Proceedings of the Workshop on Hybrid Intelligence for Natural Language Processing Tasks HI4NLP (co-located at ECAI-2020)
Santiago de Compostela, August 29, 2020, published at http://ceur-ws.org

A new approach for extracting the conceptual schema of texts based on the linguistic Thematic Progression theory

Elena del Olmo[1] and **Ana María Fernández-Pampillón**[2]

Abstract. The purpose of this article is to present a new approach for the discovery and labelling of the implicit conceptual schema of texts through the application of the Thematic Progression theory. The underlying conceptual schema is the core component for the generation of summaries that are genuinely consistent with the semantics of the text.

1. INTRODUCTION

Automatic Summary Generation was first proposed in the late 1950s. Outstanding examples of this early stage are Luhn (1958), whose method is based on sentence extraction relying on its words weightings, inferred from *TF-IDF* metrics, or Edmundson (1969), who proposed novel sentence weighting metrics, such as the presence of words from a predefined list, the presence of the words of the title of the document or its positioning at the beginning of documents and paragraphs. These are paradigmatic examples of the first extractive summarization techniques: techniques based on the verbatim extraction of the most relevant parts of a text. The generated text summary was, thus, a collection of sentences considered relevant but, often, semantically inconsistent because of the overall weakness in coherence (the text does not make overall sense) and cohesion (the sentences are connected incorrectly). The summary generated was consequently a poorly connected text with no global meaning, presumably due to the assumption of independence of the extracted sentences (Lloret *et al.* 2012).

Currently, five main approaches to extractive techniques can be distinguished: (i), statistical approaches (Luhn 1958, McCargar 2004, Galley 2006), based on different strategies for term counting, (ii), *topic-based* approaches (Edmundson 1969, Harabagiu *et al.* 2005), which assume that several topics are implicit in a text and attempt to formally represent those topics, (iii), graph-based approaches (Erkan *et al.* 2004, Giannakopoulos *et al*. 2008), based on the representation of the linguistic elements in texts judged to be relevant as nodes connected by arcs, (iv), discourse-based approaches (Marcu 2000, Cristea et al. 2005, da Cunha et al. 2007), whose target is to capture the discursive relations within texts, and, (v), machine learning approaches (Aliguliyev 2010, Hannah *et al.* 2014), intended to reduce the text summarization task to a classification task by assigning a relevance value to each sentence.

Although historically less addressed in the literature, abstractive models try to address the lack of coherence and cohesion in the summaries produced, using some source of semantic internal representation of the text (which can be merely the output of an extractive process) to generate the ultimate summary, composed of sentences not necessarily included in the original text. Although this approaches theoretically improve the consistency issue, they introduce a new complexity layer: a natural language generator module. Despite this greater complexity, nowadays text summarization research is progressively shifting towards abstractive approaches (Lin *et al.* 2019).

Traditionally, abstractive summarization techniques have been classified into *structure-based,* intended to populate predefined information structures out of the relevant sentences of the texts, and *semantic-based*, involving a wide variety of knowledge representation techniques. Regarding the former, depending on the structural schema chosen, it is possible to identify, (i), tree-based models (Kikuchi *at. al.* 2014), which perform different strategies for syntactic parsing analysis in order to codify paraphrasing information mainly by linking and reducing the syntactic sentence trees of the text, (ii), template-oriented models (Elhadad *et al.* 2015, Wu *et al.* 2018, Wang *et al.* 2019), which rely on extraction rules led by linguistic patterns matching sequences of tokens to be mapped into predefined templates, (iii), ontology-based models (Nguyen 2009, Baralis *et al.* 2013), which are highly domain-dependent and include a hierarchical classifier mapping concepts into the nodes of an ontology, and, (iv), *rule-based* models (Genest *et al.* 2011), based on extraction rules operating on categories and features representative of the content of the text. Regarding the latter, semantic-based techniques for abstractive summarization, there are interesting approaches based on the concept of *information item* (Gatt *et al.* 2009), the smallest units with internal coherence, in the format of subject-verb-object triplets obtained through semantic role labeling, disambiguation, coreference resolution and the formalization of predication. Besides, there are approaches based on discourse information (Gerani *et al.* 2014, Goyal *et al.* 2016), *predicate-argument* tuples (Li 2015, Zhang *et al.* 2016) and semantic graphs (Liu *et al.* 2019).

The aforementioned tendency towards abstractive approaches in recent years is framed at a stage when the new Deep Learning models have proved to be particularly promising for using vector spaces as a way to address the shortcomings of discrete symbols as the input for Natural Language Processing tasks, such as tokens or lemmas, which cannot represent the underlying semantics of the concepts involved. This new paradigm has provided techniques for both extractive and abstractive summarization, such as the clustering of sentence and document embeddings, or the generation of correct sentences given a sentence embedding and a language model. Remarkable examples are the contributions of Templeton *et al.* (2018), who compare different methods of sentence embeddings computing their cosine similarity, or Miller et *al.* (2019), who

[1] General Linguistics department, Complutense University of Madrid, Spain, email: elenadelolmo@ucm.es
[2] General Linguistics department, Complutense University of Madrid, Spain, email: apampi@ucm.es

proposed *k-means* clustering to identify sentences closest to the centroid for summary selection.

In addition to the distinction between extractive and abstractive approaches, there is a crucial challenge in automatic summarization which affects them both: the subjectivity of the accuracy scoring of summaries. This implies a new difficulty in the creation of objective gold datasets composed of correct summaries. In this context, unsupervised summary models, such as the one proposed in this paper, which does not require training labelled data, has become particularly relevant. Among the unsupervised approaches we can highlight, (i), approaches which are extensions of word embedding techniques, such as the *n-grams* embeddings (Mikolov et *al.* 2013), or *doc2vec* (Le *et al.* 2014), (ii), the *skip-thought* vectors (Kiros *et al.* 2015), (iii), the *Word Mover's Distance* model (Kusner *et al.* 2015), (iv), *quick-thought* vectors (Logeswaran *et al.* 2018), and, (v), models based on contextual embeddings obtained from transformers, such as SBERT (Reimers *et al.* 2019).

This paper addresses one of the weaknesses of extractive models discussed in the previous section, *i. e.* the lack of coherence in the summaries produced, especially when there are insufficient linguistic datasets in a language for applying machine or Deep Learning methods. In this respect, the solution we propose identifies implicit conceptual schemas from texts using the morpho-syntactic knowledge currently provided by NL analyzers.

The paper is organized as follows: in section 2 we define the hypothesis and objectives of the research work. In section 3 we present a review of the linguistic theories on which we base our solution: the Thematic theory and Thematic Progression theory. In section 4 we present our solution: the application of both theories for the identification of the text conceptual schema. In section 5 we study the feasibility of our solution for the automatic extraction of thematic progression in Spanish, a language with few linguistic datasets for text summarization. Finally, in section 6 we draw the conclusions of this work and present our future research lines.

2. HYPOTHESIS AND OBJECTIVES

Our hypothesis is that applying the Thematic Theory and the Thematic Progression Theory to annotate the discourse features *theme*, *rheme* and their coreferences will allow us to extract *thematic progression schemas*, which represent the implicit conceptual schemas of texts.

Therefore, our aim is obtaining an internal representation of the text informational structure as a formal representation for text summarization. The advantage of this solution is that it can be applied to any language regardless of whether or not there are enough training data for the implementation of machine learning and Deep Learning techniques. In our work we will use Spanish as the language to study the feasibility of the solution. We also hope to contribute to the generation of summaries in Spanish, a task currently performed with moderate efficiency due to the limited availability of linguistic resources.

3. REVIEW OF THEMATIC AND THEMATIC PROGRESSION

3.1. Thematic theory

The thematic theory is framed within the optics of linguistic analysis corresponding to the informational layer. The uses and applications that the authors have been giving to terms such as *theme*, *focus*, *topic* and notions such as *new information* or *emphasis* have been overwhelmingly numerous (Gutiérrez 2000). In accordance with the Thematic theory, in descriptive and narrative texts, which are the ones most typically to summarize, known information, or *theme*, is consensually described to be positioned at the beginning of sentences. By contrast, the phrases containing the informative contribution of the sentence, also known as *rheme*, tend to be located further to the right, ahead in the time of enunciation. This description is consistent with how the acquisition of new knowledge is described at the neurological level, through networking the known with the novel or by altering pre-existing relationships (McCulloch *et al.* 1943).

In order to clarify how we will use these concepts, we present here a series of examples adapted from Gutiérrez (2000: 18) and their corresponding answers:

1. Who joined Luis this morning?
   Ludwig was joined this morning by Peter.
2. When did Pedro join Luis?
   Pedro joined Luis this morning.
3. Who joined Pedro this morning?
   This morning Peter joined Ludwig.

That these statements are different is a standard judgment for any native speaker. Although they share the same representative function, *i. e.* their syntactic and semantic relations do not differ, they show different informative functions. Therefore, in spite of transmitting the same information about the world, they do not *inform* in the same way. Accordingly, the underlying assumption of our proposal is that the *thematic status* of a phrase (like *who*, *when* and *who* in the examples above, respectively) is relevant in terms of the prevalence of the concept involved for the summarization of a document. Their clustering along a document, taking into account the thematic progressions patterns found, as further explained in the next section, is expected to reveal the conceptual schema of the text.

3.2. Thematic Progression theory

Daneš (1974: 114) presents the thematic progression as the choice and arrangement of discourse themes, their concatenation, hierarchy and relation with the topics of texts. Accordingly, he argues that it is possible to infer the informational schema of a text from its *theme-rheme organization*. It is considered that there are three main typologies of thematic progressions: (i), linear progression, in which the rheme of one sentence is the theme of the subsequent sentence, (ii), constant progression, in which a theme is repeated over several sentences, and, (iii), derived progression, in which several topics are derived from the same inferred *hypertheme*. Apart from these three basic types, Daneš (1974: 120) also proposed that the combination of them can lead to thematic progressions of higher levels of abstraction, such as, (iv), the split rheme progression, which consists of the existence of a complex rheme, whose hyponyms and meronyms are themes of the subsequent sentences. Finally, he concludes (1974: 122) that the study of the thematic organization of

a text could be useful for numerous practical applications, among which outstands information retrieval, given the performance achieved nowadays by the tools available for the automatic text analysis.

4. THEMATIC PROGRESSION AS A MODEL FOR SEMANTIC TEXT REPRESENTATION

The usefulness of the thematic or rhematic roles of concepts along texts for automatic text summarization arises from two main facts. On the one hand, the theoretical validation of the concept of thematic progression enjoys consensus among researchers as a relevant description for the semantic structure of texts. On the other hand, although it has been traditionally examined through the optics of the Pragmatics layer, the thematic or rhematic status of a concept is actually embodied in the surface syntactic layer, which is prone to be represented in an easy-to-compute form.

Concerning the correlation between the theme of a sentence and its syntactic structure, which is crucial for its automatic annotation, Halliday (1985) proposed an interesting categorization based on the concept of linguistic markedness. Thus, in SVO languages, such as English or Spanish, for declarative sentences there are unmarked themes, prototypically the syntactic subjects preceding principal verbs, and marked themes, such as circumstantial attachments, complements, or sentences with predicate construction. Examples for the former are the first and second sentence of the examples provided above with *Ludwig* and *Pedro* as unmarked themes respectively, whilst the third sample sentence is an example for the latter, with *this morning* as theme. Thematic equative sentences, such as *What I want is a proper cup of coffee*, would be excluded from this categorization. For interrogative sentences, the unmarked themes are definite verbs for *yes-no* questions, such as *did deliver* in *Did the president deliver a speech*, and interrogative pronouns and similar phrases for non-polar questions, such as *where* in *Where is Berlin located?*, whilst marked themes are circumstantial adjuncts.

Besides, a constituent which is not the theme of a sentence may appear occasionally in a prototypical theme position. This phenomenon has been referred to by several names, such as *focussing* (Campos *et al.* 1990), *focus preposition* (Ward 1985) or *thematization* (Chomsky 1972). Examples of this type of informational structure are *It was Pedro who lied to me*. A number of authors (*e.g.* Gutiérrez 2000: 34) have argued that the intent of this particular information schemas is to gain the attention of the interlocutor to overcome their presumed predisposition to receive information that is at some point contrary to that which is intended to be communicated, or simply to emphasize the importance of a certain aspect in the informational process. This nuance of enunciative modality would undoubtedly be applicable for the weighing of the relevant concepts for a proper summary, especially since the syntactic structures involved are relatively easy to match with rules only out of the tokens positions, the dependency relation tags and the dependency heads.

In short, it is possible to locate the discourse elements *theme* and *rheme* using syntactic knowledge. To the extent that syntactic analysis is a task that can be considered well solved in NLP, it seems feasible to be able to automatically locate the *theme* and *rheme* in every sentence of a text. The next natural step in order to obtain the thematic progression schema, *i. e.* the conceptual schema of a text, is to connect each theme and rheme of the sentences of the text in the ultimate thematic path.

5. A FIRST STUDY OF THE FEASIBILITY OF THEMATIC PROGRESSION THEORY IN SPANISH TEXTS

Aiming to conduct a first study to verify the applicability of the Thematic Progression theory for the extraction of the underlying conceptual schema of a text, we carried out an exploratory corpus survey with Spanish descriptive texts. We analyzed the mean ratio and the ratio per text of preceding subjects, since they are the prototypically unmarked themes. The examined corpus is AnCora Surface Syntax Dependencies[3] (*AnCora GLiCom-UPF 1.1*), published in 2014 at Pompeu Fabra University, which contains 17,376 sentences manually annotated with the lemmas, the *PoS* tags plus other morphological features and both dependency heads and relations for every token. The analysis was based on a symbolic rule-based grammar expressed as sub-tree extraction operations from the *dependency tree* of the sentences. In order to ensure the generality of the grammar, the elements obtained through rules for sampling preceding subjects were compared with the corresponding results in a second version of the corpus, resulting from its automatic annotation with the Freeling analyzer. The *thematic progression analyzer* scripts used to process the grammar and to generate the outputs of the corpus rendering are publicly accessible from github[4].

Basically, the grammar and the analysis algorithm to extract the thematic progression schema of the text were designed in three consecutive steps: (i), first, the automatic identification and labelling of themes for every sentence; (ii), second the subsequent identification of their rheme; and, (iii), finally, the identification of concepts corresponding to the same theme or rheme in a text. Each step was carried out using a symbolic syntactic-semantic rule-based grammar expressed as sub-tree extraction operations. For the grammar definition, an approach based on the transformation of dependency trees has been applied. Thus, for the simplest scenario of finding unmarked subjects in SVO languages, such as Spanish, two categories of rules were defined: (i), matching rules of child dependencies from a selected head token, consisting of the identification of a dependency relation as the name of a relation of arity two with a first argument as the key and the value expected for the selected parent and a second argument with the options for matching, being *ALL* if all children nodes from the head of a scope should be matched or *ONE* if only the immediate child should be matched. For example, the *SUB(deprel:ROOT, ALL)* rule would match subtrees consisting of all child nodes of the *SUB* children of a token tagged with a *ROOT* dependency relation (as shown in figure 1, obtained from the Freeling 4.1 demo[5]); and, (ii), matching rules of head dependencies from a selected child token, consisting also of

[3] http://clic.ub.edu/corpus/es/ancora-descarregues
[4] https://github.com/eelenadelolmo/HI4NLP/tree/master
[5] http://nlp.lsi.upc.edu/freeling/demo/demo.php

the identificator of a dependency relation as the name of a relation of arity two, whose arguments are the same as for the child dependency rules but in the opposite order. The second type of rules can apply, for example, for sentence compression when several propositions are involved.

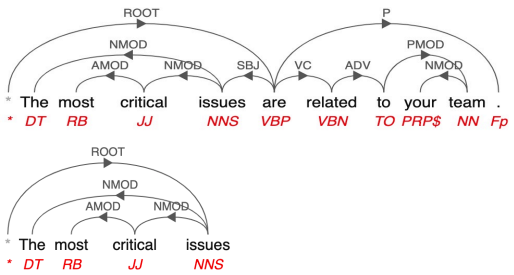

Figure 1. Input sentence and output (subtree) for a rule of the first type.

The analyzer accepts various corpus formats as input and transforms them into the universal *CoNLL-U* format, where the additional theme and rheme features are added for every token in the main proposition of sentences. We found that, with our first version of the rules, theme and rheme annotation was correct in roughly half of the cases, as shown in table 1. Through a careful manual review of the data, these results have enabled us to confirm a significant correlation between the syntactic-semantic and discourse layers outlined by the Thematic theory, and, consequently, the feasibility of automating identification of, at least, half of the themes.

Table 1. Ratio of preceding subjects in AnCora corpus.

	GLiCom-UPF 1.1	Freeling version
Sentences with preceding subjects as themes	50.4 %	46.2 %

With the syntactically annotated *GLiCom-UPF 1.1* version of the AnCora corpus we seek to provide objective metrics, not prone to major annotation errors, as the corpus annotation has been manually reviewed. The qualitative analysis of this data is intended to assess whether or not the preceding subjects match the main themes of sentences in order to ultimately detect the underlying thematic progression template of texts. By contrast, with the Freeling version of the corpus we aim to assess the accuracy in applying the rule for the extraction of unmarked themes with no dependence on manually annotated data, because Freeling is the best option for syntax annotation in Spanish at this stage. According to this objective, we have generated two different files, one for each version of the corpus, with the suspected overmatched and undermatched sentences, as shown in table 2.

Table 2. Ratio of suspected annotation errors.

	Freeling version
Suspected overmatches	1283 (7.2 %)
Suspected undermatches	3550 (20.1 %)

As the figures suggest, a pervasive tendency for undermatching sentences with preceding subjects by Freeling has been confirmed through careful examination. We found that the vast majority of mismatched annotations involve some type of coordinated or juxtaposed clauses. These syntactic structures are analyzed by Freeling with a highly fluctuating dependency structure, which is quite different from the analysis in *GLiCom-UPF 1.1*. This high variability in the syntactic tree accounts for the vast majority of both the undermatched and overmatched sentences ratios.

A qualitative analysis of unmatched sentences has also been conducted, revealing a strong presence of thematization as the most relevant finding. This sentence pattern has been referred to by the Thematic theory as a relevant discourse feature, indicating a break in the information flow of the text, as further discussed above. A promising conclusion of the analysis of the patterns found is their suitability for being implemented in the rule formalism designed. Besides, as a synthesis of the findings obtained from the qualitative analysis of the manually annotated version, there is a strong presence of subordinate clauses in the corpus, which implies the necessity of more complex rules to select the most informative proposition in the sentence. The observed patterns have been categorized into three main categories (the most informative clauses appear in bold and the selected theme is underlined):

1. Sentences whose root clause is the most relevant (*e. g. Since pharmacists work with a high profit margin,* ***the business opportunity is huge***).
2. Sentences whose root clause is not the most relevant. (*e. g. The main factor is that* ***electricity consumption during the summer is now not much lower than it used to be***).
3. Sentences whose root clause is not the most relevant but provides a crucial modality feature for information retrieval (*e. g. Investigators are convinced that* ***someone deliberately cut that rubber***).

6. CONCLUSIONS AND FUTURE WORK

As observed in the study conducted, this first approach to rule-based theme annotation seems to claim the theoretically hypothesized correlation between the syntactic-semantic and discourse layers required by our proposal. However, the qualitative analysis of both matched and unmatched sentences revealed the need for more complex tree rewriting rules to achieve a more accurate theme selection in order to obtain thematic progression schemas from texts.

Regarding subordination, *i. e.* sentences with several propositions with different syntactic status, we are working on two feasible options for sentence compression: (i), the choice of the most relevant proposition for every sentence, and, (ii), the choice of the ordered subset of its *n* more relevant clauses. In addition, this study shows the necessity to implement an algorithm to infer the modality from the main verb. We also found that the rules should be refined in order to capture the various ways in which coordinated and juxtaposed clauses could be analyzed, given the high variability observed in the automatic syntactic annotation. Finally, the study revealed that it will be necessary to design lexicon-based rules to capture lexical semantic generalizations. With all this, in principle, it seems possible to apply this new linguistic approach for the extraction of implicit conceptual schemas from texts.

REFERENCES

1. A. Gatt and E. Reiter, 'SimpleNLG: A realisation engine for practical applications', *Proceedings of the 12th European Workshop on Natural Language Generation*, 90-93, (2009).
2. A. Templeton and J. Kalita. 'Exploring Sentence Vector Spaces through Automatic Summarization', *Proceedings of the 17th IEEE International Conference on Machine Learning and Application*s, (2018).
3. C.Q. Nguyen and T.T. Phan. 'An ontology-based approach for key phrase extraction', *Proceedings of the ACL-IJCNL*, 181-184, (2009).
4. D. Cristea, O. Postolache and I. Pistol. 'Summarisation through discourse structure', *Lecture Notes in Computer Science,* 3406, 632-644, (2005).
5. D. Marcu. *The Theory and Practice of Discourse and Summarization*, The MIT Press, Cambridge, 2000.
6. D. Miller. 'Leveraging BERT for Extractive Text Summarization on Lectures', *Proceedings of arXiv*, (2019).
7. E. Baralis, L. Cagliero, S. Jabeen, A. Fiori and S. Shah. 'Multi-document summarization based on the Yago ontology', *Expert Systems with Applications*, 9, 6976-6984, (2013).
8. E. Lloret and M. Palomar. 'Text summarization in progress: a literature review', *Artificial Intelligence Review*, 37, 1-41, (2012).
9. F. Daneš, *Papers on functional sentence perspective,* Mouton, The Hague,1974.
10. G. Erkan and D.R. Radev. 'LexRank: Graph-based lexical centrality as salience in text summarization', Journal of Artificial Intelligence Research, 22, 457-459, (2004).
11. G. Giannakopoulos, V. Karkaletsis and G.A. Vouros. 'Testing the Use of N-gram Graphs in Summarization Sub-tasks', *Proceedings of the Text Analytics Conference*, (2008).
12. G. Ward, *The Semantics and Pragmatics of Preposing*, Universidad de Pennsylvania, 1985.
13. H. Campos and M. Zampini. 'Focalization Strategies in Spanish', *Probus*, 2, 47-64, 1990.
14. H. Lin and V. Ng. 'Abstractive Summarization: A Survey of the State of the Art', *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence*, 9816-9822, (2019).
15. H.P Edmundson. 'New methods in automatic extracting', *Journal of the ACM*, 16(2), 264–285, (1969).
16. H.P. Luhn. ' The Automatic Creation of Literature Abstracts', *IBM Journal of research and development*, 2(2), 159-165, (1958).
17. I. da Cunha, S. Fernández, P.V. Morales, J. Vivaldi, E. San Juan and J.M. Torres-Moreno. 'A new hybrid summarizer based on vector space model, statistical physics and linguistics', *Lecture Notes in Computer Science*, 4827, 872-882, (2007).
18. J. Zhang, Y. Zhou and C. Zong. 'Abstractive Cross-Language Summarization via Translation Model Enhanced Predicate Argument Structure Fusing', IEEE/ACM, 24, (2016).
19. K. Wang, X. Quan and R. Wang. 'BiSET: Bi-directional Selective Encoding with Template for Abstractive Summarization', *The 57th Annual Meeting of the Association for Computational Linguistics*, (2019).
20. L. Logeswaran and H. Lee. 'An efficient framework for learning sentence representations', *Proceedings of the 6th International Conference on Learning Representations*, (2018).
21. M.A.K. Halliday, *An Introduction to Functional Grammar*, Arnold, London, 1985.
22. M.E. Hannah and S. Mukherjee. 'A classification-based summarisation model for summarising text documents', *International Journal of Information and Communication Technology*, 6, 292-308, (2014).
23. M. Galley. 'Skip-chain Conditional Random Field for ranking meeting utterances by importance', *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, (2006).
24. M.J. Kusner, Y. Sun, N.I. Kolkin and K.Q. Weinberger. 'From Word Embeddings To Document Distances', *Proceedings of the 32nd International Conference on International Conference on Machine Learning*, 37, 957–966, (2015).
25. N. Elhadad, K. McKeown, D. Kaufman, and D. Jordan, 'Facilitating physicians' access to information via tailored text summarization', *Proceedings of the AMIA Annual Symposium*, 226-300, (2005).
26. N. Chomsky, *Studies on Semantics in Generative Grammar*, Mouton, La Haya, 1972
27. N. Goyal and J. Eisenstein. 'A Joint Model of Rhetorical Discourse Structure and Summarization', *Proceedings of the Workshop on Structured Prediction for NLP*, 25-34 , (2016).
28. N. Reimers and I. Gurevych. 'Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks', *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing*, (2019).
29. P.E. Genest and G. Lapalme. 'Framework for Abstractive Summarization using Text-to-Text Generation', *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, 64-73, (2011).
30. P. Wu,, Q. Zhou, Z. Lei, W. Qiu and X. Li, 'Template Oriented Text Summarization via Knowledge Graph', *International Conference on Audio, Language and Image Processing (ICALIP)*, 79-83, (2018).
31. Q. Le and T. Mikilov. 'Distributed Representations of Sentences and Documents', *Proceedings of the 31st International Conference on Machine Learning*, (2014).
32. R. Kiros, Y. Zhu, R. Salakhutdinov, R.S. Zemel, A. Torralba, R. Urtasun and S. Fidler. 'Skip-Thought Vectors', *Proceedings of the 28th International Conference on Neural Information Processing Systems*, 2, 3294–3302, (2015).
33. R.M. Aliguliyev. 'Clustering techniques and discrete particle swarm optimization algorithm for multi-document summarization', *Computational Intelligence*, 26, 420-448, (2010).
34. S. Gerani, Y. Mehdad, G. Carenini, R.T. Ng and B. Neja. 'Abstractive summarization of product reviews using discourse structure', *Proceeding of the Conference on Empirical Methods in NLP*, 1602-1613, (2014).
35. S. Gutiérrez, *Temas, remas, focos, tópicos y comentarios*, Arco Libros, Madrid, 2000.
36. S. Harabagiu and F. Lacatusu. 'Topic themes for multi-document summarization', *Proceedings of the 28th Annual International ACM SIGIR,* (2005).
37. T. Mikolov, I. Sutskever, K. Chen, G. Corrado and J. Dean. 'Distributed Representations of Words and Phrases and their Compositionality', *Advances in neural information processing systems*, 26, (2013).
38. V. McCargar. 'Statistical approaches to automatic text summarization', *Bulletin of the American Society for Information Science and Technolog*, 30, (2004).
39. W. Li. 'Abstractive multi-document summarization with semantic information extraction', *Proceedings of the Conference on Empirical Methods in NLP*, 1908-1913, (2015).
40. W.S. McCulloch and W. Pitts. 'A logical calculus of the ideas immanent in nervous activity', *Bulletin of Mathematical Biophysics*, 5, 115-133, (1943).
41. Y. Kikuchi, T. Hirao, H. Takamura, M. Okumura and M. Nagata. 'Single Document Summarization based on Nested Tree Structure', *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, 315–320, (2014).
42. Y. Liu, I. Titov and M. Lapata. 'Single Document Summarization as Tree Induction', *Proceedings of NAACL-HLT*, 1745-1755, (2019).