The LIA at QA@CLEF-2006

Laurent Gillard, Laurianne Sitbon, Eric Blaudez, Patrice Bellot and Marc El-Bèze
LIA, University of Avignon
339 ch. des Meinajaries, BP 1228, F-84911 Avignon Cedex 9, France
{laurent.gillard, laurianne.sitbon, eric.blaudez,
patrice.bellot, marc.elbeze} @univ-avignon.fr

Abstract

This article presents the first participation of the Laboratoire Informatique d'Avignon (LIA) to the Cross Language Evaluation Forum (CLEF). LIA participated to the monolingual Question Answering (QA) track dedicated to French language, and to the cross-lingual English to French QA track. Two runs for each track were submitted. English questions were first translated and then answered by using the monolingual French system. This QA System (QAS) already participated to the French Technolangue QA campaign (EQueR) but some improvements needed to be evaluated: definition questions answering module; or were developed specifically for CLEF: integration of Lucene search engine, and re-ranking based on redundancy for factoid answer candidates. The CLEF-QA provided an opportunity to evaluate these modules. Also, English conversion of the QAS was started, even if, only the Question Analysis module was adapted to English this year.

The generic factoid QAS is based on an extraction of answer candidates in the form of Named Entities (NE) by using keywords density measures. Few factoid answers were also provided by a knowledge base module. Lastly, definition questions were answered by a module based upon detection of frequent appositive informational nuggets appearing near the focus to define.

The system obtained reasonable results in all runs except for Temporal and List questions that were not recognized as such and wrongly handled as simple factoid one.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.1 Content Analysis and Indexing; H.3.3 Information Search and Retrieval; H.3.4 Systems and Software

General Terms

Measurement, Performance, Experimentation

Keywords

Question answering, Questions beyond factoids, Definition questions, Keywords density metrics, Multilingual question answering.

1 Introduction

Question Answering (QA) systems aim at retrieving precise answers to questions expressed in natural language rather than list of documents that may contain an answer. They have been

particularly studied since 1999 and the first large scale QA evaluation campaign held as a track of the Text REtrieval Conference [10]. Since 2003, the Cross Language Evaluation Forum (CLEF) studies multilingual issues of QA and provide an evaluation platform for QA dedicated to many languages.

It is the first participation of the Laboratoire Informatique d'Avignon (LIA) to CLEF (we already participated to monolingual English, TREC-11, and monolingual French, the EQueR Technolangue QA campaign [3]). This year LIA participated in two tracks: monolingual French and English to French. For both submissions, the main system was quite the same and inherited of the one we built for our EQueR participation [5]. Moreover, CLEF-2006 allowed us to evaluate some improvements: English question analysis module, integration of Lucene as a search engine, a module to handle definition questions, and another one to experiment on re-ranking by using redundancy for answer candidates. For the two tracks, two run were submitted, the second run for each language used redundancy re-ranking module.

Our QA system (QAS) follows the typical QAS architecture, and involves pipelined main components. A Question Analysis (describes in section 2) is first done to extract the semantic type(s) of the expected answer(s) and keywords but also to decide which subsystem, factoid or definition questions, must be used. The factoid QAS (section 3) performs Document Retrieval (section 3.2) to restrict the amount of processed data by next components; Passage Retrieval (section 3.3) to choose the best answering passages from documents; and finally Answer Extractions to determine the best answer candidate(s) drawn from the previously selected passages. This answer extraction is mainly done by using a density (section 3.4.1) of the keywords appearing around an answer candidate, but may also involve knowledge bases (section 3.4.2). The definition subsystem (section 4) use frequent appositive nuggets of information appearing near the focus to define it. We also experiments briefly with redundancy to re-rank answer candidates but this module was broken. Lastly result of our participation will be presented and discussed with perspectives of future improvements.

2 Question Analysis

Definition questions are firstly recognized with a simple pattern matching process, which also extracts the focus of the definition. All questions which don't fit these definition patterns are considered as factual questions, including list questions (and consequently, they were wrongly answered as factoid questions with only one instance, this behaviour must be corrected).

Then, the analysis of factual questions contains two main independent steps which are: question classification and keywords extraction. For the questions written in English, the keywords extraction step is done after a translation of the whole question by Google Translator[1].

2.1 Expected Answer Types Classification

The answer nugget expected for a factoid question is considered to be a Named Entity (NE) based upon Sekine's hierarchy [8]. Thus, questions are classified according to this hierarchy. This determination is done with patterns and rules for questions in French; and, complemented, but only for English questions, with semantic decision trees (a machine learning approach introduced by [6] and [1]).

Semantic decision trees are adapted as described in [2] to automatically extract generic patterns from questions in a learning corpus. The corpus *CLEF multinine* composed of 900 questions was used for learning. As all questions are available in both English and French languages, French questions were first automatically classified with the French rules based module, then, checked manually, and lastly, paired with their English translation to be used as a learning corpus. The learning features are words, part-of-speech tags[2], lemmas, and the number of words.

[1] `http://www.google.fr/language_tools`

[2] All the part-of-speech tags and lemmas used in our QAS are obtained with the help of the TreeTagger [9].

Language	OK	Wrong	Unknown
French	86%	2%	12%
English	78%	13%	5%

Table 1: Evaluation of the classification process on CLEF-2006 questions

The table 1 shows the results of a manual post-evaluation of the classification for the CLEF-2006 questions. Wrong and unknowns tags prevent an extraction of the answer in the downstream components. In French, only 8 of the 19 unknown tagged questions were classifiable (as being *Person* name or *Company* name, other are difficult to map to Sekine's taxonomy). Due to its pipelined architecture, the percentage of OK tags fixed the maximum score our QAS can finally achieve. Some of the correct classification in English are actually more generic than they could be, for example *President* is tagged as *Person*.

After assigning questions to classes from the hierarchy, a mapping between these classes and available Named Entities (NE) is done. However, the question hierarchy is much more exhaustive than the named entities that our NE system is able to recognize.

2.2 Keywords Extraction

We call keywords all words or expressions extracted from a question, and that are likely to appear near the answer. Keywords can be words, named entities, or noun phrases; indeed they are actually a lemmatized form of each one.

Keywords extraction is done on French questions as, at this step, English questions were previously translated. Lemmas are obtained with the help of the TreeTagger. Keywords set is only composed of nouns, adjectives and verbs.

Noun phrases may help in the ranking of passages if they appear. For example, in question Q0178, the expression *"world champion"* is more significant than *"world"* and *"champion"* taken separately. Such expressions are extracted in French with the help of SxPipe deep parser described in [7].

Named Entities encountered in question are detected as described later in section 3.1.

3 Answering factoid questions

Answering factoid questions was done by using the QA System (QAS) we built for our participation to EQueR. This QAS mainly relies on density measures to select answers which are sought as Named Entities (NE) paired with expected answer types. Two new modules were developed for this years' participation: Lucene was integrated as the Document Retrieval (DR) search engine (for EQueR, like for TREC, Topdoc lists were available, and so, this DR step was facultative) and we also experimented with redundancy to re-rank candidate answers.

3.1 Named Entity Recognition

Named Entities detection is one of the key elements in our QAS. Each answer that will be provided must first be located (and bounded) as a semantic information nugget in the form of a Named Entity. Our named entities hierarchy was created and is a subset of the Sekine's taxonomy [8].

NE detections are done by using automata; most of them are implemented by using GATE platform, but also by direct mapping of many gazetteers gathered from the Web.

3.2 Document Retrieval

The Document Retrieval (DR) step aims at identifying documents that are likely to contain an answer to the question posed and, thus, restrict search space for next steps.

Indexation and retrieval were done by using the Lucene search engine with its default similarity scoring. Each document of the collection is considered as a whole document (without any pre-processing) for indexing purposes, and only their lemmas are indexed after a stop-listing based upon their TreeTagger's part-of-speech. Disjunctive queries are formulated with the question keywords. No query words relaxation was done for CLEF experiments. At retrieval time, only the (at most) first 30 documents (this is an empirically fixed limit) returned by Lucene were considered and passed to the Passage Retrieval component.

3.3 Passage Retrieval

Since our first participation in a QA exercise [2], our passage retrieval approach changed from a conventional cosine based similarity to a density measure. Our passage retrieval component considers a question as a set of several kinds of items: lemmas, Named Entity tags, and expected answer types.

First, a density score s is computed for each occurrence o_w of each item w in a given document d. This score measures how far are the items of the question from the other items of the document. This process focuses on areas where the items of the question are most frequent. It takes into account the number of different items $|w|$ in the question, the number of question items $|w, d|$ occurring in the document d and a distance $\mu(o_w)$ that represents the average number of items from o_w to the other items in d (in case of multiple occurrences of an item, only the nearest occurrence to o_w is considered).

Let $s(o_w, d)$ be the density score of o_w in document d:

$$s(o_w, d) = \frac{\log\left[\mu(o_w) + (|w| - |w, d|)\,.p\right]}{|w|}$$

where p is an empirically fixed penalty. The score of each sentence S is the maximum density score of the question items it contains:

$$s(S, d) = \max_{o_w \in S} s(o_w, d)$$

Passages are then composed of a maximum of three sentences: the previous sentence (if it exists), the sentence S, and the following one (if it exits). The first (and at most) 1000 best passages are then considered by the Answer Extraction component.

3.4 Answer Extraction

3.4.1 Answer extraction by using a density metric

To choose the best answer to provide to a question, another density score is calculated inside the previously selected passages for each answer candidate (a Named Entity) adequately paired with an expected answer type for current question. This density score (called *compacity*) is centred on each candidate and involved keywords extracted (let $QSet$ be this set) from the question at the question analysis step.

The assumption behind our *compacity* score is that the best candidate answer is closely surrounded by the important words of the question. Any word not seen in the question can disturb the relation between a candidate answer and its responsiveness to a question. Moreover, in QA, term frequencies are not as useful as for Document Retrieval: an answer word can appear only once, and it is not guaranteed that words of the question will be repeated in the passage, particularly in the sentence containing the answer. A score improvement can come from incorporating an inverse document frequency or linguistic features for non $QSet$ encountered words to further take into account any variation of closeness.

For each *answer candidate* AC_i, compacity score is computed as follow:

$$compacity(AC_i) = \frac{\sum_{y \in QSet} p_{y_n, AC_i}}{|QSet|}$$

with y_n is the nearest occurrence of the keyword y from AC_i and:

$$\begin{aligned} p_{y_n,AC_i} &= \frac{|W|}{2R+1} \\ R &= distance(y_n, AC_i) \\ W &= \{z | z \in QSet, \text{distance}\,(z, AC_i) \leq R\} \end{aligned}$$

For the two runs submitted this year, only one word-length keywords were considered for inclusion in $QSet$. Other experiments are planned which will use compound words, named entities and noun phrases.

All answers candidates are then ranked by a product between the passage density score containing it and their compacity score. Top N best scoring distinct answers are provided as final answer, N was equal to 1 for CLEF-2006 experiments.

3.4.2 Answer extraction by using knowledge databases

For some questions, answers are *quite* invariable over time. This is the case, for example, for questions asking about capital of a country, authors of book or famous past events. To answer these questions one may use (static) knowledge databases (KDB). We had built such KDB for our participation to TREC-11 [2], and translated them to French equivalent for our participation to EQueR (they were particularly tuned on CLEF-2004 questions). For CLEF-2006, we used the same unchanged KDB module.

This module provided answer patterns, which are used to assess the reliability of our Passage Retrieval and extraction based on *compacity* score.

The table 2 show the coverage of these databases for CLEF from 2004 to 2006 (*ne* stand for "not evaluated"), the number of passages coming from PR matching a KDB pattern, and number of right supported answers. EN-FR results were lower due to translation errors (9R or 7R).

runs	2004	2005	2006 FR-FR
Q covered (/200)	51	26	24
Passages matching pattern	48	*ne*	11
Right answers	38-40	*ne*	11

Table 2: Knowledge databases coverage

3.5 Re-ranking using a redundancy criterion

One of the drawbacks of the extraction of answers by using our *compacity* measure is that excessive closeness, from others interesting keywords, of a NE (of the adequate expected answer type) may lead to a wrong extraction, particularly if the passage contains many occurrences of the interesting keywords (due to high passage density score and high candidate compacity score). To smooth this flaw, we experimented with a redundancy criterion to re-rank the answer candidates found. But we did not notice significant improvements: number of wrong answers changed to right one was compensated by number of right answers changed to wrong (12R vs. 15W for FR-FR and 9R vs. 9W for EN-FR). After more analysis, it was due to a bug in the weighted-vote mechanism we used.

4 Answering definition questions

For our participation to EQueR [5], most of the definition questions were unanswered by the processing chain described in the previous sections. Indeed, this chain mainly rely on the detection of an entity paired with a question type, but, for definition questions, such pairing to an entity

is far more difficult as the expected answer can be anything qualifying the focus of the question (even if, when asking about a person definition, the answer sought is often its main occupation - which may constitute a common Named Entity - it can still be tricky to recognize all possible occupations or reasons why someone maybe "famous").

Also, one can notice that while for TREC-QA campaign all vital nuggets should be retrieved, for the EQueR and CLEF exercises, retrieving only - one - vital was sufficient.

So, the problem we wanted to address was to find one of the best definitions available inside the corpora. Therefore, we developed a simple approach based on appositive and redundancy to deal with these questions, and it was bundled in an independent component (as it was too different from our classical QA chain):

- The focus to define is extracted from the question and is used to filter and keep all sentences of the corpora containing it.
- Then, appositives nuggets such as *< focus , (comma) nugget , (comma) >* , *< nugget , (comma) focus , (comma) >* , or *< definite article ("le" or "la") nugget focus >* are sought.
- The nuggets are divided in two sets: the first, and preferred one, contains nuggets that can be mapped, by using their TreeTagger part-of-speech tags, on a minimal noun phrase structure while the second set contains all others (it can be seen as a kind of last chance definition set).
- Both set are ordered by theirs nuggets frequency in the corpora. But the "noun phrase set" is also ordered by taking into account its head noun frequency, the overlapping count of nouns inside it among the most frequent nouns appearing inside all the nuggets retrieved and its length. All these frequency measures are aimed to choose what is expected to be "the most common informative definition".
- Best nugget of the first set or of the second, if first is empty, or by default NIL is answered.

This component was also in charge of the acronyms/expanded acronyms definitions, as the same syntactic punctuation clues can be used for these questions such as the frequent *< acronym ((opening parenthesis) expanded acronym) (closing parenthesis) >* or *< expanded acronym ((opening parenthesis) acronym) (closing parenthesis) >* - and it performed very well as all answers of this year set were retrieved (even if, due to a bug in the re-matching process between answer and justification, the "TDRS"/Q0145 wasn't finally answered, while the correct answer was found).

5 Results

Two run for monolingual French (FR-FR) and two run for multilingual English to French (EN-FR) were submitted. As required, only one answer per question was provided. For steps such as DR, PR, or re-ranking, deeper analysis will be done when all the correct participants' answers will be available after the CLEF workshop. But, we evaluate one of the main bottlenecks in our pipelined QAS: missing pairing between adequate NE and expected answer types. So, for factual and temporally restricted FR-FR questions, QAS could not extract more than 131 answers.

On the 200 test questions, and for our best run (FR-FR1), 93 right answers (88 + 5 lists) were provided. Table 3 presents the ventilation of our two runs without re-ranking (our second runs, which were buggy; the + or - are due to answers that we think are misclassified).

Knowledge Databases: The knowledge databases best contribution was +11 right answer, however, without using KDB, 5 of these correct answers would have been also found by the "generic QAS". Final best KBD contribution is +6R.

IneXact answers: if we examine all the inexact answers from all our 4 runs, 8 of 20 where *Date* and 3 of 20 were *Number of people*. For *Date*, 6 of 8 where correct date answers but missed

runs	Right	Fact.	Def.	Temp.	R NIL	NIL answered	ineXact	Unsupported
FR-FR1	88	56 (+1)	32	0	2	30	7 (-1)	2
EN-FR1	67	40 (+1)	27	0	5	34	7 (-1)	2

Table 3: CLEF-QA 2006 results for our best FR-FR and EN-FR runs

the year. By the way, the year was not present inside the justification, but it was the directly preceding "19[0-9][0-9]" year for 3 and the year of the current document for the 3 others (with no other year date cited). So, simple rule of detection, with default value to the year of the document, for these questions could have helped to answer them. Lastly, one other *Date* question was judged as an inexact but is actually a wrong one: Australia will probably never be a European Union Member (Q0180). For *Number of people* questions, all 3 where missed due to bad detection of NE boundaries, but 2 because of the lack of approximations "prs d[e']/around".

Temporally restricted factoid questions: Were handled as simple factoid one, without any particular effort to justify date information (actually we were able to extract time constraints from question but not to build the checking module). From the results provided by CLEF staff none of them were correctly answered by our system.

Definition questions: Results for the definition module were quite good: of the 41 definitions that we manually identified (if the question "Décrire le World Trade Center./Describe WTC."/Q0187 can also be added to this set, none of our module could be able to handle it, and even for a human, the answer to provide is not clear), our best run (FR-FR) answered 32 (not NIL) Right and 1 NIL Right ("Linux"/Q0003, the word Linux never appear inside corpora).

Among the 5 incorrect answers provided: 2 of them were due to bugs inside the sentence and justification matching processing and should have been answered correctly; 1 expected a NIL answer ("T-shirt"/Q0144), and cannot be answered by this approach (which tends to always provide an answer unless the object to define did not appear inside corpora, it miss some confidence or validation measure to discard very improbable answer); and the last 2 needed more complex resolutions and were not located at all: answers were before or after a relative clause but also separated from the focus by using more than one punctuation (a dependency tree could have help to provide them).

For the 3 inexact definition answers provided: "Boris Becker"/Q0090 is hard to define without using any anaphora resolution, the string "Boris Becker" never co-occurs inside a sentence with a pattern "Tennis" ("joueur/player" or "man") or any "vainqueur/winner". The 2 others also need to extract a subpart of a relative clause, but the corresponding passages were located in the top five definitions selected.

Results for the EN-FR runs were lower (best one was 27 + 1 NIL/41) due to translation errors.

List questions: Our QAS wrongly recognized the 6 list questions[3] of this year test set as simple factoid question (and answered with only one answer). For our best FR-FR run, we provided one right answer for 4 questions, and for 3 of these, the justification contained all other needed right answers. For our best EN-FR run, 2 right answers were provided but only one of the justifications contained all the other correct instances.

6 Conclusion and future works

We have presented the Question Answering system used for our first participation to the QA@CLEF-2006 Evaluation. We participated in two tracks: monolingual French (FR-FR), and cross-lingual English to French (EN-FR). Results obtained were acceptable with an accuracy of 46% for FR-FR, and 35% for EN-FR.

However, from the quick analysis done in the previous sections, performance can also be improved at every step. For example, redundancy is still something we must evaluate. It could be a

[3](4 questions were misclassified by the CLEF staff and are not actual list questions: Q0133, Q0195, Q099, Q0200)

redundancy inside corpora, extracted answers and/or using the Web as a statistical judge. Indeed, we noticed that our broken criterion allowed us to answer some questions that were previously wrongly answered.

Inadequate Named Entities (NE) were also a bottleneck in the processing chain, and many of them should still be added (or improved). By the way, to compensate this known weakness, we tried to include in our system a simple reformulation module, notably for the "What/Which (something)" questions. It would have been in charge of verifying answers already extracted by the generic extraction system and, particularly, of completing it when NE was unknown. But our preliminary experiments showed us that such question rewriting was tricky. To our opinion, questions were (intentionally or not) formulated such that it is difficult by changing words order to match an answering sentence (synonyms might help here). Another problem is, when using the Web to match such reformulation, results obtained are noisy: we experiment it for presidents or events (Olympic Games) as CLEF corpora is dated from 1994 to 1995 while Web emphasizes on recent era.

With an accuracy of 76%, performance on definition questions was quite good even if it could still benefit from anaphora resolution. Moreover, we think that our actual methodology is language independent as it doesn't involve any language knowledge (other than detecting appositives mark up, a similar approach was used by [4] to answer spanish definition questions). We are interested in improving it in many ways: first, to locate more difficult definitions (as the one we missed); but also to choose better qualitative definitions (by combining better statistics), and finally to synthesize these qualitative definitions. For example, "Bill Clinton" was defined as an "'American president", "Democrat president", "democrat", "president", he can be best defined as an "American democrat president". "Airbus" was defined as "European consortium", or "Aeronautic consortium", so defining it as a "European Aeronautic consortium" could be even better.

Concerning date and temporally restricted questions, we could not achieve the development of a complete processing chain, so this is a future work (actually, by the time of the CLEF evaluation, we only have built a module to extract time constraints from the question). We also need to complete our post-processing to select or to synthesize complete date (with the difficulties one can expect for relative dates).

Lastly, our QAS still miss confidence scores after each pipelined components be able to do some constraint relaxations (for example on keywords), any loop back or simply to decide when a NIL answer must be provided.

References

[1] F. Béchet, A. Nasr, and F. Genet. Tagging unknown proper names using decision trees. In *Proceedings of the ACL 2000*, pages 77–84, Hong-Kong, China, 2000.

[2] P. Bellot, E. Crestan, M. El-Bèze, L. Gillard, and C. de Loupy. Coupling named entity recognition, vector-space model and knowledge bases for trec-11 question-answering track. In *Proceedings of The Eleventh Text REtrieval Conference (TREC 2002), NIST Special Publication 500-251*, 2003.

[3] B. Grau C. Ayache and A. Vilnat. Equer: the french evaluation campaign of questions answering systems. In *Proceedings of The Fifth International Conference on Language Resources and Evaluation (LREC 2006)*, Genoa, Italy, 2006.

[4] Daniel Ferrés, Samir Kanaan, Edgar González, Alicia Ageno, Horacio Rodríguez, and Jordi Turmo. The talp-qa system for spanish at clef-2005. In Carol Peters, editor, *Working Notes for the CLEF 2005 Workshop*, 2005.

[5] Laurent Gillard, Patrice Bellot, and Marc El-Bèze. Le lia à equer (campagne technolangue des systèmes questions-réponses). In *Actes de TALN-Recital 2005*, volume 2, pages 81–84, Dourdan, France, 2005.

[6] R. Kuhn and R. De Mori. The application of semantic classification trees to natural language understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(5):449–460, 1995.

[7] Benoît Sagot and Pierre Boullier. From raw corpus to word lattices: Robust pre-parsing processing with sxpipe. *Archives of Control Sciences*, 15(4):653–662, 2005.

[8] Kiyoshi Sudo Satoshi Sekine and Chikashi Nobata. Extended named entity hierarchy. In *Proceedings of The Third International Conference on Language Resources and Evaluation (LREC 2002)*, pages 1818–1824, Las Palmas, Canary Islands, 2002.

[9] H. Schmid. Probablistic part-of-speech tagging using decision trees. In *Proceedings of The First International Conference on New Methods in Natural Language Processing (NemLap-94)*, pages 44–49, Manchester, U.K., 1994.

[10] E.M. Voorhees and D. Harman. *TREC Experiment and Evaluation in Information Retrieval*, chapter 10, pages 233–257. MIT Press, 2005.