Language Features in Extractive Summarization: Supplementary Materials

Ignacio Arroyo-Fernández, Arturo Curiel,
Carlos-Francisco Méndez-Cruz

October 2018

1 Selected Summarizers

An assumption of our study is the fact that different qualities of summaries could exist among machine-made ones. Therefore, we included two groups of summarizers. The first one is composed by five baseline systems showed in Table 1. The second group is composed by the seven state-of-the-art systems showed in Table 2. The separation of these groups was made according to the ranking provided in [1]. In that study, the authors compared the ROUGE-{1-2-SU4} scores all these summarizers reached via their produced summaries. The lower scores are provided by baseline systems and the higher scores by state-of-the-art systems. Thus, we briefly describe these systems with emphasis in the features they use to summarize.

Baseline summarizers are either classical summarization methods or relatively simple approaches to build an automatic summary. For example, *Continuous LexRank* is a classical graph based approach for multidocument summarization following the concept of *sentence centrality*, i.e., to generate a summary by extracting the *central* sentences of the source documents. LexRank determines central sentences by using the PageRank algorithm over a graph of sentences that is built using the cosine similarity among them[2]. As cosine

Table 1: ROUGE scores (%) for baseline summarizers according to [1]

System	Rouge-1	Rouge-2	Rouge-SU4
Continuous LexRank	35.95	7.47	0.82
Centroid	36.41	7.97	1.21
FreqSum	35.3	8.11	1.00
TsSum	35.88	8.15	1.03
Greedy-KL	37.98	8.53	1.26
Average	36.30	8.04	1.064

Table 2: ROUGE scores (%) for state-of-the-art summarizers according to [1]

System	Rouge-1	Rouge-2	Rouge-SU4
CLASSY 04	37.62	8.96	1.51
CLASSY 11	37.22	9.2	1.48
Submodular	39.18	9.35	1.39
DPP	**39.79**	9.62	1.57
RegSum	38.57	9.75	1.6
OCCAMS_V	38.5	9.76	1.33
ICSISumm	38.41	**9.78**	**1.73**
AV	38.47	9.48	1.51

similarity is calculated over the $N-$dimensional vector representation of the sentences, where N is the number of all different words in the source documents. The value of each dimension is the *TF-IDF weight* of the corresponding word, which is calculated as the frequency of the word times the percentage of documents containing that word. Thus, this approach only takes into consideration information of lexical items (word frequency) to summarize.

Centroid is also a sentence centrality approach for multidocument summarization that only considers information of lexical items. First, source documents are transformed into vector representation of TF-IDF weights to be clustered according to certain similarity measure [3]. Conceptually, these clusters correspond to different topics the source documents talk about. The approach determines one centroid for each cluster, which is a set of relevant words for the topic. Finally, central sentences are obtained using these centroids.

FreqSum is a classical approach based on word frequency as the main feature for determining relevance. Then, following some criteria, sentences containing relevant words are selected as relevant sentences [4]. Similarly, *TsSum* compares distribution of words of the source documents and of a large corpus to weight words with a log-likelihood ratio [5].

The last baseline method considered in our study is called *Greegy-KL*. It minimizes the Kullback–Leibler divergence between the probability distribution of words from the automatic summary and the source documents [6].

State-of-the-art summarizers utilize more elaborate approaches to summarize. *CLASSY 04* employs a Hidden Markov Model (HMM), using the probability of topic words, to get the probability of a sentence is part of the automatic summary [7]. *CLASSY 11* is a successor of this method. It also uses an HMM, but it estimates the probability that a pair of words (a bigram) occur in a manual summary [8].

Submodular treats the problem of multidocument summarization as a monotone submodular function maximization using word frequency as a main feature [9]. The method called Determinantal Point Process *DPP* is a probabilistic model that optimizes the ROUGE score and learn parameters for several fea-

tures extracted from a dataset of human-made summaries [10]. These features are: length of sentence in characters, sentence position in source document, personal pronouns, mean cluster similarity and LexRank scores. The lengths and the positions of sentences were quite important features to improve the final score.

RegSum makes use of a supervised approach to predict word importance taking into consideration several features: position, part-of-speech, name entities, subjectivity (polarity and intensity), topic categories and word context [11]. In that work, authors analyzed abstract-article pairs to identify preserved words in order to predict those from the source documents that appear in the abstracts. This work is specially interesting for us because the authors characterized human-made summaries with respect to preserved features from source documents. First, the characterization was performed from the point of view of lexical items. Second, from the point of view of both language features and sentiment analysis. The authors found that abbreviations, Persons, Organizations and Locations tend to be selected for abstracts whereas Time and Date words do not. Also, there are more nouns (NN, NNS, NNPS) in abstracts, but fewer verbs (VBG, VBP, VB) and cardinal numbers compared to the source document. Concerning subjectivity, words with strong polarity, either positive or (especially) negative, do not tent to appear in abstracts. In addition, those authors report that events about conflict, death, anger, achievements, money and negative emotions frequently appear in abstracts. Those that infrequently appear include auxiliary verbs, hearing (heard, listen, sound), pronouns, negation, function words, social words (friend, family), swear words and adverbs. It is remarkable that although this summarizer takes into account a number of interesting aspects, its performance is not much better than that of much less complex summarizers.

OCCAMS_V combines a latent semantic analysis (LSA) to learn the latent distribution of words in topics from the source documents and greedy heuristics to select relevant sentences based on two combinatorial problems [12]. Differently, *ICSISumm* employs a global linear optimization approach to find the globally optimal summary instead of greedily select relevant sentences. This approach focuses on summarize the relevant concepts expressed by word n-grams.

2 The general set of features

Here we show a table including the feature labels used in the paper and their descriptions. This is a general set features (See table 3), which was pruned via the Feature Spectrum to obtain the final set of pertinent features. This latter set of features was studied deeper via boxplots and hypothesis testing.

Table 3: Features showing tendency to help the discrimination between human and machine summaries in our first analysis

Feature description	Label
Preposition, average	IN_AVG
Preposition, median	IN_MED
Preposition, minimum	IN_MIN
Preposition + Proper noun	IN_NNP
Adjective, average	JJ_AVG
Named entity mentions	named_entities
Proper Noun	NNP
Proper Noun average	NNP_AVG
Proper Noun + Preposition	NNP_IN
Proper Noun, maximum	NNP_MAX
Proper Noun median	NNP_MED
Proper Noun, minimum	NNP_MIN
Proper Noun + Noun	NNP_NN
Proper Noun + Proper Noun	NNP_NNP
Proper Noun + Possessive ending	NNP_POS
Proper Noun Plural	NNPS
Proper Noun Plural, average	NNPS_AVG
Proper Noun + Verb 3rd person singular present	NNP_VBZ
Noun plural median	NNS_MED
Noun plural + period	NNS_PERIOD
Number of coreferences	no_corefs
Number of openIE triplets	no_openie
Number of tokens	no_tokens
Positive sentiment of sentence	sent_sentim_3
Neutral sentiment of sentence	sent_sentim_2
Negative sentiment of sentence	sent_sentim_1
Positive sentiment of token	tok_sentim_3
Neutral sentiment of token	tok_sentim_2
Negative sentiment of token	tok_sentim_1
Total RST tags	TotalRST
Verb past tense	VBD
Verb past tense, average	VBD_AVG
Verb past tense, maximum	VBD_MAX
Verb past tense, median	VBD_MED
Verb past tense + Proper noun	VBD_NNP
Coordinating conjunction	CC
Cardinal number, maximum	CD_MAX
Cardinal number, median	CD_MED

3 The Feature Spectrum

The Feature Spectrum is a method taken from applied engineering (acoustics and communications). In this area, engineers frequently need of comparing a measured level of power/voltage against a reference power/voltage level [13]. This comparison is defined as a base-10 log ratio, so it is said that such ratio is expressed in dBs (decibels). In our extrapolation of the definition of this ratio, we measured the ratio between a statistic on the frequency of occurrence of a feature in two sets of summaries. In this sense, it is difficult to talk about of power/voltage levels. Rather, we propose uniquely a reinterpretation of the ratio and the measurements or levels it involves, i.e., (i) the reference level is a location statistic of the distribution of the human-made summaries, (ii) the measured level is assigned to the same statistic of a set of machine-made summaries, and (iii) the *feature ratio* is the base-10 log difference between human-made summaries and a given machine-made summary under interest. Formally, given a language feature f_i, this feature is extracted from a set of human-made summaries $h_1, \ldots, h_{N_h} \in H$, as well as from a set of machine-made summaries $m_1, \ldots, m_{N_m} \in M$. Thus, the feature ratio F_i can be defined as 20 times the log ratio between the reference level $\varphi_H(f_i)$ of the feature f_i in the human-made summaries and the measured level $\varphi_M(f_i)$ of the same feature of the summaries produced by a machine under interest [14], i.e.:

$$F_i = 20 \log \frac{1 + \varphi_M(f_i)}{1 + \varphi_H(f_i)}. \tag{1}$$

In the definition (1) of the feature ratio, both the measured level $\varphi_M(f_i)$ and the reference level $\varphi_H(f_i)$ are functions of the frequency of occurrence of f_i in the sets of summaries M and H, respectively. It is not a requirement that M to come from summaries generated with a unique machine. Grouping machine-made summaries according to any criterion to build M and then comparing the resulting groups with the reference level is also a valid approach (this is what we followed in this work).

If we have a feature histogram both on some M and some H, then we define the Feature Spectrum as the set:

$$\mathcal{F} = \{F_1, \ldots, F_N\}, \tag{2}$$

where N is the number of features under observation. Overall, there are three main interpretations of the ratios F_i of the Feature Spectrum $\mathcal{F}$ in terms of positive, negative and (nearly-)zero values. Namely,

1. **Equal use** if the machine-made summaries hold no difference with the human-made summaries with respect to f_i, then $\varphi_M(f_i) = \varphi_H(f_i)$ and thus $F_i = 20 \log 1 = 0$.

2. **Lack of use** if $F_i < 0$, then the machine-made summaries show *lack of use* of a feature f_i with respect to the human-made summaries.

3. **Excessive use** if $F_i > 0$, then the machine-made summaries show *excessive use* of feature f_i with respect to the human-made summaries.

To determine the values of the levels $\varphi_H(f_i)$ and $\varphi_M(f_i)$, which are functions of the feature frequencies, we selected the median of the frequency of occurrence of each f_i in each group of samples of human- and machine-made summaries. That is,

$$\begin{aligned} \varphi_H(f_i) &= \text{median}\{h^i_1, \ldots, h^i_{N_h}\} \\ \varphi_M(f_i) &= \text{median}\{m^i_1, \ldots, m^i_{N_m}\}, \end{aligned} \tag{3}$$

where $h^i_{(\cdot)}$ and $m^i_{(\cdot)}$ are the frequencies of occurrence of f_i in each of the human-made and machine-made summaries contained in H and M, respectively. The medians of Eq. (3) are suited for nonparametric tests and help us keeping $\varphi_M(f_i)$ and $\varphi_H(f_i)$ insensitive to outliers.

Let us show a couple of specific examples for the interpretation of the values of ratios F_i. By isolating $\varphi_M(f_i)/\varphi_H(f_i)$ in Eq. (1) we revert the log scale as $10^{\frac{F_i}{20}}$. Thus, in the case we have the ratio $F_i = 6$ it means that feature f_i is used twice more by machines than by humans, e.g. $10^{\frac{6}{20}} \approx 2$. Negative values are interpreted contrarily. For instance, if $F_i = -6$ it means that f_i is used twice less ($10^{\frac{-6}{20}} \approx 1/2$) by machines than by humans. These statements do not depend directly on the exact levels being compared, $\varphi_H(f_i)$ and $\varphi_M(f_i)$, but rather on their ratio $\varphi_M(f_i)/\varphi_H(f_i)$.

It is needed to say that the ambiguity in $F_i = 0$ does not affect the generality of our results. There are two main reasons for that. In our experiments, we used randomly sampled human-made summaries coming from different text domains and from multiple human annotators. Furthermore, we performed the measurements on two independent groups of random summaries: fitting and development groups. Thus, we computed the median of the frequency of occurrence of each f_i in all these randomly sampled human-made summaries (the same computations was made for machine-made summaries), i.e. $\varphi_H(f_i) = \text{median}\{h^i_1, \ldots, h^i_{N_h}\}$ and $\varphi_M(f_i) = \text{median}\{m^i_1, \ldots, m^i_{N_m}\}$, where h^i_1 and m^i_1 are the frequencies of occurrence of f_i in the human-made and machine-made summaries H and M, respectively. The second reason is that the aim of using the Feature Spectrum is to observe the general differences in language use of humans and machines. Therefore, given the randomness, variety, text domains of summaries and annotators, we adopt the uniqueness of the meaning of $\mathcal{F}_i = 0$ so as to be the equality between humans and machines with respect to their use of f_i.

References

[1] K. Hong, J. M. Conroy, B. Favre, A. Kulesza, H. Lin, A. Nenkova, A repository of state of the art and competitive baseline summaries for generic news summarization., in: LREC, 2014, pp. 1608–1616.

[2] G. Erkan, D. R. Radev, Lexrank: Graph-based lexical centrality as salience in text summarization, Journal of Artificial Intelligence Research 22 (2004) 457–479.

[3] D. R. Radev, H. Jing, M. Styś, D. Tam, Centroid-based summarization of multiple documents, Information Processing & Management 40 (6) (2004) 919–938.

[4] A. Nenkova, L. Vanderwende, K. McKeown, A compositional context sensitive multi-document summarizer: exploring the factors that influence summarization, in: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval, ACM, 2006, pp. 573–580.

[5] C.-Y. Lin, E. Hovy, The automated acquisition of topic signatures for text summarization, in: Proceedings of the 18th conference on Computational linguistics-Volume 1, Association for Computational Linguistics, 2000, pp. 495–501.

[6] A. Haghighi, L. Vanderwende, Exploring content models for multi-document summarization, in: Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Association for Computational Linguistics, 2009, pp. 362–370.

[7] J. M. Conroy, J. D. Schlesinger, J. Goldstein, D. P. O'leary, Left-brain/right-brain multi-document summarization, in: Proceedings of the Document Understanding Conference (DUC 2004), 2004.

[8] J. M. Conroy, J. D. Schlesinger, J. Kubina, P. A. Rankel, D. P. O'Leary, Classy 2011 at tac: Guided and multi-lingual summaries and evaluation metrics., TAC 11 (2011) 1–8.

[9] H. Lin, J. Bilmes, A class of submodular functions for document summarization, in: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1, Association for Computational Linguistics, 2011, pp. 510–520.

[10] A. Kulesza, B. Taskar, et al., Determinantal point processes for machine learning, Foundations and Trends® in Machine Learning 5 (2–3) (2012) 123–286.

[11] K. Hong, A. Nenkova, Improving the estimation of word importance for news multi-document summarization, in: Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, 2014, pp. 712–721.

[12] S. T. Davis, J. M. Conroy, J. D. Schlesinger, Occams–an optimal combinatorial covering algorithm for multi-document summarization, in: Data

Mining Workshops (ICDMW), 2012 IEEE 12th International Conference on, IEEE, 2012, pp. 454–463.

[13] R. F. Coughlin, F. F. Driscoll, Operational Amplifiers and Linear Integrated Circuits, 3rd Ed., Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1987.

[14] I. Arroyo-Fernández, J.-M. Torres-Moreno, G. Sierra, L. A. Cabrera-Diego, Automatic text summarization by non-topic relevance estimation, in: Proceedings of the 8th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management - Volume 1: KDIR, (IC3K 2016), INSTICC, ScitePress, 2016, pp. 89–100.