Robust Question Answering for Speech Transcripts: UPC Experience in QAst 2009

Pere R. Comas and Jordi Turmo
TALP Research Center
Technical University of Catalonia (UPC)
{pcomas,turmo}@lsi.upc.edu

Abstract

This paper describes the participation of the Technical University of Catalonia in the CLEF 2009 Question Answering on Speech Transcripts track. We have participated in the English and Spanish scenarios of QAst. For both manual and automatic transcripts we have used a robust factual Question Answering that uses minimal syntactic information. We have also developed a NERC designed to handle automatic transcripts. We perform a detailed analysis of our results and draw conclusions relating QA performance to word error rate and the difference between written and spoken questions.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.1 Content Analysis and Indexing; H.3.3 Information Search and Retrieval; H.3.4 Systems and Software;; H.2.3 [**Database Managment**]: Languages—*Query Languages*

General Terms

Measurement, Performance, Experimentation

Keywords

Question Answering, Spoken Document Retrieval, Oral Question Answering

1 Introduction

The CLEF 2009 Question Answering on Speech Transcripts (QAst) track [7] consists of four Question Answering (QA) tasks for three different languages: T1 English, T2 Spanish and T3 French. Task *m* is QA in manual transcripts of recorded European Parliament Plenary Sessions (EPPS). Tasks *a*, *b*, and *c*, use three different transcripts of the recorded audio using three Automatic Speech Recognizers (ASR). This transcriptions have an increasing percentage of errors. There are two sets of questions for each language: set B contains oral questions spontaneously asked by several human speakers, while set A consists of grammatically corrected transcriptions of the questions in set B. The questions are divided in two sets of development (50 questions) and test (100 questions). Given the languages, questions and transcripts, there is a total of 24 possible scenarios in the QAst evaluation. For example, we will refer as T2B-*a* to the scenario taking the best automatic transcripts of the Spanish EPPS using spontaneous questions. The automatic transcripts have different levels of word error rate (WER). WERs for T1 are 10.6%, 14%, and

T1-*m*: *"Abidjan is going going the way of Kinshasa Kinshasa which was of course a country in the past with skyscrapers and boulevards and now a country a city in ruins"*

T1-*a*: "average down is going to go in the way of Kinshasa other at Kinshasa which was of course a country in the past of skyscrapers and poorer parts and our country as a city in ruins"

Figure 1: Sample of manual and automatic transcripts.

24.1%. For T2 WERs are 11.5%, 12.7% and 13.7%. Figure 1 shows a text sample extracted from T1 corpus.

This paper summarizes our methods and results in QAst. We have participated in scenarios T1 and T2 with all transcripts and question sets. Our QA system is based on our previous work in [1] and [6]. We have used the same system architecture for all the tasks, having interchangeable language–dependant parts and different passage retrieval algorithms for automatic transcripts.

2 Overview of the System Architecture

The architecture of our QA system follows a commonly-used schema which splits the process into three phases performed in a sequential pipeline: Question Processing (QP), Passage Retrieval (PR), and Answer Extraction (AE) This QA system is designed to answer to factoid questions, those whose answer is a named entity (NE).

2.1 Question Processing and Classification (QC)

The main goal of this component is to detect the type of the expected answer. We currently recognize the 53 open-domain answer types from [4]. The answer types are extracted using a multi-class Perceptron classifier and a rich set of lexical, semantic and syntactic features. This classifier obtains an accuracy of 88% on the corpus of [4]. Additionally, the QP component extracts and ranks relevant keywords from the question

For scenario T2, he have developed an Spanish question classifier using human translated questions from the corpus of [4] following the same machine learning approach. This classifier obtains an accuracy of 74%.

2.2 Passage Retrieval (PR)

This component retrieves a set of relevant passages from the document collection, given the previously extracted question keywords. The PR algorithm uses a query relaxation procedure that iteratively adjusts the number of keywords used for retrieval and their proximity until the quality of the recovered information is satisfactory (see [6]). In each iteration a Document Retrieval application (IR engine) fetches the documents relevant for the current query and a subsequent passage construction module builds passages as segments where two consecutive keyword occurrences are separated by at most t words.

When dealing with automatic transcripts, the incorrectly transcribed words may create a problem of word recognition to the IR engine, introducing false positives and false negatives to its input.

To overcome such drawbacks, we have used an IR engine relying on phonetic similarity for the automatic transcripts. This tool is called PHAST (after PHonetic Alignment Search Tool) and uses pattern matching algorithms to search for small sequences of phones (the keywords) into a larger sequence (the documents) using a measure of sound similarity. Then the PR algorithm may be applied to the words found with PHAST. A detailed description of PHAST can be found in [2].

T1: English					T2: Spanish				
Set	WER	Precision	Recall	$F_{\beta=1}$	Set	WER	Precision	Recall	$F_{\beta=1}$
m	$\sim 0\%$	70.63%	68.19%	69.39	m	$\sim 0\%$	76.11%	71.19%	73.57
a	10.6%	63.57%	55.26%	59.13	a	11.5%	72.40%	62.03%	66.81
b	14%	61.51%	52.28%	56.52	b	12.7%	64.33%	55.95%	59.85
c	24.1%	58.62%	43.92%	50.22	c	13.7%	67.61%	55.60%	61.02

Table 1: NERC performance

2.3 Answer Extraction (AE)

Identifies the exact answer to the given question within the retrieved passages. First, answer candidates are identified as the set of named entities (NEs) that occur in these passages and have the same type as the answer type detected by QP. Then, these candidates are ranked using a scoring function based on a set of heuristics that measure keyword distance and density [5]. These heuristic measures use approximated matching for AE in automatic transcripts as shown in the passage retrieval module from the previous section. The same measure is used for T1 and T2.

3 Named Entity Recognition and Classification (NERC)

As described before, we extract candidate answers from the NEs that occur in the passages retrieved by the PR component. We detail below the strategies used for NERC in both manual and automatic transcripts.

We have taken a machine learning approach to this problem. First we apply learning at word level to identify NE candidates using a BIO tagging scheme. Then these candidates are classified into NE categories. Each function is modeled with *voted perceptron* [3]. As learning data we have manually labeled the NEs that occur in the QAst corpora T1 and T2 with their types (i.e. date, location, number, organization, person and time).

Our NERC uses a rich set of lexical and syntactic features which are standard to state-of-the-art NERCs. This features include: words, lemmas, POS tags, word affixes, flags regarding presence of numerals and capitalization, use of gazetteers, and n-grams of this features within a certain window of words. New features specially designed for automatic transcripts have been added to the sets *a*, *b* and *c*. These features use phonetic transcription of the words:

- Prefixes and suffixes of phones.
- Phonetic similarity with words in the gazetteer.
- A clustering of the transcriptions of the words has been done by grouping words with similar pronunciation. This clustering reduces the sparseness of the word–based features by mapping the words in several smaller subsets of different coarseness.
- Features capturing the possibility of splitting or merging adjacent words due to ASR recognition errors.

The addition of this phonetic features improves the results by no more than 2 points of $F_{\beta=1}$ score in datasets *a*, *b* and *c*.

Given that all there is no specific datasets for development, and we don't have more automatic transcripts of EPPS data, it is not possible to train our NERC in a dataset other than the test set. Therefore we have relabeled both corpora through a process of cross–validation. Both corpora have been randomly split in 5 segments, a NERC model has been learned for all subsets of 4 segments and the remaining segment has been labeled using this model. Thus we can train our NERC with documents from the same domain but test it on unseen data.

Task, System	#Q	MRR	TOP1	TOP5	Task, System	#Q	MRR	TOP1	TOP5
T1A-*m* 1	75	0.27	14	32	T1A-*m* 2	75	0.31	17	35
T1B-*m* 1	75	0.15	7	19	T1B-*m* 2	75	0.15	7	18
T1A-*a* 1	75	0.27	14	29	T1A-*a* 2	75	0.26	14	30
T1B-*a* 1	75	0.08	4	11	T1B-*a* 2	75	0.09	4	12
T1A-*b* 1	75	0.24	13	26	T1A-*b* 2	75	0.26	15	29
T1B-*b* 1	75	0.08	3	11	T1B-*b* 2	75	0.08	3	12
T1A-*c* 1	75	0.21	12	22	T1A-*c* 2	75	0.24	13	26
T1B-*c* 1	75	0.08	4	10	T1B-*c* 2	75	0.08	3	11

Table 2: Overall factoid results for our sixteen English runs.

Task, System	#Q	MRR	TOP1	TOP5	Task, System	#Q	MRR	TOP1	TOP5
T2A-*m* 1	44	0.24	7	16	T2A-*m* 2	44	0.29	8	20
T2B-*m* 1	44	0.34	8	20	T2B-*m* 2	44	0.33	12	20
T2A-*a* 1	44	0.15	6	8	T2A-*a* 2	44	0.20	8	12
T2B-*a* 1	44	0.14	6	6	T2B-*a* 2	44	0.24	10	12
T2A-*b* 1	44	0.18	6	12	T2A-*b* 2	44	0.20	7	13
T2B-*b* 1	44	0.20	7	12	T2B-*b* 2	44	0.20	7	12
T2A-*c* 1	44	0.22	9	12	T2A-*c* 2	44	0.20	8	11
T2B-*c* 1	44	0.13	5	8	T2B-*c* 2	44	0.21	9	10

Table 3: Overall factoid results for our sixteen Spanish runs.

Table 3 shows the overall results of our NERC for the four datasets of both T1 and T2. As the transcript WER increases, the scores consequently drop.[1]

4 Experimental Results

UPC participated in 2 of the 3 scenarios, English (T1) and Spanish (T2) ones. We submitted two runs for each task, run number 1 uses the standard NERC described in Section 3 and run number 2 uses the hand–annotated NEs. Each scenario included 100 test questions, from which 20 do not have an answer in the corpora (these are *nil* questions). In T1 75 question are factoids for 44 in T2. Our QA system is designed to answer only factual questions, therefore the our experimental analysis will refer only to factual questions.

We report two measures: (a) TOPk, which assigns to a question a score of 1 only if the system provided a correct answer in the top k returned; and (b) Mean Reciprocal Rank (MRR), which assigns to a question a score of $1/k$, where k is the position of the correct answer, or 0 if no correct answer is found. The official evaluation of QAst 2008 uses TOP1 and TOP5 measures [7]. An answer is considered correct by the human evaluators if it contains the complete answer and nothing more, and it is supported by the corresponding document. If an answer was incomplete or it included more information than necessary or the document did not provide the justification for the answer, the answer was considered incorrect.

Tables 2 and 3 summarizes our overall results for factual questions in English and Spanish. It shows MRR, TOP1 and TOP5 scores for each track and run as defined previously.

Table 4 contains a statistical error analysis of our system covering the QC, PR and AE parts. It deals only with factoid questions with non–nil answers. The meaning of each column is the following. Q: number of factual question. QC: number of questions with answer type correctly detected by QP. PR: number of question where at least on passage with the correct answer war retrieved. QC&PR: number of questions with correct answer type and correct passage retrieval.

[1] Consider manual transcripts $m$having a WER of almost 0.

Track	Run	Q	QC	PR	QC&PR	C.NE	TOP5 non-Null	TOP1 non-Null
T1A-*m*	1	64	49	50	41	41	28	11
T1A-*a*	1	64	49	48	38	38	26	11
T1A-*b*	1	64	49	46	34	29	23	11
T1A-*c*	1	64	49	40	30	30	20	10
Avg. Loss			-23%	-28%	-44%	-3%	-29%	-55%
T1B-*m*	1	64	22	49	18	18	18	6
T1B-*a*	1	64	22	50	17	17	11	4
T1B-*b*	1	64	22	49	17	17	11	3
T1B-*c*	1	64	22	45	15	15	10	4
Avg. Loss			-65%	-24%	-73%	0%	-25%	-66%
T1A-*m*	2	64	49	50	41	41	28	12
T1A-*a*	2	64	49	48	38	38	27	12
T1A-*b*	2	64	49	46	34	34	26	13
T1A-*c*	2	64	49	40	30	30	23	11
Avg. Loss			-23%	-28%	-44%	0%	-27%	-53%
T1B-*m*	2	64	22	49	18	18	17	6
T1B-*a*	2	64	22	50	17	11	11	4
T1B-*b*	2	64	22	49	17	17	12	3
T1B-*c*	2	64	22	45	15	15	10	3
Avg. Loss			-65%	-24%	-74%	0%	-25%	-68%
T2A-*m*	1	44	39	35	30	24	10	3
T2A-*b*	1	44	39	33	28	21	11	5
T2A-*a*	1	44	39	36	31	24	9	5
T2A-*c*	1	44	39	36	31	24	11	8
Avg. Loss			-11%	-20%	-31%	-22%	-56%	-48%
T2B-*m*	1	44	27	36	19	19	10	4
T2B-*a*	1	44	27	36	19	19	6	4
T2B-*b*	1	44	27	33	17	16	9	4
T2B-*c*	1	44	27	36	20	18	6	3
Avg. Loss			-38%	-19%	-46%	-4%	-57%	-51%
T2A-*m*	2	44	39	35	30	27	15	5
T2A-*a*	2	44	39	36	31	28	12	7
T2A-*b*	2	44	39	33	28	26	13	6
T2A-*c*	2	44	39	36	31	29	11	8
Avg. Loss			-11%	-20%	-31%	-8%	-53%	-49%
T2B-*m*	2	44	27	36	19	19	13	7
T2B-*a*	2	44	27	36	19	19	11	8
T2B-*b*	2	44	27	33	17	17	9	5
T2B-*c*	2	44	27	36	20	20	8	6
Avg. Loss			-38%	-19%	-46%	-0%	-45%	-36%

Table 4: Error analysis of the QA system components.

C.NE: number of questions where the retrieved passages contain the correct answer tagged as a NE of the right type (specified by the QC module), so it is a candidate answer for the AE module. TOP5 non-nil: number of question with non-nil answer correctly answered by our system in the TOP5 candidates. There is an "Avg. Loss" row for each task and run that shows the performance loss (averaged in all transcripts) introduced by each module in relation to the previous step. Note that this numbers have been gathered using an automatic application and some disagreement between in the selection of factoid questions may exist, therefor the TOP5 scores in this table may differ slightly from the official QAst scores.

In Table 2 we can see that moving from transcript m to a implies a loss of 10 points in TOP5

score for T1. For T2 this loss is as much as 50 points. But subsequent increases of WER in transcripts b and c have a low impact in our performance. According to the QAst 2009 overview paper [7], the incidence of WER rates in our system is less severe than in other participants but our initial results in m track are also lower.

We should note an important loss of performance in T1 scenario when using the question set B. Table 2 shows that TOP5 and TOP1 scores decrease by 50% or more when compared to question set A. Table 4 shows that the number of correctly classified questions drops from 49 to 22 in T1 (44% of the original), therefore QA&PR drops about 30%. The AE module has comparable performance with both sets thus this poor performance is due solely to our QC module. In scenario T2 there is a smaller loss in QC (just 30%) and it has different repercussion. The specific distribution of errors among QC and PR leads a higher QC&PR count in all T2 tracks than in T1 tracks, although T1 has 20 more questions than T2, and this yields a smaller performance drop in T2. This must be blamed on our machine learning question classifier. Although this is based on shallow textual analysis, the model doesn't generalize well to spontaneous questions. Probably it has an strong dependency on the usual well–formed question structure. Additionally, we note that the classification of written questions is better for T2 question set than T1 question set. This suggests that in this evaluation T1 questions are more domain specific than the others.

The difference between runs number 1 and 2 is that number 2 uses hand–tagged NEs instead of our automatic NERC. The results show that it has little impact on performance. In Table 4 we can see that most of the correct answers retrieved by our PR module are annotated with the correct entity. It is shown by the small difference between QC&PR and C.NE columns. Using hand-tagged NEs improves slightly the results for TOP5 and TOP1, probably because it filters out incorrect candidates and the AE process becomes easier. As we have seen in Table 3, the $F_{\beta=1}$ score of our NERC models is below 70% but this poor performance doesn't reflect in the final QA results. We think that hand-tagged NEs doesn't improve the results due to two facts. On one hand, the NERC we have developed is useful enough for this task even having poor $F_{\beta=1}$ scores, and on the other hand, there is a probable disagreement between the humans who tagged the NEs and the humans who wrote the questions.

One of the main issues of our QAst 2009 system is the poor performance of the AE module. More than 25% of the correctly retrieved and tagged answers aren't correctly extracted in T1, and more than 50% are lost in T2. In fact, AE is the main source of errors in T2, more than the combination of both PR and QC. This is a big difference with the results achieved in 2008 evaluation, where AE was of high accuracy. It shows that our answer selecting heuristics may be more domain–dependant than we knew and they should be tuned for this task.

5 Conclusions

This paper describes UPC's participation in the CLEF 2009 Question Answering on Speech Transcripts track. We submitted runs for all English and Spanish scenarios. In this evaluation we analyzed the impact of using *gold–standard* NEs with using a far from perfect NERC.

We have developed a new NERC designed for speech transcripts that shows results competitive with *gold-standard* NEs when used in Question Answering.

The results achieved in the different scenarios and tasks are not the top ones. But there is little degradation due to ASR effects thus showing that our QA system is highly robust to transcript errors, being this one of the main focuses of the QAst evaluation.

Acknowledgements

This work has been partially funded by the Spanish Ministry of Science and Technology (TEXTMESS project).

References

[1] P.R. Comas and J. Turmo. Robust question answering for speech transcripts: Upc experience in qast 2009. *Proceedings of the CLEF 2008 Workshop on Cross-Language Information Retrieval and Evaluation*, 2008.

[2] P.R. Comas and J. Turmo. Spoken document retrieval based on approximated sequence alignment. *11th International Conference on Text, Speech and Dialogue (TSD)*, 2008.

[3] Y. Freund and R. Schapire. Large margin classification using the perceptron algorithm. In *COLT' 98: Proceedings of the eleventh annual conference on Computational learning theory*, 1998.

[4] X. Li and D. Roth. Learning question classifiers: The role of semantic information. *Journal of Natural Language Engineering*, 2005.

[5] M. Pacsca. *High-performance, open-domain question answering from large text collections.* PhD thesis, Southern Methodist University, Dallas, TX, 2001.

[6] M. Surdeanu, D. Dominguez-Sal, and P.R. Comas. Design and performance analysis of a factoid question answering system for spontaneous speech transcriptions. *Proceedings of the International Conference on Spoken Language Processing (INTERSPEECH 2006)*, 2006.

[7] J. Turmo, P.R. Comas, S. Rosset, O. Galibert, N. Moreau, D. Mostefa, P. Rosso, and D. Buscaldi. Overview of QAST 2009. *Proceedings of the CLEF 2009 Workshop on Cross-Language Information Retrieval and Evaluation*, 2009.