB	0.283	0.241	0.258	0.114	0.098	0.104	0.118	0.101	0.108
DT	0.274	0.230	0.248*	0.106	0.089	0.096*	0.110	0.092	0.099*
RF	0.288	0.246	**0.263**	0.119	0.102	**0.109**	0.121	0.104	**0.111**
Unsupervised ILP-based approach									
ELSA [19]	0.265	0.224	0.241*	0.097	0.083	0.089*	0.105	0.089	0.096*
Centroid	0.261	0.214	0.233*	0.086	0.072	0.078*	0.097	0.080	0.087*
LexRank	0.261	0.214	0.233*	0.086	0.072	0.078*	0.098	0.080	0.087*
TextRank	0.260	0.213	0.232*	0.086	0.072	0.078*	0.097	0.080	0.087*
LSARank	0.246	0.207	0.223*	0.071	0.060	0.065*	0.085	0.071	0.077*
Unsupervised standard (summarization-only) approach									
ELSA [19]	0.241	0.202	0.218*	0.066	0.056	0.060*	0.083	0.070	0.075*
Centroid	0.257	0.212	0.230*	0.077	0.064	0.070*	0.092	0.076	0.082*
LexRank	0.248	0.197	0.218*	0.066	0.053	0.058*	0.084	0.066	0.073*
TextRank	0.240	0.197	0.214*	0.068	0.055	0.060*	0.083	0.068	0.074*
TextRank BM25	0.253	0.207	0.226*	0.069	0.057	0.062*	0.087	0.071	0.078*
LSARank	0.242	0.205	0.220*	0.070	0.060	0.064*	0.083	0.070	0.075*
CoreRank	0.254	0.218	0.233*	0.079	0.069	0.073*	0.092	0.079	0.084*

TABLE II

ROUGE F1-SCORE COMPARISON BETWEEN SUPERVISED AND UNSUPERVISED METHODS. STATISTICALLY SIGNIFICANT VARIATIONS WITH RESPECT TO THE BEST PERFORMING ALGORITHM (P-VALUE < 0.05) ARE STARRED.

	Introduction			Method			Results			Discussion		
Approach	R1-F	R2-F	RSU4-F	R1-F	R2-F	RSU4-F	R1-F	R2-F	RSU4-F	R1-F	R2-F	RSU4-F
Supervised ILP-based approach												
MLP	0.035*	0.000*	0.007*	0.035*	0.001	0.006	**0.044**	**0.001**	**0.008**	0.053	**0.003**	0.011
SVR	0.037*	0.003	0.007	0.038*	**0.002**	0.007	**0.044**	**0.001**	**0.008**	0.048*	0.002	0.009*
GB	0.042	0.003	0.008	0.040	0.001	**0.008**	**0.044**	**0.001**	**0.008**	0.053	**0.003**	0.011
DT	0.037*	0.003	0.007*	**0.045**	0.001	**0.008**	**0.044**	**0.001**	**0.008**	0.051*	0.002	0.010
RF	0.032*	0.003	0.006*	0.038*	**0.002**	0.007	0.043	0.000	**0.008**	0.046*	0.002	0.009*
Unsupervised ILP-based approach												
ELSA [19]	**0.045**	0.003	**0.009**	0.041	0.001	0.008	0.041	0.000	0.007	**0.06**	**0.003**	**0.012**
Centroid	0.026*	0.000*	0.005*	0.032*	0.000*	0.006	**0.044**	0.000	**0.008**	0.052*	**0.003**	0.010
LexRank	0.026*	0.000*	0.005*	0.032*	0.000*	0.006	**0.044**	0.000	**0.008**	0.052*	**0.003**	0.010
TextRank	0.026*	0.000*	0.005*	0.032*	0.000*	0.006	**0.044**	0.000	**0.008**	0.052*	**0.003**	0.010
LSARank	0.042	0.003	0.008	0.036*	0.001	0.006	0.024*	0.000*	0.004*	0.059	**0.003**	**0.012**
Unsupervised summarization-only approach												
ELSA [19]	0.033*	0.001	0.006*	0.026*	0.000*	0.005*	0.032*	0.000	0.006*	0.033*	0.001*	0.006*
Centroid	0.041	**0.004**	0.008	0.020*	0.000*	0.004*	0.026*	0.000*	0.005*	0.040*	0.001	0.008*
CoreRank	0.041	0.003	0.008	0.025*	0.000*	0.005*	0.030*	0.000	0.006*	0.028*	0.000*	0.005*
TextRank BM25	0.036*	0.001*	0.006*	0.023*	0.000*	0.004*	0.029*	0.000	0.005*	0.043*	0.002	0.008*
LexRank	**0.045**	0.001*	**0.009**	0.024*	0.000*	0.004*	**0.044**	**0.001**	**0.008**	0.054	0.000*	0.010*
TextRank	0.029*	0.001*	0.005*	0.025*	0.000*	0.005*	**0.044**	**0.001**	**0.008**	0.029*	0.000*	0.005*
LSARank	0.026*	0.001*	0.005*	0.026*	0.000*	0.005*	0.036*	0.000	0.007*	0.032*	0.001*	0.006*

[14] R. Mihalcea and P. Tarau, "Textrank: Bringing order into text," in *Proceedings of the 2004 conference on empirical methods in natural language processing*, 2004, pp. 404–411.

[15] J. Steinberger and K. Ježek, "Using latent semantic analysis in text summarization and summary evaluation," in *In Proc. ISIM'04*, 2004.

[16] S. Lamsiyah, A. El Mahdaouy, B. Espinasse, and S. El Alaoui Ouatik, "An unsupervised method for extractive multi-document summarization based on centroid approach and sentence embeddings," *Expert Systems with Applications*, p. 114152, 2020.

[17] A. Rajaraman and J. D. Ullman, *Mining of Massive Datasets*. New York, NY, USA: Cambridge University Press, 2011.

[18] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[19] L. Cagliero, P. Garza, and E. Baralis, "Elsa: A multilingual document summarization algorithm based on frequent itemsets and latent semantic analysis," *ACM Transactions on Information Systems (TOIS)*, vol. 37, no. 2, pp. 1–33, 2019.

[20] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019, pp. 3973–3983.

[21] F. Barrios, F. López, L. Argerich, and R. Wachenchauzer, "Variations of the similarity function of textrank for automated summarization," *arXiv preprint arXiv:1602.03606*, 2016.