Adapting QA Components to Mine Answers in Speech Transcripts⋆

[1] LT lab, DFKI, Saarbrücken, Germany
neumann@dfki.de
[2] Saarland University, Saarbrücken, Germany
rwang@coli.uni-sb.de

Abstract. The paper describes QAST-V1 a robust question answering system for answering factoid questions in manual and automatic transcriptions of speech. The system is an adaptation of our text–based crosslingual open–domain QA system that we used for the CLEF main tasks.

1   Introduction

The focus of the new Question Answering on Speech Transcripts (QAst) track within CLEF 2007 is on extracting answers to written factoid questions in manual and automatic transcripts of records of spoken lectures and meetings. Although the basic functionality of a QAst–based system is similar to that of a textual QA–system the nature of the different scenarios and answer sources provoke new challenges.

The answer sources for CLEF and TREC–like systems are usually text documents like news articles or articles from Wikipedia. In general, an article of such a corpora describes a single topic using a linguistically and stylistically well–formed short text which has been created through a number of revision loops. In this sense, such an article can be considered as being created off–line for the prospective reader. By contrast, transcripts from lectures or meetings are live records of spontaneous speech produced incrementally or on–line in human–human interactions. Here, revisions (of errors or refinements) of utterances take place explicitly and immediately or not at all. Thus, speech transcripts also have to encode such properties of incremental language production, like word repetition, error corrections, refinements or interruptions. Consequently, transcripts are less well–formed, stylistic and fluent as written texts. Furthermore, in case of automatic transcripts errors and language gaps caused by the used automatic speech recognition system also make things not easier for a QAst–based system.

It seems that QA on speech transcripts demands a high degree of robustness and flexibility from the QA components and its architecture.

Nevertheless, the component architecture of a QAst–based system is similar to that of a textual QA–system and consists of the following core functionality: NL question analysis, retrieval of relevant snippets from speech transcripts, answer extraction, and answer selection. Therefore, we decided to develop our initial prototype QAST-V1 following the same underlying design principles that we used for our textual QA system and by the adaptation of some of its core components, cf. [3,4].

2   System Overview

The current information flow is as follows: In an off–line phase we firstly generate an inverted index for the speech corpora such that each sentence is considered as a single document and indexed by its word forms and named entities. In the question answering phase, a list of NL questions is passed to the system. Each NL question is analyzed by the named entity recognizer and by the question analysis component. The main output is a question object which represents the expected answer type (EAT) of the question and its relevant keywords. For example, the EAT of the question "Where is Southern Methodist University?" is LOCATION and the relevant keywords are "Southern Methodist University". From the question object an IR–query expression is created in order to access the indexed document space. The IR–query for the example question is {+neTypes:LOCATION AND + "southern methodist university"} which can be read as "select only documents (in our case only sentences) which contain at least one location entity and the phrase Southern Methodist University ". In the answer extraction step all found location names are considered as answer candidates and the most frequent answer candidates are selected as answers to the question, e.g., "Dallas" and "Texas" are found as possible answers in the manual transcript of the lecture corpus. For each question a list of its N–best answers is returned. In the next sub–sections, we describe some of the core components in more detail.

2.1   Named Entity Recognition

Named Entity Recognition (NER) plays a central role in a factual QA architecture: Named entities are the answers of factual questions and as such define the range for the expected answer types. The answer types directly corresponds to the type of named entities.

There exists already a number NER components, but with different coverage of types. For that reason, we developed a hybrid NER approach where we combined three different NER components:

- LingPipe[1]: It mainly covers PERSON, LOCATION, and ORGANIZATION names for English and co–references between pronouns and corresponding named entities. It realizes a supervised statistical based approach to NER.

---

[1] http://www.alias-i.com/lingpipe/

- Opennlp[2]: Its name finder is also based on a supervised statistical approach and covers mainly seven types of NEs for English, viz. PERSON, LOCATION, ORGANIZATION, DATE, TIME, MONEY, and PERCENTAGE.
- BiQueNER developed by our group. It is based on the semi–supervised approach developed by [1] and handles the following NE types: LANGUAGE, SYSTEM/METHOD, MEASURE, COLOUR, SHAPE, and MATERIAL.

All three NERs run in parallel on an input text. The individual results are combined via the IR–query construction process and the answer extraction process. In this way, also conflicting cases are handled like different NE readings and (implicit) partial or overlapping annotations.

2.2   Document Preprocessing

A sentence-oriented preprocessing determining only sentence boundaries, named entities (NE) and their co-references turned out to be a useful level of offline annotation of written texts, at least for the CLEF–kind of factual questions, cf. [3] for a detailed discussion. For that reason we decided to apply the same off–line preprocessing approach also to the QAst collections. In particular the following steps are performed: 1) Extracting lines of words from the automatic speech transcripts so that both the manual and automatic transcript are in the same format. 2) Identification of sentence boundaries using the sentence splitter of the Opennlp tool which is based on maximum entropy modeling. We are currently using the language model the sentence splitter comes with which is optimized for written texts. 3) Annotation of the sentences with recognized named entities.

The preprocessed documents are further processed by the IR–development engine Lucene, cf. [2]. We are using Lucene in such a way that for all extracted named entities and content words, Lucene provides indexes which point to the corresponding sentences directly. Especially in the case of named entities type–based indexes are created which support the specification of type constraints in an IR–query. This will not only narrow the amount of data being analyzed for answer extraction, but will also guarantee the existence of an answer candidate.

2.3   Question Processing and Sentence Retrieval

In the current QAst 2007 task setting natural language questions are specified in written form. For this reason we were able to integrate the question parser from our textual QA–system into QAST-V1. The question parser computes for each question a syntactic dependency tree (which also contains recognized named entities) and semantic information like question type, the expected answer type, and the question focus, cf. [3] for details.

In a second step the result of the question parser is mapped to an ordered set of alternative IR–queries following the same approach as in our textual QA system, cf. [3].

---

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

3   Results and Discussion

We took part in the tasks:

– T1: Question-Answering in manual transcriptions of lectures;
– T2: Question-Answering in automatic transcriptions of lectures;

In both cases the CHIL corpus was used which was adapted by the organizers for the QAst 2007 track. It consists of around 25 hours (around 1 hour per lecture) both manually and automatically transcribed. The language is European English, mostly spoken by non–native speakers.

We submitted only one run to each task and the table below shows the results we obtained:

Run	task	Questions returned (#) [98]	Correct answers (#)	MRR	Accuracy
dfki1_t1	T1	98	19	0.17	0.15
dfki1_t2	T2	98	9	0.09	0.09

where MRR is the Mean Reciprocal Rank that measures how well ranked is the right answer in the list of 5 possible answers in average. Accuracy is the fraction of correct answers ranked in the first position in the list of 5 possible answers.

The currently low number of returned correct answers has two main error sources. On the one hand side, the coverage and quality of the named entity recognizers are low. This is probably due to the fact that we used the languages models that were created from written texts. One possible solution is to improve the corpus preprocessing step, especially the sentence splitter and the repairment of errors like word repetition. Another possible source of improvement is the development of annotated training corpus of speech transcripts for named entities. Both activities surely demand further research and resources.

On the other hand side, the performance of the answer extraction process strongly depends on the coverage and quality of the question analysis tool. We will improve this by extending the current coverage of the English Wh–grammar, especially by extending the mapping of general verbs and nouns to corresponding expected answer types and by exploiting strategies that validate the semantic type consistency between the relevant nouns and verbs of a question.

References

1. Collins, M., Singer, Y.: Unsupervised models for named entity classification (1999)
2. Hatcher, E., Gospodnetic, O.: Lucene in Action. In Action series. Manning Publications Co., Greenwich (2004)
3. Neumann, G., Sacaleanu, S.: Experiments on robust nl question interpretation and multi-layered document annotation for a cross-language question/answering system. In: Peters, C., Clough, P., Gonzalo, J., Jones, G.J.F., Kluck, M., Magnini, B. (eds.) CLEF 2004. LNCS, vol. 3491, pp. 411–422. Springer, Heidelberg (2005)
4. Sacaleanu, B., Neumann, G.: Dfki-lt at the CLEF 2006 multiple language question answering track. In: Working notes of CLEF 2006, Alicante, Spain (August 2006)