Gathering Definition Answers by Information Gain

Carmen Martínez-Gil[1] and A. López-López[1]

[1] Instituto Nacional de Astrofísica, Óptica y Electrónica,
Luis Enrique Erro #1, Santa María Tonanzintla, Puebla, 72840, México
{Carmen, allopez}@ccc.inaoep.mx

Abstract. A definition question is a kind of question whose answer is a complementary set of sentence fragments called nuggets, which define the target term. Since developing general and flexible patterns with a wide coverage to answer definition questions is not feasible, we propose a method using information gain to retrieve the most relevant information. To obtain the relevant sentences, we compared the output of two retrieval systems: JIRS and Lucene. One important feature that impacts on the performance of definition question answering systems is the length of the sentence fragments, so we applied a parser to analyze the relevant sentences in order to get clauses. Finally, we observed that, in most of the clauses, only one part before and after the target term contains information that defines the term, so we analyzed separately the sentence fragments before (*left*) and after (*right*) the target term. We performed different experiments with the collections of questions from the *pilot* evaluation of definition questions 2002, *definition* questions from TREC 2003 and *other* questions from TREC 2004. F-measures obtained are competitive when compared against the participating systems in their respective conferences. Also the best results are obtained with the general purpose system (Lucene) instead of JIRS, which is intended to retrieve passages for factoid questions.

1 Introduction

Question Answering (QA) is a computer-based task that tries to improve the output generated by Information Retrieval (IR) systems. A definition question [9] is a kind of question whose answer is a complementary set of sentence fragments called nuggets.

After identifying the correct target term (the term to define) and context terms, we need to obtain useful and non redundant definition nuggets. Nowadays, patterns are obtained manually as surface patterns [5, 6, 12]. These patterns can be very rigid, leading to the alternative soft patterns [2], which are even extracted in an automatic way [5]. Then, once we have the patterns, we apply a matching process to extract the nuggets. Finally, we need to perform a process to determine if these nuggets are part of the definition; where a common criterion employed is the frequency of appearance of the nugget.

According to the state of the art, the highest F-measure in a pilot evaluation [9] for definition questions in 2002 is 0.688 using the nugget set supplied by author, taking

Advances in Natural Language Processing and Applications
Research in Computing Science 33, 2008, pp. 225-234

Received 15/10/07
Accepted 07/12/07
Final Version 11/01/08

β=5. For the TREC 2003 [10], the best F-measure was 0.555 also with β=5, and the TREC 2004 [11] F-measure was 0.460, now with β=3.

In contrast to the traditional way to extract nuggets, we propose a method that uses information gain to retrieve the most relevant information. First, we obtain passages from the AQUAINT Corpus using the retrieval system Lucene[1]. Next, from the passages, we extract the relevant sentences, these are further parsed (using Link Grammar [4]) to obtain clauses. Then, from the clauses, we select four kinds of sentence fragments, these are: noun phrases containing an appositive phrase, noun phrases containing two noun phrases separated by comma, embedded clauses, and main or subordinate clauses without considering embedded clauses. Finally, the sentence fragments are separated in two kinds of fragments, i.e. the fragments to the *left* and *right* of the target term. We then assess the information gain of sentence fragments to decide which are the most relevant, and in consequence select them as part of the final answer.

For this task, we work with the questions of the *pilot* evaluation of definition questions 2002 [9], *definition* questions from TREC 2003 [10] and *other* questions from TREC 2004 [11]. First, we test the output of two retrieval systems JIRS[2] and Lucene. In the second experiment, we test balanced and non-balanced sets of sentence fragments from the *right* and *left* sets. Finally, we compare the F-measure obtained with our system DefQuestions_IG against the participating systems in the TREC conferences.

The paper is organized as follows: next section describes the process to extract sentence fragments; Section 3 describes the approaches used and the method to retrieve only definition sentence fragments; Section 4 reports experimental results; finally, some conclusions and directions for future work are presented in Section 5.

2 Sentence Fragments Extraction

Official definition of F-measure used in the TREC evaluations [9] is:

Let

r : be the number of vital nuggets returned in a response

a : be the number of non-vital nuggets returned in a response

R : be the total number of vital nuggets in the assessors list

l : be the number of non-whitespace characters in the entire answer string

Then

$$recall(\Re) = r / R \tag{1}$$

$$allowance(\alpha) = 100 \times (r + a) \tag{2}$$

$$precision(\mathrm{P}) = \begin{cases} 1 & if \quad l < \alpha \\ 1 - \dfrac{l - \alpha}{l} & otherwise \end{cases} \tag{3}$$

[1] http://lucene.apache.org/

[2] http://jirs.dsic.upv.es/

Finally, for a given β, F-measure is: $F(\beta = 3) = \frac{(\beta^2 + 1) \times P \times \Re}{\beta^2 \times P + \Re}$ (4)

Thus, a reason to extract sentence fragments is that we need to retrieve only the most important information from relevant sentences. Other reason to extract short sentence fragments is related to the performance F-measure applied to definition systems in the TREC evaluation; this measure combines the recall and precision of the system. The precision is based on length (in non-white-space characters) used as an approximation to nugget precision. The length-based measure starts from an initial allowance of 100 characters for each (vital or non-vital) nugget matched. Otherwise, the measure value decreases as the length the sentence fragment increases.

After our experiments comparing two retrieval systems (and detailed later on), we decide to use Lucene as main system to extract candidate paragraphs from the AQUAINT Corpus of English News Text. From these candidate paragraphs, we extract the relevant sentences, i.e. the sentences that contain the target term. Then, to extract sentence fragments we propose the following process:

1) **Parse the sentences.** Since we need to obtain information segments (phrases or clauses) from a sentence, the relevant sentences were parsed with Link Grammar [6]. We replace the target term by the label **SCHTERM**. As an example, we get the following sentence for the target term **Carlos the Jackal**:

   ```
   The man known as Carlos the Jackal has ended a hunger
   strike after 20 days at the request of a radical Pales-
   tinian leader, his lawyer said Monday.
   ```

   The Link Grammar the produces the following output with the target term replaced as detailed above:

   ```
   [S [S [NP [NP The man NP] [VP known [PP as [NP
   SCHTERM NP] PP] VP] NP] [VP has [VP ended [NP a
   hunger strike NP] [PP after [NP 20 days NP] PP]
   [PP at [NP [NP the request NP] [PP of [NP a radi-
   cal Palestinian leader NP] PP] NP] PP] VP] VP] S]
   , [NP his lawyer NP] [VP said [NP Monday NP] . VP]
   S]
   ```

2) **Resolve co-references.** We want to obtain main clauses without embedded clauses or only embedded clauses, so we need to resolve the co-reference, otherwise important information can be lost. To resolve co-reference the relative pronouns WHNP are replaced with the noun phrase preceding it.

3) **Obtain sentence fragments.** An information nugget or an atomic piece of information can be a phrase or a clause. We analyzed the sentences parsed with Link Grammar and we identified four main kinds of sentence fragments directly related to the target and with a high possibility that their information define the target. These fragments are:

   Noun phrase (NP) containing an appositive phrase.
   Noun phrase (NP) containing two noun phrases separated by comma [NP, NP].
   Embedded clauses (SBAR).
   Main or subordinate clauses (S) without considering embedded clauses.

To retrieve the four kinds of sentence fragments, we analyze the parse tree, according to the following procedure:

i) Look for the nodes which contain the target, in our case, the label SCHTERM.

ii) Find the initial node of the sentence fragment. The process analyzes the path from the node with the SCHTERM label towards the root node. The process stops when a NP with appositive phrase, NP with [NP, NP], an embedded clause SBAR, or a clause S is reached.

iii) Retrieve the sentence fragment without embedded clauses.

iv) Mark as visited the parent node of the second phrase. For the case [NP1, NP2], we mark as visited the parent node of NP2. For appositive phrase, SBAR or S, the second phrase can be NP, VP or PP.

The steps ii to iv are repeated for the same node with a SCHTERM label until a visited node is found in the path to the node towards the root, or the root node is reached. Also the steps ii to iv are repeated for each node found in step i.

The next module of our definition question system selects definition sentence fragments. In order to select only definition nuggets from all of sentence fragments, we analyze separately, the information that is to the left of SCHTERM and the information that is to the right of SCHTERM, so we form two data sets.

Now, we present some sentence fragments of two sets (right and left) obtained using the process for the target term **Carlos the Jackal**:

Right sentence fragments

SCHTERM , a Venezuelan serving a life sentence in a French prison

SCHTERM , nickname for Venezuelan born Ilich Ramirez Sanchez

SCHTERM , is serving a life sentence in France for murder

SCHTERM as a comrade in arms in the same unnamed cause

SCHTERM refused food and water for a sixth full day

SCHTERM , the terrorist imprisoned in France

Left sentence fragments

the friendly letter Chavez wrote recently to the terrorist SCHTERM

The defense lawyer for the convicted terrorist known as SCHTERM

he was harassed by convicted international terrorist SCHTERM

an accused terrorist and a former accomplice of SCHTERM

Ilich Ramirez Sanchez , the terrorist known as SCHTERM

Ilich Ramirez Sanchez , the man known as SCHTERM

We found that analyzing separately the sentence fragments before and after the target term is an advantage since in many candidate sentences, only one part contains

information defining the target term. When a fragment appears in both sides, this serves to affirm its informative feature, as assessed by information gain.

3 Nuggets Selection

In order to obtain only the most informative nuggets from the *left* and *right* sentence fragments, we use information gain.

3.1 Information Gain

The information gain [1] for each word or term l is obtained using the following definition:

Given a set of sentence fragments D, the entropy H of D is calculated as:

$$H(D) \equiv \sum_{i=1}^{c} - p_i \log_2 p_i \tag{5}$$

In this expression, P_i is the probability of i word, and c is the size of the vocabulary. Then, for each term l, let D^+ be the subset of sentence fragments of D containing l and D^- denotes its complement. The information gain of l, $IG(l)$, is defined by:

$$IG(l) = H(D) - \left[\frac{|D^+|}{|D|} H(D^+) + \frac{|D^-|}{|D|} H(D^-) \right] \tag{6}$$

3.2 Method to Select Nuggets

The process to obtain informative nuggets using information gain is the following:

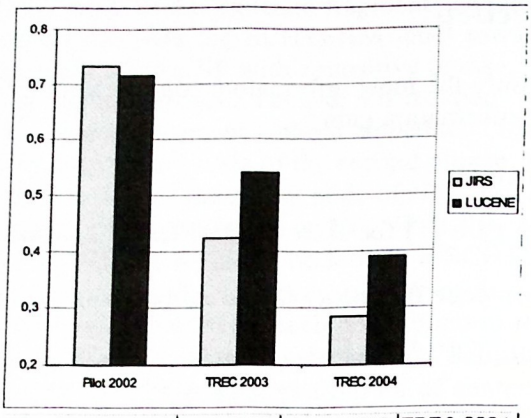
I) Obtain the vocabulary of all the sentence fragments (*left* and *right* sets).
II) Calculate the information gain for each word of the vocabulary using the definition of section 3.1.
III) Using the value of the information gain of each word (except stop words), calculate the sum of each sentence fragment.
IV) Rank the sentence fragments according to the value of the sum.
V) Eliminate redundant sentence fragments.

To eliminate redundancy, we compare pairs (X, Y) of sentence fragments using the following steps:

a) Obtain the word vector without empty words for each sentence fragment.
b) Find the number of similar words (*SW*) between the two sentence fragments.
c) If $\frac{SW}{|X|} \geq \frac{2}{3}$ or $\frac{SW}{|Y|} \geq \frac{2}{3}$, remove the sentence fragment with lower sum of information gains of the vector.

For example, if we have the following sentence fragments for the target **Carlos the Jackal,** with their corresponding sums:

```
2.290 nickname for Venezuelan born Ilich Ramirez Sanchez
```



	Pilot 2002	TREC 2003	TREC 2004
JIRS	0,733	0,424	0,284
LUCENE	0,715	0,54	0,39
% of Improvement	-2,456	27,358	37,324

Fig. 1. Comparison of the F-measures obtained with two different retrieval systems JIRS and Lucene.

```
2.221 Ilich Ramirez Sanchez , the Venezuelan born former
guerrilla
2.157 Ilich Ramirez Sanchez , the terrorist
1.930 Ilich Ramirez Sanchez , the man
1.528 Illich Ramirez Sanchez
```

If we compare the first and the second sentences, the result of the step a) is:

```
[nickname, Venezuelan, born, Ilich, Ramirez, Sanchez]
[Ilich, Ramirez, Sanchez, Venezuelan, born, former, guerrilla]
```

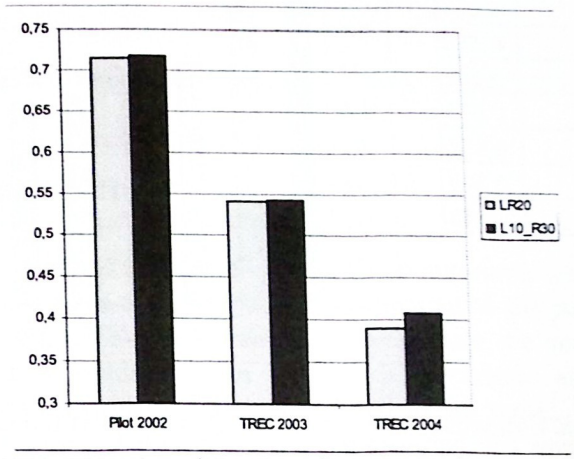
In the step b) we obtained that the number of similar words is SW=5.

Finally, in the step c) we remove the second sentence fragment since it has a lower sum of information gains. Applying the procedure with the remaining sentence fragments, the result is that we keep only the first:

```
2.290 nickname for Venezuelan born Ilich Ramirez Sanchez
```

4 Experiment Results

We performed experiments with three sets of definition question, the questions from the *pilot* evaluation of definition question 2002 [9], *definition* questions from TREC 2003 [10], and *other* questions from TREC 2004 [11]. (We did not compare our results with the collections of the TREC 2005 and 2006 since in these years the list of nuggets was not readily available). To obtain passages, first we test the output of two retrieval systems JIRS and Lucene, since the overall performance of definition question system depends on the resources and tools used for answer finding [7,8]. Figure 1 shows the comparisons of the F-measure, the best results are obtained with the general propose system (Lucene) instead of JIRS, which is intended to retrieve passages for factoid questions. These results led to the decision to keep using Lucene in further experiments (instead of JIRS), since it provides an improved set of paragraphs.



	Pilot 2002	TREC 2003	TREC 2004
LR20	0,715	0,54	0,39
L10_R30	0,717	0,542	0,408
% of Improvement	0,28	0,37	4,615

Fig. 2. Comparison of the F-measures obtained with balanced sets LR20 and non-balanced sets L10_R30.

In the second experiment, we try to identify which set (*left* or *right*) contributes more for the identification (since we found that the *right* set is usually larger than *left* set). So we set the experiment comparing the results of taking the first 20 sentence fragments from the *left* and the first 20 fragments from *right* sets against taking a ratio of 1:3 between *left* and *right* sets, i.e. we take 10 sentence fragments from the left set and 30 from the right set. We obtained the best results with non-balanced sets, as presented in figure 2.

Thus, we built a system using Lucene to obtain the passages. From these passages we retrieve relevant sentences. Then, we applied a parser (Link Grammar [4]) to analyze the relevant sentences in order to get clauses. Next, from the clauses we ob-

tained the four kinds of sentence fragments detailed above, in section 2. Finally, the sentence fragments were separated in two kinds of fragments, the fragment to the *left* and *right* of the target term. The approach of information gain is then applied to these sentence fragments to obtain the most relevant fragments. Also, we used non-balanced sets of sentence fragments, as the results of the second experiment suggested. Figure 3 displays the F-measure obtained with our system (DefQuestion_IG) compared against the systems of the pilot version of definition questions proposed in 2002. Figure 4 shows the comparison of the F-measures of our system with the systems that participated in the TREC 2003. Finally, figure 5 presents the comparison of the F-measures of the systems in the TREC 2004 and our system. From the figures 3 to 4, we can observe that our system DefQuestions_IG showed very competitive results.

1	**DefQuestion_IG**	**0,717**
2	F	0,688
3	A	0,606
4	D	0,568
5	G	0,562
6	E	0,555
7	B	0,467
8	C	0,349
9	H	0,33

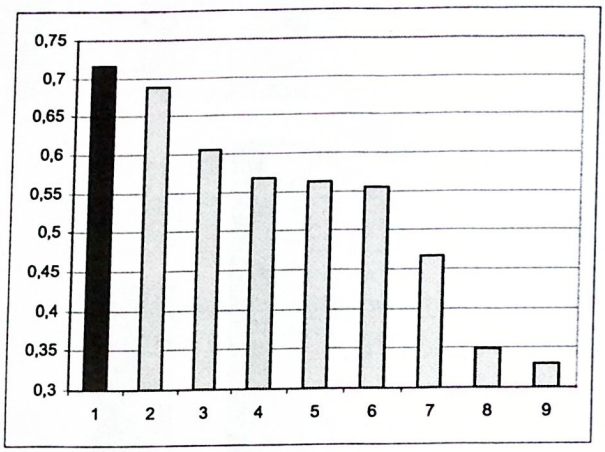


Fig. 3. Comparison of F-measure values of pilot evaluation of definition questions using the AUTHOR list of nuggets.

1	BBN	0,555
2	**DefQuestion_IG**	**0,543**
3	National University of Singapore	0,473
4	University of Southern Calif. ISI	0,461
5	Language Computer Corp.	0,442
6	Univ. Of Colorado/Columbia Univ.	0,338
7	ITC-irst	0,318
8	Univ. Of Amsterdam	0,315
9	MIT	0,309
10	Univ. Of Sheffield	0,236
11	Univ. Of Iowa	0,231
12	Carnegie Mellon University	0,216
13	Fundan University	0,192
14	Univ. Of Pisa	0,185
15	IBM Research	0,177
16	NTT Communication Science Labs	0,169

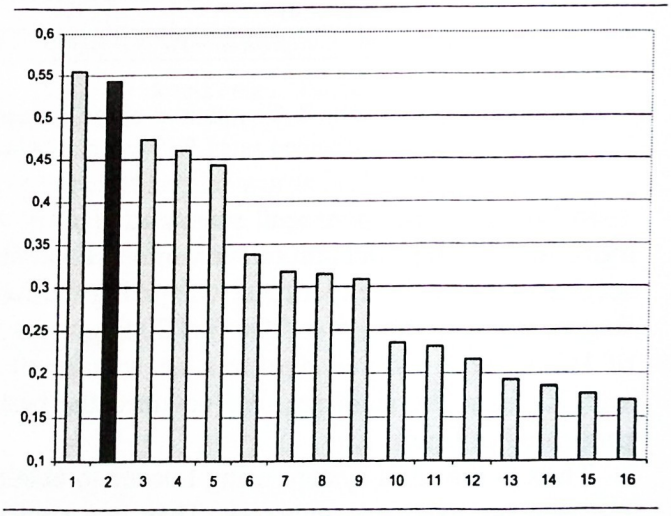


Fig. 4. Comparison of F-measure values of TREC 2003.

1	National Univ. Of Singapore	0,46
2	**DefQuestion_IG**	**0,408**
3	Fudan University	0,404
4	National Segurity Agency	0,376
5	University of Sheffield	0,321
6	University of North Texas	0,307
7	IBM Research	0,285
8	Korea University	0,247
9	Language Computer Corp.	0,24
10	CL Research	0,239
11	Saarland University	0,211

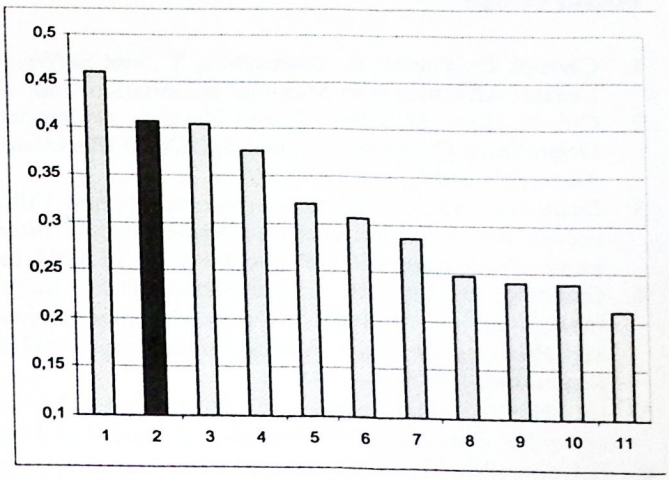


Fig. 5. Comparison of F-measure values of TREC 2004.

5 Conclusions and Future Work

We have presented a method to extract nuggets in an automatic and flexible way and the results obtained are quite competitive when compared to the participating systems in the TREC whose sets of nuggets were used to evaluate the output of our system. The sentence fragments obtained with the process presented are acceptable since these contain only the information directly related to the target. Other advantage is that these sentence fragments present a short length, and this improves the precision of our definition question system.

Future work includes combining Machine Learning algorithms with Information Gain to identify definition sentence fragments since we have showed previously [7] that the combination can improve the performance of the system. Also we plan to categorize the targets in three classes: ORGANIZATIONS, PERSON and ENTITIES and then train three different classifiers. We expect that in this way we can exploit the peculiarities of each kind of entity.

Acknowledgements. We would like to thank CONACyT for supporting this work under scholarship grant 157233 for the first author, and the second author was partially supported by SNI, CONACyT México.

References

1. Carmel, D., Farchi, E., Petruschka, Y., and Soffer, A.: Automatic Query Refinement using Lexical Affinities with Maximal Information Gain. *SIGIR* (2002): 283-290.
2. Cui, H., Kan, M. Chua, T. and Xiao, J.: A Comparative Study on Sentence Retrieval for Definitional Questions Answering. *SIGIR Workshop on Information Retrieval for Question Answering (IR4QA)*, (2004) 90-99.
3. Denicia-Carral, C., Montes-y-Gómez, M., and Villaseñor-Pineda, L.: A Text Mining Approach for Definition Question Answering. *5th International Conference on Natural Language Processing, Fin Tal. Lecture Notes in Artificial Intelligence, Springer* (2006).
4. Grinberg, D., Lafferty, J., and Sleator, D.: A Robust Parsing Algorithm for Link Grammars. Carnegie Mellon University Computer Science technical report CMU-CS-95-125, and *Proceedings of the Fourth International Workshop on Parsing Technologies*, Prague, September, (1995).
5. Harabagiu, S., and Lacatusu, F., Strategies for Advanced Question Answering. *In Proceeding of the Workshop on Pragmatics of Questions Answering at HLT-NAACL.* (2004): 1-9.
6. Hildebranddt, W., Katz, B. and Lin, J.: Answering Definition Question Using Multiple Knowledge Sources. *In Proceeding of HLT/NAACL*, Boston (2004): 49-56.
7. Martínez-Gil, C., and López-López, A.: Answering Definition Questions using a Statistical Method Combined with a Classifier Ensemble, in *Proceedings of the Seventh International Symposium on Natural Language Processing, SNLP2007*, Thailand, December, (2007): 189-194.
8. Moldovan, D., Pasca, M., Harabagiu,S., and Surdeanu, M.: Performance Issues and Error Analysis in an Open-Domain Question Answering System. *Proceeding of the 40th Annual Meeting of the Association for Computational Linguistics (ACL).* Philadelphia, July, (2002): 33-40.
9. Voorhees, E.: Evaluating Answers to Definition Questions. *NIST* (2003) 1-3.
10. Voorhees, E.: Overview of the TREC 2003 Question Answering Track. *NIST* (2003): 54-68.
11. Voorhees, E.: Overview of the TREC 2004 Question Answering Track. *NIST* (2004): 12-20.
12. Xu, J., Licuanan A., and Weischedel R.: TREC 2003 QA at BBN: Answering Definitional Questions. In the *Twelfth Text Retrieval Conference* (2003): 28-35.