Automatic extractive summaries using supervised learning approach

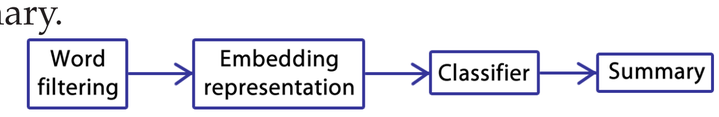
SANDRA J. GUTIÉRREZ-HINOJOSA; HIRAM CALVO; MARCO A. MORENO-ARMENDÁRIZ
sandyguh04@gmail.com, hiramcalvo@gmail.com, mam.armendariz@gmail.com

INTRODUCTION

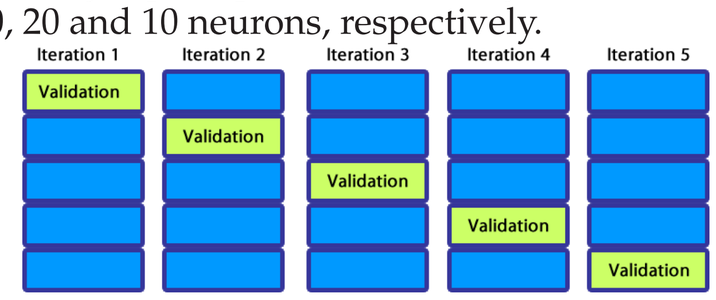
The main goal of a summary is to find the main ideas in a document reducing the original documents size; algorithms created for solving this task have a relevant application given the exponential growth of textual information online, and the need to find the main ideas of documents in a shorter time. In order to perform this task automatically there are two different approaches: extracting the main sentences from the documents or paraphrasing the main ideas. Abstractive methods are highly complex as they need to simulate human cognitive process for generating summaries (Gupta y Lehal, 2010). Therefore, research community is focusing more on extractive summaries, trying to achieve more coherent and meaningful summaries (Gambhir y Gupta, 2017). In this work multi-document, extractive summaries have been obtained using supervised learning algorithms. The supervised learning has three phases: training, validation and testing, each one requires labeled data, i.e., examples of documents and the sentences which belong to the summary. Also, we used cross-validation method that divides the data in n folds and in each iteration the validation fold is a different one and the remaining folds are for training. It is important to note that testing data is used neither in training nor in the validation (Mohri M. y Talwalkar, 2018). Furthermore, we used sentence embeddings as text representation; these models are a spatial representation of word meaning and rely on two ideas: *the geometric metaphor of meaning* and *the distributional hypothesis*. The core idea of the geometric metaphor of meaning is that semantic similarity can be represented as proximity in n-dimensional space and the distributional hypothesis is referred to distributional methodology where a set of facts between the basic entities of language (phonemes, morphemes and syntactic units) and their relations are established; the members of the basic classes of these entities behave distributionally similar, and therefore can be grouped according to their distributional behavior (Sahlgren, 2008).

METHODOLOGY

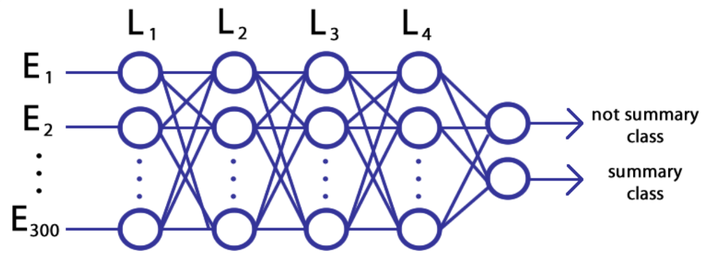
In this work multi-document, extractive summaries have been obtained using supervised learning algorithms in the well-known DUC 2002 corpus. The methodology has three steps: the pre-processing step which filters irrelevant words and reduces vocabulary using stemming, the representation step which transforms sentences into vectors and the classification step which select sentences for the summary.



We used a pre-trained sentence embeddings, for further details in the embedding settings refer to (Lau y Baldwin, 2016). In the supervised learning 80 % of the data was used for training and the remaining for testing. Also, we used k-fold cross validation with $k = 5$ in the training phase, which means that the validation fold changes every iteration and the algorithms for classifications were Gaussian Naïve Bayes (GNB), Bernoulli Naïve Bayes (BNB), multilayer perceptron with hyperbolic tangent (MLP-tanh), logistic sigmoid (MLP-log) and rectified linear unit (MLP-relu), each multilayer perceptron has four layers with 100, 50, 20 and 10 neurons, respectively.



The performance measures were recall, precision, accuracy and F1 score for the classifiers, which contains information about the number of misleading classification on each class, whilst ROUGE-n measure, based on n-gram overlapping, was used as intrinsic quality of the summaries.



EXPERIMENTAL RESULTS

The following tables show the performance of five classifiers and the summary quality of each one. We select the classifier with best accuracy performance in the validation phase and then verify its performance measuring the summary quality.

Summary quality performance:

Classifier	ROUGE-1	ROUGE-2	F1-score
GNB	0.2789	0.0831	0.3386
BNB	0.4297	0.1737	0.4389
MLP-tanh	0.3854	0.1160	0.3887
MLP-log	***0.4342***	***0.1530***	***0.4095***
MLP-relu	0.3834	0.1497	0.3963

Accuracy performance classifiers:

Classifier	Training	Validation	Testing
GNB	0.6980	0.6977	0.6995
BNB	0.8432	0.8410	0.8424
MLP-tanh	***0.9928***	0.9297	0.9240
MLP-log	0.9592	***0.9592***	***0.9522***
MLP-relu	0.9963	0.9444	0.9399

Recall performance classifiers:

Classifier	Training	Validation	Testing
GNB	0.5317	0.5316	0.5312
BNB	0.5434	0.5361	0.5314
MLP-tanh	0.9736	0.5495	0.5411
MLP-log	0.4796	0.4796	0.4761
MLP-relu	***0.9848***	***0.5664***	***0.5533***

Precision performance classifiers:

Classifier	Training	Validation	Testing
GNB	0.6748	***0.6746***	***0.6463***
BNB	0.6420	0.6149	0.5819
MLP-tanh	0.9319	0.5487	0.5327
MLP-log	0.5000	0.5000	0.5000
MLP-relu	***0.9678***	0.5362	0.5203

F1 score performance classifiers:

Classifier	Training	Validation	Testing
GNB	0.5948	***0.5946***	***0.5831***
BNB	0.5886	0.5727	0.5555
MLP-tanh	0.9523	0.5491	0.5369
MLP-log	0.4896	0.4896	0.4877
MLP-relu	***0.9762***	0.5508	0.5363

CONCLUSIONS

We find that the classifier performance is not related to the summary quality because different measures were used to quantify summary quality and classifier performance; sentence embeddings and word overlapping were used in classifier task and summary quality, respectively. It is important to highlight that n-gram overlapping is the basis of measures for intrinsic summary quality, such as ROUGE-n, which is relevant while comparing performance between different works in the state of the art. We believe that using word embedding combined with n-grams as inputs to the classifiers is an interesting direction for further research.

ACKNOWLEDGEMENTS

The authors would like to thank government of México, Instituto Politécnico Nacional, SNI, SIP-IPN, COFAA-IPN, BEIFI-IPN and CONACyT.

BIBLIOGRAPHY

Gambhir, Mahak y Vishal Gupta (2017). "Recent automatic text summarization techniques: a survey". En: *Artificial Intelligence Review* 1, págs. 1-66.

Gupta, Vishal y Gurpreet Singh Lehal (2010). "A survey of text summarization extractive techniques". En: *Journal of emerging technologies in web intelligence* 2.3, págs. 258-268.

Lau, Jey Han y Timothy Baldwin (2016). "An empirical evaluation of doc2vec with practical insights into document embedding generation". En: *arXiv preprint arXiv:160 7.05368*.

Mohri M., Rostamizadeh A. y A. Talwalkar (2018). *Foundations of machine learning*. MIT Press.

Radev, Dragomir R y col. (2004). "Centroid-based summarization of multiple documents". En: *Information Processing & Management* 40.6, págs. 919-938.

Sahlgren, Magnus (2008). "The distributional hypothesis". En: *Italian Journal of Disability Studies* 20, págs. 33-53.