Dealing with Spoken Requests in a Multimodal Question Answering System

Roberto Gretter, Milen Kouylekov, Matteo Negri

Fondazione Bruno Kessler
Via Sommarive, 18 - Povo, Trento, Italy
{gretter,kouylekov,negri}@fbk.eu

Abstract. This paper reports on experiments performed in the development of the QALL-ME system, a multilingual QA infrastructure capable of handling input requests both in written and spoken form. Our objective is to estimate the impact of dealing with automatically transcribed (*i.e.* noisy) requests on a specific question interpretation task, namely the extraction of relations from natural language questions. A number of experiments are presented, featuring different combinations of manually and automatically transcribed questions datasets to train and evaluate the system. Results (ranging from 0.624 to 0.634 F-measure in the recogniton of the relations expressed by a question) demonstrate that the impact of noisy data on question interpretation is negligible with all the combinations of training/test data. This shows that the benefits of enabling speech access capabilities, allowing for a more natural human-machine interaction, outweight the minimal loss in terms of performance.

Keywords: Question Answering, Textual Entailment, Speech Recognition

1 Introduction

Recent years have seen an increasing interest towards advanced information access applications, motivated by the huge market potential of systems providing natural human-machine interaction capabilities. Question Answering (QA) research plays an important role in this direction, focusing on the development of systems that return actual *answers* in response to *natural language questions*. In the same direction, enabling users to express their needs in the most natural way, the increased reliability of Automatic Speech Recognition (ASR) systems offers new opportunities for a simplified and more effective access to information.

The combination of QA technology and multimodal interaction capabilities is among the challenges addressed by the EU funded project QALL-ME[1]. The project aims at developing a distributed infrastructure for multilingual QA over structured data, in the domain of cultural events in a town. The foreseen multimodal capabilities of the QALL-ME system include the possibility of access by

[1] http://qallme.itc.it/

means of mobile devices (*e.g.* mobile phones), to pose natural language questions either in *textual* form (*e.g.* sms), or in *speech* modality. From a research perspective, the speech access modality is particularly interesting, since it raises the need for robust methods capable of handling noisy and sub-optimal inputs. Often, in fact, spoken requests are more complex than written ones (*e.g.* they contain hesitations and repetitions), and their automatic transcription may contain errors. Investigating these aspects, which are currently out of the scope of traditional QA research, is the main purpose of this paper.

Our work builds on top of [1], which addresses QA over structured data reformulating the problem as a *Textual Entailment Recognition* (RTE) problem. Textual Entailment (TE) has been recently proposed as a unifying framework for applied semantics [2], where the need for an explicit representation of a mapping between linguistic objects and data objects can be, at least partially, bypassed through the definition of semantic inferences at the textual level. In this framework, a text (T) is said to entail a hypothesis (H) if the meaning of H can be derived from the meaning of T. According to the TE framework, [1] proposes that the interpretation of a given question can be addressed as a Relation Extraction task based on TE, where the text (T) is the question, and the hypothesis (H) is a relational pattern, which is associated to instructions for retrieving the answer to the question. Given a question q and a set of relational patterns $P=\{p_1, ..., p_n\}$, the basic operation is to select those patterns in P that are entailed by q. Instructions associated to patters may be viewed as high precision procedures for answer extraction, which are dependent on the specific data source accessed for answer extraction. For instance, in case of QA over structured data, instructions would be SQL queries to a database.

Building on the positive results reported in [1], we adopt a similar TE-based approach to question interpretation, to investigate the impact of handling noisy data obtained from automatic transcriptions. For this purpose, different experiments are carried out running the system under different training/test conditions. In the optimal situation, the system is trained and tested over datasets of manually transcribed Italian questions. The results achieved in this first setting are then compared with: *i)* those achieved by a system trained over manually transcripted questions, and tested over automatic transcriptions (to reproduce a situation that is closer to a real on-field evaluation), *ii)* those achieved by a system trained and tested over automatic transcriptions (to verify if the system can "learn" from systematic errors produced by the ASR), and *iii)* those achieved by training the system over both manual and automatic transcriptions, and testing it over automatic transcriptions (to check if the two sources used together give an added value at a training stage).

The paper is organized as follows. Section 2 introduces TE-based Relation Extraction as a question interpretation task. Section 3 describes the dataset used for experiments. Section 4 describes our automatic speech recognition system. Sections 5 and 6 respectively report experiment results, and concluding remarks.

2 TE-based Relation Extraction

The TE-based approach to question interpretation has been defined in [1] as a classification problem, where a question q has to be assigned to all the relations R_1,...,R_n it expresses, selected from a predefined set R (in this work we focus on binary relations, although extensions to n-ary relations are expected). For instance, given the question *"What can I see today at cinema Astra?"*, the following relations represent the expected system's output:
R1: HasMovieSite(Movie:?, Site:"Astra")
R2: HasDate(Movie:?, Date:"today")

The classification (see Figure 1 for a schematic representaton of the overall process) is carried out by means of a RTE engine, which is in charge of discovering entailment relations between the input question q, and a set of textual patterns stored in a *Pattern Repository (P)*. P contains n sets of textual patterns, each set representing possible lexicalizations of one relation R_i in R.

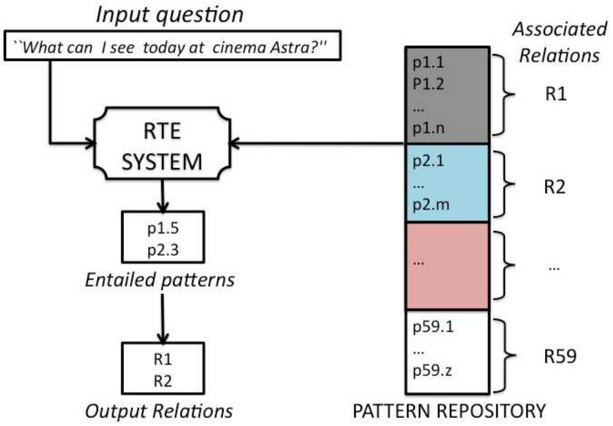

Fig. 1. Entailment-based Relation Extraction process

Given a question q, the RTE engine attempts to verify whether an entailment relation holds between q and each pattern in P. All the relations associated to the patterns entailed by q are output by the system. If none of the patterns in P is entailed by q, this is interpreted as evidence that the question is out of domain. Sections 2.1 and 2.2 overview two crucial aspects of the proposed approach to question interpretation, namely: *i)* the *type of the textual patterns stored in the repository P*, and *ii)* our *TE recognition algorithm.*

2.1 Minimal Relational Patterns (MRPs)

According to our formulation of the task, we say that a relational pattern p expresses a relation $R(arg1, arg2)$ in a certain language L if speakers of L agree that the meaning of p expresses the relation R between $arg1$ and $arg2$, given their knowledge about the entities. For instance, all the examples in Table 1 represent relational patterns for the relation HasMovieSite(Movie, Site).

(1)	<ARG2:MOVIE:X> *is shown at cinema* <ARG1:CINEMA:Y>
(2)	*What* <ARG2:*movie*> *is on at* <ARG1:CINEMA:Y>*?*
(3)	*Is there any* <ARG2:*movie*> *that I can see at* <ARG1:CINEMA:Y>*?*
(4)	*Can I see* <ARG2:MOVIE:X> *at cinema* <ARG1:CINEMA:Y> *on* <ARG?:DATE:Z>*?*

Table 1. *Examples of relational patterns*

In order to be profitably used in the proposed entailment framework valid patterns have the additional property of representing only *one* relation. Pattern representing multiple relations, in fact, would be entailed only by questions containing all these relations, thus resulting limited in their usage. To describe one-relation patterns [1] introduces the notion of *Minimal Relational Pattern* (MRP), which can be formally defined in terms of TE. Given a set $P=\{p_1, p_n\}$ of relational patterns for a relation R, a pattern p_k belonging to P is a MRP for the relation R if condition (1) holds.

$$\forall p_i \in P, p_k \mapsto p_i = \emptyset \tag{1}$$

In other words, a pattern p_k is minimal if none of the other relational patterns contained in P can be derived from p_k (*i.e.* is logically entailed by p_k). According to such definition, patterns (1)-(3) in Table1 are MRPs for the relation HasMovieSite(Movie, Site), while (4) is not, since it entails the others.

2.2 Distance Based Entailment Recognition

For each relation we train a RTE engine, which is an adaptation of our English system evaluated in the framework of the Pascal-RTE Challenge [3]. The system has been implemented within a distance-based framework, and is based on *Levenshtein Distance* (LD) or *Linear Distance* [4]. The intuition is that, given a question q and a pattern p, the probability of an entailment relation between q and p is related to the possibility of mapping the whole content of q into the content of p. The more straightforward the mapping can be established, the more probable is the entailment relation.

Algorithm. Edit distance approaches for RTE, such as the one proposed in [5], assume that the distance between T and H is a characteristic that separates the positive pairs, for which entailment holds, from the negative pairs, for which entailment does not hold. Such distance is computed as the cost of the editing operations (*i.e.* insertion, deletion and substitution) which are required to transform T into H. Each edit operation on two text fragments A and B (denoted as

$A \rightarrow B$) has an associated cost (denoted as $\gamma(A \rightarrow B)$). The entailment score for a T-H pair is calculated on the minimal set of edit operations that transform T into H. An entailment relation is assigned to a T-H pair only if the overall cost of the transformation is below a certain threshold empirically estimated over training data. The entailment score function is defined in the following way:

$$score_{entailment}(T, H) = 1 - \frac{\gamma(T, H)}{\gamma_{nomap}(T, H)} \tag{2}$$

where $\gamma(T, H)$ is the function that calculates the edit distance between T and H, and $\gamma_{nomap}(T, H)$ is the *no mapping* distance equivalent to the cost of inserting the entire text of H, and deleting the entire text of T. The entailment score function has a range from 0 (when T is identical to H), to 1 (when T is completely different from H).

LD is calculated by converting both the text T and the hypothesis H into sequences of words. Accordingly, edit operations have been defined as as follows:

- **Insertion** ($\Lambda \rightarrow A$): insert a word A from H into T.
- **Deletion** ($A \rightarrow \Lambda$): delete a word A from T.
- **Substitution** ($A \rightarrow B$): substitute a word A in T with a word B from H.

Cost schemes for edit operations. The core of the edit distance approach is the mechanism for the definition of the cost of edit operations. This mechanism is defined apart from the distance algorithm, and should reflect the knowledge of the user about the processed data. The principle behind it is to capture certain phenomena that facilitate the algorithm to assign small distances to positive T-H pairs, and high distances to negative pairs. For instance, since our task consists in comparing questions (T) with MRPs (H) usually composed by few terms[2], for our experiments we adopted the following simple cost calculation scheme:

$$\gamma(\Lambda \rightarrow A) = length(T)$$
$$\gamma(B \rightarrow \Lambda) = length(H)$$
$$\gamma(A \rightarrow B) = \begin{cases} 0 & A = B \\ \gamma_{i+d}(A \rightarrow B) & otherwise \end{cases}$$

In this scheme the cost of inserting a text fragment from H in T is equal to the length of T, and the cost of deleting a text fragment from T is equal to the length of H. The cost of the substitution of two fragments is set to the sum of the insertion and the deletion of the text fragments, if they are not equal. This means that the algorithm would prefer to delete and insert text fragments rather than substituting them, in case they are not equal. During system development we discovered that this cost calculation scheme performs better than considering fixed costs for insertion and deletion operations.

[2] In the experiments reported in Section 5 we compare questions and MRPs of respective average lengths of around 12.5 and 4.5 words.

3 The QALL-ME Benchmark

In order to experiment with the proposed TE-based approach to question interpretation, we used a dataset of 1487 Italian questions extracted from the Italian part of the QALL-ME benchmark[3] [6]. The benchmark contains several thousand questions, in the four languages involved in the project (English, German, Italian and Spanish), concerning cultural events in a town. Questions have been acquired at the telephone, then manually transcribed and annotated with all the relevant information necessary to train/test the core components of a QA system. The availability of both the original recorded questions and their manual transciptions, together with the annotation of part of the acquired data with the relations of interest in the selected domain[4], provides all the data necessary for our evaluation purposes. Sections 3.1 and 3.2 respectively report additional information about questions acquisition and annotation.

3.1 Data acquisition

To obtain a reasonable linguistic variability in the acquired questions, more than 100 speakers for each language were involved in the data collection process. Each speaker was given a list of scenarios presented on a computer screen, and describing possible information needs in the selected domain. Scenarios were designed to allow the formulation of useful queries, without providing too many suggestions about "how" to formulate them. For this purpose, each scenario was presented as a request template, containing a limited amount of textual material.

Every speaker performed 30 spoken questions, based on 15 scenarios randomly chosen out of a set of 90. For each scenario speakers first generated a spontaneous request, and then read a written one previously prepared. As far as the Italian language is concerned, 161 speakers (93 females and 68 males), 12 of which non-native, were involved in data acquisition. The resulting database contains 4768 Italian utterances (2316 read + 2452 spontaneous), for a total speech duration of about 9 hours and 20 minutes. The average utterance duration is about 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). The average word lengths of read and spontaneous utterances are respectively 11.2 and 14.1 words.

3.2 Data annotation

Besides the original questions in the four languages, their orthographic transcription[5], and their translations into English, different annotation levels have been

[3] The QALL-ME benchmark is freely available on the project's website.

[4] Relation annotation is still work in progress: in the current version of the benchmark only 1487 questions out of 4664 are annotated at this level.

[5] Transcriptions were manually produced using Transcriber, a tool for assisting the manual annotation of speech signals freely downloadable from http://trans.sourceforge.net.

considered in the creation of the QALL-ME benchmark. These include pragmatic (speech acts) and semantic (Named Entities) annotations, the Expected Answer Type, the Expected Answer Quantifier, the Question Topical Target, and relations between entities appearing in the utterance.

As far as relation annotation is concerned, in the current version of the benchmark questions have been manually marked as containing one or more relations chosen from a set of 59 binary relations defined in the QALL-ME ontology. As an example, the annotation of the question *"What is the name of the director of 007 Casino Royale, which is shown today at cinema Modena?"* contains three relations, namely:

HasDate(Movie,Date)

HasMovieSite(Movie,Site)

HasDirector(Movie,Director).

On average, spontaneous and read questions have been respectively annotated with 1.94 and 2.26 relations (ranging from 1 to 6 relations per question). A Kappa value of 0.94 (*almost perfect agreement*) was measured for the agreement between two annotators over part of the dataset (150 questions), demonstrating the reliability of relations annotation.

4 Automatic Speech Recognition (ASR)

ASR over the original recorded questions was carried out using the speech recognizer described in [7]. The system is based on a set of phonetic units represented by continuous density Hidden Markov Models (HMMSs). The acoustic features used are LPC Cepstral coefficients and log-energy, with the corresponding first and second order time derivatives. This feature vector is computed every 10 ms on overlapping windows of lenght 20 ms. HMMs were trained on a set of audio data completely disjoint from the QALL-ME benchmark.

Each word of the lexicon was transcribed and foreign words were hand-checked and possibly corrected. A class-based trigram language model was trained on the benchmark manual transcriptions, following the 10-fold cross validation paradigm (*i.e.* splitting the speech data into 10 blocks). All the utterances pronounced by the same speaker are contained in the same block, to avoid that utterances of the same person will appear both in train and in test data. In this way, each block contains utterances from 16 speakers, except one block that contains data from 17 speakers. After this division, for each block we defined two lists: a test list, which corresponds to the audio files of the given block, and a training list, which corresponds to the union of the other 9 blocks. For each test list, the manual transcriptions of the corresponding training list were used to train the class-based trigram language model. Hence, to evaluate speech recognition performance, 10 different recognizers were trained and 10 different tests were performed. After that, results were merged.

The main drawback of the resulting speech recognition system lies in the language model, which is trained on a very small amount of data. Anyway, despite this potential problem, the speech recognizer performed in a relatively satisfactory way. Following the cross validation paradigm, we evaluated the coverage of the test data in terms of Out-Of-Vocabulary (OOV) words :

- *running words*: 98.9% - about 1 word every 100 words is an OOV word;
- *sentences*: 89.4 % - a sentence is covered if and only if all its words are known;
- *lexicon*: 92.7 % - most of the OOV occur only rarely.

It is worth noting that, since the scenarios considered a fixed time interval, with a fixed list of entities (*e.g.* movie titles, persons, theatre names), if pronounced correctly they will never appear in the OOV lists.

Speech recognition was computed following the cross validation paradigm. Results are reported in Table 2 in terms of Sentence and Word Accuracy. *Sentence Accuracy* measures the percentage of sentences recognized without any error, so that the interpretation will be exactly the same of the corresponding manual transcription. *Word Accuracy* gives the percentage of correctly recognized words and is a better indicator of the quality of ASR performance. These results, altough not very high if compared to state of the art speech recognizers for a small task like this, are quite interesting because, as we will see in the experiments, they do not affect very much the interpretation of the sentences. In fact, many of the errors concern either functional words or similar words carrying the same meaning.

	# sentence accuracy	# word accuracy
bigrams	24.1 %	74.3 %
trigrams	35.4 %	77.4 %

Table 2. Sentence and Word Accuracy for two language models. These results merge the results obtained in the 10-fold cross validation paradigm.

5 Experiments and Results

A number of experiments have been carried out to evaluate our TE-based approach to Relation Extraction under different training/test conditions, depending on the use of *clean* data (*i.e* those resulting from manual transcriptions), or *noisy* data (*i.e.* question transciptions output by the ASR system).

Training/test sets. The benchmark questions annotated with the 59 selected relations have been used to create the *training* and *test* sets for our experiments. For this purpose, the question corpus was randomly split in two sets, respectively containing 999 and 448 questions. Such separation was carried out guaranteeing that, for each relation R, the questions marked with R are distributed in the two sets in proportion 2/3-1/3. In addition, the two sets contain a balanced random mixture of spontaneous and read questions, due to some differences noticed between the two types. On average, in fact, spontaneous questions are longer than the read ones, (the respective average lengths are 14.1, and 11.2 words), and involve more relations (2.26 vs. 1.94).

The larger set of 999 questions is used for the manual creation of MRPs and, together with the resulting Pattern Repository, is used to train our RTE system (*i.e.* to empirically estimate an entailment threshold for each relation, considering positive and negative examples). The smaller set of 448 questions (which remained "unseen" in the MRP acquisition phase) is used as test set to evaluate system's performance.

Pattern Repository. According to the definition given in Section 2.1, for each of the 59 relations R we manually[6] extracted a set of MRPs from the training questions annotated with R. The resulting Pattern Repository contains a total of 226 patterns, with at least 1 MRP per relation (4 on average).

Experiments. Evaluation has been carried out under the following conditions (*i.e.* using different combinations of *clean* and *noisy* training/test data):

- **Experiment 1: Clean/Clean (C/C).** In this configuration the system is trained and tested over manually transcribed questions. This is the same evaluation setting proposed in [1], and is used for comparison with the other configurations involving noisy data.
- **Experiment 2: Clean/Noisy (C/N).** In this configuration the system is trained over manually transcribed questions, and tested over automatic transcriptions. The idea is to check performance variations with non-homogeneous training/test data.
- **Experiment 3: Noisy/Noisy (N/N).** In this configuration both training and test are carried out over automatically transcribed questions. The idea is to verify if the errors produced by the ASR system can be "learned" by the system.
- **Experiment 4: Clean+Noisy/Noisy (C+N/N).** In this configuration the system is trained over the combination of manual and automatic transcriptions, and tested over automatic transcriptions. The idea is to check if the two sources used together give an added value at a training stage.

Results. For each configuration, system performance has been calculated comparing the relations recognized by the system, with those manually marked in the reference test questions. Precision, Recall, and F-measure scores are reported in Table 3.

Quite surprisingly, the impact of dealing with noisy data is negligible under all the training/test combinations. This can be explained by the fact that most of the ASR errors typically concern functional words (articles and prepositions) which are not really important in determining the meaning of an input question.

[6] Even though automatic pattern extraction (either from local corpora or from the Web) is a very active research area, this particular aspect falls beyond the scope of this work.

	C/C	C/N	N/N	C+N/N
Precision	0.543	0.546	0.534	0.55
Recall	0.76	0.763	0.75	0.744
F-measure	0.634	0.637	0.624	0.633

Table 3. Results obtained with different combinations of *clean* (C) and *noisy* (N) data

A crucial point could also be the way in which relations are learned, which demonstated to be quite robust. This also motivates the fact that little differences are observed when clean, noisy or clean + noisy data are used.

6 Conclusions

This paper addressed the problem of extracting relations from a natural language question, focusing on a comparative evaluation of system's performance under different training/test conditions, which depend on the type of data used (manual vs. automatic transcriptions of the same questions). A number of experiments have been described, showing that the use of noisy data has a negligible impact under all the training/test combinations. Such positive result demonstrates that, at least in the proposed task, the benefits of enabling speech access capabilities providing a more natural human-machine interaction outweight the minimal loss in terms of performance.

References

1. Negri, M., Kouylekov, M., Magnini, B.: Detecting Expected Answer Relations through Textual Entailment. Lecture Notes in Computer Science, Computational Linguistics and Intelligent Text Processing **4919** (2008)
2. Dagan, I., Glickman, O.: Probabilistic Textual Entailment: Generic Applied Modeling of Language Variability. In: Proceedings of the PASCAL Workshop on Learning Methods for Text Understanding and Mining, Grenoble, France (2004)
3. Dagan, I., Glickman, O., Magnini, B.: The PASCAL Recognising Textual Entailment Challenge. Lecture Notes in Computer Science **3944** (2006) 177–190
4. Levenshtein, V.: Binary Codes Capable of Correcting Deletions, Insertions, and Reversals. Doklady Akademii Nauk SSSR **163** (1965)
5. Kouylekov, M., Magnini, B.: Combining Lexical Resources with Tree Edit Distance for Recognizing Textual Entailment. Lecture Notes in Computer Science, Machine Learning Challenges **3944** (2006)
6. Cabrio, E., Coppola, B., Gretter, R., Kouylekov, M., Magnini, B., Negri, M.: Question Answering Based Annotation for a Corpus of Spoken Requests. In: Proceedings of the Workshop on Semantic Representation of Spoken Language (SRSL07), Salamanca, Spain, (2007)
7. Falavigna, D., Gretter, R.: Telephone Speech Recognition Applications at IRST. In: Proceedings of IVTTA, Turin, Italy (1998)