Detecting summary-worthy sentences: the effect of discourse features

Maximilian Droog-Hayes
School of Electronic Engineering
and Computer Science
Queen Mary University of London
London, United Kingdom
Email: m.droog-hayes@qmul.ac.uk

Geraint A. Wiggins
AI Lab, Vrije Universiteit Brussel
Belgium & School of
Electronic Engineering and Computer Science
Queen Mary University of London
London, United Kingdom
Email: geraint@ai.vub.ac.be

Matthew Purver
School of Electronic Engineering
and Computer Science
Queen Mary University of London
London, United Kingdom
Email: m.purver@qmul.ac.uk

***Abstract*—We examine the benefit of a variety of discourse and semantic features for the identification of summary-worthy content in narrative stories. Using logistic regression models, we find that the most informative features are those that relate to the narrative structure of a text. We show that automatic methods for feature extraction perform significantly worse than full manual annotation, but that with optimization, a fully automatic approach can outperform a variety of existing extractive approaches to summarization.**

***Keywords*-Discourse Structure, Natural Language Processing, Semantic Analysis, Summarization**

I. Introduction

All automatic text summarization systems need to be able to distinguish the most relevant content from the least relevant, whether they are *extractive* (selecting and concatenating text material to form a summary) or *abstractive* (interpreting a document into an intermediate representation from which a concise and original summary can be generated).

This paper is part of a larger body of work on *abstractive* summarization, and examines the benefits of various discourse features – in particular, knowledge of discourse structure – for the selection of summary-worthy content. We focus on Russian Folktales, due to the extensive work on narrative structure in this genre [1]; but our approach can apply to any genre for which such structural information can be obtained. We show the relative importance of a variety of different discourse features and how their use can greatly improve the detection of summary-worthy content over existing extractive algorithms. The use of automatic methods to detect features such as coreference and narrative structure information significantly reduces performance; but our best automatic system still outperforms extractive alternatives.

II. Related Work

A. Automatic summarisation methods

Most approaches to content selection for summarization are based on surface heuristics and statistics, e.g. word frequency, cue words [2], [3], Term Frequency-Inverse Document Frequency [4] and Latent Semantic Analysis [5]. Other methods go beyond observable surface features to consider meaning or structure in text. One example is *sentiment* analysis, which can help identify major plot points [6]; another is *lexical chains*, sequences of semantically related words, which help identify salient concepts [7]; a third is the use of theories of text structure such as Rhetorical Structure Theory (RST, [8]) [9].

Regardless of the methods used, extractive summarization approaches cannot match human capabilities [10]: even human-formed extractive summaries perform poorly in comparison to natural abstractive human summaries, and the outputs of the best automatic systems from several years ago are already close to the best of what human-extractors produce [11]. To achieve more abstractive summarization, we instead consider the underlying structure of a text. This aims to aid both recognition of key events and generation of summaries.

B. Discourse structure

Aside from explicit approaches such as RST, discourse structure is also implicitly used in many approaches to summarization, abstractive or extractive. Research concerning the structure of stories has repeatedly highlighted the importance of story events and the goals of characters [12], [13], and the relations between characters [14]. One framework that captures these is that of Vladimir Propp [1], who analyzed and annotated the structure of a corpus of Russian Folktales. The primary component of Propp's analysis is a sequence of 31 character-based narrative units, which we call 'Propp functions', general descriptions of key events in a tale and the types of character involved. These functions cover events such as *Villainy* or *Victory*, and many are paired, such as the pursuit of the hero, and the hero's subsequent rescue. Seven distinct roles are enumerated for the characters of the tale (e.g., *Hero* or *Villain*). A folktale can then be represented by the subset of functions which cover the events of that narrative.

Propp's work has been applied to computational story *generation* [15], and there is potential to learn the narrative units [16], the characters who fulfil particular roles (e.g. *Hero*, *Villain*) [17], and to automatically determine both the sequence of Propp functions and the characters who fill the roles [18].

III. Approach

We first discuss the features examined, then describe our dataset and how it was annotated with these features and with ground-truth values for summary-worthiness. Finally, we explain how we determine the relative utility of the features.

A. Annotated Features

Our overall approach to summarization is a semantic, abstractive approach (see [18]). Given this, while we take our task as content selection at a sentence level, our interest here is in features that give a deeper, more human-like, insight into the meaning of a document. Of 43 features examined, 35 were based around Propp's character roles and narrative units; these carry more semantic information than those used by conventional summarization systems. It is worth emphasizing however, that our analysis is not focused on the merits of a particular feature, or the work of Propp. Instead, it focuses on the usage of semantic and structural features in general, and comparing them to the capabilities of more traditional approaches. We consider the following features of each sentence:

a) Sentence Position: The relative position of the sentence within the story, normalised for story length.

b) Number of character mentions: The number of noun-phrases in the sentence that refer to the characters of the story.

c) Number of unique characters mentioned: The number of unique story-characters mentioned in the sentence.

d) Number of lexical chains: The number of unique lexical chains that have items present in the sentence. We implement a lexical chainer as described in [7].

e) Speech: A binary flag indicating whether the sentence contains direct speech or not.

f) Hero and *Villain:* Two binary flags indicating whether the sentence contains a reference to the hero/villain.

g) Propp function weights: 33 features, each corresponding to one of Propp's narrative functions:[1] the story-level score (1 if the story contains the function, else 0) is divided equally between every sentence which represent that function.

h) Number of Propp functions: The total number of Propp functions of which the sentence is part. (Propp functions may overlap, and each span multiple sentences; some sentences may be part of no Propp functions).

i) Desiring verb: A binary feature indicating whether the sentence contains a verb indicating desire (from the VerbNet synsets for verbs such as 'want' and 'long'). Statements of desire by hero/villain often motivate subsequent story events, and so may indicate summary-worthy information.

j) Goal phrase: A binary feature indicating whether or not the sentence contains a goal marker, e.g., 'in order…' or 'so that…'. Like desiring verbs, these phrases indicate purpose and so may signify summary-worthy sentences.

B. Training and Testing Data

Using the features in III-A above, we annotated every sentence in the first 10 Russian folktales annotated by [16]. Stories were manually annotated with Propp's narrative functions according to his own labels [1], using sentence boundaries provided by [19]. Coreference resolution was carried out manually, allowing all other features to be annotated automatically (e.g. determining hero/villain and character mentions).

As well as these manual gold-standard features, we derived automatic annotations for both coreference information (using Stanford's CoreNLP [20] and the neuralcoref [21] extension to spaCy), and Propp's narrative functions (using our own method [18], which derives all valid Propp function assignments and creates a probability distribution over functions for each sentence). This provided multiple versions of our dataset to examine the impact of obtaining these features automatically, given the errors that automatic approaches introduce:

a) Manual: both coreference resolution and the assignment of Propp functions were carried out manually.

b) Auto-CoreNLP: CoreNLP coreference resolution; manual Propp function assignment.

c) Auto-spaCy: neuralcoref [21] coreference resolution; manual Propp function assignment.

d) Auto-Propp: manual coreference resolution; automatic Propp function assignment.

e) Propp-CoreNLP and *Propp-spaCy:* entirely automatic: automatic coreference resolution (CoreNLP or neuralcoref respectively), and automatic Propp function assignment.

C. Ground-Truth Data

We created two sets of ground-truth summary-worthiness labels, one for short summaries and another for longer ones. Summary-worthy sentences were marked by three human assessors, first working separately and then agreeing a consensus. Annotators first read each story in its entirety before annotating its sentences. For the short summary ground-truth, they marked sentences they considered essential to convey the main events of the story. For the long summary ground-truth, they also marked sentences containing information about noteworthy events in the narrative chain of the story.

885 sentences were annotated. Of these, 115 sentences were marked as summary-worthy for short summaries, and 223 for long summaries. The summary-worthy sentences for short summaries are a proper subset of those for long summaries.

D. Method

To evaluate our features, we set up our goal as a classification task: to predict which sentences are marked as summary-worthy by our annotators. We used a logistic regression classifier,[2] as implemented in Weka [22], and evaluated performance via 10-fold class-balanced cross-validation.

To measure performance we use Cohen's Kappa coefficient, which measures agreement taking into account the level expected by chance. (A Kappa score of 1 indicates perfect agreement; -1 indicates complete disagreement; and 0 indicates the agreement expected by chance).

For comparison, we implemented four existing extractive summarization algorithms. Although our overall approach to

[1] We make two additions to the 31 character functions defined by Propp [18]. The first represents his description of an 'Initial Situation'. The second comes from splitting of 'Villainy' and 'Lack' into two distinct functions.

[2] We also tested ZeroR, Naive Bayes, Sequential Minimal Optimization, K-Nearest Neighbours, the PART rule based algorithm, and the decision tree algorithms REPTree and J48. Logistic regression performed best.

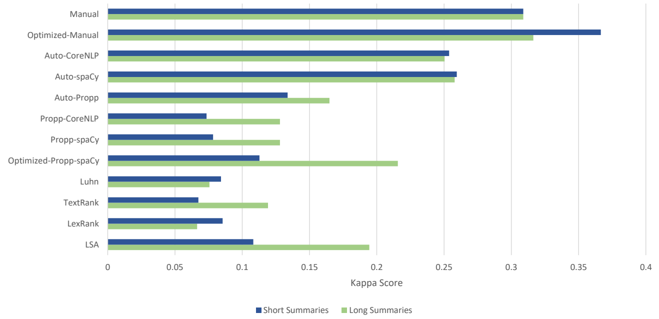

Fig. 1. A comparison of the Kappa scores obtained by all experiments, across both short and long summary data

summarization is abstractive, this work focuses on the prerequisite step of determining summary-worthy content (before using it to generate an original summary), and therefore can be compared to extractive systems (which select summary-worthy content and present it directly). We compare our method against the following extractive approaches: Luhn's original summarization algorithm [2], TextRank [23], LexRank [24], and a Latent Semantic Analysis (LSA) approach [5].

IV. Results

First, we examine the relative utility of our features, using the *Manual* dataset (see Section III-B). We then discuss the effects of automatic feature extraction. Finally, we compare the performance of our approaches with a range of existing extractive summarization algorithms. It is important to understand that the focus of this analysis is *comparative*, and not about the absolute values of our results.

A. Feature Values

To assess the importance of our features, we inspect the coefficient values from our logistic regression models. We are aware of the shortcomings of bald comparisons of such numerical weights, so we use them only to aid a qualitative discussion of the relative utility of the features. Space limitations prevent exhaustive listing of the values.

The coefficients with the greatest absolute values correspond to Propp's narrative functions. For short summaries, the highest valued coefficient is the *Violation* function: this corresponds to (and is paired with) a character disobeying a previous command (the *Interdiction* function), such as going to a forbidden location. We observe that for paired functions the coefficient for the later function always has the higher value. The first function of a pair may even have a negative coefficient; i.e. its presence is an indicator *against* a sentence being summary-worthy. The *Pursuit* and *Rescue* functions are another instance of this. We attribute this to the fact that the meaning of the first function is implicit in the second half, and can therefore be superfluous: for example, rescue implies a prior chase; violation of a command implies the prior stating thereof. This is especially true for our short summary data.

Coefficients for *Villainy* and *Lack* are also relatively high. Propp describes these as very important aspects of the tale, being the means by which the actual movement of the tale is created. The act of villainy or a character's lack (typically expressed as desire for wealth or marriage) are the only elements that Propp explicitly says must always be present.

Certain coefficients with low, or negative, values for the short summary data have higher values for the longer summary data, e.g. Propp functions such as *Initial situation* and *Absentation*. These are background to a tale: unnecessary for summarizing the main points in a short summary, but suitable as additional information in longer summaries. This also holds for features such as *Phrase Goal* and number of noun-phrase mentions: a short summary will focus most on the protagonist and antagonist, while a longer summary will introduce more explanatory sentences and more characters.

Aside from Propp's functions, the benefit of our other discourse features is also evident. The importance of phrase goals is particularly evinced by its coefficient for the prediction of long summary data. This could be due to the inclusion of more explanatory sentences. Summaries generally contain little, if any, speech. This is reflected by the low-negative speech coefficient for both short and long summaries.

B. Automatic Performance

To compare performance based on manual annotations with varying degrees of automatic annotation, we examine the Kappa scores obtained by logistic regressions performed over the various versions of our dataset (see Figure 1; Short and Long bars correspond to accuracy for short summaries and long summaries respectively).

As Figure 1 shows, the best Kappa scores for both short and long summary data are achieved with manually annotated features. As we are most interested in the correct prediction of the small class of positive instances of summary-worthy sentences in this imbalanced dataset, we examined the use of a weighted cost function in our logistic regression classifier. This penalizes the incorrect classification of summary-worthy sentences more harshly in training. We optimized the cost weight parameter for each version of the dataset, to give the best Kappa score over cross-validation. This is shown by the entries prepended with *Optimized* in Figure 1.[3] The 'Auto-spaCy' and 'Auto-CoreNLP' rows show that the use of automatic coreference resolution information causes a reduction in performance, with spaCy performing marginally better than CoreNLP. The 'Auto-Propp' row shows the effect of automatic Propp function assignment: the effect of inaccurate information here has a greater impact than poor coreference information. Finally, 'Propp-spaCy' and 'Propp-CoreNLP' show the results using fully automatic information. As can be seen by the poor performance of fully automatic approaches, some manual annotation is still desirable for this task until the accuracy of the requisite systems has improved.

It is evident from these results that the automatic assignment of Propp functions is more damaging to the formation of short summaries than long summaries. We believe this is

[3] The following numbers indicate the cost that was applied to penalize the misclassification of summary-worthy sentences: Manual+Short 1.7, Manual+Long 2.3, Propp-spaCy+Short 2.3, Propp-spaCy+Long 1.9.

because knowledge about the discourse structure of a story is more critical in the creation of shorter summaries. As shorter summaries are necessarily more condensed and must only cover the most key aspects of a text, it is more important to know where in the text the key narrative events occur.

Several factors help to explain the poor predictive abilities of the automated annotations. One of the stories in our dataset had no valid assignments of narrative functions. This is due to the very strict interpretation of Propp's constraints in order to produce a feasible number of interpretations [18]. As a result, the Propp features for every sentence in this story were given a score of 0 due to the absence of data. Furthermore, errors stemming from the automatic coreference resolution systems can propagate and affect multiple annotated features. Coreference information is used to determine the characters that fulfill the roles of the hero and villain, as well as information about the number of characters mentioned in each sentence of a story.

C. Extractive Comparisons

Figure 1 also shows the performance of the four extractive algorithms. For each algorithm, the cutoff percentage of highest-ranking sentences was optimized over our entire dataset in order to obtain the highest Kappa scores. These were determined separately for each extractive algorithm, and for the prediction of both short and long summary data, in order to show their best possible results.

The results show that the most semantically driven algorithm, LSA, performs best for the prediction of both short and long summaries. It is interesting to note, however, that the relatively simple approach of Luhn compares well against these far more complex and computationally intensive approaches. This supports the claim that there is little further to be gained by research in extractive summarization.

The performance differences between our system and the extractive systems are shown in Figure 1. The benefit of annotating sentences with semantic and discourse features is evident. When using an optimized cost weight in the classifier training, even our automatically annotated data (Optimized Propp-spaCy) outperforms every extractive method examined for the prediction of both short and long summaries. This particular result surprises us, because errors can propagate through our annotations via incorrect coreference information and assignment of Propp's functions. The benefit of these discourse features is even more evident when comparing the results of our optimized manual annotations with the scores for the extractive algorithms.

V. Conclusions and Future Work

Our results show the value of discourse features—in particular, knowledge about discourse structure—for the recognition of summary-worthy content, and the improvements this affords over standard extractive approaches, even when using errorful automatic annotation. Note however that decisions about sentence summary-worthiness are made in isolation; but we expect that results could be improved by considering the context in which a sentence appears. The use of Propp's narrative functions and lexical chains, which both capture aspects of structure which span multiple sentences, mitigate this somewhat; but this effect can be weak, particularly with Propp functions which span multiple sentences and give low values for any individual sentence. We leave this open as an area of future research, along with investigating improvements in automatic coreference resolution and the detection of Propp functions.

References

[1] V. Propp, "Morphology of the folktale. 1928," 1968.

[2] H. P. Luhn, "The automatic creation of literature abstracts," *IBM Journal of Research and Development*, vol. 2, no. 2, pp. 159–165, 1958.

[3] H. P. Edmundson, "New methods in automatic extracting," *Journal of the ACM (JACM)*, vol. 16, no. 2, pp. 264–285, 1969.

[4] J. L. Neto, A. D. Santos, C. A. Kaestner, N. Alexandre, D. Santos *et al.*, "Document clustering and text summarization," 2000.

[5] J. Steinberger and K. Jezek, "Using latent semantic analysis in text summarization and summary evaluation," *Proc. ISIM*, vol. 4, 2004.

[6] A. J. Reagan, L. Mitchell, D. Kiley, C. M. Danforth, and P. S. Dodds, "The emotional arcs of stories are dominated by six basic shapes," *EPJ Data Science*, vol. 5, no. 1, p. 31, 2016.

[7] H. G. Silber and K. F. McCoy, "An efficient text summarizer using lexical chains," in *Proc. Intl. Conf. Natural Language Generation*, 2000.

[8] W. C. Mann and S. A. Thompson, *Rhetorical structure theory: A theory of text organization*. U. Southern Calif., Information Sci. Inst., 1987.

[9] D. Marcu, "Improving summarization through rhetorical parsing tuning," in *The 6th Workshop on Very Large Corpora*, 1998, pp. 206–215.

[10] K. Spärck Jones, "Automatic summarising: The state of the art," *Information Processing & Management*, vol. 43, no. 6, 2007.

[11] P.-E. Genest and G. Lapalme, "Absum: a knowledge-based abstractive summarizer," *Génération de résumés par abstraction*, vol. 25, 2013.

[12] R. C. Schank and R. P. Abelson, *Scripts, Plans, Goals, and Understanding*. Lawrence Erlbaum, 1977.

[13] B. J. Grosz and C. L. Sidner, "Attention, intentions, and the structure of discourse," *Computational Linguistics*, vol. 12, no. 3, 1986.

[14] W. G. Lehnert, "Plot units and narrative summarization," *Cognitive Science*, vol. 5, no. 4, pp. 293–331, 1981.

[15] P. Gervás, "Reviewing Propp's story generation procedure in the light of computational creativity," in *AISB Symposium on Computational Creativity*, 2014.

[16] M. A. Finlayson, "ProppLearner: Deeply annotating a corpus of Russian folktales to enable the machine learning of a Russian formalist theory," *Digital Scholarship in the Humanities*, vol. 32, no. 2, pp. 284–300, 2015.

[17] J. Valls-Vargas, S. Ontanón, and J. Zhu, "Toward character role assignment for natural language stories," in *Proc. 9th Artificial Intelligence and Interactive Digital Entertainment Conference*, 2013, pp. 101–104.

[18] M. Droog-Hayes, G. Wiggins, and M. Purver, "Automatic detection of narrative structure for high-level story representation," in *The 5th AISB Computational Creativity Symposium*, 2018, pp. 26–33.

[19] M. A. Finlayson *et al.*, "Supplementary materials for "ProppLearner: Deeply annotating a corpus of Russian folktales to enable the machine learning of a Russian formalist theory"," 2015.

[20] C. D. Manning, M. Surdeanu, J. Bauer, J. R. Finkel, S. Bethard, and D. McClosky, "The Stanford CoreNLP Natural Language Processing Toolkit." in *ACL (System Demonstrations)*, 2014, pp. 55–60.

[21] T. Wolf, "State-of-the-art neural coreference resolution for chatbots," Jul 2017. [Online]. Available: https://medium.com/huggingface/state-of-the-art-neural-coreference-resolution-for-chatbots-3302365dcf30

[22] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: an update," *SIGKDD Explorations*, vol. 11, no. 1, pp. 10–18, 2009.

[23] R. Mihalcea and P. Tarau, "Textrank: Bringing order into text," in *Proc. Empirical Methods in Natural Language Processing*, 2004.

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