The REG summarization system at QA@INEX track 2010

Jorge Vivaldi[1], Iria da Cunha[1] and Javier Ramírez[2]

[1] Instituto Universitario de Linguistica Aplicada - UPF
Barcelona
[2] Universidad Autonoma Metropolitana-Azcapotzalco
Mexico
{iria.dacunha,jorge.vivaldi}@upf.edu;jararo@correo.azc.uam.mx
http://www.iula.upf.edu

Abstract. In this paper we present REG, a graph approach to study a fundamental problem of Natural Language Processing: the automatic summarization of documents. The algorithm models a document as a graph, to obtain weighted sentences. We applied this approach to the INEX@QA 2010 task (question-answering). To do it, we have extracted the terms from the queries, in order to obtain a list of terms related with the main topic of the question. Using this strategy, REG obtained good results with the automatic evaluation system FRESA.

Key words: INEX, Automatic Summarization System, Question-Answering System, REG.

1 Introduction

Nowadays automatic summarization is a very prominent research topic. We can define summary as "a condensed version of a source document having a recognizable genre and a very specific purpose: to give the reader an exact and concise idea of the contents of the source" (Saggion and Lapalme, 2002: 497). Summaries can be divided into "extracts", if they contain the most important sentences extracted from the original text (ex. Edmunson, 1969; Nanba and Okumura, 2000; Gaizauskas et al., 2001; Lal and Reger, 2002; Torres-Moreno et al., 2002) and "abstracts", if these sentences are re-written or paraphrased, generating a new text (ex. Ono et al., 1994; Paice, 1990; Radev, 1999). Most of the automatic summarization systems are extractive. These systems are useful in several domains: medical (ex. Johnson et al., 2002 Afantenos et al., 2005; da Cunha et al., 2007; Vivaldi et al., 2010), legal (ex. Farzindar et al., 2004), journalistic (ex. Abracos and Lopes, 1997; Fuentes et al., 2004), etc. One of the tasks where these extractive summarization systems could help is question-answering. The objective of the INEX@QA 2010 track is to evaluate a difficult question-answering task, where questions are very precise (expecting short answers) or very complex (expecting long answers, including several sentences). In this second task is where

automatic summarization systems could help. The used corpus in this track contains all the texts included into the English Wikipedia. The expected answers are automatic summaries of less than 500 words exclusively made of aggregated passages extracted from the Wikipedia corpus. The evaluation of the answers will be automatic, using the automatic evaluation system FRESA (Torres-Moreno et al., 2010a, 2010b, Saggion et al., 2010), and manual (evaluating syntactic incoherence, unsolved anaphora, redundancy, etc.). To carry out this task, we have decided to use REG (Torres-Moreno and Ramirez, 2010; Torres-Moreno et al., 2010), an automatic summarization system based on graphs. We have performed some expansions of the official INEX@QA 2010 queries, detecting the terms they contain automatically, in order to obtain a list of terms related with the main topic of all the questions.

This paper is organized as follows. In Section 2 we show REG, the summarization system we have used for our experiments. In Section 3 we explain how we have carried out the terms extraction of the queries. In Section 4 we present the experimental settings and results. Finally, in Section 5, we expose some conclusions.

2 The REG system

REG (Torres-Moreno and Ramirez, 2010; Torres-Moreno et al. 2010) is an Enhanced Graph summarizer (REG) for extract summarization, using a graph approach. The strategy of this system has two main stages: a) to carry out an adequate representation of the document and b) to give a weight to each sentence of the document. In the first stage, the system makes a vectorial representation of the document. In the second stage, the system uses a greedy optimization algorithm. The summary generation is done with the concatenation of the most relevant sentences (previously scored in the optimization stage).

REG algorithm contains three modules. The first one carries out the vectorial transformation of the text with filtering, lemmatization/stemming and normalization processes. The second one applies the greedy algorithm and calculates the adjacency matrix. We obtain the score of the sentences directly from the algorithm. Therefore, sentences with more score will be selected as the most relevant. Finally, the third module generates the summary, selecting and concatenating the relevant sentences. The first and second modules use CORTEX (Torres-Moreno et al., 2002), a system that carries out an unsupervised extraction of the relevant sentences of a document using several numerical measures and a decision algorithm.

3 Terms extraction

The first procedure for obtaining the query terms has been to found the main topic of the questions. This has been obtained by finding the terms candidate present in every query. Terms are usually defined as lexical units to designate concepts in a domain. The detection of these units is a complex task mainly

because terms adopt all word formation rules in a given language [22]. Also, as mentioned in the term definition itself, it is necessary to confirm that a given lexical unit belong to the domain of interest. Due to the difficulties to verify this condition it is usual to refer the results obtained by an extractor as term candidates instead of just "terms". In this context we have used the basic procedure for obtaining term candidates in the field of term extraction. Such candidates are typically obtained by using the morphosyntactic terminological patterns for any given language (see [23, 24]), English in this case.

As the queries do not belong to any specific domain it is not possible determine the termhood of the retrieved candidates.

Considering that questions are very short, only a few candidates are obtained by such procedure; therefore, they have a high probability to be the main topic of the question.

For example, for the query "How does GLSL unify vertex and fragment processing in a single instruction set?", we consider the terms "glsl", "vertex processing", "fragment processing" and "single instruction set". But for the query "Who is Eiffel?", there are not any term, only the proper name "Eiffel?".

4 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[1], in particular on the INEX 2010 QA Track (QA@INEX). These sets of documents where provided by the search engine Indri.[2] REG has produced multidocument summaries using sets of 30, 40 and 50 of the documents provided by Indri using all the queries of the track.

To evaluate the efficiency of REG over the INEX@QA corpus, we have used the FRESA package.

Table 1 shows an example of the results obtained by REG using 50 documents as input. The query that the summary should answer in this case was the number 2009006. This table presents REG results in comparison with an intelligent baseline (Baseline summary), and two simple baselines, that is, summaries including random n-grams (Random unigram) and 5-grams (Random 5-gram). We observe that our system is always better than these two simple baselines, but in comparison with the first one the performance is variable.

5 Conclusions

We have presented the REG summarization system, an extractive summarization algorithm that models a document as a graph, to obtain weighted sentences. We applied this approach to the INEX@QA 2010 task, extracting the terms from

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

[2] Indri is a search engine from the Lemur project, a cooperative work between the University of Massachusetts and Carnegie Mellon University in order to build language modelling information retrieval tools: `http://www.lemurproject.org/indri/`

Table 1. Example of REG results using 50 documents as input.

Distribution type	unigram	bigram	with 2-gap	Average
Baseline summary	22.64989	31.70850	32.07926	28.81255
Random unigram	18.18043	28.25213	28.44528	24.95928
Random 5-gram	17.47178	26.33253	27.03882	23.61437
Submitted summary	22.77755	32.06325	32.53706	29.12595

the queries, in order to obtain a list of terms related with the main topic of the question.

Our experiments have shown that the system is always better than the two simple baselines, but in comparison with the first one the performance is variable. We think this is due to the fact that some queries are long and they have several terms we could extract, but there are some queries that are very short and the term extraction is not possible or very limited. Nevertheless, we consider that, over the INEX-2010 corpus, REG obtained good results in the automatic evaluations, but now it is necessary to wait for the human evaluation and the evaluation of other systems to compare with.

References

1. Abracos, J.; Lopes, G. (1997). *Statistical methods for retrieving most significant paragraphs in newspaper articles*. In Proceedings of the ACL/EACL'97 Workshop on Intelligent Scalable Text Summarization. Madrid. 51-57.
2. Afantenos, S.; Karkaletsis, V.; Stamatopoulos, P. (2005). *Summarization of medical documents: A survey*. Artificial Intelligence in Medicine 33 (2). 157-177.
3. da Cunha, I.; Wanner, L.; Cabré, M.T. (2007). *Summarization of specialized discourse: The case of medical articles in Spanish*. Terminology 13 (2). 249-286.
4. Edmunson, H. P. (1969). *New Methods in Automatic Extraction*. Journal of the Association for Computing Machinery 16. 264-285.
5. Farzindar, A.; Lapalme, G.; Desclés, J.-P. (2004). *Résumé de textes juridiques par identification de leur structure thématique*. Traitement automatique des langues 45 (1). 39-64.
6. Fuentes, M.; Gonzalez, E.; Rodriguez, H. (2004). *Resumidor de noticies en catala del projecte Hermes*. In Proceedings of II Congrés d'Enginyeria en Llengua Catalana (CELC'04). Andorra. 102-102.
7. Gaizauskas, R.; Herring, P.; Oakes, M.; Beaulieu, M.; Willett, P.; Fowkes, H.; Jonsson, A. (2001). *Intelligent access to text: Integrating information extraction technology into text browsers*. En Proceedings of the Human Language Technology Conference. San Diego. 189-193.
8. Johnson, D.B.; Zou, Q.; Dionisio, J.D.; Liu, V, Z.; Chu, W.W. (2002). *Modeling medical content for automated summarization*. Annals of the New York Academy of Sciences 980. 247-258.
9. Lal, P.; Reger, S. (2002). *Extract-based Summarization with Simplication*. In Proceedings of the 2nd Document Understanding Conference at the 40th Meeting of the Association for Computational Linguistics. 90-96.

10. Nanba, H.; Okumura, M. (2000). *Producing More Readable Extracts by Revising Them*. In Proceedings of the 18th International Conference on Computational Linguistics (COLING-2000). Saarbrucken. 1071-1075.
11. Ono, K.; Sumita, K.; Miike, S. (1994). *Abstract generation based on rhetorical structure extraction*. In Proceedings of the International Conference on Computational Linguistics. Kyoto. 344-348.
12. Paice, C. D. (1990). *Constructing literature abstracts by computer: Techniques and prospects*. Information Processing and Management 26. 171-186.
13. Radev, D. (1999). Language Reuse and Regeneration: Generating Natural Language Summaries from Multiple On-Line Sources. New York, Columbia University. [PhD Thesis]
14. Saggion, H.; Lapalme, G. (2002). *Generating Indicative-Informative Summaries with SumUM*. Computational Linguistics 28 (4). 497-526.
15. Torres-Moreno, J-M.; Saggion, H. da Cunha, I. SanJuan, E. Velázquez-Morales, P. SanJuan, E.(2010a). *Summary Evaluation With and Without References*. Polibitis: Research journal on Computer science and computer engineering with applications 42.
16. Saggion, H.; Torres-Moreno, J-M.; da Cunha, I.; SanJuan, E.; Velázquez-Morales, P.; SanJuan, E. (2010b). *Multilingual Summarization Evaluation without Human Models*. In Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010). Pekin.
17. Torres-Moreno, J.-M.; Saggion, H.; da Cunha, I.; Velázquez-Morales, P.; SanJuan, E. (2010b). *Ealuation automatique de résumés avec et sans référence*. In Proceedings of the 17e Conférence sur le Traitement Automatique des Langues Naturelles (TALN). Université de Montréal et Ecole Polytechnique de Montréal: Montreal (Canada).
18. Torres-Moreno, J-M.; Ramírez, J. (2010). *REG : un algorithme glouton appliqué au résumé automatique de texte*. JADT 2010. Roma, Italia.
19. Torres-Moreno, J-M.; Ramírez, J.; da Cunha, I. (2010). *Un resumeur a base de graphes, indépendant de la langue*. Workshop African HLT 2010. Djibouti.
20. Torres-Moreno, J. M.; Velázquez-Morales, P.; Meunier, J. G. (2002). *Condensés de textes par des méthodes numériques*. En Proceedings of the 6th International Conference on the Statistical Analysis of Textual Data (JADT). St. Malo. 723-734.
21. Vivaldi, J.; da Cunha, I.; Torres-Moreno, J.M.; Velázquez, P. (2010). "Automatic Summarization Using Terminological and Semantic Resources". En actas del 7th International Conference on Language Resources and Evaluation (LREC 2010). Valletta, Malta.
22. Pearson J. (1998). Terms in context. John Benjamin. Amsterdam.
23. Cabré, M.T.; R. Estopà; Vivaldi, J. (2001). *Automatic term detection: a review of current systems*. In Bourigault, D., C. Jacquemin and M.C. L'Homme (eds.). Recent Advances in Computational Terminology. 53-87. Amsterdam: John Benjamins.
24. Pazienza, M.T.; Pennacchiotti, M.; Zanzotto, F.M. (2005). *Terminology Extraction: An Analysis of Linguistic and Statistical Approaches*. In: Studies in Fuzziness and Soft Computing. Volume 185/2005. 255-279.