Reconstructing DIOGENE: ITC-irst at TREC 2006

Matteo Negri, Milen Kouylekov, Bernardo Magnini, Bonaventura Coppola
ITC-irst, Centro per la Ricerca Scientifica e Tecnologica
{negri,kouylekov,magnini,coppolab}@itc.it

Abstract

Our participation in the TREC 2006 QA task is the first step on the way of developing a new and improved DIOGENE system. The leading principles of this re-engineering activity are: *i*) to create a modular architecture, based on a pipeline of modules which share common I/O formats, open to the insertion/substitution of new components; *ii*) to allow for the capability of configuring the settings of the different modules with external configuration files; *iii*) to provide the capability of performing fine-grained evaluation cycles over the individual processing modules which compose a QA system. Another long-term objective of our work on QA, is to make the core components of the system freely available to the QA community for research purposes. This paper overviews the work done up to date to achieve these objectives, focusing on the description of a prototype module designed to handle the anaphoric questions often contained into TREC QA series. Preliminar evaluation results of the new module are presented, together with those achieved by DIOGENE at TREC 2006.

1 Introduction

This year the ITC-irst activity on Question Answering (QA) has been concentrated on system re-engineering. Our objective is to transform our DIOGENE QA system into a more modular architecture whose basic components are completely separated, easily testable with rapid turnaround of controlled experiments, and easily replaceable with new implementations. In this direction, our long-term ambition is to make all these components (or even the entire system) freely available in the future to the QA community for research purposes.

Up to date we are halfway in our roadmap, and unfortunately the results of our participation in TREC 2006 reflect this situation. While most of the original system's components are now separated from each other, the work done to improve their performance has not led to concrete results yet. Moreover, as it emerged from the analysis achieved by our TREC 2006 submitted run, a number of bugs still affect the overall system's behavior.

Besides the effort in the modularity direction, the main directions of improvement pursued this year concern all the basic component of the architecture of DIOGENE.

As for **Question Processing**, the first version of a module to handle follow-up questions has been implemented. Such module, on which this paper is focused, is in charge of tracking the correct referent of anaphoric questions contained into a TREC series.

As for **Document Retrieval**, this year's novelty is the substitution of the MG search engine [10] with Lucene [3], an open-source, cross-platform search engine, which allows for customizable document ranking schemes. The many advantages of Lucene now do enable us to implement more refined strategies for indexing, ranking, and querying the target document collection.

As for **Answer Extraction**, we improved the candidate answers selection process by developing a more effective way to handle "*generic*" factoid questions (*i.e.* questions whose expected answer type does not belong to the six broad named entity categories currently handled by DIOGENE). Generic questions are now handled extracting, as possible answer candidates, all the hyponyms of the question focus that are found in the text passages returned by the search engine. For instance, given the question Q: *"What instrument did Jimi Hendrix play?"*, the answer extraction module will select *"guitar"* as a possible answer candidate as it is an hyponym of *"instrument"*, which is the focus of the question.

As for **Answer Validation** (AV), new strategies to select the best answer among a set of possible candidates have been implemented. Besides our well tested statistical approach to the problem, described in [6], we plan to include in our AV package a new method based on textual entailment recognition, allowing the user for experimentations with different AV settings (see [4], [5], and [7] for further details) .

Since the improvements on document retrieval, answer extraction, and answer validation have already been documented in previous works, the rest of this paper concentrates on the newest extension of the system (*i.e.* the improved question processing component), specifically designed for this year's participation in TREC. A preliminary solution to the problem of tracking the questions' referent in a QA series is reported, together with the evaluation results achieved both by the implemented module (over the TREC 2005 test set), and the overall DIOGENE system (in the TREC 2006 Main QA task).

2 Dealing with QA series

In the framework of the main evaluation campaigns (*i.e.* TREC[1], CLEF[2], and NTCIR[3]), the complexity of the input questions is constantly increasing. Adhering to the roadmap for QA research [1], for instance, the test set of the TREC QA task moved from a list of simple isolated *factoid* questions (TREC 2000 and 2001), to the heterogeneous question typologies used in the last editions. Since 2004 [8], systems participating in TREC are presented with "factoid", "list", and "other" questions grouped into series. Each series has the target of a definition associated with it (including people, organizations, events, and other entities), which provides the initial context of the session. Each question in a series asks for additional information either about the target, or about the context originated by answers to the preceding questions in the session. The main objective of this evaluation setting is to provide an abstraction of an information dialog in which the user is trying to define the target. In a real-world scenario, in fact, questions are not asked in isolation, but rather in a cohesive manner that often involves a sequence of related questions to meet the users information needs.

One of the complexity factors in the proposed evaluation scenario is related to the presence of anaphoric expressions within the input questions. These may refer either to the target of the series, or to the answer given to a previous question. The example reported in Figure 1 provides a clear picture of the situation. While some questions in the series (*i.e.* questions 136.1, 136.4, 136.6, and 136.7) are non anaphoric and relatively easy to handle, others (*i.e.* questions 136.2, 136.3, 136.5) show a shift of the referent, realized by means of anaphoric expressions (*i.e.* the pronouns "*his*" and "*he*", and the nominal construct "*this person*"), which makes them more difficult to manage. In this framework, where question interpretation has to be situated in a particular context as the QA session proceeds, anaphora resolution and referent tracking capabilities emerge as key features of a QA system.

2.1 Our approach

In the past TREC editions the typical approach to anaphoric questions was a simple rewriting procedure, based on the substitution of question's pronouns with the target of the series [8]. Following this approach, once the input has been transformed into a question whose interpretation

[1] http://trec.nist.gov
[2] http://www.clef-campaign.org/
[3] http://research.nii.ac.jp/ntcir/

target id="136" text="Shiite"

- q id="136.1" *Who was the first Imam of the Shiite sect of Islam?*
- q id="136.2" *Where is his tomb?*
- q id="136.3" *What was this person's relationship to the Prophet Mohammad?*
- q id="136.4" *Who was the third Imam of Shiite Muslims?*
- q id="136.5" *When did he die?*
- q id="136.6" *What portion of Muslims are Shiite?*
- q id="136.7" *What Shiite leaders were killed in Pakistan?*
- q id="136.8" *Other*

Figure 1: A series from the TREC-2005 main QA task.

does not depend on the surrounding context, the analysis is carried out in the same way as it is done with non-anaphoric questions. Unfortunately, this solution is no longer feasible under the conditions of a "natural" dialogue-like series. In this situation (as it is reflected by the TREC 2005 and 2006 test sets), we must take into account that:

1. The target may be a complex or vague concept, which is not easy to describe with a single noun or named entity (*e.g. "Russian submarine Kursk sinks"*, *”France wins World Cup in soccer"*, *"first 2000 Bush-Gore presidential debate"*);

2. Follow-up questions do not necessarily refer to the series target, but also to answers given to previous questions.

On the one side, the complexity of the target makes the question reformulation procedure a hard and error-prone task (consider, for instance, the target/question pair: T=*"Plane clips cable wires in Italian resort"*, Q= *"When did the accident occur?"*). On the other side, as the series proceeds, a mechanism to track the actual referent of each question becomes necessary.

In light of these considerations, our solution adopts a bag-of-words approach which gets round the question reformulation problem, still remaining capable to track the referent as a series proceeds. The underlying assumption is that combining the right keywords to search the document collection is safer that trying to rewrite the input question in a "complete" form. In practice, once the referent of an input question has been chosen (between the target and the answer to an earlier question), its terms are combined with the question keywords into a query to the document collection. In case of reference shifts, a new referent remains valid until a non-anaphoric question is found. Following this idea, our question processing component carries out the analysis of each question (*i.e.* answer type recognition, keywords extraction, keywords expansion, query composition) as it is (*i.e.* without resolving anaphora), leaving to the document retrieval component the task of finding the best matching text portions.

Adopting this approach the problem of dealing with anaphoric questions in a TREC-like QA series is "reduced" to the problem of finding the right referent of each question. The next section describes how this task is accomplished by means of a classifier opportunely trained.

2.2 Question referent selection

Training corpus. To develop our referent tracking module, a training corpus has been manually annotated starting from the test set of the TREC 2005 main QA task (75 series, for a total of 530 questions).

Each question q_i, in a series S having T as a target, has been classified with respect to the three typologies we want to deal with, namely:

Type0: non anaphoric questions;
Type1: anaphoric questions referring to T;
Type2: anaphoric questions referring to the answer given to the preceding question in S (referred to as q_{i-1}).

Unfortunately, the distribution over the three typologies in the training corpus is rather unbalanced, with 129 questions assigned to *Type0*, 385 assigned to *Type1*, and only 16 assigned to *Type2*. Even though this unbalanced distribution is likely to reflect a real-world situation, the possible impact on the learning process should not be neglected. In fact, unless very discriminative features will be selected, the large amount of *Type1* questions will probably influence the classifier's performance, determining a bias towards this class.

Learning features. A classifier has been trained considering the following features of the target T, the current question q_i, and the previous question q_{i-1} in a series S:

- T features:
  T-has-verbs (0, 1): set to 1 if T contains at least one verb, 0 otherwise.
  T-capitalized (0, 1): set to 1 if all the words in T are capitalized.
  T-contains-LifeForms (0, 1): set to 1 if T contains at least one hyponym of "life_form" in WordNet [2].
  T-first-LifeForm-position (integer): position of the first hyponym of "life_form" in T.

- q_i Features:
  q_i**-number-in-Session** (integer): q_i position in S.
  q_i**-Type** (1,..., 3): factoid=1, list=2, other=3.
  q_i**-StartsWith** (0,..., 7): "who"=0, "where"=1, "when"=2, "which"=3, "what"=4 "how+adj"=5, "how+verb"=6 other=7.
  q_i**-contains-Target-tokens** (0, 1): set to 1 if q_i contains at least one token present in T.
  q_i**-contains-Target-lemmas** (0, 1): set to 1 if q_i contains at least one lemma of a term in T.
  q_i**-contains-prons** (0, 1): set to 1 if q_i contains at least one pronoun.
  q_i**-contains-pers-prons** (0, 1): set to 1 if q_i contains at least one personal pronoun.
  q_i**-contains-ThisPlusTargetWord** (0, 1): set to 1 if q_i contains the word "this" followed by a term present in T.
  q_i**-contains-ThisPlusWord** (0, 1): set to 1 if q_i contains the word "this" followed by any term non present in T.
  q_i**-contains-ThisPlusQ1Noun** (0, 1): set to 1 if q_i contains at least one noun present in q_{i-1}.
  q_i**-contains-LifeForms** (0, 1): set to 1 if q_i contains at least one hyponym of "life_form" in WordNet.
  q_i**-first-LifeForm-position** (integer): position of the first hyponym of "life_form" in q_i.

- q_{i-1} Features:
  q_{i-1}**-Type** (1,..., 3): factoid=1, list=2, other=3.
  q_{i-1}**-StartsWith** (0,..., 7): "who"=0, "where"=1, "when"=2, "which"=3, "what"=4 "how+adj"=5, "how+verb"=6 other=7.
  q_{i-1}**-contains-pers-pron** (0, 1): set to 1 if q_{i-1} contains at least one personal pronoun.

q_{i-1}**-contains-LifeForms** (0, 1): set to 1 if q_{i-1} contains at least one WordNet hyponym of "life_form".
q_{i-1}**-first-LifeForm-position** (integer): position of the first hyponym of "life_form" in q_{i-1}.

2.3 Evaluation

A preliminary experiment has been carried out to evaluate the viability of the proposed solution. For this purpose, our classifier has been trained using a decision tree algorithm (namely the Weka J48 algorithm [9]). At first glance, the results of this evaluation are encouraging. The 10-fold cross-validation over the training set resulted in fact in 505 out of 530 instances correctly classified (corresponding to an accuracy of 95.28%). However, as can be observed from the confusion matrix reported in Table 1, the performance over the three classes is not uniform, reflecting the unbalanced distribution of the examples in the training. In particular, unsurprisingly, only 4 instances out of 16 (25%) of the class less represented in the training set (*Type2*) are correctly recognized.

	Type0	Type1	Type2
Type0	82	12	0
Type1	0	344	1
Type2	0	12	4

Table 1: Confusion matrix

Apart from these very preliminary results, further conclusions on the suitability of our approach to TREC-like QA series will be drawn after a thorough analysis of the results of our run submitted to TREC 2006. Even though, as foreseen, the overall performance of the system (see Table 2) is rather poor, the possibility of separately evaluating the behavior of each single module will provide us with additional evidence about the effectivenes of our approach.

Statistics over 403 questions	
Number wrong (W)	325
Number unsupported (U)	10
Number inexact (X)	12
Number locally correct (L)	2
umber globally right (R)	54
Accuracy	0.134
Precision of recognizing no answer	7/114 = 0.061
Recall of recognizing no answer	7/17 = 0.412

Table 2: DIOGENE at TREC 2006

3 Conclusions

DIOGENE, the QA system developed at ITC-irst, is currently subject to an heavy re-implementation process. This activity aims at creating a more modular system, easier to modify and evaluate under different parameters settings, and open for a future distribution for research purposes.

In this paper we reported the main improvement directions we are pursuing, and described one of the newest modules specifically designed to participate in TREC 2006. Such module is in charge of determining the correct referent of anaphoric questions in a TREC-like QA series. One of the specific issues raised by follow-up questions, in fact, is to track the target shifts that often occur as the series proceeds. Focusing on this problem, we presented a very preliminary experiment with a classifier trained over the TREC 2005 question set.

We also reported the results achieved at TREC 2006. As foreseen, the evaluation reflects the work-in-progress situation and, unfortunately, the impact of a number of bugs which was not worth describing here.

In the next future, after a thorough error analysis, the re-implementation process will continue following the criteria mentioned above. As for the question analysis component, since there is still room for improvement, future developments will follow two complementary directions. More precisely: *i)* a more accurate feature selection, trying to find a set of more discriminative features; and *ii)* learning the model from a larger set of training examples (*e.g.* including the TREC 2006 series), trying to balance the number of examples of the question typologies we want to deal with.

References

[1] John Burger, Claire Cardie, Vinay Chaudhri, Robert Gaizauskas, Sanda Harabagiu, David Israel, Christian Jacquemin, Chin-Yew Lin, Steve Maiorano, George Miller, Dan Moldovan, Bill Ogden, John Prager, Ellen Riloff, Amit Singhal, Rohini Shrihari, Tomek Strzalkowski, Ellen Voorhees, and Ralph Weishedel. Issues, Tasks and Program Structures to Roadmap Research in Question & Answering (Q&A). In *Document Understanding Conferences Roadmapping Documents*, 2001.

[2] Christiane Fellbaum. *WordNet: An Electronic Lexical Database.* MIT Press, 1998.

[3] Erik Hatcher and Otis Gospodnetic. *Lucene in Action (In Action series).* Manning Publications, December 2004.

[4] Milen Kouylekov and Bernardo Magnini. In *Proceedings of PASCAL Workshop on Recognizing Textual Entailment*, Southampton, UK, 2005.

[5] Milen Kouylekov and Bernardo Magnini. In *Proceedings of PASCAL Workshop on Recognizing Textual Entailment*, Venezia, Italy, 2006.

[6] Bernardo Magnini, Matteo Negri, Roberto Prevete, and Hristo Tanev. Is it the right answer? exploiting web redundancy for answer validation. In *In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, ACL-2002*, pages 1495–1500, Philadelphia (PA), 7-12 July 2002.

[7] Matteo Negri, Milen Kouylekov, Bernardo Magnini, and Bonaventura Coppola. Towards Entailment-based Question Answering: ITC-irst at CLEF 2006. In *CLEF-2006 Working Notes*, Alicante, Spain, 2006.

[8] Ellen Voorhees. Overview of the TREC 2004 Question Answering Track. In *TREC 2004 Proceedings*, 2005.

[9] Ian H. Witten and Eibe Frank. *Data Mining: Practical machine learning tools and techniques.* Morgan Kaufmann, 2005.

[10] Ian H. Witten, Alistair Moffat, and Timothy C. Bell. *The second edition of Managing Gigabytes: Compressing and Indexing Documents and Images.* Morgan Kaufmann Publishing.