FIDJI @ ResPubliQA'10

Xavier Tannier, Véronique Moriceau

LIMSI-CNRS
Univ. Paris-Sud, Orsay, France
xtannier, moriceau@limsi.fr

Abstract. In this paper, we present the results obtained by the system FIDJI for both French and English monolingual evaluations, at ResPubliQA 2010 campaign. In this campaign, we focused on carrying on our evaluations concerning the contribution of our syntactic modules with this specific collection.

1 Introduction

FIDJI (Finding In Documents Justifications and Inferences) is an open-domain question-answering (QA) system for French [1] and, more recently, English. It combines syntactic information with traditional QA techniques such as named entity recognition and term weighting in order to validate answers through different documents.

This paper focuses on the results obtained by FIDJI at ResPubliQA 2010 evaluation. It presents first a brief overview of the system and of its adaptation to English. Then, the specific choices made for the campaign are detailed, and some results are finally given.

2 FIDJI

Figure 1 presents the architecture of FIDJI. The system relies on a syntactic analysis and named entity tagging of the question and of a limited number of documents for each question. This analysis is performed by the parser XIP [2] enriched with some additional specific rules.

The document collection is indexed by the search engine Lucene[1]. The index contains raw text only. First, the system analyses the question and submits the keywords of the question to Lucene (module A): the first 15 documents are then processed (module B). We decided to reduce the number of documents because they are rather long and their parsing would take too much time. The reason we perform this analysis online is that we aim at avoiding as much preprocessing as possible (the system is designed to explore Web collections [1]). Among these documents, FIDJI looks for sentences containing the highest number of syntactic relations of the question (module C1). Finally, answers are extracted from these

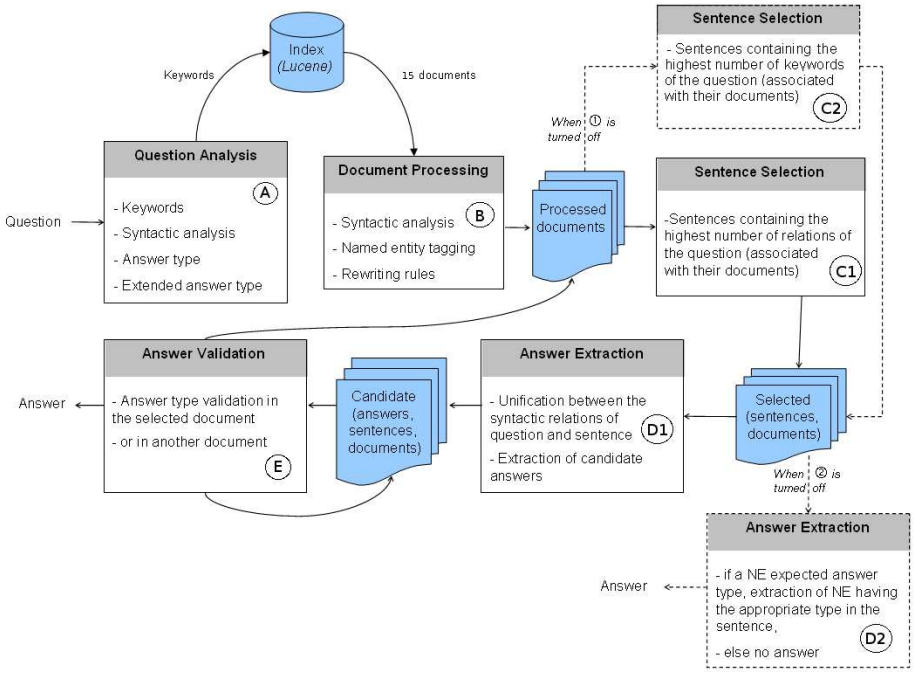

Fig. 1. Architecture of FIDJI

sentences (module D1) and the answer type, when specified in the question, is validated (module E).

The main objective of FIDJI is to produce answers which are fully validated by a supporting text (or passage) with respect to a given question. The difficulty is that an answer (or some pieces of information composing an answer) may be validated by several documents.

Our approach consists in checking if all the characteristics of a question (namely the dependency relations and the answer type) may be retrieved in one or several documents. In this context, FIDJI has to detect syntactic implications between questions and passages containing the answers and to validate the type of the potential answer in this passage or in another document.

Since the last evaluation campaign in 2009, FIDJI has been adapted to English. Specific rules have been developped for question analysis (module A) and document processing (module B). The other modules are common to both English and French.

The following examples illustrate how FIDJI extracts answers, and more details concerning the system can be found in [1].

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

2.1 Example 1

Question analysis provides lemmatisation, POS tagging and dependency relations, as well as the question type and the expected answer type. For example:

Question: Quel premier ministre s'est suicidé en 1993 ?
(*Which Prime Minister committed suicide in 1993?*)
Dependencies: `DATE(1993)`
`PERSON(ANSWER)`
`SUBJ(se suicider, ANSWER)`
`attribut(ANSWER, ministre)`
`attribut(ministre, premier)`
Question type: factoid
Expected answer type: person (specific answer type: prime minister)

The question is turned into a declarative sentence where the answer is represented by the 'ANSWER' lemma. The following sentence is selected because it contains the highest number of dependency relations:

Pierre Bérégovoy s'est suicidé en 1993.
(*Pierre Bérégovoy committed suicide in 1993.*)
Dependencies:
`DATE(1993)`
`PERSON(Pierre Bérégovoy)`
`SUBJ(se suicider, Pierre Bérégovoy)`

Pierre Bérégovoy instantiates the ANSWER slot of the question dependencies and becomes a candidate answer. The named entity type (person) and the first three dependencies of the question are validated in this sentence. In order to fully validate the candidate answer, the system searches the missing dependencies (`attribut(Pierre Bérégovoy, ministre)` and `attribut(ministre, premier)`) in a single sentence of the whole document collection. These dependencies will be found in any sentence speaking about "*le premier ministre Pierre Bérégovoy*" (*Prime Minister Pierre Bérégovoy*) and the answer will be validated.

2.2 Example 2

For complex questions, it is obvious that answers are not always short phrases. For this reason, FIDJI provides a full passage as an answer. On these kinds of questions, the system behaves as a classical passage retrieval system, except that candidate passages are retrieved through syntactic relations and relevant discourse markers (about 100 nouns, verbs, prepositions and adjectives, manually compiled) instead of keywords only. Here is an example of a complex question:

Question: Why is the sky blue?
Dependencies: `attribut(sky, blue)`

Question type: complex (why)
Expected answer type: reason[2]

The following passage is selected because it contains all the dependency relations of the question and a causal marker:

And if the sky is blue, it is ***because of*** *Rayleigh scattering ...*

```
attribut(sky, blue)
VMOD(be, scattering)
PREPOBJ(scattering, because of)
...
```

3 ResPubliQA'10 experiments

In 2009, ResPubliQA results learned us a lot about the behavior of our system.

Other evaluations (former CLEF and Quaero campaigns) had shown that using syntactic analysis modules for retrieving documents and extracting the answers significantly improved the results [1]. However, with ResPubliQA evaluation set, passage extraction turned out to be much better by replacing syntax by traditional bag-of-words techniques [3]. This is done by turning off modules C1 and D1 in Figure 1.

Passage extraction is then performed by a classical selection of sentences containing a maximum of question significant keywords (module C2), and answer extraction is achieved without slot instantiation within dependencies (module D2).

The new guidelines in ResPubliQA 2010 offered us the possibility to carry on our experiments in this way. Indeed, two different tasks were allowed this year:

- Paragraph selection (PS), similar to 2009 task, where only the full paragraph containing the exact answer were to be returned. Passages are not indefinite parts of texts of limited length, but predefined paragraphs identified in the corpus by XML tags `<p>`.
- Answer selection (AS), closer to traditional QA tasks, where systems were required to demarcate also the exact answer, supported by a full paragraph.

In this latter task, judged answers can be "INEXACT" (good support but bad boundaries for short answer), "MISSED" (good support but wrong short answer), "RIGHT" (good support and good answer) or "WRONG".

Two runs per language were allowed. In order to continue testing our plug/unplug strategies, and to experiment them for the first time in English, we chose the following procedure for our two runs:

[2] "Reason" is not a named entity, as "person" in the first example, but this answer type points out that a text explicitely explaining a reason should be prefered (in our case, using discourse markers).

1. **PS task**, syntactic modules **turned off**, leading to an approach closer to passage retrieval, that had the best results of the system last year.
2. **AS task**, syntactic modules **turned on**, in order to test whether answer extraction was effective or not on this collection. Moreover, by adding answers with "INEXACT", "MISSED" and "RIGHT" status from our AS run, we can obtain a "PS" run with modules turned on, which allows us to evaluate modules on the same task.

4 Results

We present the results of 5 experiments for both French and English. The first three come from official ResPubliQA runs:

- ①: AS task with syntactic modules turned on (exact answers judged as "RIGHT"),
- ②: PS task with syntactic modules turned on (exact answers of ① judged as "RIGHT", "INEXACT", "MISSED"),
- ③: PS task with syntactic modules turned off.

To complete the evaluation, we also ran unofficial configuration and achieved the assessment by ourselves:

- ④: AS task with passage retrieval turned off but answer extraction turned on (modules C2 and D1, with exact answers judged as "RIGHT"),
- ⑤: PS task with passage retrieval C1 turned off but answer extraction turned on (exact answers of ④ judged as "RIGHT", "INEXACT", "MISSED").

In order to evaluate the performance of the question analysis module, we manually identified the types of question. As FIDJI cannot process opinion questions, we decided to consider them as factoid. Although questions in French and English are translations of each other and their respective answer should be extracted from the same paragraph, we noticed that, for a given question, its type is not always the same in English as in French. For example, in English, the type of question 169 is *reason/purpose* while in French, it is *factoid*:

(EN) *Why is the trade in ammonium nitrate fertilizers hampered within the European Economic Community?*
(FR) *Qu'est-ce qui a entravé le commerce d'engrais à base de nitrate d'ammonium dans la Communauté Économique Européenne?* (*What has hampered the trade in ammonium nitrate fertilizers...?*)

This is not only an issue of syntactic differences due to translation paraphrasing; the target of the question is different. Strictly speaking, the French question might accept a noun phrase like "*les réglementations régissant la commercialisation des engrais à base de nitrate d'ammonium*" (*the different regulations controlling the marketing of ammonium nitrate based fertilizers*), while such an

answer would be odd with the English question. We identified 7 questions raising this issue[3].

Tables 1 and 2 presents FIDJI's results for runs ①, ② and ③, as well as experiments ④ and ⑤, by types of questions (manually identified). In French, 86% of question types were correctly identified by FIDJI (we found 9 questions that were ill-formed or with misspellings and which FIDJI could not correctly analyse) whereas in English, only 69.5% were correctly identified.

Concerning our official runs, as we can see in Tables 1 and 2, answer extraction performance (①) is very low (0.25 for both English and French). Results are better for passage selection (② and ③) for every type of questions and even better when syntactic modules are switched off (③). Results are globally better for English than for French so the performance of the question analysis module cannot explain these results.

In both languages, correct answers to definition questions dramatically decrease with D1 turned off. This is because we do not have any non-syntactic way to extract the answer for many of these questions (definitions not expecting a named entity, as *What is maladministration?*, can only be answered by definition patterns in FIDJI). Turning off syntactic modules necessarily leads to a *NOA* answer in these cases.

We can notice that for both English and French, the results follow the same trend and that results for passage selection are better for "complex" questions (reason/purpose and procedure), probably because FIDJI selects passages containing discourse markers for this type of questions. Also, for these questions, we always returned the full paragraph as exact "short" answer, considering that trying to focus even more inside the paragraph was not useful for such questions. As the assessors did consider that shorter answers can be better, the system often gets an "INEXACT" status for.

Finally, our additional runs ④ and ⑤ show a small improvement, showing that best results are obtained when turning off syntactic passage retrieval, but turning on syntactic answer extraction (using modules C2 and D1). This is at least clear concerning non-factoid questions. This finding is important and will help us in the future to choose our search strategies according to different corpora and question types.

Last year, the "pure information retrieval" baseline [4] which consisted in querying the indexed collection with the exact text of the question and returning the paragraph retrieved in the first position, had the best results for French and ranked 5 out of 14 in English [5]. Even if a subset of the Europarl corpus has been added to the document collection in 2010, we can see that our c@1 measures (see Table 3) are still lower than the 2009 baseline (0.53 for English and 0.45 for French).

In 2009, we noted that our results were due to ACQUIS corpus specificities: different register of language, more constrained vocabulary, texts having a particular structure, with an introduction followed by long sentences extending on sev-

[3] Questions 3, 11, 134, 169, 175, 197, 199.

Type of questions		Factoid	Definition	Reason/Purpose	Procedure	TOTAL
Number of questions		110	29	29	32	200
①	Correct answers	10 (9.1%)	3 (10.3%)	1 (3.5%)	3 (9.4%)	17 (8.5%)
②	Correct passages	33 (30%)	10 (34.5%)	10 (34.5%)	14 (43.8%)	67 (33.5%)
③	Correct passages	51 (46.3%)	3 (10.3%)	18 (**62%**)	17 (**53.1%**)	89 (44.5%)
Unofficial runs						
④	Correct answers	13 (11.8%)	3 (10.3%)	2 (6.9%)	4 (12.5%)	22 (11%)
⑤	Correct passages	47 (42.7%)	9 (31.0%)	19 (65.5%)	18 (56.3%)	93 (46.5%)

Table 1. Results by question type (English).

Type of questions		Factoid	Definition	Reason/Purpose	Procedure	TOTAL
Number of questions		117	29	26	28	200
①	Correct answers	11 (9.4%)	2 (6.9%)	0 (0%)	1 (3.6%)	14 (7%)
②	Correct passages	35 (29.9%)	6 (20.7%)	8 (30.8%)	8 (28.6%)	57 (28.5%)
③	Correct passages	30 (25.6%)	6 (20.7%)	13 (**50%**)	13 (**46.4%**)	62 (31%)
Unofficial runs						
④	Correct answers	12 (10.3%)	3 (10.3%)	0 (0%)	2 (6.3%)	17 (8.5%)
⑤	Correct passages	31 (28.2%)	7 (24.1%)	14 (53.8%)	15 (50.0%)	67 (33.5%)

Table 2. Results by question type (French).

eral paragraphs, etc. Table 4 shows that FIDJI found correct answers/passages mainly in the ACQUIS collection. As FIDJI has difficulty with selecting passages in the ACQUIS collection, FIDJI's low results could be explained if a majority of correct answers are in the ACQUIS collection.

The main difference between FIDJI architecture used for ResPubliQA and the one used for other evaluation campaigns (CLEF, Quaero) is the number of documents returned by Lucene: 15 documents for ResPubliQA and 100 for other campaigns. We have to evaluate if selecting more documents would improve the results.

Campaign	FIDJI 2010		FIDJI 2009	
Language	English	French	English	French
①	0.09	0.08	-	-
②	0.35	0.30	-	0.30
③	0.48	0.36	-	0.42
④	0.11	0.08	-	-
⑤	0.47	0.34	-	-

Table 3. c@1 measure for French and English.

Language	English		French	
Corpus	Europarl	Acquis	Europarl	Acquis
①	3	14	6	8
②	24	43	22	36
③	33	56	21	41

Table 4. Number of correct answers/passages per corpus.

5 Conclusion

We presented in this paper our participation to the campaign ResPubliQA 2010 in French and English. We evaluated two strategies: plugging or unplugging the syntactic modules for document selection and answer extraction. As in 2009, the system got low results and even lower when syntactic modules are turned off. Different experiments on the collection confirmed that the use of syntactic analysis decreased results, whereas it proved to help when used in other campaigns. We still have to evaluate if a higher number of documents selected by the search engine can improve the results.

6 Acknowledgements

This work has been partially financed by OSEO under the Quaero program.

References

1. Moriceau, V., Tannier, X.: FIDJI: Using Syntax for Validating Answers in Multiple Documents. Information Retrieval, Special Issue on Focused Information Retrieval **10791** (2010)
2. Aït-Mokhtar, S., Chanod, J.P., Roux, C.: Robustness beyond shallowness: Incremental deep parsing. Natural Language Engineering **8** (2002) 121–144
3. Tannier, X., Moriceau, V.: Studying Syntactic Analysis in a QA System: FIDJI @ ResPubliQA'09. In: Proceedings of CLEF 2010. Number LNCS 6241 in Lecture Notes in Computer Science, Springer-Verlag, New York City, NY, USA (2010)
4. Pérez, J., Garrido, G., Álvaro Rodrigo, Araujo, L., Peñas, A.: Information Retrieval Baselines for the ResPubliQA Task. In: Working Notes for the CLEF 2009 Workshop, Corfu, Greece (2009)
5. Peñas, A., Forner, P., Sutcliffe, R., Rodrigo, A., Forăscu, C., Alegria, I., Giampiccolo, D., Moreau, N., Osenova, P.: Overview of ResPubliQA 2009: Question Answering Evaluation over European Legislation. In: Working Notes for the CLEF 2009 Workshop, Corfu, Greece (2009)