Towards Storytelling Automatic Textual Summerized

Yazid Bounab, Joshua Muyiwa Adeegbe, Mourad Oussalah
University of Oulu, Finland
yazid.bounab, adeegbejoshua@oulu.fi, mourad.oussalah@oulu.fi

Abstract—**This paper examines the development of an automatic summarizer for storytelling. The approach makes use of semantic role labelling and co-referencing in natural language processing. The developed approach has been tested and evaluated using BBC news summary dataset and Grimm Hansel & Gretel novel dataset where a comparison with a state of art TextRank summarizer has been carried out. The results demonstrate the feasibility of the proposal and pave the way for new enhancement of the model to take into account additional constraints that can be imposed on the summarizer**

I. Introduction

With the advances in natural language processing (NLP) technologies, automatic (multi) document summarization has achieved new milestones where several applications in question–answering systems, machine translation systems and information extraction / summarization systems [2], [1]. On the other hand, automatic summarization has also been motivated by the drastic increase in size and number of online documents as a result of emergence of social media platforms, various online user-generated contents, and substantial improvement in storage capacity and performance. This renders the ability of users to perform simple reading tasks very challenging and time–consuming. The need for automatic summarizers has also seen increase from areas like email summary, short message news on mobile devices and information summary for business and governmental official research. Therefore, automatic summarization, with its subsequent development in personalized automatic summarizers strives to answer such challenge[9].

Typically, the goal of Text Summarization as a field of NLP is to produce a condensed and a shortened version of a long text or series of sentences that contain relevant information for users, hence providing a simpler version for users to comprehend the original document. One distinguishes both single–document and multi–document depending on the number of sources documents used to generate the summary document, which enables the user to track, for instance, a given event / topic from a series of news stories / web–pages. On other hand, one also distinguishes generic summarizer and query–focused summarizer where the summarization task is constraint by some ontology or user text query. Since the pioneering work of Luhn's [8] on statistical text summarization in early sixties based on word/phrase frequencies, several methodological and practical milestones have been achieved in the field. Notably, two streams have been acknowledged: extractive and abstraction summarization. In the former, the summary is constituted of selected sentences from the original source (s), which forms the vast majority of summarization literature, while in the case of abstractive summarization; a deeper semantic analysis of the text is performed to generate potentially new sentences distinct from original source's sentences. Extractive summarization is based on the concept of sentence scoring where the sentences of highest score are chosen to constitute the summarizer. Nevertheless, the scoring function is often subject to several parameters that govern the construction of the underlined summarizer. This includes the type of feature employed, nature of preprocessing, feature weighting, mathematical model for aggregation function at sentence level and sentence scoring as well as any inconsistency and uncertainty handling [14], [13]. One may distinguish for instance, features that rely on selected dictionary, cue phrases, tf–idf like features; those that take into account the title and headings of the document; those that account for the location of the terms in the text assuming that terms appearing in the beginning or at the end of the document have higher probability of being relevant than others. Likewise, several proposals have been put forward to mathematical model governing the scoring functions ranging from pure greedy algorithms for selecting k–sentences among total sentences to advanced optimization based techniques that maximize the diversity and minimize the redundancy [4], [12]. In this course, several prototypes have emerged. This includes Microsoft News2, Google1 or Columbia Newsblaster3 [2] for news article summarization; Base Line, Freq Dist, Sum Basic, MEAD, AutoSummarize for generic text summarizer; SWESUM for biomedical text[1]. Some open source tools such as Open Text summarizer, Classifier4J, N Classifier, CNGL Summarizer and Text Rank [9] emerged. Text Compacter, Sumplify, Tools4Noobs, Free Summarizer, Wiki Summarizer & Summarize Tool are among popular online summarization tools. Despite the attractiveness and appealing features of the aforementioned summarizer prototypes, the results are far from satisfactory, which makes the research into a universally accepted summarizer widely open.

This paper contributes to a special summarization task referred to as a storytelling summarizer. In essence, storytelling summarizer attempts to maintain the coherence of the story conveyed by the source document. Although one acknowledges the complexity and multi-disciplinary aspect of digital storytelling scheme that identifies historical events and establishes the corresponding link in a way that captures the key-events while maintaining reader's interest to the story. This paper advocates a rather holistic approach that relies on the identification of key-actors and sustain their semantic roles. For this purpose, an approach that combines Semantic Role Labelling (SRL), Named-Entity Recognition (NER) and statistical based analysis is devised. In the sequel, our approach pays special attention to Co-referencing Textual Analysis (CoTA). CoTA occurs when two or more expressions in a text refer to the same entity [11], [3]. The goal is therefore to build co-reference chains that comprise nouns or pronouns, part-whole relations between query and document, title and sentences within documents.

II. Semantic role labelling

A semantic role is the underlying relationship, also known as semantic case or thematic role that a participant has with the main verb in the clause [17]. Exemplary roles used in SRL are labels such as Agent, Patient, and Location for the entities participating in an event, and Temporal and Manner for the characterization of other aspects of the event or participant relations. More specifically, given a sentence containing a target verb referred to as a frame, SRL aims to label the semantic arguments or roles of that verb, where each verb is associated with modifiers, like temporal, locational, or an adverb, which are considered as parameters of the respective function representing the verb. Therefore, SRL helps finding the meaning of the sentence and the associated actions. PropBank semantic annotation [15] is among popular frameworks employed by many SRL implementations, for discovering the predicate argument structure of each predicate in a given input sentence. In this respect, PropBank defines semantic role for each verb and sense in the frame files. The core arguments are labelled by numbers, where, e.g., A0 stands for the subject of the verb, A1 for its object, A2 for indirect object. Furthermore, a set of adjunct-like arguments or modifiers (i.e., AM-LOC for location, TMP for time, ADV for adverb, DIR for direction, DIS for discourse marker, etc.) were defined.

This type of role labeling thus yields a first- level semantic representation of the text that indicates the basic event properties and relations among relevant entities that are expressed in the sentence [6].

Example: The police officer detained the suspect at the crime scene.

$$\underbrace{\text{The police officer}}_{\text{Who}}\underbrace{\text{detained}}_{\text{Did what}}\underbrace{\text{the suspect}}_{\text{To whom}}\underbrace{\text{at the crime scene.}}_{\text{At where}}$$

SRL is often extended for the events characterization task that answer simple questions such as "who" did "what" to "whom", "where", "when", and "how" the action happens.

III. Textual co-referencing

In linguistics co-referencing appears when two expressions in text refer to the same reference person or thing. This includes standard use of pronouns (i.e., it, he, she, and they) but also metaphor-like expression, which rises further challenges to computational linguistic community. There are basically four types of co-referencing which are explained below:

- *Anaphora:* It follows the expression to which it refers (its antecedent), e.g. ***the music*** was so loud that ***it*** couldn't be enjoyed [11].
- *Cataphora:* It precedes the expression to which it refers (its postcedent), e.g. If ***they*** are angry about the music, ***the neighbors*** will call the cops[11].
- *Split antecedent:* The anaphor has a split antecedent, referring to more than one referent, e.g. ***Carol*** told ***Bob*** to attend the party. ***They*** arrived together[11].
- *Co-referring noun phrase:* This occurs where the second noun phrase is a predication over the first noun phrase, e.g. the project leader is refusing to help. The selfish thinks only of himself [11].

Often, the fourth type of co- referencing is not considered as one of the major types of co- referencing because it refers to the same referent [6]. For example, we have two noun phrases, whereby the second phrase refers to the first one and the first one refers to the same referent.

For the purpose of our study, co-referencing is used to replace the pronouns (He, She, It, They and We) in the text with their referents. This makes it easy to figure out the most important chunks in the paragraphs which contribute significantly in the processes of textual summarization.

Among many co-reference resolution tools, we have used Stanford CoreNLP. It is a natural language software written in Java that provides a set of human language technology tools such named-entity recognition, co-reference, pat of speech tagging, dependency parsing, sentiment analysis and other basic operations [10].
Interestingly, the SRL can also be utilized to guide the co-referencing process. *Example*: John caught the thief.

TABLE I. Semantic Role Labeling Using Senna for Co-reference Resolution Using Stanford Coref

Tokens	**PropBank tags**	**Semantic Roles**
John	(3, [John, -,])	Who
caught	(3, [caught, caught, B-V])	Did what
the	(3, [the, -, B-A1])	To whom
thief	(3, [thief, -, I-A1])	

From Table I, We have three arguments each one represents a specific semantic role.

IV. Methodology

The proposed approach for storytelling-based summarizer is based on the following principles. First, we hypothesize that a summary is constituted by the main actors mentioned in the original document. This implicitly assumes the existence of entities playing the role of actors in the original document. Second, these actors are person-like entities so that the use of named-entity recognition system enables us to identify such actors. Third, one hypothesizes that the summary of the source documents is primarily conveyed by the main actors, in statistical sense, identified from the original source document. Therefore, the summary is constituted of sentences that are somehow connected to such actors. Fourth, in the sequel of maintaining consistency and coherence, the use of SRL and co-referencing enables the identification and association of appropriate sentences that are directly relevant to these actors. More specifically, the following steps have been implemented:

- Preprocessing of raw textual data in order to filter out sentences with removal of digits, extra spaces and punctuations.
- Identifying the main character, generally the most frequent Term that has the POS tag of PERSON.
- Identifying sentences that contain the main character, by scoring all the sentences and ranking them.
- Simplifying the sentences that contain the main character by removing the unnecessary parts with preserving the meaning.
- Combine the simplified sentences to make the text summary.

Algorithm 1 Story Summarization(Text T)

```
$Sentences \leftarrow Sentence\_Tokenizer(T)$;
$Tokens \leftarrow Word\_Tokenizer(T)$;
$TFs = dict()$;
for $Token\ \ in\ \ Tokens$ do
    if $Token\ in\ StopWords$ then
        $Tokens.Romove(Token)$;
    else
        if $Token\ not\ in\ TFs.keys()$ then
            $TFs[Token] = 1$;
        else
            $TFs[Token] = +1$;
        end if
    end if
end for
$MC \leftarrow MainCharacter(TFs, POS =' PERSON')$;
for $Sent\ \ in\ \ Sentences$ do
    if $MC\ not\ in\ Sent$ then
        $Sentences.remove(Sent)$;
    else
        $Simplify(Sent)$;
    end if
end for
$Summary \leftarrow Join(Sentences)$;
return $Summary$;
```

From an implementation perspective, Fig.1 outlines the main milestones in this framework.

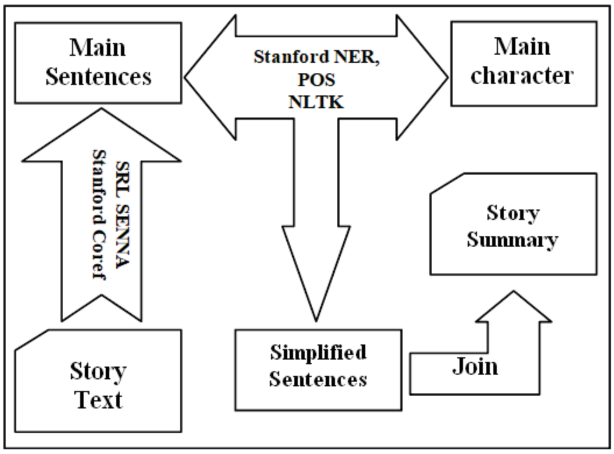

Fig. 1. Co-reference Story Summarization

V. Results

A. Dataset:

In this study, two dataset were employed to test the developed storytelling summarizer; namelely, BBC News Summary and Grimm Hansel and Getel novel. BBC News Summary dataset is downloaded from Kaggle (https://www.kaggle.com/pariza/bbc-news-summary). It contains two main directories. The first one is named' News Articles', and contains a set of sub-directories, where each one describes a specific topical area; namely, business, politics, entertainment, sport and tech. For each sub-directory, a set of articles is stored as textual files. The second directory 'Summaries' contains the summary of every textual file in every topical area. The dataset is also provided in our GitHub repository (https://github.com/bounabyazid/Co-reference-Summarization-NLP). The brothers Grimm Hansel and Gretel novel is available at (https://www.storynory.com/hansel-and-gretel-2/).

B. Metrics:

The most common metrics employed for evaluating textual summarization systems are:

ROUGE This is an abbreviation for Recall-Oriented Understudy for Gisting Evaluation. Its importance is to evaluate text summarizations derived automatically by machine tools via a set of metrics[7]. This is done through the comparison of already made reference summary usually written by humans and summary automatically produced (Machine Summary, MS) by a machine. These metrics therefore quantify the extent to which the automated summary (Machine Summary) and the human summary (ies) considered as the Reference Summaries match to each others. ROUGE-N - This is understood to be the scale or level of detail in texts being compared between the summary automatically produced (Machine Summary, MS) and reference summaries[7]. For instance, the case of ROUGE- 1 denotes the overlap of unigrams between the automated summarization and reference summary. ROUGE-2 refers to the overlap of bigrams between the automated summarization and reference summaries. This leads to the concept of Precision and Recall in the context RGOUE.

Precision and Recall- These are needed in the computation of ROUGE. It is important to mention that recall is the degree of reference summary (RS) which the automated summarization is improving or capturing[5]. Hence, to compute the recall the parameters; number of overlapping words and total words in reference summary are needed[5], hereby, Recall is measured as:

$$Recall = \frac{Overlapping\ Ngrams\ MS\ vs\ RS}{Total\ Ngrams\ of\ RS} \tag{1}$$

$$Precision = \frac{Overlapping\ Ngrams\ MS\ vs\ RS}{Total\ Ngrams\ of\ MS} \tag{2}$$

Where:

- ***MS***:Machine Summary.
- ***RS***:Reference Summary.

BLUE This stands for the Bilingual Evaluation Understudy. It is a score used in making comparisons between a Candidate Summary of texts and Reference Summary [16], [16].

Finally, F1 measure combines both Blue and Rouge metric together as follows:

$$F1 = \frac{2 \times BLUE1 \times ROUGE}{BLUE1 + ROUGE} \tag{3}$$

Especially, ROUGE and BLUE metrics use the concept of n-gram in their formulating expressions, where the metric estimate increases with the number of common n-grams between machine summary and reference summary.

C. Results and Discussions

BBC News Summary dataset:

This dataset has four hundred and seventeen political news articles of BBC from 2004 to 2005 in the News Articles folder. For each article, five summaries were provided in the Summaries folder. The first clause of the text of articles is the respective title.

Hence, for every sub-directory of the above dataset corresponding to a specific topic, evaluation in terms of Blue1, Blue2, Rouge1, Rouge2, Precision, Recall and F-score as well as the compressed rate are highlighted. These estimations are provided using minimum, maximum and average value. It is worth emphasizing that the Tables [II,VI] exhibit relatively acceptable performance of our summarization system.

TABLE II. STATISTICS FOR BUSINESS SUB-DIRECTORY

Statistics of Business				
Metrics		**Min**	**Max**	**Avg**
Rouge1	**Recall**	0.14	0.88	0.5
	Precision	0.22	0.72	0.48
	F1	0.21	0.75	0.47
Blue1		0.03	0.85	0.46
Rouge2	**Recall**	0.08	0.98	0.57
	Precision	0.18	0.91	0.55
	F2	0.21	0.92	0.55
Blue2		0.03	0.85	0.46
Compressed Rate		0.14	0.96	0.46

TABLE III. STATISTICS FOR ENTERTAINMENT SUB-DIRECTORY

Statistics of Entertainment				
Metrics		**Min**	**Max**	**Avg**
Rouge1	**Recall**	0.03	0.85	0.51
	Precision	0.18	0.74	0.47
	F1	0.06	0.7	0.47
Blue1		0.0	0.79	0.45
Rouge2	**Recall**	0.06	0.96	0.57
	Precision	0.11	0.97	0.54
	F2	0.06	0.91	0.55
Blue2		0.0	0.79	0.45
Compressed Rate		0.03	0.91	0.49

TABLE IV. STATISTICS FOR POLITICS SUB-DIRECTORY

Statistics of Politics				
Metrics		**Min**	**Max**	**Avg**
Rouge1	**Recall**	0.04	0.87	0.41
	Precision	0.28	0.9	0.51
	F1	0.08	0.7	0.44
Blue1		0.0	0.84	0.41
Rouge2	**Recall**	0.06	0.98	0.48
	Precision	0.18	0.95	0.62
	F2	0.08	0.88	0.49
Blue2		0.0	0.84	0.41
Compressed Rate		0.03	0.89	0.35

TABLE V. STATISTICS FOR SPORT SUB-DIRECTORY

Statistics of Sport				
Metrics		**Min**	**Max**	**Avg**
Rouge1	**Recall**	0.16	0.95	0.51
	Precision	0.24	0.78	0.5
	F1	0.21	0.76	0.48
Blue1		0.0	0.78	0.43
Rouge2	**Recall**	0.01	1.0	0.55
	Precision	0.01	0.92	0.56
	F2	0.21	0.95	0.53
Blue2		0.0	0.78	0.43
Compressed Rate		0.1	0.96	0.46

TABLE VI. STATISTICS FOR TECH SUB-DIRECTORY

Statistics of Tech				
Metrics		**Min**	**Max**	**Avg**
Rouge1	**Recall**	0.07	0.77	0.39
	Precision	0.23	0.81	0.53
	F1	0.12	0.72	0.43
Blue1		0.0	0.78	0.4
Rouge2	**Recall**	0.09	0.93	0.48
	Precision	0.15	0.93	0.64
	F2	0.12	0.84	0.49
Blue2		0.0	0.78	0.4
Compressed Rate		0.05	0.76	0.32

Grimm Hansel and Gretel novel dataset: Due to the availability of reference summary and the multiplicity of actors, for further testing the developed summarizer, the authors have also chosen the brothers Grimm Hansel and Gretel novel https://www.storynory.com/hansel-and-gretel-2/). The folowing table shows some ROUGE results with a summary (http://www.comedyimprov.com/music/schmoll tales.html):

In addition to the metrics employed for BBC News dataset, a comparison with TextRank summarizer, which is considered as baseline is represented in TABLE VII.

TABLE VII. STATISTICS FOR HANSEL AND GRETEL STORY SUMMARY VS MACHINE SUMMARY

Statistics of Tech			
Metrics		**Our System**	**TextRank**
Rouge1	**Recall**	0.13	0.19
	Precision	0.45	0.45
	F1	0.2	0.27
Blue1		0.02	0.09
Rouge2	**Recall**	0.09	0.15
	Precision	0.3	0.29
	F2	0.11	0.17
Blue2		0.02	0.09
Compressed Rate		0.03	0.04

VI. CONCLUSION

In this paper, a novel method for extractive textual summarization for storytelling is put forward. Extractive textual summarization is based on the reuse of some of the existing sentences in the original text. Therefore, the extraction of the main sentences is based on the most highly ranked sentences. More specifically, the storytelling based summarizer starts by identifying the main characters / actors using standard named-entity recognition and statistical reasoning. Besides, a co-reference based Semantic role labeling SRL approach has been devised and implemented in order to preserve the coherence among the identified actors. The approach has been implemented using BBC news dataset. The results great efficiency for identifying the most informative and important sentences in the given text. Nevertheless, more work is still required in order to minimizes redundandancy and maximizes diversity in case of large scale sentences containing the main actors.

ACKNOWLEDGMENT

The authors very much appreciate the support by our colleges from CMVS Laboratory University of Oulu, for the fruitful discussions. This work is partly supported by CBC Karelia IoT Business Creation (2018-2020) and EU YoungRes (#823701) projects.

REFERENCES

[1] Steven Bird, Ewan Klein, and Edward Loper. *Natural language processing with Python: analyzing text with the natural language toolkit.* " O'Reilly Media, Inc.", 2009.

[2] Gobinda G Chowdhury. Natural language processing. *Annual review of information science and technology*, 37(1):51–89, 2003.

[3] David Crystal. A dictionary of linguistics and phonetics. blackwells, 1997.

[4] Rafael Ferreira, Luciano de Souza Cabral, Rafael Dueire Lins, Gabriel Pereira e Silva, Fred Freitas, George DC Cavalcanti, Rinaldo Lima, Steven J Simske, and Luciano Favaro. Assessing sentence scoring techniques for extractive text summarization. *Expert systems with applications*, 40(14):5755–5764, 2013.

[5] Eduard H Hovy, Chin-Yew Lin, Liang Zhou, and Junichi Fukumoto. Automated summarization evaluation with basic elements. In *LREC*, volume 6, pages 604–611. Citeseer, 2006.

[6] Atif Khan, Naomie Salim, and Yogan Jaya Kumar. A framework for multi-document abstractive summarization based on semantic role labelling. *Applied Soft Computing*, 30:737–747, 2015.

[7] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.

[8] Hans Peter Luhn. The automatic creation of literature abstracts. *IBM Journal of research and development*, 2(2):159–165, 1958.

[9] Inderjeet Mani. *Advances in automatic text summarization*. MIT press, 1999.

[10] Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. The stanford corenlp natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, pages 55–60, 2014.

[11] James H Martin and Daniel Jurafsky. *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition*. Pearson/Prentice Hall Upper Saddle River, 2009.

[12] Ani Nenkova and Kathleen McKeown. A survey of text summarization techniques. In *Mining text data*, pages 43–76. Springer, 2012.

[13] Ani Nenkova, Kathleen McKeown, et al. Automatic summarization. *Foundations and Trends® in Information Retrieval*, 5(2–3):103–233, 2011.

[14] Constantin Orăsan. Comparative evaluation of term-weighting methods for automatic summarization. *Journal of Quantitative Linguistics*, 16(1):67–95, 2009.

[15] Martha Palmer, Daniel Gildea, and Paul Kingsbury. The proposition bank: An annotated corpus of semantic roles. *Computational linguistics*, 31(1):71–106, 2005.

[16] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics, 2002.

[17] Thomas E Payne and Thomas Edward Payne. *Describing morphosyntax: A guide for field linguists*. Cambridge University Press, 1997.