The QALL-ME Benchmark: a Multilingual Resource of Annotated Spoken Requests for Question Answering

Elena Cabrio[1], Milen Kouylekov[1], Bernardo Magnini[1], Matteo Negri[1], Laura Hasler[2], Constantin Orasan[2], David Tomás[3], José L. Vicedo[3], Günter Neumann[4], Corinna Weber[4]

[1] FBK- irst
{cabrio, kouylekov, magnini, negri}@fbk.eu

[2]University of Wolverhampton
{L.Hasler, c.orasan}@wlv.ac.uk

[3]University of Alicante
{tomas, vicedo}@dlsi.ua.es

[4]DFKI
{neumann, cowe01}@dfki.de

Abstract

This paper presents the QALL-ME benchmark, a multilingual resource of annotated spoken requests in the tourism domain, freely available for research purposes. The languages currently involved in the project are Italian, English, Spanish and German. It introduces a semantic annotation scheme for spoken information access requests, specifically derived from Question Answering (QA) research. In addition to pragmatic and semantic annotations, we propose three QA-based annotation levels: the *Expected Answer Type*, the *Expected Answer Quantifier* and the *Question Topical Target* of a request, to fully capture the content of a request and extract the sought-after information. The QALL-ME benchmark is developed under the EU-FP6 QALL-ME project which aims at the realization of a shared and distributed infrastructure for Question Answering (QA) systems on mobile devices (e.g. mobile phones). Questions are formulated by the users in free natural language input, and the system returns the actual sequence of words which constitutes the answer from a collection of information sources (e.g. documents, databases). Within this framework, the benchmark has the twofold purpose of training machine learning based applications for QA, and testing their actual performance with a rapid turnaround in controlled laboratory setting.

1. Introduction

This paper presents the QALL-ME benchmark, a multilingual resource of annotated spoken requests in the tourism domain, freely available for research purposes. The QALL-ME benchmark is developed under the EU-FP6 QALL-ME project[1] which aims at the realization of a shared and distributed infrastructure for Question Answering (QA) systems on mobile devices (e.g. mobile phones). Within this framework, our main motivation is to deliver a human-annotated resource for QA systems development and evaluation.

Since the project deals with both textual and spoken requests, the annotation of the resource pays particular attention to the information needed in the QA and speech processing research areas. Annotation levels include both pragmatic and semantic annotations. Moreover, additional layers specifically referring to QA processing have been considered to fully capture the relevant information for more general applications. In particular, we introduced three QA-based annotation levels: Expected Answer Type, Expected Answer Quantifier, and Question Topical Target. To the best of our knowledge, none of the currently available annotated corpora of spoken language dealing with information requests contains QA specific labels. Therefore, our contribution aims at improving the proposed annotation schemes by considering specific information broadly and successfully exploited in QA. The benchmark currently includes 14645 questions in four different languages (Italian, Spanish, English, and German), related to the domain of cultural events in a town (e.g. cinema, theatre, exhibitions, etc.).

The paper is structured as follows: Section 2 provides an overview of the QALL-ME benchmark as inserted in the project frame; Section 3 presents the data collection, with respect to the spoken language acquisition, the transcription and the translations of the data into English; Section 4 introduces the annotation layers, in particular the annotation of the speech acts; Section 5 presents the three QA-based annotation levels: the Expected Answer Type, the Expected Answer Quantifier and the Question Topical Target; Section 6 describes the annotation of the relations contained in the questions; Section 7 concerns related work; finally, Section 8 discusses future work and draws some conclusions.

2. The QALL-ME Benchmark

This Section shortly introduces the QALL-ME benchmark under the perspective of the QALL-ME project. The general objective of the project is to establish a shared infrastructure for multilingual and multimodal open domain Question Answering for mobile phones. The scientific and technological objectives pursue three crucial directions: multilingual open domain QA, user-driven and context-

[1]http://qallme.fbk.eu

aware QA, and Machine Learning technologies for QA. The specific research objectives of the project include state-of- the art advancements in the complexity of the questions handled by the system (e.g. how questions); the development of a web-based architecture for cross-language QA (i.e. question in one language, answer in a different language); the realization of real time QA systems for concrete applications; the integration of the temporal and spatial context both for question interpretation and for answer extraction; the development of a robust framework for applying minimally supervised machine learning algorithms to QA tasks; and the integration of mature technologies for automatic speech recognition within the open domain question answering framework.

The selected domain is represented by local events in a town, usually available either through specialized web sites or local newspapers and publications. Experimentations will be carried out in four cities (one for each language involved in the project), using constantly updated information provided by a number of selected data providers.

In the project context, we have been developing two strategic resources: the QALL-ME Ontology, a formal representation of the domain of cultural events, and the QALL-ME benchmark, a corpus of multilingual annotated questions. The two resources are strictly connected as far as semantic annotations are concerned, as the Ontology provides semantic labels for the annotation of Expected Answer Type, Question Topical Target and questions relations. The use of the QALL-ME benchmark as a training for machine learning based algorithms for question interpretation is reported in (Negri et al., 2008).

Both the QALL-ME benchmark and the QALL-ME ontology are being made incrementally available at the Project website (http://qallme.fbk.eu), where new updated versions in any of the four languages are published once a new annotation layer is completed.

3. Data Collection

3.1. Spoken Requests Acquisition

For data acquisition, a large number of speakers has been presented with a graphical interface, describing possible information needs in the selected domain. For each scenario two utterances were collected: the first one is spontaneous, while the second one is a pre-defined question that is simply 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 to them on a computer screen as a list containing: the context in which the question has to be posed (e.g. "Cinema/Movie" or "Concert"); 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); a list of items that must be present in the question in order to ensure its validity (e.g. the name of the cinema is "Astra", the title of the opera is "La Boheme"); a list of additional items that the speaker can use to make the question (e.g.the cinema is located in "Via Manci", the concert venue is "Teatro Sociale").

Each question was acquired using a telephone, and recorded together with information for identifying the corresponding scenario.

3.2. Transcription

After the acquisition, all the audio files acquired from a speaker were joined together and orthographically transcribed using the tool Transcriber[2]. For each session, a dedicated transcription file was initialized, which includes time markers, the text of the read sentences, and the gender and accent of the speaker.

Being domain-restricted, our scenarios often led to the same utterance (matching word sequence). However, the number of repetitions is actually small and concentrated within the read utterances; the repetitions are well documented in the resource, where the repeated utterances have been clustered. The number of distinct utterances, i.e. non repeated ones is: 3289 for Italian, 2427 for Spanish, 3472 for English and 796 for German.

Data concerning the total speech duration, and the distribution of the speakers with respect to their language, gender and mother tongue is reported in Table 1[3][4].

Data concerning the resulting database are reported in Table 2.

3.3. Translations

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.). One of the future goals is to have all the data collected for one language translated into the other three languages (using English as an interlingua, if necessary). The study on the portability of annotation layers from one language to another is in the pipeline.

4. Speech Acts Annotation

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 introduces relevant elements for the semantic interpretation of the request, including Question Topical Target, Expected Answer Type and Expected Answer Quantifier.

Transcribed files were annotated using CLaRK, an XML-based System for Corpora Development [5].

On 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. Request labels identify all the utterances used by the speaker to ask for information. Requests are marked either as DIRECT or INDIRECT. 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 its end (punctuation is actually not present in our corpus). Conversely, INDIRECT requests include requests formulated in indirect or in implicit ways, as shown

[2] http://trans.sourceforge.net

[3] The gender of 4 English speakers is unknown.

[4] Since there is no speech processing foreseen in the QALL-ME project for German, at present the main focus is on written questions. Nevertheless, the creation of audio files from a subset of the questions is in progress.

[5] http://www.bultreebank.org/clark/index.html

	# speakers	males	females	non-native	tot. speech dur.	avg. utt. dur.
ITALIAN	161	68	93	12	9h 20'	7"
SPANISH	150	109	41	8	16h 4'	5.14"
ENGLISH	113	46	63	21	7h 35'	6.1"
GERMAN	9	4	5	2	1h 21'	4.9"

Table 1: *Data acquisition features.*

		# words	# utterances	avg. len. (words)
ITALIAN	**read utterances**	25715	2290	11.2
	spontaneous utterances	33492	2374	14.1
	total utterances	59207	4664	12.7
SPANISH	**read utterances**	25919	2250	11.52
	spontaneous utterances	26327	2250	11.70
	total utterances	52246	4500	11.61
ENGLISH	**read utterances**	26626	2215	12
	spontaneous utterances	36000	2286	15.8
	total utterances	62626	4501	13.9
GERMAN	**read utterances**	10990	903	12.17
	spontaneous utterances	985	77	12.79
	total utterances	11975	980	12.22

Table 2: *Features of the valid utterances in the collected database.*

in Example 2. For non-request acts (utterances used by the speaker to introduce or contextualize the request), we use the label GREETINGS, THANKS, ASSERT (usually referred to as "declarative clause" as in (Soria and Pirrelli, 1999)), and OTHER, which includes non request utterances such as *"well", "hallo"*, and *"listen"*. To date, this level of annotation has been completed only for Italian (see (Cabrio et al., 2007)) and Spanish.

The inter-annotator agreement has been calculated for Italian using the Dice coefficient, over 1000 randomly picked sentences annotated by two annotators. The Dice coefficient is computed as 2C/(A+B), where C is the number of common annotations, while A and B are respectively the number of annotations provided by the first and the second annotator. The 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 (direct requests).

```
<direct>what is the name of the pharmacy
located in via San Pio X 77 in Trento
</direct>
```

Example 2: Speech acts (indirect requests).

```
<greetings> good morning </greetings>
<indirect>I would like to know the address
of the church of Santissima Trinita' in
Trento </indirect> <thanks> thanks </thanks>
```

4.1. Speech Acts Annotation for the English section

For speech acts annotation on the English section of the QALL-ME benchmark, a slightly different scheme is applied using the multipurpose annotation tool PALinkA (Orasan, 2003). Labels such as GREETING, THANKING, THANK-BYE and REQUEST-INFO are adapted from existing speech acts theories and dialogue annotation projects (see, e.g. (Larsson, 1998) for an overview/comparison) to suit our data. First, utterances are marked as suitable (<utterance>) or unsuitable (<interrupted>, <trash>, <nonsense>); then, C-units (Biber et al., 1999) are marked within suitable utterances. The C-UNIT tag takes the attributes CLAUSAL and NON-CLAUSAL; NON-CLAUSAL is further split into PRAGMATIC and SPECIFY-INFO.

Next, the speech acts themselves are annotated. There are two general tags, PRIMARY_SPEECH_ACT and SECONDARY_SPEECH_ACT, the attributes assigned to which determine the attribute given to the final tag, SPEECH_ACT_TYPE, which is marked as a relation between the two Speech Acts tags. PRIMARY_SPEECH_ACT can take any of the attributes REQUEST, QUESTION, STATE, INTRODUCE, END, depending on the surface form. As we are concerned with requests/questions requiring a response, only primary Speech Acts which are labelled as REQUEST, QUESTION, STATE are assigned a secondary speech act tag. The attributes of SECONDARY_SPEECH_ACT are REQUEST, QUESTION, STATE, depending on the underlying, or 'real' function of the utterance (e.g., a statement or a question can function to request information). SPEECH_ACT_TYPE is DIRECT if the primary and secondary Speech Acts take the same attribute, and INDIRECT if they do not (as shown in Example 3).

Example 3: Speech acts .

```
<clausal><question><request><indirect>
would you be able to tell me
<non-clausal:specify-info>the bus 5 4 3
</non-clausal:specify-info>the start
```

```
hours for the bus</indirect></request>
</question></clausal>
```

5. Question Answering Annotation

This section describes the QA-based annotation levels, in particular the Expected Answer Type, the Expected Answer Quantifier, and the Question Topical Target.

5.1. Expected Answer Type (EAT).

The EAT has been defined by (Prager, 2007) as the class of object (or rhetorical type of sentence) sought by the question; in other words, it is the semantic category associated with the desired answer, chosen from a predefined set of labels. For EAT annotation, we extracted our EAT taxonomy from Graesser's taxonomy (Graesser et al., 1988), adding two other levels: one is based on the QALL-ME ontology, a domain specific ontology developed specifically for the project purposes; the other one is based on Sekine's Named Entity Hierarchy (ENE).[6] In detail, the level related to Graesser's taxonomy is domain independent and includes labels as FACTOID, PROCEDURAL, VERIFICATION, and DEFINITION/DESCRIPTION. Deeper levels (referring only to FACTOID questions tend to be more domain dependent, e.g. FACTOID EATs take semantic labels such as PERSON, LOCATION, ORGANIZATION, and TIME, referring to the QALL-ME ontology (see Example 4).

Concerning VERIFICATION, the definition of the question type is not enough, since the speaker implicitly needs more information than simply a yes/no answer (e.g. *"is there a web-site of the police headquarters?"*). These questions have thus been annotated both with the tag VERIFICATION and the appropriate tags of the QALL-ME ontology (e.g. CONTACT). The choice of the correct EAT is not always straightforward and it is difficult to define unambiguous guidelines. For PROCEDURAL, and DEFINITION/DESCRIPTION types, no deeper levels have been defined.

To enhance a broad use of the benchmark also for open-domain QA applications, we annotated the EAT also based on Sekine's ENE, a shared EAT taxonomy. For the annotation task we used Sekine's tagging tool FuuTag.

5.2. 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. For EAQ annotation, the possible values are: one, at least one, all, n.

Example 4 (EAT, and EAQ).

```
what are the address and the telephone number
of Venezia hotel in Trento
<eats>
<eat type="FACTOID" sekine="ADDRESS_OTHER"
qallme="PostalAddress" eaq="one"/>
<eat type="FACTOID" sekine="PHONE_NUMBER"
qallme="Contact" eaq="one"/></eats>
```

[6] http://nlp.cs.nyu.edu/ene/

5.3. Question Topical Target (QTT).

The QTT (sometimes referred to as question *focus* (Monz, 2003), or question *topic* (Prager, 2007)) is the part of text, within the question, that describes the entity about which the request has been made. We define the extension of the QTT as the whole syntactic phrase (noun or verb phrase) whose head is the entity about which something is asked, as in: *"How much does it cost to get to Santa Chiara hospital by taxi?"*(QTT is underlined). Especially in the Document Retrieval phase of the QA process, QTT identification becomes useful: 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 (Monz, 2003). However, especially when dealing with complex queries, more than one candidate QTT can be found, and their identification is not always straightforward. Since more than one QTT may appear in the same utterance, we introduced a QTT identifier to allow for EAT references, as shown in Example 5.

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).

Example 5 (QTT, EAT, and EAQ).

```
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 Alpi-
ne</QTT>
<eat type="FACTOID" sekine="ADDRESS"
qallme="PostalAddress" eaq="one" QTT="1"/>
<eat type="FACTOID" sekine="ADDRESS"
qallme="PostalAddress" eaq="one" QTT="2"/>
```

6. Annotation of Relations

In order to enhance a richer semantic interpretation of the questions, also the annotation of the relations that they contain has been addressed. Such annotation is work in progress, and will be completed in future releases of the QALL-ME benchmark. Detecting relations among entities is often crucial, especially in QA applications, as they convey and complete the context in which a specific request has to be interpreted. Often, in fact, discovering relations is necessary to capture all the constraints that define the actual information need expressed by the request, thus defining and narrowing the search space of potential answers. For instance, the relations between a MOVIE and the DATE of its projection, the MOVIE and the STARTINGHOUR of a specific show, and a MOVIE and the CINEMA where it is projected must be taken into account while interpreting the question: "*at what time is the movie il grande capo beginning tomorrow afternoon at Vittoria cinema*".

At this stage the annotation focuses only on binary relations. For this purpose, a total of 75 relations defined in the QALL-ME ontology have been selected.

number of questions in the Cinema/Movie domain	367
number of possible relations	12
average relations per question	2.43
min relations per question	1
max relations per question	6

Table 3: *Annotation of relations in the Italian Cinema/Movie questions*

These include relations such as HASDATE(EVENT,DATE), ISINDESTINATION(SITE,DESTINATION), and HASPHONENUMBER(SITE,PHONENUMBER), which respectively connect an event (*e.g.* of the type MOVIE, CONCERT, MATCH, etc.) and the site (*e.g.* of the type CINEMA, MUSEUM, PHARMACY, etc.) where it takes place, a site and the city where it is located, and a site and its telephone number. As an example, the relation HASDATE(MOVIE,DATE) represents a relation which has MOVIE as domain and DATE as range. Possible lexicalizations of the relation are:

- *"when will Eragon be on in Trento"*
- *"what is the name of the director of dreamgirls today at Nuovo Roma cinema"*
- *"which dramatic movie directed by Gabriele Muccino is now showed"*

As can be seen from the previous examples (specifically the first and the second) relation annotation is related to EAT annotation, with partial overlaps. Often, in fact, the EAT of a question (*e.g.* TIME) can be mapped to the range of one of the annotated relations (*e.g.* STARTINGHOUR).
In the current version of the QALL-ME benchmark around 10% of the Italian questions (367 out of 3289), namely those referring to the Cinema/Movie domain, have been annotated with the 12 (out of 75) relations that hold in such domain. Even though this is a relatively small subset of the whole benchmark, it's worth noting that all the relations annotated for a specific question are portable across languages, being our translations strictly literal. As an example, all the translations of the Italian question *"what is the name of the director of 007 Casino Royale on today at cinema Modena"* can be assigned to the three relations:

HASDATE(MOVIE,DATE)
HASMOVIESITE(MOVIE,CINEMA)
HASDIRECTOR(MOVIE,DIRECTOR).

Table 3 provides some relevant figures about the annotation completed to date.
A Kappa value of 0.94 (*almost perfect agreement*) was measured for the agreement between two annotators over the same dataset, showing that relation annotation, at least in the Cinema/Movie domain, is a well defined task.

7. Related Work

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 corpora[7] 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.
Spoken dialogue material collected within the MATE project[8] 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

[7] http://www.ldc.upenn.edu/Catalog
[8] http://mate.nis.sdu.dk

	audio	transcr.	transl.	speech acts	EAT Sekine	EAT ontol.
ITALIAN	X	X	X	X	X	X
SPANISH	X	X	X	X	X	May 08
ENGLISH	X	X	–	April 08	April 08	April 08
GERMAN	June 08	June 08	undefined	undefined	undefined	undefined

Table 4: *Present situation and tentative scheduling of the availability of the resource.*

be considered to fully capture the relevant information for more general applications.

8. Conclusions and Future Work

This paper presented the QALL-ME benchmark, a multilingual resource (for Italian, Spanish, English and German) of annotated spoken requests in the tourism domain. The benchmark takes into account the importance of annotation layers specifically referring to the QA area.

The present situation is summarized in Table 4. According to the QALL-ME project agenda, the above mentioned annotation layers will be completed, for all languages involved, during the second year of the project (due to technical problems, the scheduling of the availability of the resource for German is still undefined). 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.

9. Acknowledgements

The present work has been partially supported by the QALL-ME EU Project - FP6 IST-033860 (http://qallme.fbk.eu).

10. References

Douglas Biber, Stig Johansson, Geoffrey Leech, Susan Conrad, and Edward Finnegan. 1999. Longman grammar of spoken and written english. In *Technical report*, Longman, Harlow.

Elena Cabrio, Bonaventura Coppola, Roberto Gretter, Milen Kouylekov, Bernardo Magnini, and Matteo Negri. 2007. Question answering based annotation for a corpus of spoken requests. In *Proceedings of the workshop on the Semantic Representation of Spoken Language*, Salamanca, Spain, November.

A.C. Graesser, K. Lang, and D. Horgan. 1988. A taxonomy for question generation. *Questioning Exchange*.

Staffan Larsson. 1998. Coding schemas for dialogue moves. In *Technical report*, Göteborg University, Department of Linguistics.

Christof Monz. 2003. *From Document Retrieval to Question Answering*. Ph.D. thesis, University of Amsterdam.

Matteo Negri, Milen Kouylekov, and Bernardo Magnini. 2008. Detecting expected answer relations through textual entailment. In *Proceedings of Cicling 2008*, Haifa, Israel, February.

Constantin Orasan. 2003. PALinkA: A highly customisable tool for discourse annotation. In *4th SIGdial Workshop on Discourse and Dialogue, ACL'03*, Sapporo, Japan, July.

John Prager. 2007. Open-Domain Question-Answering. In *Foundations and Trends in Information Retrieval*. Now Publishers.

Claudia Soria and Vito Pirrelli. 1999. A Recognition-Based Meta-Scheme for Dialogue Acts Annotation. In Marilyn Walker, editor, *Towards Standards and Tools for Discourse Tagging: Proceedings of the Workshop*, pages 75–83. ACL, Somerset, New Jersey.