The Impact of Rule-Based Text Generation on the Quality of Abstractive Summaries

Tatiana Vodolazova, Elena Lloret
Dept. of Software and Computing Systems
University of Alicante
Apdo. de Correos 99
E-03080, Alicante, Spain
{tvodolazova,elloret}@dlsi.ua.es

Abstract

In this paper we describe how an abstractive text summarization method improved the informativeness of automatic summaries by integrating syntactic text simplification, subject-verb-object concept frequency scoring and a set of rules that transform text into its semantic representation. We analyzed the impact of each component of our approach on the quality of generated summaries and tested it on DUC 2002 dataset. Our experiments showed that our approach outperformed other state-of-the-art abstractive methods while maintaining acceptable linguistic quality and redundancy rate.

1 Introduction

Rapid growth of digital information increases the need for automatic text summarization methods that can digest large amounts of textual data, such as scientific articles, blogs and news articles to extract concise and relevant information from them. Text summarization methods can be classified into abstractive and extractive ones (Nenkova and McKeown, 2012). Extractive methods compose summaries from the most salient sentences of the original document. In contrast, abstractive methods generate novel or partially novel text using such techniques as sentence compression, fusion, calculation of path scores in graphs or, natural language generation tools such as SimpleNLG (Gupta and Gupta, 2018). They involve an intermediate step of deep linguistic analysis and an abstract semantic representation of the data. Extractive techniques have been intensively researched for over half a century and, according to some studies, "have more or less achieved their peak performance" (Mehta, 2016).

Over the past few years interest in the field of text summarization has shifted towards abstractive methods and quickly produced a large variety of approaches. Gupta and Gupta (2018) classify them broadly into methods based on the structure, semantics and deep learning with neural networks.

The main advantage of semantic-based approaches over deep learning ones lies in their independence from a large training corpus. Most of the available datasets for deep learning belong to the domain of news text that further restricts the application of these methods to other domains. However, semantic-based approaches rely on a parser to transform text to its semantic representation and, therefore, a poor parser performance will reduce the quality of generated summaries. Another limitation of the deep learning methods comes from the fact that they rely on statistical co-occurrence of words and are prone to semantic and grammatical errors. This is something that a reliable parser could help to avoid.

Structure-based methods, such as template and ontology based ones reveal other weaknesses. Template-based methods lack diversity. At the same time, ontology based ones rely on a time-consuming task of creating an ontology by a human expert. However, they provide highly coherent summaries and can handle uncertainties respectively. Semantic-based approaches that rely on handcrafted rules to transform text into semantic representation may be criticized for the same reason related to the human effort and time required to solve the laborious task of creating transformation rules.

It becomes clear that each abstractive approach can reliably handle only some aspects of the summarization process while revealing weaknesses in the remaining ones. Thus far, none of the approaches has been capable of offering a broad-based solution. Research in this field is mak-

Proceedings of Recent Advances in Natural Language Processing, pages 1275–1284,
Varna, Bulgaria, Sep 2–4, 2019.
https://doi.org/10.26615/978-954-452-056-4_146

ing headway, each time with more elaborate algorithms and combining techniques from a number of different methods. However, Chen et al. (2016) have shown via their analysis of the reading comprehension task – another natural language processing task that requires interpretation of the text – that a straightforward approach designed around a small set of carefully selected features can obtain high, state-of-the-art accuracy.

Therefore, this study has a threefold objective. First, to design a broad-based abstractive text summarization method. Second, to evaluate whether the proposed method is capable of delivering concise and informative summaries while maintaining above-average linguistic quality and redundancy rate. Third, to compare it against other state-of-the-art abstractive methods.

The approach that we propose in this work falls into the previously mentioned semantic-based group of abstractive summarization approaches and has been inspired by the ideas of Genest and Lapalme (2011) and Lloret et al. (2015). Our contribution takes their abstractive models one step further by scoring abstract information representation without taking into account its surface representation. The proposed method incorporates syntactic text simplification, subject-verb-object concept frequency scoring, and a set of rules that transform text into its semantic representation.

This paper is structured as follows: Section 2 discusses related semantic-based abstractive summarization approaches. Section 3 describes in detail the architecture of our method. Evaluation methods and results are presented in Section 4. Section 5 describes the effect of individual components of our approach upon the quality of generated summaries. Section 6 provides a summary of the conclusions and areas for future work.

2 Related Work

All the methods in the semantic-based abstractive summarization group include the initial step of converting texts into an abstract semantic representation. For example, Genest and Lapalme (2011) introduced the concept of information item that was defined as a smallest element of coherent information and represented as a dated and located subject-verb-object triplet. Lloret et al. (2015) also base their concept representation on subject-verb-object triplets. Alshaina et al. (2017) use predicate-argument structure as their underlying information representation and extract a number of features from it that are later used for ranking. Li (2015) define the concept of Basic Semantic Unit (BSU) where each BSU is an actor-action-receiver triplet with its obligatory arguments, namely, actor and receiver of the action. The BSUs are used to construct a BSU semantic link network representation for each text.

Abstract Meaning Representation (AMR) graphs is the most recent approach to abstract semantic representation of texts (Vilca and Cabezudo, 2017). AMR nodes are represented by either words or PropBank[1] frames, and edges define relationships between them. Both the AMR graph representation and the subject-verb-object (SVO) representation depend on the efficiency of the parser. However, AMR graphs also rely on PropBank framework whose limitations pose additional constraints on AMR graphs. Furthermore, the problem of text generation from AMR graphs is still a challenge and it has not yet been solved (Li, 2015).

The summarization method based on BSUs proposed by Li (2015) overcomes the limitation of text generation faced by AMR graphs, and produces informative, coherent and compact summaries. However, as the authors state, the BSU network cannot yet handle data that express opinions rather than facts and actions, since these cases involve verbs that lack meaningful actions, such as 'be', and the underlying representation of actor-action-receiver cannot be appropriately computed.

Alshaina et al. (2017) use K-means and agglomerative hierarchical clustering algorithms to group similar predicate-argument structures (PAS) based on semantic similarity measures, and to eventually select the most representative PAS based on a weighted set of 12 features. The PAS proposed by this approach are classified into simple and complex ones. Complex PAS are derived from sentences with multiple verbs, otherwise they are considered to be simple. Nested PAS are eliminated. One of the features that determines whether to include a PAS into the summary or not is the "number of verbs and nouns" that gives preference to complex PAS as crucial to summary generation.

Lloret et al. (2015) propose an abstractive semantic-based approach to ultra-concise opinion summarization. It involves a syntactic sentence simplification in the preprocessing step and

[1] https://propbank.github.io/

semantic representation based on subject-verb-object triplets. Their scoring heuristics relies on subject-verb-object term frequencies.

The approaches closest to ours are those of Lloret et al. (2015); Genest and Lapalme (2011, 2010). However, the difference between them is twofold. First, the aforementioned systems use term or document frequencies for scoring. We integrate word sense disambiguation to identify similarities between subject-verb-object triplets on the conceptual level that allows us to introduce concept frequencies for scoring. Second, the architecture of our approach is characterized by a higher level of abstraction. Namely, our approach scores abstractive concepts represented in the form of enriched subject-verb-object triplets and not their surface representation. Their surface representation is integrated in the final step when all the triplets have already been assigned their score.

Unlike the approach of Alshaina et al. (2017) who give preference to sentences with more than one verb, our approach integrates syntactic sentence simplification in the preprocessing step in order to split complex sentences into simpler ones and ideally reduce syntactic structure to a single main verb. This allows us to generate various subject-verb-object triplets from a single sentence and to manipulate them in a more precise manner.

3 Abstractive Summarization Framework

The architecture of our proposed abstractive text summarization approach is illustrated in Figure 1. This section describes the role and the implementation of each of its components.

Simplification. We begin by applying syntactic simplification to the original document as a pre-processing step. Simplification targets only complex sentences, splitting their syntactic trees into simpler ones. Each newly created sentence is a fully grammatical construction that, not always but in most cases, contains one main verb and covers one single concept[2]. In the next stages our method generates an information item from each simplified sentence. Simplifying the syntactic structure of the input text allows us to have fewer, less recursive and less error-prone rules for information item extraction. And capturing as many concepts as possible benefits the process of information item selection: only the most salient bits of information are selected while the irrelevant ones are discarded. We use the Factual Statement Extractor to carry out the simplification task (Heilman and Smith, 2010).

Analysis. In this stage, we perform a linguistic analysis decomposing each supplied simplified sentence into lemmas, stems, parts of speech, senses, named entities, syntactic roles and noun phrases. This is done mainly with the help of Stanford CoreNLP (Manning et al., 2014). Additionally we use Porter stemmer for stemming (Porter, 1997), Freeling for word sense disambiguation (Padró and Stanilovsky, 2012) and Java DOM parser for noun phrase chunking.

Information Items Generation. Once the data have been analyzed we proceed to build an abstract representation of each of the sentences. We adopt the same naming convention as Genest and Lapalme (2011) and refer to them as information items (`InIts`). At the core of each `InIt` lies the main verb of the sentence accompanied by its subject and object, if they are present. Contrary to Genest and Lapalme (2011) we do not incorporate any manual rules to reject candidate `InIts`. However, a small portion of them will be lost during the surface realization stage if SimpleNLG fails to generate a sentence from an `InIt`. It happens at most to 1-2 simplified sentences per document. Preserving all `InIts` may introduce a higher rate of grammatically incorrect sentences due to the incorrect sentence parses[3]. However, since no clear pattern between syntactic linguistic phenomena and incorrect parses was observed, we could not discard such cases. Additionally, we extend the core subject-verb-object structure to include open clausal complements and prepositional phrases. Since the Stanford CoreNLP configuration that we used implements Universal Dependencies [4] for dependency parsing, our rules for transforming text into `InIts` are also designed around this annotation scheme. We implemented 5 transformation rules:

1. `ccomp` rule retains a clausal complement of a verb or adjective, rejecting the initial part. *He says that [you like to swim].*
2. `subject` and `verb` rule identifies them in the remaining sentence. It also handles copula and passive voice.

[2] Table 9 provides an example of a simplified sentence.

[3] Common mistakes provoked by this decision can be found in Section 4.2.

[4] https://universaldependencies.org/

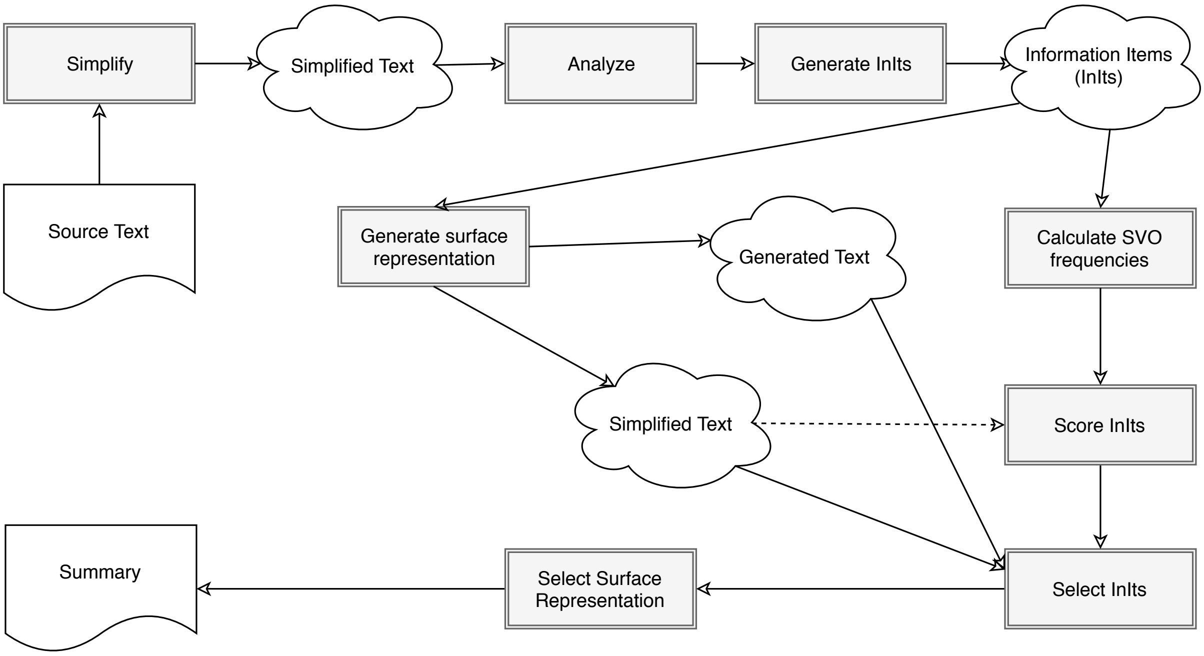

Figure 1: Our Abstractive Summarization Framework.

3. `direct` and `indirect object` rule sets corresponding objects if they exist.
4. `xcomp` rule handles open clausal complements of verbs and adjective. *She looks [very beautiful]. I consider him [a fool]. He tried [to run].*
5. `pp` rule identifies remaining prepositional phrases. *They talked [about London].*

All the `InIts` are stored internally as an ordered list.

Calculation of frequencies. In this stage we analyze `InIts` and calculate concept frequencies of all the verb, subject and object phrase heads of the input. For the purpose of evaluating effectiveness of concept frequencies, we also incorporated their term frequency scoring for comparative purposes. Our scoring strategy is based on the idea that there is "a very strong correlation between concepts of topic and subject in English." (Foley, 1994). And it has also been shown in previous research on text summarization that subjects, verbs and objects play a crucial role in content selection and cannot be dropped (Harabagiu and Lacatusu, 2010). Along with the SVO frequencies we also calculate term frequencies of named entities that represent subject or object phrase heads.

Information Items Scoring. Unlike the approaches of Genest and Lapalme (2011); Lloret et al. (2015) in our approach, `InIt` scoring and surface realization are independent from each other. We apply extracted SVO and named entity head frequencies from the previous step to score `InIts` directly. This gives us the flexibility to choose which parts of `InIts` to use for scoring. Our scoring is based on the idea that `InIts` that cover the main topic of the document contain the most frequent SVO concepts and named entities in any of their components. Given the flexibility to work with `InIts` directly and not the raw text, we experimented with scoring on SVO components and also combined them with open clausal complements and prepositional phrases. While scoring, we calculate matches not only between candidate noun phrase heads, but other phrase constituents es well.

For testing purposes we also integrate a modification of this step that, instead of scoring `InIts` directly, applies SVO and named entity frequencies to the simplified text. This configuration is indicated with the dashed arrow in Figure 1. It allows us to compare how much information is lost during the transformation and generation stages.

Text Generation. We generate sentences from `InIts` with the help of SimpleNLG realization engine (Gatt and Reiter, 2009). The order of text generation rules is defined mainly by functionalities of SimpleNLG and follows these steps:

- generate a noun phrase (NP) to represent the subject if present;
- generate the main verb;

- generate an NP to represent direct object if present;
- generate an NP for indirect object if present;
- generate prepositional phrases
- generate open clausal complements if present (xcomp transformation rule); and,
- assemble all the components and generate the verb phrase (VP).

We do not do any other modifications apart from syntactic simplification of long sentences in the preprocessing step of our approach. This means, for example, that we do not convert passive constructions into active ones. However, since we always use the same order for the text generation rules, the original order of constituents may be changed, i.e. prepositional phrases will always be generated after subject, verb or objects, despite the fact that in the original sentence they may be in a different position. Generated sentences play no role in InIts scoring or InIt selection. They remain on hold until the selection of InIts and surface representation stage.

Information Items Selection. At this stage, we inspect all the InIts and reject the ones with empty text representation generated by SimpleNLG.

Selection of Surface Representation. For all the remaining ranked InIts, starting from the highest ranked one, we add each InIt's surface representation to the final summary until the maximum allowed size has been reached. Once we reach it we reorder sentences to preserve the original order of simplified sentences that each InIt originated from and deliver the summary. For surface representation our approach allows the selection of either a representation generated with SimpleNLG or the simplified sentence. In this final stage we do not integrate additional date or location information as Genest and Lapalme (2011), but if an InIt contained them among its prepositional phrases, they are included into the generated sentence by SimpleNLG.

4 Evaluation

Our approach is evaluated on DUC 2002 dataset for the single document summarization task[5]. After discarding duplicates, the dataset consists of 530 newswire articles. Each article is accompanied by one or more manually created abstractive model summaries of approximately 100 words.

At this development phase, our approach generates summaries operating exclusively with the words present in the original text. However, as a result of the syntactic simplification, they are likely to be reorganized into shorter sentences. Moreover, some of the words are ordered differently or not included into generated sentences as a consequence of the implemented translation and surface realization rules. These operations create summaries that go beyond the literal extraction of original text fragments.

We evaluate the content selection part of our approach with ROUGE toolkit and use human evaluation to assess the linguistic quality of generated summaries as described in Sections 4.1 and 4.2 respectively.

4.1 Informativeness

Following the example of recent works on abstractive text summarization we used ROUGE toolkit (Lin, 2004) to evaluate generated summaries (Vilca and Cabezudo, 2017; Hsu et al., 2018). ROUGE-1 and ROUGE-2 are used to assess informativeness and together with ROUGE-SU4 they have been found to correlate well with human judgement. The longest common subsequence ROUGE-L is used to assess fluency. We compared our summaries to the human summaries provided for DUC 2002 corpus, and each text can be evaluated against at least 2 of them.

We also calculated average pairwise ROUGE values for human summaries to identify the highest score that an abstractive summary can obtain with ROUGE (see Table 1).

The selected baseline was implemented with the help of our method such that each original text passes through all the stages specified in Section 3, including sentence simplification and surface realization stages but avoiding the SVO and named entity scoring. To produce the baseline summary we applied tf-scoring to such regenerated sentences. This ensures that the baseline is an abstractive summary only differing in the scoring method.

We compared our approach to two state-of-the-art approaches for abstractive text summarization of a different nature: 1) Vilca and Cabezudo's (2017) approach based on AMR graphs and Rhetorical Structure Theory; and, 2) the approach proposed by Hsu et al. (2018) based on deep learning and combines abstractive and extractive components. To compare our approach with the latter

[5] http://duc.nist.gov/

	R-1	**R-2**	**R-L**	**R-SU4**
Human	0.507	0.218	0.460	0.239
Ours	**0.410**	0.154	**0.378**	**0.180**
Baseline	0.378	0.138	0.351	0.163
Hsu'18 abs	0.266	0.116	0.239	0.126
Vilca'17	0.244	**0.231**	-	0.033

Table 1: ROUGE scores for different summarization methods.

one, we used their abstractive model pre-trained on CNN/Daily Mail dataset of newswire articles.

Table 1 shows that our approach outperforms both the abstractive baseline and the approach of Hsu et al. (2018) on all the ROUGE metrics. It also outperforms Vilca and Cabezudo's (2017) approach on 3 of the 4 metrics.

To illustrate how our approach and the approach of Hsu et al. (2018) modify original sentences, we contrast an extractive term-frequency based summary with the abstractive summaries generated by both of the approaches (see Table 2). For convenience, the common chunks between the summaries are numbered and surrounded by square brackets, while the unique chunks are italicized.

Our approach: [More than 4,000 workers at a coal mine in the southern city of Jastrzebie went to demand legalization of Solidarity and higher wages on strike]^{1}. [Workers on the overnight shift at the Manifest Lipcowy mine stayed outside the mine shaft]^{2}. [The miners are demanding the legalization of Solidarity]^{4}. *The workers are calling for higher wages and better working conditions. The workers are requesting two lawyers and two economists. Workers at the Rudna copper mine near the city of Wroclaw staged a protest rally.*[Workers at factories around the northern port of Gdansk joined striking shipyard workers.]^{5}.
Extractive TF summary: *Solidarity spokeswoman Katarzyna Ketrzynska said* [[workers on the overnight shift at the Manifest Lipcowy mine stayed outside the mine shaft]^{2} all night and were joined by workers arriving for the morning shift]^{3}. *The strike began at noon today, according to Katrzynska. She said* [the miners are demanding the legalization of Solidarity]^{4} *and reinstatement of workers fired for union activities. Three members of Solidarity were barred Saturday from working. On Aug. 16, 1980,* [workers at factories around the northern port of Gdansk joined striking shipyard workers]^{5} *to form Solidarity, the first and only independent trade federation in the Soviet bloc.*
Hsu'18: [more than 4,000 workers at a coal mine in the southern city of jastrzebie went on strike today to demand legalization of solidarity and higher wages]^{1}. [[workers on the overnight shift at the manifest lipcowy mine stayed outside the mine shaft]^{2} all night and were joined by workers arriving for the morning shift]^{3}.

Table 2: An comparison of abstractive summaries with an extractive summary.

Grammaticality: 1. TAS **gave not** details of Gorbachev 's suggestion. 2. Six bodies were *founded* in the hull of the ferry by Police. 3. The Lone Star Statuette *were* built by Chicago 's Creative House Promotions.
Redundancy: 1. Martin Nelson was another meteorologist at the center *at center*. 2. Dullah Omar was an activist and family friend of the Mandelas *of Mandelas*. 3. A resolution promises reforms. A resolution promises reforms.
Completeness: 1. A quake of 6 on the scale is capable. 2. Reunification mishandled. 3. Arthur Andersen wanted.

Table 3: Examples of some of the mistakes produced by our approach.

4.2 Human Evaluation

For our preliminary human evaluation of generated summaries, we used the statistical formula to calculate the correct size of a representative sample that was proposed by Pita-Fernández (1996) and successfully applied to different NLP tasks (Vázquez et al., 2010; Lloret et al., 2019). For DUC 2002 dataset, a representative sample consists of 77 documents that we randomly chose from the corpus. They were evaluated according to the following criteria based on the DUC guidelines, but adapted to the specific task and errors:

- *grammaticality* - grammatical correctness of the summary (i.e. number agreement);
- *non-redundancy* - no unnecessary repetitions; and,
- *completeness* - completeness of grammatical construction (i.e. a missing direct object of a transitive verb).

The generated abstractive summaries were assessed on a five-point Likert scale by 3 external annotators without any knowledge about how the summaries were produced. A grammatically correct, non-redundant and complete summary would receive a score of 5-5-5 respectively. The results in Table 4 show that the summaries produced by our approach scored above the average on the three criteria.

Measure	**Score**
Grammaticality	3.60
Non-redundancy	3.71
Completeness	3.81

Table 4: Average scores for human evaluation.

Table 3 shows examples of such mistakes. Upon closer inspection we detected that the completeness errors are often caused by incorrect parses. Some of the grammatical errors are produced by SimpleNLG, whereas others refer to the cases not covered by information item extraction rules. Contrary to our predictions, the non-redundancy rate was above the average. The overall linguistic quality looks promising and reveals areas for improvement. However, there is a need for a deeper evaluation that is planned for future work.

5 Further Experiments and Discussion

In this section we analyze the impact of each of the components of our method on the informativeness of generated summaries.

5.1 Syntactic Constituents

We experimented with different configurations of our scoring module to test whether the subject, verb and object are enough for the scoring, or should be extended with open clausal complements and prepositional phrases to improve its performance. For this purpose we applied the scoring in three different contexts: exclusively SVO (SVO); the SVO extended with clausal complements (SVO+xComp); and, the SVO+xComp extended with prepositional phrases (SVO+xComp+PPs). This means that the scoring module checked occurrences of the most frequent SVO elements only in subject-object-object triplets or in the extended structures. The results in Table 5 show that there is some improvement in performance when additional syntactic components are included. We believe that this improvement may increase as the corpus increases, since a larger corpus will contain more cases of open clausal complements and prepositional phrases.

	R-1	**R-2**	**R-L**
SVO	0.4064	0.1522	0.3756
SVO xComp	0.4078	0.1523	0.3765
SVO xComp PPs	0.4102	0.1544	0.3776

Table 5: ROUGE scores for syntactic components

5.2 Generation and Recall

Another experimental setup addresses the question of how much important information is lost during the generation stage. As described in Section 3, we integrated a setting (signaled with the dashed arrow in Figure 1) that, instead of scoring `InIts`, applied the SVO frequencies to the simplified text and delivered it in the final summary. This setting overcomes two possible limitations of our approach: it also scores the parts of the sentence that are not included into an `InIt` and provides more text for the future recall evaluation with ROUGE. Results in Table 6 show a slight improvement over the `InIt`-based scoring, but the difference is not as high as we expected. We may conclude that our `InIt` extraction rules capture most of the information, and surface realization rules generate sufficient material for the ROUGE evaluation.

	R-1	**R-2**	**R-L**
InIt	0.4102	0.1544	0.3776
Simpl. text	0.4181	0.1668	0.3797

Table 6: ROUGE evaluation of text-based scoring

5.3 Effect of Concept Frequency Scoring

Word sense disambiguation and the resulting concept scoring should positively affect `InIt` selection as well. Table 7 shows that in this setting the difference between term and concept frequencies is almost non-existent. We believe that if we integrate the entire noun phrase when calculating SVO frequencies and not only the noun phrase head, it may lead to a more significant difference.

	R-1	**R-2**	**R-L**
SVO cf	0.4102	0.1544	0.3776
SVO tf	0.4100	0.1545	0.3777

Table 7: ROUGE scores for concept and term frequency scoring.

5.4 Simplification and Recall

Our motivation behind the integration of a syntactic simplification module was to reach a greater degree of concept granularity that would allow us to select only the most salient `InIts` while discarding the less relevant ones. We tested our approach both with and without simplification. The results revealed in Table 8 indicate that working with original text yields a slightly better recall.

Close inspection showed that our simplification module generates syntactically more simple sentences, but introduces more repetitions that are

	R-1	**R-2**	**R-L**
Simplified	0.4102	0.1544	0.3776
Original	0.4169	0.1588	0.3803

Table 8: ROUGE scores for simplification test.

picked up by the SVO and named entity scoring. Consider the example in Table 9:

Original sentence: Greek marine archaeologists focus on locating and surveying historic wrecks scattered around the Aegean and rarely carry out excavations.
Simplified: 1. Greek marine archaeologists focus on locating. 2. Greek marine archaeologists focus on surveying historic wrecks scattered around the Aegean. 3. Greek marine archaeologists carry out excavations.
Simplified summary: 1. Greek marine archaeologists focus. 2. Greek marine archaeologists carry out excavations.
Original summary: not included

Table 9: Simplification example.

When we split a long sentence into several shorter ones with the repeated subject, the scoring module gives them more importance by considering the repeated subject to be the topic of the document. If some of these split sentences are included in the final summary, the repeated subject noun phrase takes summary space that otherwise could be occupied by a different phrase. On the other hand if the subject of such a split phrase is the true topic of the document, our method generates a very topic-focused summary. We hypothesize that scoring should be performed on the original subject-verb-object distribution of the document so as to avoid scoring for repeated subjects.

5.5 Summary Readability

Readability is rarely studied in detail in the context of automatic text summarization. Our summarization approach integrates syntactic simplification that results in syntactically simpler summaries and concept frequency scoring that may yield summaries with richer vocabulary when compared to term frequency based ones. To assess readability of generated summaries we calculated their Flesch Reading Ease (FRE), Dale-Chall (DC) and depth of the parse tree (PTD) scores. These three metrics give us a quick but complete assessment of the length, vocabulary and syntactic complexity-based readability aspects. Higher FRE and lower PTD and DC values correspond more comprehensible texts.

Results in Table 10 show that human summaries include longer sentences and words, and are also more concept dense than the original texts. Human summaries also tend to consist of syntactically less complex sentences. Unlike human summaries, our approach generates more comprehensible texts in terms of sentence and word length. As expected, syntactic sentence simplification positively affects the parse tree depth metric. However, it also generates summaries with greater lexical density.

	FRE	**DC**	**PTD**
Ours	50.74	10.56	8.30
Human	42.76	10.45	10.51
Original	43.51	10.13	11.48

Table 10: Readability metrics for different methods.

6 Conclusions and Future Work

This paper presents a broad-based abstractive text summarization method that outperforms other state-of-the-art abstractive approaches while maintaining acceptable linguistic quality and redundancy rate. Our approach is based on the set of syntactic rules that transform text into its semantic representation as well as the combination of subject-verb-object concept frequency and named entity frequency for scoring.

The results show that some aspects of the proposed approach require improvement. Integration of the entire subject and object noun phrases for the calculation of frequencies may increase informativeness of the generated summaries. Coreference resolution and sentence fusion may help to lower the degree of redundancy introduced through the syntactic sentence simplification.

In future work, we plan to integrate these improvements and to evaluate our method on other datasets such as CNN/Daily Mail dataset. First, a larger dataset can provide more insights on the relative importance of open clausal complements, prepositional phrases and concept frequency for information item rating. Second, it will allow us to gauge the weaknesses and strengths of our approach, which is based on the concept of information items and handcrafted syntactic transformation rules, via a comparative analysis with state-of-the-art deep learning and semantic graph approaches.

Acknowledgments

This research work has been partially funded by the University of Alicante (Spain), Generalitat Valenciana and the Spanish Government through the projects SIIA (PROMETEU/2018/089), LIVING-LANG (RTI2018-094653-B-C22), INTEGER (RTI2018-094649-B-I00) and Red iGLN (TIN2017-90773-REDT).

References

S. Alshaina, A. John, and A. G. Nath. 2017. Multi-document abstractive summarization based on predicate argument structure. In *2017 IEEE International Conference on Signal Processing, Informatics, Communication and Energy Systems (SPICES)*. pages 1–6. https://doi.org/10.1109/SPICES.2017.8091339.

Danqi Chen, Jason Bolton, and Christopher D. Manning. 2016. A thorough examination of the CNN/daily mail reading comprehension task. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Berlin, Germany, pages 2358–2367. https://doi.org/10.18653/v1/P16-1223.

William A Foley. 1994. Information structure. In Roland E. Asher and Joy M. Y. Simpson, editors, *The Encyclopedia of Language and Linguistics*, Pergamon Press, Oxford, volume 3, pages 1678–1685.

Albert Gatt and Ehud Reiter. 2009. Simplenlg: A realisation engine for practical applications. In *Proceedings of the 12th European Workshop on Natural Language Generation*. Association for Computational Linguistics, Stroudsburg, PA, USA, ENLG '09, pages 90–93.

Pierre-Etienne Genest and Guy Lapalme. 2010. Text generation for abstractive summarization. In *Proceedings of the Third Text Analysis Conference*. National Institute of Standards and Technology, Gaithersburg, Maryland, USA.

Pierre-Etienne Genest and Guy Lapalme. 2011. Framework for abstractive summarization using text-to-text generation. In *Proceedings of the Workshop on Monolingual Text-To-Text Generation*. Association for Computational Linguistics, Portland, Oregon, pages 64–73. https://www.aclweb.org/anthology/W11-1608.

Som Gupta and S.K. Gupta. 2018. Abstractive summarization: An overview of the state of the art. *Expert Systems with Applications* 121:49–65. https://doi.org/10.1016/j.eswa.2018.12.011.

Sanda Harabagiu and Finley Lacatusu. 2010. Using topic themes for multi-document summarization. *ACM Trans. Inf. Syst.* 28(3):13:1–13:47. https://doi.org/10.1145/1777432.1777436.

Michael Heilman and Noah A Smith. 2010. Extracting simplified statements for factual question generation. In *Proceedings of QG2010: The Third Workshop on Question Generation*. pages 11–20.

Wan Ting Hsu, Chieh-Kai Lin, Ming-Ying Lee, Kerui Min, Jing Tang, and Min Sun. 2018. A unified model for extractive and abstractive summarization using inconsistency loss. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*. ACL, Melbourne, Australia, volume 1, pages 132–141. https://aclanthology.info/papers/P18-1013/p18-1013.

Wei Li. 2015. Abstractive multi-document summarization with semantic information extraction. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Lisbon, Portugal, pages 1908–1913. https://doi.org/10.18653/v1/D15-1219.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*. Association for Computational Linguistics, Barcelona, Spain, pages 74–81. https://www.aclweb.org/anthology/W04-1013.

Elena Lloret, Ester Boldrini, Tatiana Vodolazova, Patricio Martínez-Barco, Rafael Muñoz, and Manuel Palomar. 2015. A novel concept-level approach for ultra-concise opinion summarization. *Expert Systems with Applications* 42(20):7148–7156. https://doi.org/10.1016/j.eswa.2015.05.026.

Elena Lloret, Tatiana Vodolazova, Paloma Moreda, Rafael Muñoz, and Manuel Palomar. 2019. Are better summaries also easier to understand? Analyzing text complexity in automatic summarization: Challenges, models, and approaches. In Marina Litvak and Natalia Vanetik, editors, *Multilingual text analysis challenges, models, and approaches*, World Scientific, New Jersey, pages 337–369. https://doi.org/10.1142/9789813274884_0010.

Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. Association for Computational Linguistics, Baltimore, Maryland, pages 55–60. https://doi.org/10.3115/v1/P14-5010.

Parth Mehta. 2016. From extractive to abstractive summarization: a journey. In *Proceedings of the ACL 2016 student research workshop*. pages 100–106. https://www.aclweb.org/anthology/P16-3015.

Ani Nenkova and Kathleen McKeown. 2012. A survey of text summarization techniques. In *Mining text data*, Springer, pages 43–76. https://doi.org/10.1007/978-1-4614-3223-4_3.

Lluís Padró and Evgeny Stanilovsky. 2012. Freeling 3.0: Towards wider multilinguality. In *Proceedings of the Language Resources and Evaluation Conference (LREC 2012)*. ELRA, Istanbul, Turkey. https://www.aclweb.org/anthology/papers/L/L12/L12-1224/.

Salvador Pita-Fernández. 1996. Determinación del tamaño muestral. *Cadernos de atención primaria* 3(3):138–141.

M. F. Porter. 1997. An algorithm for suffix stripping. In Karen Sparck Jones and Peter Willett, editors, *Readings in Information Retrieval*, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, pages 313–316.

Yoan Gutiérrez Vázquez, Antonio Fernández Orquín, Andrés Montoyo Guijarro, and Sonia Vázquez Pérez. 2010. Integración de recursos semánticos basados en wordnet. *Procesamiento del lenguaje natural* 45:161–168.

Gregory César Valderrama Vilca and Marco Antonio Sobrevilla Cabezudo. 2017. A study of abstractive summarization using semantic representations and discourse level information. In Kamil Ekštein and Václav Matoušek, editors, *Text, Speech, and Dialogue*. Springer International Publishing, Cham, pages 482–490. https://doi.org/10.1007/978-3-319-64206-2_54.