The Cortex Automatic Summarization System at the QA@INEX Track 2010

Juan-Manuel Torres-Moreno[1,2] and Michel Gagnon[2]

[1] Université d'Avignon et des Pays de Vaucluse - LIA
339, chemin des Meinajariès, Agroparc BP 91228, 84911 Avignon Cedex 9 France
[2] École Polytechnique de Montréal - Département de génie informatique
CP 6079 Succ. Centre Ville H3C 3A7 Montréal (Québec), Canada
juan-manuel.torres@univ-avignon.fr, michel.gagnon@polymtl.ca
http://www.lia.univ-avignon.fr

Abstract. The Cortex system is constructed of several different sentence selection metrics and a decision module. Our experiments have shown that the Cortex decision on the metrics always scores better than each system alone. In the INEX@QA 2010 task of Long Questions, Cortex strategy system obtained very good results in the automatic evaluations FRESA.

Keywords: INEX, Automatic summarization system, Question-Answering system, Cortex.

1 Introduction

Automatic summarization is indispensable to cope with ever increasing volumes of valuable information. An abstract is by far the most concrete and most recognized kind of text condensation [1]. We adopted a simpler method, usually called *extraction*, that allow to generate summaries by extraction of pertinence sentences [2,3]. Essentially, extracting aims at producing a shorter version of the text by selecting the most relevant sentences of the original text, which we juxtapose without any modification. The vector space model [4,5] has been used in information extraction, information retrieval, question-answering, and it may also be used in text summarization. Cortex[1] is an automatic summarization system, recently developed [6] which combines several statistical methods with an optimal decision algorithm, to choose the most relevant sentences.

An open domain Question-Answering system (QA) has to precisely answer a question expressed in natural language. QA systems are confronted with a fine and difficult task because they are expected to supply specific information and not whole documents. At present there exists a strong demand for this kind of text processing systems on the Internet. A QA system comprises, *a priori*, the following stages [7]:

[1] *COndensés et Résumés de TEXte* (Text Condensation and Summarization).

S. Geva et al. (Eds.): INEX 2010, LNCS 6932, pp. 290–294, 2011.

- Transform the questions into queries, then associate them to a set of documents;
- Filter and sort these documents to calculate various degrees of similarity;
- Identify the sentences which might contain the answers, then extract text fragments from them that constitute the answers. In this phase an analysis using Named Entities (NE) is essential to find the expected answers.

Most research efforts in summarization emphasize generic summarization [8,9,10]. User query terms are commonly used in information retrieval tasks. However, there are few papers in literature that propose to employ this approach in summarization systems [11,12,13]. In the systems described in [11], a learning approach is used (performed). A document set is used to train a classifier that estimates the probability that a given sentence is included in the extract. In [12], several features (document title, location of a sentence in the document, cluster of significant words and occurrence of terms present in the query) are applied to score the sentences. In [13] learning and feature approaches are combined in a two-step system: a training system and a generator system. Score features include short length sentence, sentence position in the document, sentence position in the paragraph, and tf.idf metrics. Our generic summarization system includes a set of eleven independent metrics combined by a Decision Algorithm. Query-based summaries can be generated by our system using a modification of the scoring method. In both cases, no training phase is necessary in our system.

This paper is organized as follows. In Section 2 we explain the methodology of our work. Experimental settings and results are presented in Section 3. Section 4 exposes the conclusions of the paper and the future work.

2 The CORTEX System

Cortex (***CO**ndensation et **R**ésumés de **T**extes*) [14,15] is a single-document extract summarization system using an optimal decision algorithm that combines several metrics. These metrics result from processing statistical and informational algorithms on the document vector space representation.

The INEX 2010 Query Task evaluation is a real-world complex question (called long query) answering, in which the answer is a summary constructed from a set of relevant documents. The documents are parsed to create a corpus composed of the query and the the multi-document retrieved by Indri.

The idea is to represent the text in an appropriate vectorial space and apply numeric treatments to it. In order to reduce complexity, a preprocessing is performed to the question and the document: words are filtered, lemmatized and stemmed.

The Cortex system uses 11 metrics (see [16] for a detailed description of these metrics) to evaluate the sentence's relevance.

- The frequency of words (F).
- The overlap between the words of the question (R).
- The entropy the words (E).

- The shape of text (Z).
- The angle between question and document vectors (A).
- The sum of Hamming weights of words per segment times the number of different words in a sentence.
- The sum of Hamming weights of the words multiplied by word frequencies.
- ...

The system scores each sentence with a decision algorithm that relies on the normalized metrics. Before combining the votes of the metrics, these are partitionned into two sets: one set contains every metric $\lambda^i > 0.5$, while the other set contains every metric $\lambda^i < 0.5$ (values equal to 0.5 are ignored). We then calculate two values α and β, which give the sum of distances (positive for α and negative for β) to the threshold 0.5 (the number of metrics is Γ, which is 11 in our experiment):

$$\alpha = \sum_{i=1}^{\Gamma} (\lambda^i - 0.5); \quad \lambda^i > 0.5$$

$$\beta = \sum_{i=1}^{\Gamma} (0.5 - \lambda^i); \quad \lambda^i < 0.5$$

The value given to each sentence is calculated with:

```
if($\alpha > \beta$)
    then  $Score^{cortex}(s,q) = 0.5 + \alpha/\Gamma$
    else  $Score^{cortex}(s,q) = 0.5 - \beta/\Gamma$
```

In addition to this score, two other measures are used: the question-document similarity and the topic-sentence overlap. The Cortex system is applied to each document of a topic set and the summary is generated by concatenating higher score sentences.

The specific similarity measure [17] between the question and the corpus allows us to re-scale the sentence scores according to the relevance of the document from which they are extracted. This measure is the normalized scalar product of the tf.idf vectorial representations of the document and the question q. Let $\boldsymbol{d} = (d_1 \ldots d_n)$ and $\boldsymbol{q} = (q_1 \ldots q_n)$ be the vectors representing the document and the question, respectively. The definition of the measure is:

$$Similarity(\boldsymbol{q}, \boldsymbol{d}) = \frac{\sum_{i=1}^{n} d_i q_i}{\sqrt{\sum_{i=1}^{n} {d_i}^2 + \sum_{i=1}^{n} {q_i}^2}}$$

The last measure, the topic-sentence overlap, assigns a higher ranking for the sentences containing question words and makes selected sentences more relevant. The overlap is defined as the normalized cardinality of the intersection between the uestion word set T and the sentence word set S.

$$Overlap(T, S) = \frac{card(S \cap T)}{card(T)}$$

The final score of a sentence s from a document d and a question q is the following:

$$Score = \alpha_1 \; Score^{cortex}(s,q) + \alpha_2 \; Overlap(s,q) + \alpha_3 \; Similarity(d,q)$$
$$\text{where} \sum_i \alpha_i = 1$$

3 Experiments Settings and Results

In this study, we used the document sets made available during the Initiative for the Evaluation of XML retrieval (INEX) 2010[2], in particular on the INEX 2010 QA Track (QA@INEX) `http://www.inex.otago.ac.nz/tracks/qa/qa.asp`.

To evaluate the efficacity of Cortex on INEX@QA corpus, we used the FRESA package[3].

INEX queries. No pre-processing or modification was applied on queries set. Cortex used the query as a title of a big document retrieved by Indri. Table 1 shows an example of the results obtained by Cortex system using 50 documents as input. The query that the summary should answer in this case was the number 2010111:

What is considered a beautiful body shape?.

This table presents Cortex results in comparison with an the INEX baseline (Baseline summary), and three baselines, that is, summaries including random n-grams (Random unigram) and 5-grams (Random 5-gram) and empty baseline. We observe that our system is always better than Baseline summary and empty baseline.

Table 1. Example of Cortex Summarization results

Summary type	1-gram	2-gram	SU4-gram	FRESA average
Baseline summary	26.679	34.108	34.218	31.668
Empty baseline	31.715	39.452	39.534	36.901
Random unigram	25.063	32.822	32.852	30.246
Random 5-gram	23.168	30.644	30.838	28.217
Cortex summary	**26.421**	33.935	34.030	**31.462**

4 Conclusions

We have presented the Cortex summarization system that is based on the fusion process of several different sentence selection metrics. The decision algorithm obtains good scores on INEX-2010, indicating that the decision process is a good strategy for preventing overfitting on the training corpus. In the INEX-2010 corpus, Cortex system obtained very good results in the automatic FRESA evaluations.

[2] `http://www.inex.otago.ac.nz/`

[3] FRESA package is disponible at web site: `http://lia.univ-avignon.fr/fileadmin/axes/TALNE/downloads/index_fresa.html`

References

1. ANSI. American National Standard for Writing Abstracts. Technical report, American National Standards Institute, Inc., New York, NY (ANSI Z39.14.1979) (1979)
2. Luhn, H.P.: The Automatic Creation of Literature Abstracts. IBM Journal of Research and Development 2(2), 159 (1958)
3. Edmundson, H.P.: New Methods in Automatic Extracting. Journal of the ACM (JACM) 16(2), 264–285 (1969)
4. Salton, G.: The SMART Retrieval System - Experiments un Automatic Document Processing, Englewood Cliffs (1971)
5. Salton, G., McGill, M.: Introduction to Modern Information Retrieval. McGraw-Hill, New York (1983)
6. Torres-Moreno, J.M., Velazquez-Morales, P., Meunier, J.-G.: Condensés automatiques de textes. Lexicometrica. L'analyse de données textuelles: De l'enquête aux corpus littéraires, Special (2004), `www.cavi.univ-paris3.fr/lexicometrica`
7. Jacquemin, C., Zweigenbaum, P.: Traitement automatique des langues pour l'accès au contenu des documents. Le Document en Sciences du Traitement de l'information 4, 71–109 (2000)
8. Abracos, J., Lopes, G.P.: Statistical Methods for Retrieving Most Significant Paragraphs in Newspaper Articles. In: Mani, I., Maybury, M.T. (eds.) ACL/EACL 1997-WS, Madrid, Spain, July 11 (1997)
9. Teufel, S., Moens, M.: Sentence Extraction as a Classification Task. In: Mani, I., Maybury, M.T. (eds.) ACL/EACL 1997-WS, Madrid, Spain (1997)
10. Hovy, E., Lin, C.Y.: Automated Text Summarization in SUMMARIST. In: Mani, I., Maybury, M.T. (eds.) Advances in Automatic Text Summarization, pp. 81–94. The MIT Press, Cambridge (1999)
11. Kupiec, J., Pedersen, J.O., Chen, F.: A Trainable Document Summarizer. In: Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 68–73 (1995)
12. Tombros, A., Sanderson, M., Gray, P.: Advantages of Query Biased Summaries in Information Retrieval. In: Hovy, E., Radev, D.R. (eds.) AAAI1998-S, Stanford, California, USA, March 23-25, pp. 34–43. The AAAI Press, Menlo Park (1998)
13. Schlesinger, J.D., Backer, D.J., Donway, R.L.: Using Document Features and Statistical Modeling to Improve Query-Based Summarization. In: DUC 2001, New Orleans, LA (2001)
14. Torres-Moreno, J.M., Velazquez-Moralez, P., Meunier, J.: CORTEX, un algorithme pour la condensation automatique de textes. In: ARCo, vol. 2, p. 365 (2005)
15. Torres-Moreno, J.M., St-Onge, P.-L., Gagnon, M., El-Bèze, M., Bellot, P.: Automatic summarization system coupled with a question-answering system (qaas). CoRR, abs/0905.2990 (2009)
16. Torres-Moreno, J.M., Velazquez-Morales, P., Meunier, J.G.: Condensés de textes par des méthodes numériques. JADT 2, 723–734 (2002)
17. Salton, G.: Automatic text processing, 9th edn. Addison-Wesley Longman Publishing Co. Inc., Amsterdam (1989)