The REG summarization system with question expansion and reformulation at QA@INEX track 2011

Jorge Vivaldi and Iria da Cunha

Institut Universitari de Lingüística Aplicada - UPF
Barcelona
{iria.dacunha,jorge.vivaldi}@upf.edu
http://www.iula.upf.edu

Abstract. In this paper, our strategy and preliminary results for the INEX@QA 2011 question-answering task are presented. In this task, a set of 50 documents is provided by the search engine Indri, using some queries. The initial queries are titles associated wih tweets. A reformulation of these queries is carried out using terminological and names entities information. To design the queries to obtain the documents with INDRI, the full process is divided into 2 steps: a) both titles and tweets are POS tagged, and b) queries are expanded or reformulated, using: terms and name entities included in the title, terms and name entities found in the tweet related to those ones, and Wikipedia redirected terms and name entities from those ones included in the title. In our work, the automatic summarization system REG is used to summarize the 50 documents obtained with these queries. The algorithm models a document as a graph, to obtain weighted sentences. A single document is generated, considered as the answer of the query. This strategy, combining summarization and question reformulation, obtains preliminary good results with the automatic evaluation system FRESA.

Key words: INEX, Question-Answering, Terms, Name Entities, Wikipedia, Automatic Summarization, REG.

1 Introduction

The Question-Answering (QA) task can be related to two types of questions: very precise questions (expecting short answers) or complex questions (expecting long answers, including several sentences). The objective of the QA track of INEX 2011 (Initiative for the Evaluation of XML retrieval) is oriented to the second one. Specifically, the QA task to be performed by the participating groups of INEX 2011 is contextualizing tweets, i.e. answering questions of the form "what is this tweet about?" using a recent cleaned dump of the Wikipedia (WP). The general process involves: tweet analysis, passage and/or XML elements retrieval and construction of the answer. Relevant passages segments should contain relevant information but contain as little non-relevant information as possible. The

used corpus in this track contains all the texts included into the English WP. The expected answers are short documents of less than 500 words exclusively made of aggregated passages extracted from the WP corpus.

Thus, we consider that automatic extractive summarization systems could be useful in this QA task, taking into account that a summary can be defined 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.

To carry out this task, we have decided to use REG (Torres-Moreno and Ramírez, 2010; Torres-Moreno et al., 2010), an automatic extractive summarization system based on graphs. We have performed some expansions and reformulations of the initial INEX@QA 2011 queries, using terms and name entities, in order to obtain a list of terms related with the main topic of all the questions.

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.).

This paper is organized as follows. In Section 2, the summarization system REG is shown. In Section 3, queries expansions and reformulations are explained. In Section 4, experimental settings and results are presented. Finally, in Section 5, some preliminary conclusions are exposed.

2 The REG System

REG (Torres-Moreno and Ramírez, 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.

The complexity of REG algorithm is $O(n^2)$. Nevertheless, there is a limitation, because it includes a fast classification algorithm which can be used only for short instances; this is the reason it is not very efficient for long texts.

3 Terms and Name Entity Extraction

The starting point of this work is to consider that the terms and name entities (T&NE) included into the titles and the associated tweets are representative of the main subject of these texts. If this assumption is true, the results of quering the search engine with an optimized list of T&NE should be better that simply to use the title of the tweet as search query.

In order to demonstrate such hypothesis, we have decided to generate 3 different queries to Indri:

a) Using the initial query string (the title of the tweet).

b) Enriching the initial query with a list of those T&NE from the tweet that are related to the T&NE already present in the initial query. Redirections from WP are also considered.

c) Using only the above mentioned list of T&NE obtained from the previous step.

The procedure for obtaining this list from the tweet may be sketched as follows:

1. To find, in both query and tweet strings, for T&NE and verify that such strings are also present in WP. This procedure is again splitted in two stages: first finding the T&NE, and then looking for such unit in WP. The last step is close to those presented in Milne and Witten (2008), Strube and Ponzetto (2006) or Ferragina and Scaiella (2010).
2. To compare each unit in the tweet with all the units found in the query. Such comparison is made using the algorithm described in Milne and Witten (2007).
3. To choose only those units whose relatedness value are higher than a given threshold.

Figure 1 shows how the enriched query is built. From the query string we obtain a number of terms: (t_{tm}); we repeat the procedure with the tweet string (t_{tm}). We look for such terms in the WP; only the terms (or a substring of them) that have an entry in WP are considered. Then, we calculate the semantic relatedness among each term of the tweet (t_{tn}) with each term of the query. Only those terms of the tweets whose similarity with some of the term of the query is higher that a threshold value are taken into account. Assuming a query and tweet string as shown in Figure 1, each t_{tm} is compared with all t_{qn}. As a result of such comparisons, only t_{t2} and t_{t4} will be inserted in the enriched query because t_{t1} and t_{t3} will be rejected.

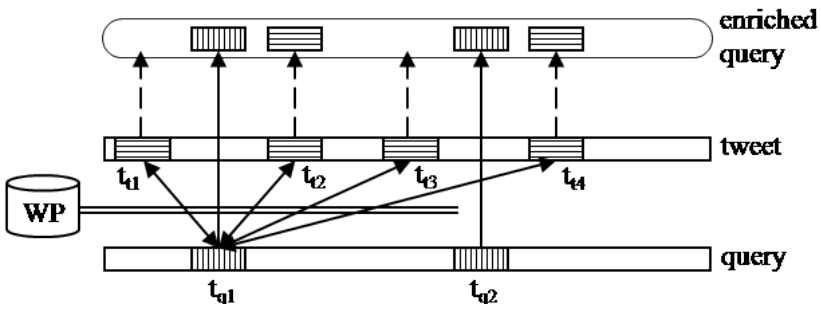

Fig. 1. Advanced query terms selection.

As mentioned above, the comparison among WP articles is done by using the algorithm described in Milne and Witten (2007). The idea is pretty simple and it is based in the links extending each article: higher is the number of number of such links shared by both article higher is their relatedness. Figure 2 shows an outline about how to calculate the relatedness among the WP pages "automobile" and "global warming". It is clear that some outgoing links ("air pollution", "alternative fuel", etc.) are shared by both articles while other links not ("vehicle", "Henry Ford", "Ozone"). From this idea it is possible to build such relatedness measure (see Milne and Witten, 2007 for details).

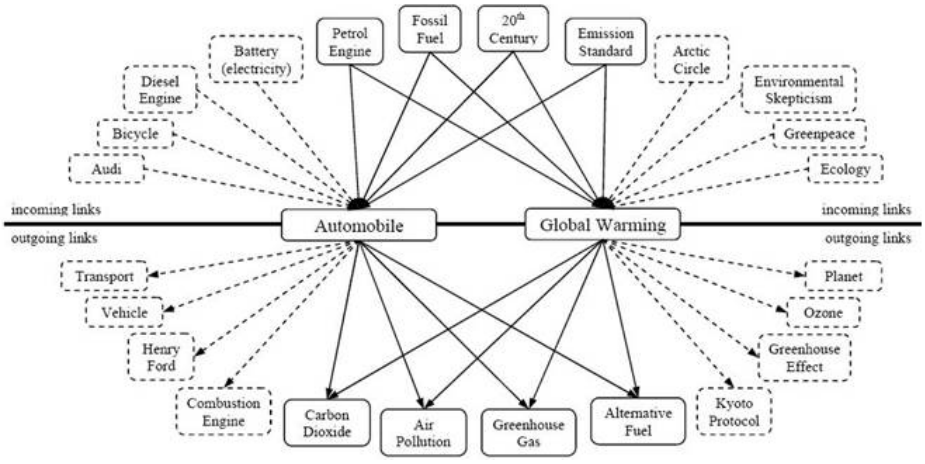

Fig. 2. Looking for the relation among WP articles (reprinted from Milne and Witten, 2007).

Let's see an example of some queries generated in our experiment. For the initial query (the title of the tweet) "Obama to Support Repeal of Defense of

Marriage Act", we extract the term "defense" and "marriage act", and the name entity "Obama". Moreover, we add the name entity "Barack Obama", since there is a redirection link in WP from "Obama" to "Barack Obama". Finally, some terms ("law", "legal definition", "marriage union", "man", "woman", "support" and "gay rights") and name entities ("President Obama" and "White House") semantically related with the units of the title are selected.

The process of building of the 3 different queries for this same example is the following:

1. The initial query is the title of the tweet:
   Obama to Support Repeal of Defense of Marriage Act.
   In this title the following T&NEs have been found: "Obama", "Defense of Marriage Act".
2. The expanded query is built from the body of the tweet:
   WASHINGTON - President Obama will endorse a bill to repeal the law that limits the legal definition of marriage to a union between a man and a woman, the White House said Tuesday taking another step in support of gay rights.
   The T&NE found in this string are: "Obama", "Defense of Marriage Act", "Barack Obama", "law", "legal definition", "marriage", "union", "man", "woman", "step", "support", "gay rights", "President Obama", "White House" and "Tuesday". The built expanded query contains the following query terms: "Obama to Support Repeal of Defense of Marriage Act", "Obama", "Barack Obama", "President Obama", "White House", "marriage", "union", "man", "gay rights" and "woman". Note that some terms are dropped (like "step" and "Tuesday") because they do no have any relation to the T&NE found in the title of the tweet, and some new query terms have been added ("President Obama") using WP redirection links.
3. The reformulated query is built using only the list of T&NE: "Obama", "Barack Obama", "President Obama", "White House", "marriage", "union", "man", "gay rights" and "woman".

The term and name entity extraction was carried out manually. Nowadays several term extraction systems and name entity recognition systems exist for English. Nevertheless, their performances are not still perfect, so if we employ these systems in our work, their mistakes and the mistakes of the system we present here would be mixed. Moreover, term extractors are usually designed for a specialized domain, as medicine, economics, law, etc, but the topics of the queries provided by INEX@QA 2011 are several, that is, they do not correspond to an unique domain. Also the relatedness among WP pages is manually done because our implementation of the relatedness measure relies in a relatively old dump of English WP.

4 Experiments Settings and Results

In this study, we used the document sets made available for the INEX 2011 QA Track (QA@INEX). These sets of documents where provided by the search

engine Indri.[1] REG produced multidocument summaries using the set of 50 documents provided by Indri using all the initial queries of the track and the expansions and reformulations following our strategy.

To evaluate the efficiency of REG over the INEX@QA corpus, we have used the FRESA package. This evaluation framework (FRESA –FRamework for Evaluating Summaries Automatically-) includes document-based summary evaluation measures based on probabilities distribution, specifically, the Kullback-Leibler (KL) divergence and the Jensen-Shannon (JS) divergence. As in the ROUGE package (Lin, 2004), FRESA supports different n-grams and skip n-grams probability distributions. The FRESA environment has been used in the evaluation of summaries produced in several European languages (English, French, Spanish and Catalan), and it integrates filtering and lemmatization in the treatment of summaries and documents. FRESA is available in the following link: `http://lia.univ-avignon.fr/fileadmin/axes/TALNE/Ressources.html`.

Tables 1, 2 and 3 show an example (document ID = 2011041) of the results obtained by REG with 50 documents as input and using the 3 different queries (a, b and c, respectively). These tables present REG results in comparison with an intelligent baseline (Baseline summary) and 3 other baselines: summaries including random n-grams (Random unigram), 5-grams (Random 5-gram) and empty words (Empty baseline). In this example, the reformulated query obtains better results (56.58465) than the initial query (59.04529) using FRESA.

Table 1. Example of REG results over the document 2011041 using query a.

Distribution type	unigram	bigram	with 2-gap	Average
Baseline summary	51.88447	60.37075	60.67855	57.64459
Empty baseline	74.53241	83.71035	83.95558	80.73278
Random unigram	55.86427	65.06233	65.25039	62.05900
Random 5-gram	47.71473	56.34721	56.71558	53.59251
Submitted summary	53.09912	61.89060	62.14615	59.04529

In table 4, the average of the results of 6 summaries selected at random from the 50 summaries is presented (IDs = 2011144, 2011026, 2011041, 2011001, 2011183, 2011081). In this case, the situation regarding the best query changes, and the initial query (the title) obtains the best results. With regard to the baselines, Baseline summary and Random 5-gram are always better than our system. However, in general, our system is better than Random unigram and Empty baseline. Nevertheless, we consider that this evaluation, including only 6 texts, is very preliminary and that it is necessary to wait for the final official evaluation of INEX 2011, in order to obtain a complete evaluation of the results.

[1] 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 2. Example of REG results over the document 2011041 using query b.

Distribution type	unigram	bigram	with 2-gap	Average
Baseline summary	48.75833	57.14385	57.36983	54.42401
Empty baseline	70.35451	79.43793	79.60188	76.46477
Random unigram	52.37199	61.48490	61.59465	58.48385
Random 5-gram	45.73679	54.36620	54.71034	51.60444
Submitted summary	52.08416	60.80215	61.04838	57.97823

Table 3. Example of REG results over the document 2011041 using query c.

Distribution type	unigram	bigram	with 2-gap	Average
Baseline summary	49.62939	58.06352	58.40945	55.36745
Empty baseline	73.65646	82.85850	83.14989	79.88829
Random unigram	53.94457	63.17036	63.39601	60.17031
Random 5-gram	45.36140	53.99297	54.45397	51.26945
Submitted summary	50.67031	59.34556	59.73806	56.58465

The 6 selected summaries can be divided in 2 sets of 3 summaries using: a) reformulated queries with a high quantity of terms and/or name entities (IDs = 2011144, 2011026, 2011041) and b) reformulated queries with a low quantity of terms and/or name entities (IDs = 2011001, 2011183, 2011081). The longest query cotains 11 units and the shortest includes 3 units. Table 5 includes a comparative evaluation between both results. It is interesting to observe that the summaries obtained using queries with a high quantity of terms and name entities obtain better results with the query c) (that is, using the reformulated query). However, when the queries do not include lots of terms, the best results are obtained with the initial queries (that is, the titles).

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 have applied this approach to the INEX@QA 2011 task, using 3 types of queries, the initial ones (titles of tweets) and other ones extracting T&NE from titles, and selecting those units that are semantically related to T&NE present in the associated tweets. Semantic relatedness is obtained directly from WP.

Our preliminary experiments have shown that our system is always better than the 2 simple baselines, but in comparison with the 2 more intelligent baselines the performance is variable. Morevoer, this preliminary evaluation shows that the reformulated queries obtain bestter results than the initial queries when the quantity of extracted terms and name entities is high.

Table 4. Average of results using the 3 queries and REG.

Average	Query a	Query b	Query c
Baseline summary	45.313355	48.297996	48.668063
Empty baseline	59.050726	63.367683	66.331401
Random unigram	46,563255	48.538538	49.83195
Random 5-gram	41.434218	43.78428	43.813228
Submitted summary	45.530785	48.5865	49.421386

Table 5. Comparison between summaries obtained with short and long queries.

Average	Query a	Query b	Query c
Summaries with short query	39.882106	43.721356	47.69209
Summaries with long query	51.179463	53.451656	51.150683

We consider that, over the INEX-2011 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. Edmunson, H. P. (1969). *New Methods in Automatic Extraction.* Journal of the Association for Computing Machinery 16. 264-285.
2. Ferragina, P. and Scaiella, U. (2010). *TAGME: On-the-fly Annotation of Short Text Fragments (byWikipedia Entities).* 19th International Conference on Information and Knowledge Management. Toronto, Canada.
3. 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.
4. 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.
5. Lin, C.-Y. (2004). *ROUGE: A Package for Automatic Evaluation of Summaries.* In Proceedings of Text Summarization Branches Out: ACL-04 Workshop. 74-81.
6. Milne, D.; Witten, I.H. (2007). *An effective , low-cost measure of semantic relatedness obtained from Wikipedia links Obtaining Semantic Relatedness from.* Association for the Advancement of Artificial Intelligence.
7. Milne, D.; Witten, I.H. (2008). *ALearning to link with wikipedia.* Proceedings of the 17th ACM conference on Information and knowledge mining. New York.
8. 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.
9. 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.

10. Paice, C. D. (1990). *Constructing literature abstracts by computer: Techniques and prospects*. Information Processing and Management 26. 171-186.
11. Radev, D. (1999). Language Reuse and Regeneration: Generating Natural Language Summaries from Multiple On-Line Sources. New York, Columbia University. [PhD Thesis]
12. Saggion, H.; Lapalme, G. (2002). *Generating Indicative-Informative Summaries with SumUM*. Computational Linguistics 28(4). 497-526.
13. Saggion, H.; Torres-Moreno, J-M.; da Cunha, I.; SanJuan, E.; Velázquez-Morales, P.; SanJuan, E. (2010). *Multilingual Summarization Evaluation without Human Models*. In Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010). Pekin.
14. Strube, M.; Ponzetto, S.P. (2006). *WikiRelate! Computing Semantic Relatedness Using Wikipedia*. Association for Artificial Intelligence.
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. 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).
17. Torres-Moreno, J-M.; Ramírez, J. (2010). *REG : un algorithme glouton appliqué au résumé automatique de texte*. JADT 2010. Roma, Italia.
18. Torres-Moreno, J-M.; Ramírez, J.; da Cunha, I. (2010). *Un resumeur a base de graphes, indépendant de la langue*. In Proceedings of the International Workshop African HLT 2010. Djibouti.
19. 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.