SABio: An Automatic Portuguese Text Summarizer Through Artificial Neural Networks in a More Biologically Plausible Model

Télvio Orrú[1], João Luís Garcia Rosa[2], and Márcio Luiz de Andrade Netto[1]

[1] Computer Engineering and Industrial Automation Department,
State University of Campinas - Unicamp, Campinas, São Paulo, Brazil
{telvio, marcio}@dca.fee.unicamp.br
[2] Computer Engineering Faculty - Ceatec,
Pontifical Catholic University of Campinas - PUC-Campinas,
Campinas, São Paulo, Brazil
joaoluis@puc-campinas.edu.br

Abstract. An implementation of a computational tool to generate new summaries from new source texts in Portuguese language, by means of connectionist approach (artificial neural networks) is presented. Among other contributions that this work intends to bring to natural language processing research, the employment of more biologically plausible connectionist architecture and training for automatic summarization is emphasized. The choice relies on the expectation that it may lead to an increase in computational efficiency when compared to the so-called biologically implausible algorithms.

1 Introduction

The wave of information and the lack of time to read long texts make the modern society to look for abstracts and news headlines instead of complete texts. Towards an automatic text summarization system for Portuguese language, the system SABio (A*utomatic* S*ummarizer for the Portuguese language with more* Bio*logically plausible connectionist architecture and learning*) - a connectionist system employing more biologically plausible training algorithm and architecture - is proposed. An unique architecture for this application is presented, considering important features such as: (a) neural network training with more biologically plausible treatment; (b) training set (input-output pairs) is formed by features of source texts and ideal extracts, represented by elements of a multi-valued logic.

In order to present the system SABio, this paper is organized in the following sections: 1 - *Introduction*: this section; 2 - *Automatic Summarization*: brings fundamental concepts about text summarization; 3 - SABio: the proposed system is presented; 4 - *Evaluation of* SABio: comparisons between biologically plausible and implausible connectionist training algorithms for SABio are highlighted; and 5 - *Conclusion*.

R. Vieira et al. (Eds.): PROPOR 2006, LNAI 3960, pp. 11–20, 2006.

2 Automatic Summarization

Text summarization is the process of production of a shorter version of a source text [1]. It is possible, through this process, to obtain an extract or an abstract. *Extracts* are created by the juxtaposition of source text sentences considered relevant; while *abstracts* alter the structure and/or content of original sentences, merging them and/or rewriting them, to generalize or specify information [2].

There are two points of view of a summary: the reader's (summary user) and the writer's (summary creator). The latter has the task of condensing a source text in a way that it is possible to transmit its main idea with the created summary.

From computational standpoint, three basic operations can describe the summarization process: analysis, content selection, and generalization [3]. An automatic system capable to make a condensation[1], with the preservation of the more relevant content of the source text, can be called a system for automatic text summarization.

An Artificial Neural Network (ANN) [4] is employed in the proposed system. Many systems use this computational tool, e.g.:

- Pardo et al. [5], in *NeuralSumm*, use an ANN of type SOM (self-organizing map), trained to classify sentences from a source text according to their importance degree[2] and then produce the related extract. The SOM network organizes information into similarity clusters based on presented features;
- Aretoulaki [3] uses an ANN with training considered biologically implausible (through supervised algorithm Back-propagation), differently from the approach proposed here (ANN with training considered more biologically plausible). Aretoulaki makes a detailed analysis of journalistic and scientific texts from several domains to recognize satisfactorily generic features that represent the sentence content of any textual type and that can be learned by a neural network. He considers features from several sources, from superficial to pragmatic, that can be identified in a superficial textual analysis.

There are automatic text summarizers that employ techniques to discover the more important sentences in source text [6, 5]. Such approaches are, in general, statistical because they try to organize sentences according to the frequency of their words in the text they belong. The sentences containing the more frequent words are called *gist sentences* and express the text main idea.

In order to evaluate automatic text summarizers, some rates are considered:

- *Precision Rate*: the amount of correctly selected sentences divided by the total amount of selected sentences;

[1] *Condensation* is the act of making something shorter (the *abstract*), according to Longman Dictionary.

[2] It is believed that sentences containing more frequent terms in source texts could present a greater importance in text.

- *Recall Rate*: the amount of correctly selected sentences divided by the total amount of correct sentences;
- *F-Measure*: two times the product of precision rate and recall rate divided by the sum of precision rate and recall rate [7].

Notice that the bigger the F-Measure, the better the generated extract, since it takes into consideration both recall and precision rates.

3 SABio

There are several implemented systems that propose automatic text summarization employing available corpora. These systems use symbolic as well as connectionist approaches. Implementations employing ANNs with training considered biologically implausible are often found, differently from this proposal.

Here, the system SABio (A*utomatic* S*ummarizer for the Portuguese language with more* BIO*logically plausible connectionist architecture and learning*) is proposed. The system produces extracts through a more biologically plausible training with ANNs.

The SABio proposal is partly motivated by the increasing interest of modern society in search of newspaper and magazines headlines instead of complete texts, mainly because of the lack of time of people nowadays.

3.1 The Biological Plausibility

According to O'Reilly and Munakata [8], there are evidences that the cerebral cortex is connected in a bi-directional way and distributed representations prevail in it. So, more biologically plausible connectionist (ANN) models should present some of the following characteristics:

- *Distributed representation*: generalization and reduction of the network size can be obtained if the adopted representation is distributed (several units for one concept, and similar concepts sharing units), since connections among units are able to support a large number of different patterns and create new concepts without allocation of new hardware;
- *Inhibitory competition*: the neurons that are next to the "winner" receive a negative stimulus, this way strengthening the winner neuron. In the nervous system, during a lateral inhibition, a neuron excites an inhibitory interneuron that makes a feed-back connection on the first neuron [9];
- *Bi-directional activation propagation*: the hidden layers receive stimuli from input and output layers. The bi-directionality of the architecture is necessary to simulate a biological electrical synapse, that can be bi-directional [9, 10];
- *Error-driven task learning*: in algorithm GeneRec - Generic Recirculation [11], the error is calculated from the local difference in synapses, based on neurophysiological properties, different from algorithm Error Back-propagation, which requires the back-propagation of error signals [12].

3.2 GeneRec: A Training Algorithm Considered More Biologically Plausible

The algorithm GeneRec - Generic Recirculation - was developed by O'Reilly [11] based on Back-propagation but considering properties of a more biologically plausible artificial neural network training algorithm.

GeneRec employs two phases: "minus" and "plus":

- *Minus Phase*: When units are presented to the input layer there is the propagation of this stimulus to hidden layer (bottom-up propagation). At the same time, the previous output propagates from the output layer to the hidden layer (top-down propagation). Then the "minus" hidden activation is generated (sum of bottom-up and top-down propagations). Finally, the real output is generated through the propagation of the "minus" hidden activation to the output layer. Notice that the architecture is bi-directional.
- *Plus Phase*: Units are presented again to the input layer; there is the propagation of this stimulus to hidden layer (bottom-up propagation). At the same time, the desired output propagates from the output layer to the hidden layer (top-down propagation). Then the "plus" hidden activation is generated, summing bottom-up and top-down propagations [11].

In order to make learning possible, the synaptic weights are updated, based on "minus" and "plus" hidden activations, real and desired outputs, input, and the learning rate [11, 4].

3.3 SABio Features

SABio's artificial neural network was trained with sentences[3] from a corpus called *TeMário*[4] that contains 100 journalistic texts in Portuguese language, with 61,412 words. 60 texts belong to the on-line Brazilian newspaper *Folha de São Paulo*[5], and the remaining 40 texts were published in *Jornal do Brasil* newspaper[6], also on-line version. These texts are distributed amongst distinct domains: opinions, critiques, world, politics, and foreign affairs [13].

The corpus *TeMário* is composed of source texts, manual summaries (abstracts), and ideal extracts. Manual summaries are produced by the authors of the source texts, after a rewritten process of the content "judged" more relevant.

Manual summaries are not employed in SABio because: (*a*) there is no explicit correspondence between source texts and manual summaries; and (*b*) their production is costly and long-lasting, since it requires human beings. Instead, SABio uses the ideal extracts[7] available in corpus *TeMário*. They are automatically produced by a system called GEI through the employment of the

[3] Two thirds of texts were used for training and one third for testing.
[4] http://www.nilc.icmc.usp.br/nilc/tools/corpora.htm
[5] http://www.folha.uol.com.br/
[6] http://jbonline.terra.com.br/
[7] Here, *ideal summaries* and *ideal extracts* are considered the same, although there are differences between them.

cosine measure [14]: for each sentence of the manual summary, the correspondent sentence in the most similar source text is searched. This is done by word co-occurrence: the larger the number of words from the manual summary a source text sentence has, the greater its chance of presenting the same content of the summary sentence. This way, it could be employed to compose the ideal summary [5].

SABio presents satisfactory results, although it has employed only lexical processing and cue words, and not dealt with other linguistic information, such as syntactic and semantic analysis, in order to develop a summary.

TeMário was chosen among several corpora, because it would be easier to make comparisons in the future with other summarizers that also employ this corpus. It was considered also that international automatic summarizer evaluations, like SUMMAC - text SUMMArization evaluation Conference - and DUC - Document Understanding Conference - have employed journalistic texts.

In training step, for each source text sentence[8], seven features that represent the input sentences are analysed. These coded sentences are associated to a desired output, which is classified according to the importance degree of the sentence in relation to the ideal extract. The possible values are: *None*, *Small*, *Small-Medium*, *Medium*, *Medium-Large*, and *Large* frequency. The classification of the frequency for the sentences that represent the desired outputs is obtained through the method of the *gist sentence*[9] discovery.

For every sentence all word letters are converted to upper case, for uniformity [15]. The employed features are [5]:

1. *Size of the sentence*: long sentences often present greater informative content, considered more relevant for the text [16]. The size of a sentence is calculated taking into account the amount of words that belong to it, except the words belonging to the StopList[10]. The sentences in the text are classified as: *Small*, which represents the size of the smaller sentence in the text; *Medium*, representing the average-size sentence, or *Large*, which represents the size of the larger sentence in the text;
2. *Sentence position in text*: the position of the sentence can indicate its relevance [3]. In SABio, similar to NeuralSumm [5], it is considered that a sentence can be in the beginning (first paragraph), at the end (last paragraph), or in the middle (remaining paragraphs) of the text;
3. *Sentence position in paragraph where it belongs*: the position of the sentence in the paragraph can also indicates its relevance [18]. In SABio, it is considered that a sentence can be in the beginning (first sentence), at the end (last sentence), or in the middle (remaining sentences) of the paragraph;

[8] In SABio the end of a sentence can be indicated by conventional punctuation marks: period, exclamation mark, or question mark.

[9] To know which is the *gist sentence*, the approach mentioned in GistSumm [6] is used. In this approach, a sentence is positioned according to the importance degree it represents in the text where it belongs.

[10] *StopList* is a list of very common words or words considered irrelevant to a text, mainly because of their insignificant semantic values [17].

4. *Presence of gist sentence words in the sentence*: sentences that contain words of the gist sentence, that is, a sentence that better expresses the text main idea, tends to be relevant [6];
5. *Sentence value based on the distribution of words in the text*: sentences with high value often are relevant to the text [19]. The value of each sentence is calculated by the sum of the occurrence number of each one of its words in the whole text divided by the number of the words in the sentence, and the obtained result in this operation will be related to the values mentioned in the first feature (Small, Medium, and Large);
6. *TF-ISF of the sentence*: sentences with high value of TF-ISF (Term Frequency - Inverse Sentence Frequency) are representative sentences of the text [20]. For each word of a sentence, the TF-ISF measure is calculated by the formula:

$$TF{-}ISF(w) = F(w) \times \frac{log(n)}{S(w)}$$

where
F(w) is the frequency of the word w in the sentence,
n is the number of words in sentence in which w belongs, and
S(w) is the number of sentences in which w appears.
The TF-ISF value of a sentence is the average of the TF-ISF values of each one of its words, and the obtained result of this equation will be related to the values mentioned in first feature (Small, Medium, and Large);
7. *Presence of indicative words in the sentence*: indicative words (cue words) often indicate the importance of the sentence content [21]. This feature is the only dependent on language, genre, and text domain. SABio uses the same indicative words used in NeuralSumm [5]: evaluation, conclusion, method, objective, problem, purpose, result, situation, and solution.

4 Evaluation of SABio

The comparisons between the training algorithms Back-propagation and GeneRec for SABio were conducted concerning the number of epochs necessary for convergence, processing time, and quality of the generated summary.

The first comparison aims to show the necessary time for the ANN convergence with the training algorithms Error Back-propagation [12] and GeneRec [11]. In order to achieve this, tests were conducted with learning rates 0.10, 0.25, 0.35, or 0.45 and also tests with different numbers of neurons in hidden layer: 10, 11, 12, 20, or 25.

It was observed that when SABio's ANN was trained with GeneRec, the minimum error[11] could be reached within a smaller number of epochs, when compared to Back-propagation. Figure 1 shows the best performances for both algorithms.

Although the displayed results in Fig. 1 intend to compare minimum errors in relation to the number of epochs and processing time for the algorithms

[11] The mean quadratic error formula is related to the derivative of the activation function (sigmoid) [4]. For the comparisons, it was considered that the network "converged" when the minimum error reached the value of 0.001.

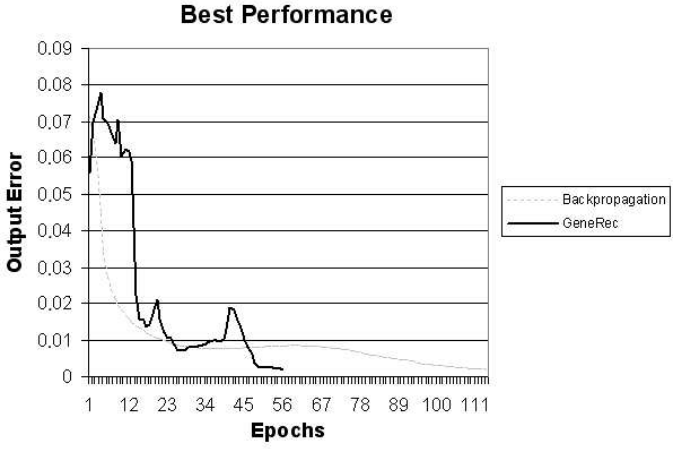

Fig. 1. Best performances in SABio tests. Back-propagation with 25 hidden neurons and learning rate of 0.35 converges in 115 epochs and 8 seconds. GeneRec with 11 hidden neurons and learning rate of 0.45 converges in 57 epochs and 3 seconds. The curve for GeneRec keeps its saw aspect for epochs over 57 (not shown in the figure). But this is irrelevant in this case, since the minimum error had already been reached.

GeneRec and Back-propagation, it is considered relevant to state that the recall rate obtained by the GeneRec training revealed greater than the recall rate by Back-propagation training (the F-Measure was respectively 33.27 and 30.97).

The comparison of SABio with other automatic text summarizers considered the recall and precision rates of summaries generated by SABio. To know which is the best architecture that SABio could employ[12] to reach greater recall and precision rates, preliminary tests with several architectures were conducted, altering the number of epochs, learning rates, and amount of hidden neurons[13]. The best architecture found in tests was used to compare SABio with other automatic text summarizers.

In order to make comparisons between SABio and other automatic summarizers, the F-Measure was employed because besides including recall and precision rates in its formula, it is often employed to compare the efficiency of automatic summarizers in relation to their generated summaries [22].

The same conditions found in Rino et al. [22] were employed, that is:

- a compression rate of 70%. Compression rates usually range from 5% to 30% of the source text content [2]. Other rates displaying similar performances were experimented;

[12] In order to make comparisons between treatments considered more biologically plausible and implausible through SABio architecture, adaptations were provided in SABio in order to be trained with Back-propagation.

[13] Preliminary tests were conducted with: a) learning rates: 0.05, 0.15, 0.25, and 0.45; b) epochs: 2,000 until 10,000 (multiples of 2,000); c) hidden neurons: 8, 16, and 22.

– a 10-fold cross validation, non-biasing process. TeMário was divided into ten distinct groups of texts so that each group contains ten different texts. For the training set, nine of these groups were employed and for the test, the remaining group. This way, the texts used for the test do not belong to the training set. Ten tests were performed (one for each group), and recall, precision, and F-Measure rates were calculated. Then, averages of these rates for the ten experiments were extracted.

This experiment made possible the comparison of SABio with other automatic text summarizers which employ exactly the same method and the same corpus. Table 1 shows the comparative frame among them.

Table 1. Performance (in %) of the systems: SABio-GR, trained by algorithm GeneRec, with 22 hidden neurons, learning rate of 0.25 and 4,000 epochs (501 seconds), SABio-EB, trained by algorithm Error Back-propagation, with 16 hidden neurons, learning rate of 0.45 and 8,000 epochs (896 seconds), among other automatic text summarizers. Adapted from Rino et al. [22].

Summarizer	*Precision rate*	*Recall rate*	*F-Measure*	*Difference to SABio-GR's F-Measure in %*
Supor	44.9	40.8	42.8	1.90
ClassSumm	45.6	39.7	42.4	0.95
SABio-GR	**43.8**	**40.3**	**42.0**	**-**
SABio-EB	**42.4**	**38.7**	**40.5**	**-3.70**
From-Top	42.9	32.6	37.0	-13.51
TF-ISF-Summ	39.6	34.3	36.8	-14.13
GistSumm	49.9	25.6	33.8	-24.26
NeuralSumm	36.0	29.5	32.4	-29.63
Random order	34.0	28.5	31.0	-35.48

SABio outcomes can be considered satisfactory, because:

– the first place - *Supor - Text Summarization in Portuguese* [23] reached performance[14] 1.90% above SABio-GR, but it employs techniques that make computational cost higher than SABio, like lexical chains and thesaurus;
– the second place - *ClassSumm - Classification System* [24] - displayed performance 0.95% above SABio-GR, but it also employs high-cost techniques like semantic analysis, similarity of the sentence with the title (its ideal extracts must have titles), anaphor occurrence analysis, and tagging.

The third place of SABio-GR can be considered a very good performance. In the fourth place, SABio appears again, but with adaptations in order to run with algorithm Back-propagation (version EB).

[14] The F-Measure is used for performance measurement.

SABio presents some similarities with NeuralSumm [5] regarding the employed training set features (7 of 8 NeuralSumm features are present in SABio - the absent feature regards the presence of keywords in the sentence, whose relevance is questionable). The search for more adequate classifiers for automatic summarization is as important as the selection of the features that better represent the focused problem [24, 5].

5 Conclusion

The increasing interest in applications of automatic text summarizers can be justified by the need of the modern society in searching of headlines instead of complete texts in magazines and newspapers, mainly because of lack of time.

Presenting the system SABio, this paper aims to show that it is possible to achieve better performance in automatic summarization when a more biologically plausible model for the ANN training is employed. It is not intention to reveal weaknesses in existent automatic summarizers neither make comparisons that can state that one summarizer is "better" than another. Of course, this "choice" would rely on the "best" features chosen for an automatic summarizer.

SABio presents several limitations, and it can be improved, but the employment of more biologically plausible models for automatic text summarization could represent the achievement of better performance, with greater recall and precision rates, and also contribute to restore principles of ANNs.

References

1. Mani, I., Maybury, M.T.: Advances in Automatic Text Summarization. MIT Press, Cambridge, MA, USA (1999)
2. Mani, I.: Automatic Summarization. John Benjamins Pub., Amsterdam (2001)
3. Aretoulaki, M.: COSY-MATS: A Hybrid Connectionist-Symbolic Approach to the Pragmatic Analysis of Texts for Their Automatic Summarisation. PhD thesis, Univ. of Manchester (1996)
4. Haykin, S.: Neural Networks: A Comprehensive Foundation. 2 edn. Prentice Hall, Upper Saddle River, New Jersey, USA (1999)
5. Pardo, T.A.S., Rino, L.H.M., Nunes, M.G.V.: NeuralSumm: Uma Abordagem Conexionista para a Sumarizacão Automática de Textos. In: Anais do IV Encontro Nacional de Inteligência Artificial - ENIA2003. Volume 1., Campinas, São Paulo, Brazil (2003) 1–10
6. Pardo, T.A.S., Rino, L.H.M., Nunes, M.G.V.: GistSumm: A Summarization Tool Based on a New Extractive Method. 6th Workshop on Computational Processing of the Portuguese Language **6** (2002) 210–218
7. van Rijsbergen, C.J.: Information Retrieval. 2 edn. Butterworth, London, England (1979)
8. O'Reilly, R.C., Munakata, Y.: Computational Explorations in Cognitive Neuroscience - Understanding the Mind by Simulating the Brain. A Bradford Book, The MIT Press, Cambridge, Massachusetts, USA (2000)
9. Kandel, E.R., Schwartz, J.H., Jessell, T.M.: Essential of Neural Science and Behavior. Appleton and Lange, Stanford, Connecticut, USA (1995)

10. Rosa, J.L.G.: A Biologically Inspired Connectionist System for Natural Language Processing. In: Proc. of the 2002 VII Brazilian Symposium on Neural Networks - SBRN2002, Recife, Brazil, IEEE Computer Society Press (2002) 243–248
11. O'Reilly, R.C.: Biologically Plausible Error-driven Learning using Local Activation Differences: The Generalized Recirculation Algorithm. Neural Computation **8:5** (1996) 895–938
12. Rumelhart, D.E., Hinton, G.E., Williams, R.J.: Learning Internal Representations by Error Propagation. In Rumelhart, D.E., McClelland, J.L., eds.: Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Volume 1. A Bradford Book, MIT Press (1986) 318–362
13. Pardo, T.A.S., Rino, L.H.M.: TeMário: Um Corpus para Sumarizacão Automática de Textos. Technical report, NILCTR-03-09 - ICMC-USP, São Carlos, São Paulo, Brazil (2003)
14. Salton, G.: Automatic Text Processing - The Transformation, Analysis, and Retrieval of Information by Computer. Addison-Wesley (1989)
15. Witten, I.H., Moffat, A., Bell, T.C.: Managing Gigabytes. Van Nostrand Reinhold, New York, NY, USA (1994)
16. Kupiec, J., Pedersen, J., Chen, F.: A Trainable Document Summarizer. ACM SIGIR **1** (1995) 68–73
17. Jurafsky, D., Martin, J.H.: Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. Prentice Hall (2000)
18. Baxendale, P.B.: Machine-Made Index for Technical Literature: An Experiment. IBM Journal of Research and Development **2** (1958) 354–365
19. Black, W.J., Johnson, F.C.: A Practical Evaluation of Two Rule-Based Automatic Abstraction Techniques. Expert Systems for Information Management **1(3)** (1988) 159–177
20. Larocca, J.N., Santos, A.D., Kaestner, C.A.A., Freitas, A.A.: Generating Text Summaries through the Relative Importance of Topics. In Monard, M.C., Sichman, J.S., eds.: Lecture Notes in Computer Science - Advances in Artificial Intelligence, Proc. of the Intl. Joint Conf. 7th. Ibero-American Conf. on AI - 15th. Brazilian Symposium on AI - IBERAMIA-SBIA 2000. Volume 1952., São Paulo, Brazil, Springer-Verlag Heidelberg (2000) 301–309
21. Paice, C.D.: The Automatic Generation of Literature Abstracts: An Approach Based on the Identification of Self-indicating Phrases. In: Proc. of the 3rd annual ACM Conf. on Research and Development in Information Retrieval, Cambridge, England, Butterworth (1981) 172–191
22. Rino, L.H.M., Pardo, T.A.S., Jr, C.N.S., Kaestner, C.A.A., Pombo, M.: A Comparison of Automatic Summarization Systems for Brazilian Portuguese Texts. In: Proc. of the XVII Brazilian Symposium on Artificial Intelligence - SBIA2004, São Luís, Maranhão, Brazil (2004) 235–244
23. Modolo, M.: SUPOR: Um Ambiente para a Exploracão de Métodos Extrativos para a Sumarizacão Automática de Textos em Português. PhD thesis, Computing Department - UFSCar, São Carlos, São Paulo, Brazil (2003)
24. Larocca, J.N., Freitas, A.A., Kaestner, C.A.A.: Automatic Text Summarization Using a Machine Learning Approach. In: Proc. of the 16th Brazilian Symposium on Artificial Intelligence, Porto de Galinhas, Pernambuco, Brazil (2002) 205–215