

Question Answering Based Annotation for a Corpus of Spoken Requests*

Elena Cabrio^{1,2}, Bonaventura Coppola^{1,2}, Roberto Gretter¹,
Milen Kouylekov¹, Bernardo Magnini¹, and Matteo Negri¹

¹ FBK-irst, ² University of Trento - 38050 Povo (Trento), Italy
{cabrio, coppolab, gretter, kouylekov, magnini, negri}@itc.it

Abstract. This paper introduces a semantic annotation scheme for spoken information access requests, specifically derived from Question Answering (QA) research. We argue that spoken requests annotations can be effectively enriched with useful information (e.g. the *Expected Answer Type* and the *Topic* of a request), which is broadly used in the QA framework to fully capture the content of a request and extract the sought-after information. The proposed annotation scheme has been adopted in the creation of the QALL-ME benchmark, a corpus of spoken requests related to cultural events of a town.

1 Introduction

Semantic annotations (e.g. named entities, co-reference relations, speech acts) of information access dialogues and requests allow automatic systems to understand the content of a dialogue, and detect relevant information for retrieving appropriate answers to users' questions. Although research and applications in Information Access dialogues/requests have been mostly driven by the speech community, in the last years, particularly fostered by evaluation campaigns organised by NIST (<http://trec.nist.gov>) and CLEF (<http://clef-qa.itc.it>), considerable progress has been achieved in open domain Question Answering (QA), where systems are asked to return short answers to natural language questions.

The main motivation of this work is that we think that significant improvements can be achieved through the integration of QA techniques and spoken dialog management approaches. In particular, techniques used for question processing may bring to domain and task independent semantic representations of the content of a request, thus filling a gap between the spoken dialogue community and the question answering community. As a first step in this direction, we address the creation of a reference benchmark of annotated spoken requests. In this corpus we introduce, in addition to already exploited annotation levels, new layers representing semantic information about a request based on QA research. More concretely, we introduce three QA-based annotation levels: the Question Topical Target (QTT), the Expected Answer Type (EAT), and the Expected

* The present work is supported by the QALL-ME EU Project - FP6 IST-033860

Answer Quantifier (EAQ). The awaited impact is the development of new, QA-based techniques for the semantic interpretation of information access dialogues. In this direction, the benchmark is intended to provide both training and test data for developing machine learning systems.

The work presented in this paper is being developed under QALL-ME, an EU-funded project aiming at the realization of a shared and distributed infrastructure for QA systems on mobile devices (e.g. mobile phones). The benchmark includes around 4,000 questions in four different languages (Italian, Spanish, English, and German), related to the domain of the cultural events in a town (e.g. cinema, theatre, exhibitions, etc.). The QALL-ME benchmark is being made incrementally available at the project web site (<http://qallme.itc.it>).

The paper is organised as follows. Section 2 shortly reviews relevant work on the semantic annotation of information access dialogues. Section 3 introduces semantic annotations derived from the QA framework. Section 4 provides details on the ongoing work on the QALL-ME benchmark. Section 5 concludes the paper, and provides indications about the forthcoming annotation steps.

2 Spoken Dialogue Annotation

In recent years, a number of research projects supported spoken dialogue annotation at different levels, with the purpose of creating language, domain, or task-specific benchmarks. Depending on the specific developers' purposes, the proposed annotation schemes cover a broad variety of information, ranging from the syntactic to the semantic and pragmatic level.

Released in the nineties, the ATIS and TRAINS corpora¹ are collections of task-oriented dialogues in relatively simple domains. The former contains speech data related to air travel information, and is partially annotated (2,900 out of a total of 7,300 utterances) with reference answers, and a classification of sentences into *i*) those dependent on context for interpretation, *ii*) those whose interpretation does not depend on context, and *iii*) the not evaluable ones. The latter includes 98 dialogs (6,5 hours of speech, 55,000 transcribed words), dealing with routing and scheduling of freight trains. Utterances are annotated with dialogue acts (or "Communicative Functions") including, among others, the types INFO-REQUEST, EXCLAMATION, EXPLICIT-PERFORMATIVE, and ANSWER.

More recently, the VERBMOBIL project (<http://verbmobil.dfki.de>) on speech-to-speech translation released large corpora (3,200 dialogs, 181 hours of speech, 1,520,000 running words) for German, English, and Japanese. Part of such material (around 1,500 dialogs) is annotated with different levels of information including: orthography, segmentation, prosody, morphology and POS tagging, semantic and dialog acts annotation. The latter annotation level has been carried out considering a hierarchy of 32 dialog acts such as GREET, THANK, POLITENESS_FORMULA, and REQUEST.

¹ <http://www ldc.upenn.edu/Catalog>

Spoken dialogue material collected within the MATE project² refers to any collection of spoken dialogue data (human-human, human-machine), including not only speech files, but also log-files or scenarios related to spoken dialogue situations. The annotation levels include prosody, morpho-syntax, co-reference, communication problems, and dialogue acts (e.g. OPENING, ASSERT, INFO_REQUEST, ANSWER).

Finally, the ongoing project LUNA (<http://www.ist-luna.eu>) is developing a multilingual and multidomain spoken language corpus, with the transcription and the semantic annotation of human-human and human-machine spoken dialogs collected for different application domains (call routing, travel information) and languages (French, Italian and Polish). At present, the completed annotation layers concern the argument structure, co-reference/anaphoric relations, and dialog acts.

Even though the proposed annotation schemes proved to be suitable for specific information access systems, we believe that additional layers referring to QA processing should be considered to fully capture the relevant information for more general applications. To the best of our knowledge, none of the currently available annotated corpora of spoken language dealing with information requests contains labels referring to the QA area of competence (such as the “Expected Answer Type” or the “Question Topical Target”). Our contribution aims at improving the proposed annotation schemes, by considering specific information broadly and successfully exploited in QA.

3 Question Answering

Building a QA system involves a complex development process, with a number of sub-tasks that have to be addressed at each stage of the loop. Typical QA systems rely on three main components, respectively in charge of Question Analysis, Document Retrieval, and Answer Extraction [1]. In the **Question Analysis** phase, questions are processed in order to: *i*) capture the actual information need they express, *ii*) extract a list of relevant search terms, and *iii*) build structured queries to a target document collection. Then, in the **Document Retrieval** phase, the collection is searched for a ranked list of relevant documents. Finally, in the **Answer Extraction** phase, the retrieved documents are analysed in order to locate the best candidate answers.

The different types of information that can be extracted from an input question in the Question Analysis phase contribute to the success of the whole process. In particular, our work concentrates on the notions of *Question Topical Target*, *Expected Answer Type*, and *Expected Answer Quantifier*.

3.1 Question Topical Target (QTT)

The QTT (sometimes referred to as question *focus*, or question *topic*) is the part of text, within the question, that describes the entity on which the re-

² <http://mate.nis.sdu.dk>

quest has been made. We define the extension of the QTT as the whole syntactic phrase (noun or verb phrase) whose head is the actual entity about which something is asked, as in (QTT is underlined): “*How much does it cost to get to Santa Chiara hospital by taxi?*”.

Effective QTT identification becomes useful in the Document Retrieval phase. More specifically, since QTT terms (or their synonyms) are likely to appear in a retrieved sentence that contains the answer, query formulation/relaxation strategies should appropriately weight such terms. However, especially dealing with complex queries, more than one candidate QTT can be found, and their identification is not always straightforward (e.g. “*At what time does the show La Boheme begin at Sociale Theatre, and what bus[?] should I take to get to the theatre[?]”).*

3.2 Expected Answer Type (EAT)

The EAT is the semantic category associated to the desired answer, chosen out of a predefined set of labels (e.g. PERSON, LOCATION, DATE). For example, the question “*How many colors are in the Italian flag?*” asks for a QUANTITY, and “*Where is the Uffizi museum?*” asks for a LOCATION.

Most QA systems described in literature heavily rely on EAT information, at least in the Answer Extraction phase, to narrow the potential answer candidates search space. The idea is that, since the selected candidates should represent instances of the EAT semantic category, all candidates that do not fall in such category can be filtered out.

3.3 Expected Answer Quantifier (EAQ)

We define the EAQ as an attribute of the EAT that specifies the number of expected items in the answer. Even though EAQ identification is usually not explicitly addressed in QA systems, the rationale behind this attribute has been implicitly asserted in the framework of the TREC and CLEF QA tasks, where test questions asking for multiple answer items are marked as “*list*” questions.

3.4 Reference Resources for Question Answering

For most QA sub-tasks, both rule-based and Machine Learning (ML) approaches have been proposed in literature. The increasing interest recently raised in supervised ML-based techniques brought the necessity of having reference linguistic resources available, previously annotated by human experts. Any module relying on such techniques, in fact, requires annotated training/test sets of reference data. This is the case, for instance, of the ML-based EAT classifier proposed by [2], the answer re-ranking module described in [3], and more in general the case of any embedded piece of linguistic analysis, as Named Entity Recognition.

In this perspective, unfortunately, few useful resources are already available to the research community. The Webclopedia project (<http://www.isi.edu/natural-language/projects/webclopedia/>) developed at USC-ISI produced a taxonomy of

over 140 Expected Answer Types (EATs), which includes lexical, syntactic, and semantic labels, and a corresponding set of answer patterns based on the analysis of several thousands of questions.

Another relevant data collection, described in [2], has been built to train classifiers on a similar EAT taxonomy. For this purpose, 6000 questions³ from the Webclopedia project and from the TREC-8, TREC-9, and TREC-10 evaluation campaigns have been manually annotated.

Both these resources have been created with the purpose of building individual QA system rather than developing a general benchmark for comparative evaluations. This is instead the case of the MULTIEIGHT corpus [4], developed in the context of the CLEF QA evaluation exercise, which provides a cross-language collection of question-answer pairs. However, such pairs completely lack any semantic annotation.

The shortage of a comprehensive human-annotated benchmark is particularly surprising if we consider that: *i*) during the years, most of QA research converged on few standard approaches considering similar features of the input questions; and *ii*) many of these techniques and features are substantially task-independent, and represent a potential richness for a variety of information-access applications. These considerations motivate our decision to develop a comprehensive reference resource, the QALL-ME benchmark, useful to train/test information access models not limited to QA.

4 The QALL-ME Benchmark

This section describes the ongoing work in building the QALL-ME benchmark, with respect to speech data acquisition, transcription, translation into English, annotation of speech acts, EAT, QTT, and EAQ. Up to now, speech act annotation is concluded, while the other annotation levels are still in progress. All data are in Italian, and parallel acquisition is being done on Spanish, German, and English. The selected domain refers to cultural events in a town.

4.1 Data Acquisition and Transcription

In order to assure linguistic variability, we collected spoken data from 161 speakers, including 68 males and 93 females. 12 were non-native. Each speaker was presented with 15 scenarios, describing possible information needs in the selected domain. For each scenario two utterances were collected: the first one was spontaneous, while the second one was previously generated and then just read by the speaker. In order to minimize the risk of influencing the speaker in the formulation of the spontaneous utterances, each scenario was presented on a computer screen as a list containing the following items:

SubDomain: the context in which the question has to be posed, such as “Cinema/Movie”, or “concert”;

³ Available at <http://l2r.cs.uiuc.edu/~cogcomp/data.php>

- DesiredOutput:** the type of information the speaker wants to obtain from the system (e.g. the telephone number of a cinema, the cost of a ticket);
- MandatoryItems:** a list of items that must be put in the question in order to ensure its soundness (e.g. the name of the cinema is “Astra”, the title of the opera is “La Boheme”);
- OptionalItems** a list of additional items that the speaker can put in the request (e.g. the cinema is located in “Via Mancini”, the concert venue is “Teatro Sociale”).

Each question was made by telephone to an automatic system, and recorded together with information for identifying the corresponding scenario. After the acquisition, all the audio files acquired from a speaker were joined together and orthographically transcribed using the tool Transcriber⁴. For each session, a dedicated transcription file was initialized, including time markers, text of the just-read sentences, gender and accent of the speaker. The resulting database currently contains 4768 utterances (2316 read + 2452 spontaneous), for a total speech duration of about 9 hours and 20 minutes. The average utterance duration is 7 seconds. 104 utterances were marked as unusable, mainly due to technical problems experienced during the acquisition. As a result, the total number of valid utterances is 4664 (2290 read + 2374 spontaneous). More data are reported in Table 1. Being domain-restricted, our scenarios often led to the same utterance (matching word sequence). However, the number of repetitions is actually small, as reported in Table 2.

	# words	# utterances	avg. len. (words)
read utterances	25715	2290	11.2
spontaneous utterances	33492	2374	14.1
total utterances	59207	4664	12.7

Table 1. Features of the valid utterances in the collected database.

4.2 Annotation Layers

Besides the translation of the collected data into English, the QALL-ME benchmark addresses two main levels of annotation. The first one refers to speech acts, while the second highlights relevant elements for the semantic interpretation of the request, including Question Topical Target (QTT), Expected Answer Type (EAT) and Expected Answer Quantifier (EAQ) as introduced in Section 3. For the annotation task we used CLARK, an XML-based System for Corpora Development (<http://www.bultreebank.org/clark/index.html>).

⁴ <http://trans.sourceforge.net>

number of repetitions	read utterances	spontaneous utterances	all
1	989	2345	3289
2	307	13	323
3	139	1	151
4	47	-	43
5	10	-	12
6	3	-	5
7	2	-	2
total	2290	2374	4664

Table 2. *Number of valid utterances with repetitions.*

Translation and Speech Acts Annotation. The collected data have been translated into English by simulating the real situation of an English speaker visiting a foreign city, i.e. with non-translated named entities (e.g. names of streets, restaurants, etc.). From the speech act side, we separate, within each utterance, what has to be interpreted as the actual request from what does not need an answer. For request speech acts, we use the labels DIRECT and INDIRECT. For non request acts (utterances used by the speaker to introduce or contextualize the request), we use GREETINGS, THANKS, ASSERT (usually referred to as “declarative clause” as in [5]), and OTHER, which includes non request utterances such as “well”, “hallo”, and “listen”.

Request labels identify all the utterances used by the speaker to require information. DIRECT requests include wh-questions (as shown in Example 1), questions introduced by e.g. “*Could you tell me*”, or “*May I know*”, or pronounced with an ascending intonation (typical of Italian spoken questions). On an intuitive level, we can say that a request is DIRECT if we can put a question mark at the end of it (punctuation is actually not present in our corpus). Conversely, INDIRECT requests include requests formulated in indirect or in implicit ways, as shown in Example 1 again. The outcome of speech acts annotation of the Italian part of the QALL-ME benchmark results in: 2350 DIRECT requests, 1545 INDIRECT requests, 556 ASSERTS, 139 THANKS, 391 GREETINGS, and 152 OTHER. The inter-annotator agreement has been calculated as the Dice coefficient on 1000 randomly picked sentences. Overall agreement is 96.1%, with the following label breakdown: ASSERT: 85.5%; DIRECT: 97.88%; GREETINGS: 99.49%; INDIRECT: 97.33%; OTHER: 76.47%; THANKS: 98.51%.

Example 1: Speech acts. (*Good morning, I would like to know the address of the church of Santissima Trinita’ in Trento, thanks*)

```
<greetings> buongiorno </greetings>
<indirect> vorrei sapere l’indirizzo della chiesa della
Santissima Trinita’ a Trento </indirect> <thanks> grazie </thanks>
```

QTT, EAT, and EAQ. According to its definition, the QTT can be a noun phrase or a verb phrase. Since more QTTs may appear in the same utterance, we

introduced a QTT identifier to allow for EAT references, as shown in Example 2. While an EAT always refers to a single QTT, a QTT can have one or more associated and possibly different EATs (e.g. when asking for both time and place of an event). For EAT annotation, we consider a taxonomy of EATs, often used in QA. The first level is domain independent and includes labels as FACTOID, PROCEDURAL, VERIFICATION, and DEFINITION. Deeper levels tend to be more domain dependent, e.g. FACTOID EATs take semantic labels such as PERSON, LOCATION, ORGANIZATION, and TIME. For EAQ annotation, the possible values are: one, many, at least one, n.

Example 2 (QTT, EAT, and EAQ): "quali sono gli indirizzi del museo Diocesano Tridentino e del museo Storico delle Truppe Alpine" (*Which are the addresses of museo Diocesano Tridentino and of museo Storico delle Truppe Alpine?*)

```
<QTT id="1">museo Diocesano Tridentino</QTT>
<QTT id="2">museo Storico delle Truppe Alpine</QTT>
<EAT cat="location" QTT="1" EAQ="one"/>
<EAT cat="location" QTT="2" EAQ="one"/>
```

5 Conclusions and Future Work

This paper presented a semantic annotation scheme for spoken information access dialogues/requests, developed considering the importance of information layers specifically brought from the QA area. Such scheme is being adopted for the annotation of the QALL-ME benchmark, a multilingual corpus of spoken questions in the tourism domain, built within the EU funded project QALL-ME. While the first annotation phase (speech acts) is concluded, and others are in progress (EAT, QTT, and EAQ), additional layers will be considered in the future. These include Multiwords, Named Entities, and normalized Temporal Expressions. The expected result is a reference resource, useful to train and test information access models not limited to QA.

References

1. Voorhees, E.M., Buckland, L.P., eds.: The Fifteenth Text REtrieval Conference Proceedings (TREC 2006), Gaithersburg, Maryland, NIST (November 14-17 2006)
2. Li, X., Roth, D.: Learning Question Classifiers. In: Proceedings of COLING 2002, Taipei, Taiwan (September 2002)
3. Moschitti, A., Quarteroni, S., Basili, R., Manandhar, S.: Exploiting Syntactic and Shallow Semantic Kernels for Question Answer Classification. In: Proceedings of ACL-2007, Prague, Czech Republic (June 2007) 776–783
4. Magnini, B., Giampiccolo, D., Forner, P., Ayache, C., Jijkoun, V., Osenova, P., nas, A.P., Rocha, P., Sacaleanu, B., Sutcliffe, R.: Overview of the CLEF 2006 Multilingual QA Track. In: CLEF 2006 Working Notes, Alicante, Spain (2006)
5. Soria, C., Pirrelli, V.: A Recognition-Based Meta-Scheme for Dialogue Acts Annotation. In Walker, M., ed.: Towards Standards and Tools for Discourse Tagging: Proceedings of the Workshop. ACL, Somerset, New Jersey (1999) 75–83