QA@L^{2}F@QA@CLEF

Ana Mendes, Luísa Coheur, Nuno J. Mamede
Luis Romão, João Loureiro, Ricardo Ribeiro, Fernando Batista, David Martins de Matos
L^{2}F - Spoken Language Laboratory, INESC-ID Lisboa
qa-clef@l2f.inesc-id.pt

Abstract

This paper introduces L2F's (INESC-ID) question/answering system and presents its results in the QA@CLEF07 evaluation task. QA@L^{2}F bases its performance on a high-quality deep linguistic analysis of the question, which is strongly based on named entity recognition. However, if a precise analysis is not possible or if no answer is found in previous processed data, the system is also capable of relaxing and tries to find an answer using a flexible pattern matching approach.

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; H.3.7 Digital Libraries; H.2.3 [**Database Management**]: Languages—*Query Languages*

General Terms

Measurement, Performance, Experimentation

Keywords

Question answering, named entity recognition

1 Introduction

At the end of 2006, we decided to build a Question-Answering (QA) system to be used at CLEF07, in the portuguese monolingual QA@CLEF task. Like many state of the art QA systems base their performance on Named Entity Recognition (NER) [10, 3] as well as on precise linguistic information [2, 4], in order to build QA@L^{2}F we profit from ongoing work on NER [5, 9] and from a (still in development) robust Natural Language Processing (NLP) chain, which are both used in corpus processing, database building and question interpretation.

Despite the fact that the system is based on high-quality deep linguistic analysis of both the question and the corpora (used to build the database), if a precise analysis is not possible, or if no answer is found in the database, the system relaxes and tries to find an answer in a more flexible way.

In the following we present QA@L^{2}F as well as the obtained results. The paper is organized as follows: section 2 describes the system general architecture, putting special emphasis on the NLP chain; section 3 shows how the knowledge database is built; section 4 presents the question interpretation and the answer extraction modules; section 5 presents and discusses the evaluation results; finally, section 6 concludes and points to future work.

2 QA@L^{2}F: overall picture

It was an option to invest in the system's architecture rather than in going deep in each one of the steps that constitute QA@L^{2}F.

In this section, the system's architecture is presented and special attention is given to its NLP processing chain.

2.1 General architecture

Figure 1 describes QA@L^{2}F general architecture.

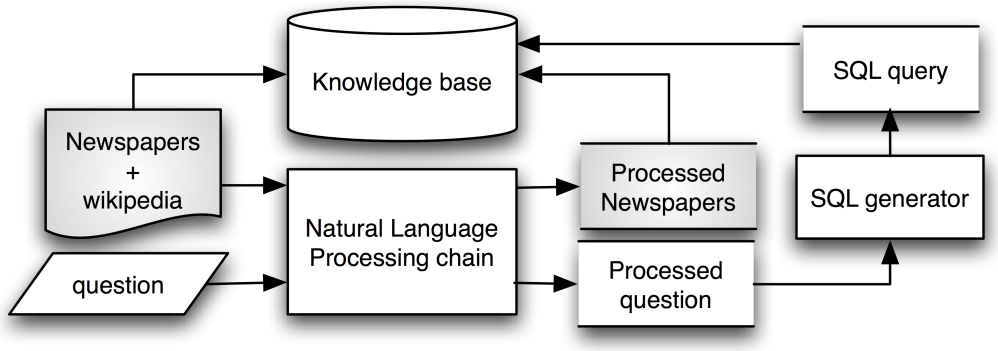

Figure 1: General Architecture.

It can be seen that at the heart of the system there is a NLP chain, that is used both to interpret the question and to pre-process information from different sources. In fact, since CLEF provides the sources where answers can be found (Publico 94, 95, Folha de São Paulo 94, 95 and Wikipedia), a database storing relevant information is built offline. Nevertheless, these sources can also go directly into the database, without any processing. This may happen for two reasons:

- we know that a certain document is a potential information source, although it was not previously processed;
- when the system will be open for the web, the set of documents where the answer might be will go directly into the database. After that, they will be processed on-line.

2.2 The Natural Language Processing Chain

The NLP chain (Figure 2) used by QA@L^{2}F both in corpus processing, database building and question interpretation, is built upon the following tools:

- Palavroso [6], responsible for the morphological analysis and Marv [8] for its desambiguation;
- Rudrico (an improved version of Pasmo [7]), applied twice, splits or concatenate tokens;
- XIP [1] returns the input organized in chunks, connected by dependency relations.

NER also uses this chain, tagging words or sequences of words as `PERSON`, `JOB`, `TIME`, `CULTURE`, among others.

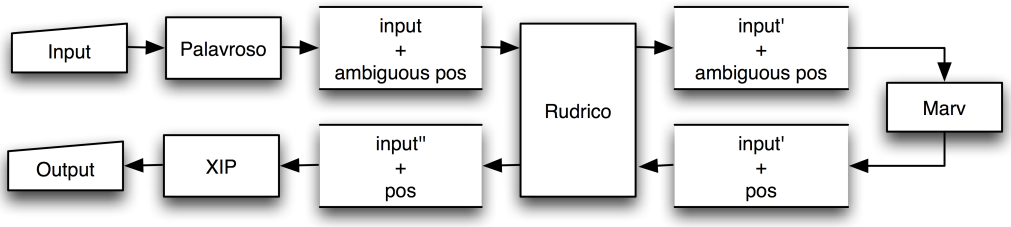

Figure 2: Natural Language Processing chain.

2.3 Example

Consider that the sentence "*O realizador Manuel de Oliveira estava no público.*" (*"The director Manuel de Oliveira was in the audience."*) was found during corpus analysis. If QA@L^{2}F is capable of classifying *Manuel de Oliveira* as the named entity `PERSON` and *realizador* as `JOB`, an entry is inserted in the database holding this information. That is, from that sentence, a relation between a person and a job is extracted. Many other relations are built in similar ways.

During question analysis, if the system considers that the question can be answered using a certain database relation, it will simply query the database. If it is asked *"Quem é Manuel de Oliveira?"* (*"Who is Manuel de Oliveira?"*), QA@L^{2}F will query the database for Manuel de Oliveira's job and the answer will be *realizador*.

If no relevant information is found in the database, QA@L^{2}F will gather every snippet in the database having the named entities present in the question. Then, it will perform a sequence of strategies in order to find a possible answer. Section 4.2 details this mechanism.

3 Building the Database

Looking again at Figure 1, it should be noticed that information sources (newspapers and Wikipedia) can go directly into the database. Two diferent databases were built in order to gather data from these two diferent information sources. Another one was created in order to hold the entire corpus, without any NLP. The first two databases will be described in the next subsections.

3.1 Newspapers Database

The QA@L^{2}F newspaper database stores all the information used by the system and can be divided into three distinct parts:

- corpus, containing raw text snippets;
- relation-concepts, containing relations between concepts, as well as a reference to the text snippet holding those relations (this information can be the answer to some specific question);
- named-entities, containing every named entity recognized by the NLP tools and the reference to the text snippets where they appear (this data can be used to locate text snippets where the answer can possibly be found).

The relation-concepts information is gathered using linguistic patterns for Portuguese. These linguistic patterns are divided into different categories, depending on the type of relation they aim to capture. The system has patterns for the categories shown in table 1. This table also presents examples of the type of questions for each category, as well as the text snippet where the relation-concept pair was found.

Category	Aimed Question	Corpora Snippet
PEOPLE	"Quem é Oscar Luigi Scalfaro?"	*O Presidente italiano, Oscar Luigi Scalfaro, iniciou ontem...*
	Who is Oscar Luigi Scalfaro?	The Italian President, Oscar Luigi Scalfaro, started yesterday...
LOCATION	"Onde se situa Times Square?"	*Este relógio da morte, instalado em Times Square (Nova Iorque)...*
	Where is Times Square?	This watch of death, located in Times Square (New York)...
CULTURE	"Quem realizou Land and Freedom?"	*...,≪Land and Freedom≫, de Ken Loach, evocação da Guerra Civil Espanhola.*
	Who directed Land and Freedom?	...,≪Land and Freedom≫, by Ken Loach, an evocation of the Spanish Civil War.
STUFF	"O que é a FIL?"	*A Feira Internacional de Lisboa (FIL) abre mais uma vez...*
	What is FIL?	Lisbon's International Fair (FIL) opened one more time...

Table 1: Examples of questions, categories and text snippets where the answers can be found.

CULTURE

id	culture	author	confidence	count
1	Land and Freedom	Ken Loach	99	4

Table 2: Entry representing the information *Ken Loach is the author of Land and Freedom.*

After finding the patterns, the corresponding information is stored in the database. For instance, the table which stores information about the category CULTURE will have an entry such as the one shown on table 2.

In the case presented in table 2, the analyser does not identify *Ken Loach* as being a `PERSON`, because it does not belong to the dictionary. Neverthless, due to the existence of an artwork's name (classified because of its position between guillemets), followed by a comma, the preposition "de" (*by*) and a proper name, the relation *Ken Loach is the author of Land and Freedom* [1] could be retrieved.

These relation-concepts tables have information concerning the confidence given to that relation. Even if we are not able to assure each concept identity, some patterns may give clues about a relation. Many examples can be given in order to ilustrate this feature, like the previous about Ken Loach and its artwork. Let's consider also the example shown in table 1, category LOCATION. If "Times Square" were not identified as a `LOCATION`, the analyzer could guess the relation LOCATION because of the existence of the first preposition "em" (*in*) and the existence of the well-identified `LOCATION` "Nova Iorque" between parenthesis [2]. These deduced relations have smaller confidence than well-identified relations.

3.2 Wikipedia Database

The WikiXML collection provided by the Information and Language Processing Systems group at the Informatics Institute, University of Amsterdam, was used, as well as its database structure [3].

[1] This relation is used in a broad sense. Ken Loach is the director of Lan and Freedom, and not its author, but our aim was just to capture the relation between the person and the artwork.

[2] It should be noticed that, for this example, we considered that Nova Iorque was included in the dictionary, but Times Square was not. This distinction makes one location be identified as `LOCATION`, and the other one not.

[3] `http://ilps.science.uva.nl/WikiXML/`

A new table containing only the XML article nodes from every Wikipedia page, with no linguistic processing, was also created. They aim to answer definition questions.

Consider, for instance, the question *"O que é Portugal?"* (*"What is Portugal?"*). In this case, having a table entry containing the information shown in table 3, the system can answer the question.

WIKIPEDIA PAGE

id	page_title	page_text
1480	Portugal	Portugal (de nome oficial República Portuguesa) fica situado no sudoeste da Europa, na zona Ocidental da Península Ibérica e é o país mais ocidental da Europa,...

Table 3: Entry representing information about Portugal.

Wikipedia's unique structure allows the retrieval of miscellaneous information (much easier than from newspapers). For instance, birth and death dates and country-related data, are easily extracted due to the quasi-standart format used in the text. However, for this experimente, we did not explore these possibilities. More processing can be done on this particular resource.

4 Question Interpretation and Answer Extraction

The question interpretation is a decisive step on a QA system. The question provides all the information used for the answer extraction (no other clues are given). This section presents QA@L^{2}F question interpretation and answer extraction modules. Some examples of QA@CLEF 2007 questions are also ilustrated and also the system's answering process.

4.1 Question Interpretation

Question interpretation comprehends all the steps responsible for the transformation of the question into a SQL query. As seen in Figure 1, the question is processed by the NLP chain described in section 2.2 and by a SQL generation. The NLP chain returns a parsed question with dependencies connecting the detected chunks; the SQL generation step comprises the stages shown in Figure 3.

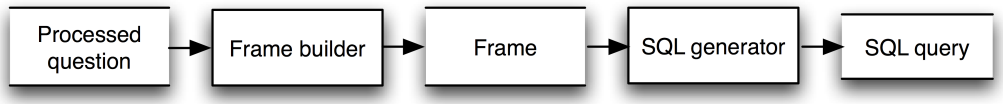

Figure 3: SQL generation.

The frame builder is responsible for identifying:

- the script to be called (this identification depends on the type of the question);
- the target entity (such as a person's name or a location);
- all the named entities identified in the question.

The SQL generation consists of a set of scripts that will map the information provided in the frames into a SQL query.

For instance, considering the question *"Quem é Boaventura Kloppenburg?"* (*"Who is Boaventura Kloppenburg?"*), after the NLP chain, besides the syntactic information, the dependency named TARGET_WHO_PEOPLE is identified:

```
<DEPENDENCY name="TARGET_WHO_PEOPLE">
<PARAMETER ind="0" num="11" word="Boaventura Kloppenburg"/>
</DEPENDENCY>
```

Then, the following frame is build:

```
SCRIPT script-who-people.pl
TARGET "Boaventura Kloppenburg"
ENTIDADES  "Boaventura Kloppenburg " PEOPLE
```

This frame is then mapped into the following MySQL query, that will *possibly* retrieve the question's answer:

```
select title,confidence,count from FACT_PEOPLE
where name="Boaventura Kloppenburg" GROUP BY confidence DESC, count DESC
```

The relation-concepts database is queried and every title (or profession) connected with Boaventura Kloppenburg is retrieved, in descendant order of confidence and quantity.

4.2 Answer Extraction

Depending on the type and information extracted from each question, the system follows diferent approaches in order to find the final answer.

If the submitted question belongs to the particular subset of those that can be answered directly using the relation-concepts database (questions like the ones presented on table 1), the system will just query that database. If the submitted question can not be answered directly using the relation-concepts database, the system will reduce the corpus to a few useful text snippets[4], using every information available from the question analysis. This step is fulfilled by merging the information both from the question and from the corpora. Thus, for the question *"Quem era rei de Portugal em 1860?"* (*"Who was the king of Portugal in 1860?"*), the question interpretation step identified "Portugal" as being a "`LOCATION`, 1860 as a `DATE` and "rei" (king) as a `TITLE`. Then, the system collected every snippet from the database having the named entity `LOCATION` Portugal, the named entity `DATE` 1860 and the named entity `TITLE` rei.

As mentioned previously, the system uses a mechanism to relax its constraints if the answer is not found. This strategy has three main reasons:

- the corpus was not entirely processed;
- the system can not determine exactly which information is important to be used later on;
- the important, although not detected, information in the corpus is not stored on any strutured database.

The system adopts the following strategies in order to return the final answer.

Linguistic Pattern Matching This method uses linguistic patterns to extract the possible answers to a specific set of questions, like those presented on table 1.

The QA@L^{2}F system used the pattern matching approach to answer questions like *"O que é a TVI?"* (*"What is TVI?"*). The answer was found in the following text snippet:

```
(...) Numa das já habituais leituras do PÚBLICO, fui surpreendido por um ligeiro
comentário feito por Maria Augusta Gonçalves. Em tom de desabafo, a dita senhora,
falando sobre a Televisão Independente (TVI) de inspiração cristã, a certa altura
solta o seguinte: [Também, as novelas hispanas, com sua dobragem em brasileiro...]
```

The relation *TVI is Televisão Independente* was caught and inserted in the relation-concepts table (category STUFF) prior to the question submission. When answering the question, the system returns the missing concept from that table.

[4]Useful text snippets are those in which the answer can be found.

Linguistic Reordering This method is used mainly for answering definition questions, like *"Quem foi Pirro?"* (*"Who was Pirro?"*) and *"O que é a Igreja Maronita?"* (*"What is the Maronite Church?"*), or list questions, like *"Diga uma escritora sarda."* (*"Mention a sardinian writer."*).

The system uses the Wikipedia in order to answer these questions. Firstly, the question analysis step recovers the question main concept ("Pirro","Igreja Marronita" and "escritora sarda", from the above examples). Then, it will perform a search over the extracted articles for linguistic patterns which can contain the answer. For definition questions, patterns like *main concept* followed by the inflected verb "to be" (e.g. Pirro foi... or Maronite Chuch é...); on the contrary, for list questions, those patterns are like the inflected "to be" followed by the *main concept.* (e.g. ...é uma escritora sarda).

This strategy is also used on those questions for which the system could not find an answer using the linguistic patterns matching. Consider, for instance, the question *"Quem foi Ésquilo?"* (*"Who was Aeschylus?"*). This one belongs to the category PEOPLE, and could have been answered using just the relation-concepts tables. However, the relation between Ésquilo and his title was not captured using linguistic patterns. Thus, the system searched on Wikipedia for the page which title is Ésquilo and returned only the information concerning his definition: a tragic greek poet.

Nevertheless, this approach was not 100% successful. The concept M31 was not found on the database and the question *"O que é M31?"* (*"What is M31?"*) was not answered correctly.

Named Entities Recognition This method uses the information available on the question recovered during the question interpretation stage (both named entities and auxiliar words), to query the named entities database. A set of text snippets, containing that information, is then retrieved.

For instance, the question analysis returns the following information for the question *"Quem sucedeu a Augusto?"* (*"Who came after Augustus?"*):

```
TARGET EMPTY
ENTIDADES  "Augusto " PEOPLE
AUXILIARES "sucedeu" ACTION "a Augusto"
```

Having this information, the system will look on the database for snippets containing the named entity PEOPLE "Augusto" and the words "sucedeu" and "a Augusto" (for these last two, and because they are not classified as named entities, the system perfoms a full-text query against the text snippets). It will then return the most frequent named entity PEOPLE or named entity PROPER NAME on those snippets. In this case, the final answer was wrong, but, in fact, the snippet supporting the returned answer had both words "Augusto " and "sucedeu".

Brute-Force plus NLP If none of the previously described strategies return the question's answer, the system will use its last chance to be sucessful in its task: it performs a full-text query against the raw text snippets database, returning the top ten best qualified snippets. Those ten snippets go through the NLP chain and the most frequent concept matching the wanted answer type is returned.

This strategy was used for the question *"Quem é Boaventura Kloppenburg?"* (*"Who is Boaventura Kloppenburg?"*). The system's answering chain did not retrieve an answer for this question using any of the above strategies. Thus, it performed a full-text query against the corpus database using "Boaventura Kloppenburg" as key. Even though the answer was incomplete, the system's blind approach returned a partial correct answer and turned out to be a good strategy.

4.3 Choosing the answer

The system uses two main approaches in order to retrieve the final answer, depending on the strategy followed during the answer extraction step. If the choosen strategy is either the linguistic patterns matching or the linguistic reordering, the system simply returns the first answer found. On the other hand, if the choosen strategy is either the named-entity recognition or the brute-force

plus NLP, the answer extraction step depends on the question target type. Having in mind that we are dealing with large corpora (564MB of newspaper text, both in European Portuguese and Brazilian Portuguese, as well as the Wikipedia pages found in the version of November, 2006), the system assumes that the correct answer is repeated on more than one text snippet. With this assumption, QA@L^{2}F searches and returns as the question final answer the most frequent named entity that matches the question's target type.

5 Evaluation

The QA@L^{2}F system was evaluated at QA@CLEF 2007. In this section the system's results are presented. Special emphasis is given to ineXact[5] and Unsupported[6] answers.

5.1 Results

Figure 4 presents the results obtained at CLEF.

Right	Wrong	ineXact	Unsupported	Total	Accuracy (%)
28	166	4	2	200	28/200 = 14%

Figure 4: QA@L^{2}F results

It should be noticed that the answer-string "NIL" was returned 152 times (being correct 11 times). It should also be noticed that, since we did not handle anaphora and ellipsis, only 150 were actually addressed.

5.2 ineXact answers

The ineXact answers were all incomplete answers. And if in the question *"Quem é George Vassiliou?"* (*"Who is George Vassiliou?"*) it is obvious that the answer *"presidente de Chipre"* is incomplete, as he was *"presidente de Chipre entre 88 e 93"*, it is not so obvious what should be the right answer to *"Quem foi Henrik Ibsen?"* (*"Who was Henrik Ibsen?"*). Considering the paragraph:

```
Meditava eu, um tanto melancólico, acerca dos dilemas, constitucionais ou não,
em que nos consumimos, e procurava para exprimi-los uma imagem viva, quando de
repente me veio à mente a cebola de Ibsen. Estou falando no Ibsen original,
norueguês, Henrik Ibsen, dramaturgo que escreveu  Peer Gynt.
```

If *"dramaturgo"* is incomplete, is *"dramaturgo norueguês"* enough? Or the right answer should be *"dramaturgo norueguês que escreveu Peer Gynt"*? It is difficult to decide.

5.3 Unsupported

One of the answers classified as unsupported was due to the fact that we did not understand that in a list-type question the page where the answer was, was not enough, but the fragment where the answer was found should also be provided. Thus, we answered the question *"Diga uma escritora sarda."* (*"Mention a sardinian writer."*) with the following:

[5]ineXact answers: the answer-string contains a correct answer and the provided text-snippets support it, but the answer-string is incomplete/truncated or is longer than the minimun amount of information required.

[6]Unsupported answer: the answer-string contains a correct answer but the provided text-snippets do not support it, or the snippets do not originate from the provided document.

```
<a score="0.0" run_id="INES072PTPT" group_id="1848" q_id="0064">
    <answer>Grazia Deledda</answer>
    <docid>Grazia_Deledda</docid>
    <support>
      <s_id>Grazia_Deledda</s_id>
      <s_string/>
    </support>
    <judgment>U</judgment>
</a>
```

Which was considered unsupported, although Grazia Deledda is a *"escritora sarda"* (*"sardanian writer"*).

6 Conclusions and future work

QA@L^{2}F represents the absolutely first steps of our Clef participation. Besides the usual difficulties in building such a system, it was particulary demanding to compete in a year where anaphoric and eliptic questions were introduced, since our system still does not handle these aspects.

The system relies on robust NLP tools, which perform a deep linguistic analysis, both on the question and on the corpus.

Our goal on developing this system was to experient different techniques to answer questions. Thus, the system's architecture was our main focus: it uses several strategies in order to answer a given question and relies on a mechanism to relax its constraints if the answer is not found.

Many improvements are yet to be done. We would like to explore in more detail the patterns matching strategy, trying to find more patterns for Portuguese, and we are also aiming to introduce syntactical clues in order to help us finding some answers.

In conclusion, we consider that QA@L^{2}F had good results in this Clef evaluation and our main goal is making it better. Next year we'll see.

References

[1] Salah Aït-Mokhtar, Jean-Pierre Chanod, and Claude Roux. A multi-input dependency parser. In *Proceedings of the Seventh IWPT (International Workshop on Parsing Technologies)*, Beijing, China, October 2001.

[2] Adán Cassan, Helena Figueira, André Martins, Afonso Mendes, Pedro Mendes, Cléudia Pinto, and Daniel Vidal. Priberam's question answering system in a cross-language environment. *Working Notes for the CLEF 2006 Workshop*, 2006.

[3] Luís Costa. Esfinge - a modular question answering system for portuguese. *Working Notes for the CLEF 2006 Workshop*, 2006.

[4] Dominique Laurent, Patrick Séguéla, and Sophie Nègre. Cross Lingual Question Answer using QRISTAL for CLEF 2006. *Working Notes for the CLEF 2006 Workshop*, 2006.

[5] João Loureiro. NER - Reconhecimento de Pessoas, Organizações e Tempo. Master's thesis, Instituto Superior Técnico, Universidade Técnica de Lisboa, Portugal, 2007. work in progress.

[6] José Carlos Medeiros. Análise morfológica e correcção ortográfica do português. Master's thesis, Instituto Superior Técnico, Universidade Técnica de Lisboa, Portugal, 1995.

[7] Joana Paulo Pardal and Nuno J. Mamede. Terms Spotting with Linguistics and Statistics, November 2004.

[8] Ricardo Ribeiro, Nuno J. Mamede, and Isabel Trancoso. Using Morphossyntactic Information in TTS Systems: comparing strategies for European Portuguese. In *Computational Processing of the Portuguese Language: 6th International Workshop, PROPOR 2003, Faro, Portugal, June 26-27, 2003. Proceedings*, volume 2721 of *Lecture Notes in Computer Science*. Springer, 2003.

[9] Luis Romão. NER - Reconhecimento de Locais e Eventos. Master's thesis, Instituto Superior Técnico, Universidade Técnica de Lisboa, Portugal, 2007. work in progress.

[10] Luís Sarmento. Hunting answers with RAPOSA (FOX). *Working Notes for the CLEF 2006 Workshop*, 2006.