# Robust Question Answering for Speech Transcripts: UPC Experience in QAst 2008

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 2008 Question Answering on Speech Transcripts track. We have participated in the English and Spanish scenarios of QAst. For the processing of manual transcripts we have deployed a robust factual Question Answering that uses minimal syntactic information. For the handling of automatic transcripts we combine the QA system with a Passage Retrieval and Answer Extraction engine based on a sequence alignment algorithm that searches for "sounds like" sequences. We perform a detailed analysis of our results and draw conclusions relating QA performance to word error rate (WER) in transcripts.

## 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.3.7 Digital Libraries; H.2.3 [**Database Management**]: Languages—*Query Languages*

## General Terms

Measurement, Performance, Experimentation

## Keywords

Question Answering, Spoken Document Retrieval, Phonetic Distance

# 1 Introduction

The CLEF 2008 Question Answering on Speech Transcripts (QAst) track [9] consists of five scenarios with several tasks: Question Answering (QA) in manual transcripts of recorded lectures (T1A) and their corresponding automatic transcripts (T1B), QA in manual transcripts of recorded meetings (T2A) and their corresponding automatic transcripts (T2B), QA in manual transcripts of french European Parliament Sessions (T3A) and three different automatic transcripts (T3B-A, T3B-B, T3B-C), QA in manual transcripts of English European Parliament Sessions (T4A) and three different automatic transcripts (T4B-A, T4B-B, T4B-C), QA in manual transcripts of Spanish European Parliament Sessions (T5A) and three different automatic transcripts (T5B-A, T5B-B, T5B-C). The automatic transcripts for tasks T3, T4 and T5 have different levels of word error rate (WER). WERs for T4 are 10.6%, 14%, and 24.1%. For T5 WERs are 11.5%, 12.7% and 13.7%. This paper summarizes our methods and results in QAst. We have participated in all the scenarios except the french language one (T3).

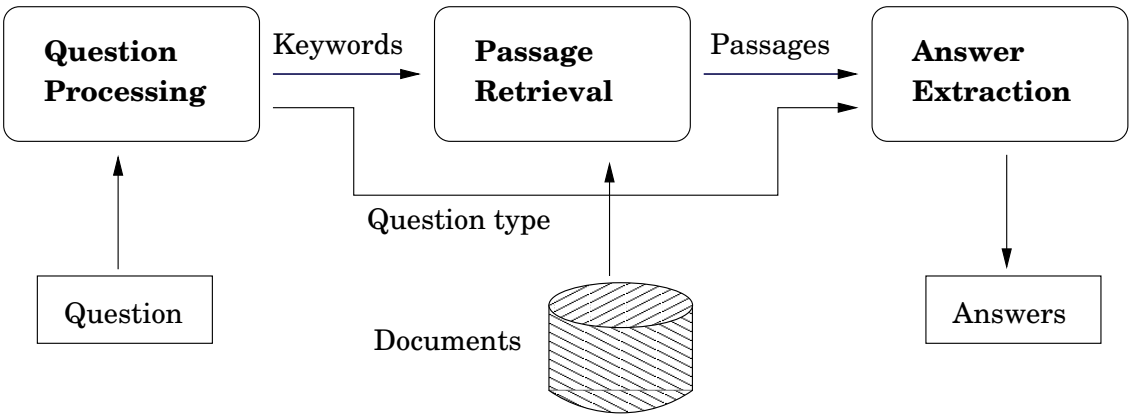


Figure 1: Overview of QA architecture

Our QA system is based on our previous work in [4, 7] and [8]. 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 sequentially: Question Processing (QP), Passage Retrieval (PR), and Answer Extraction (AE), as shows Figure 1. These three phases are described in the following sections.

### 2.1 Question Processing and Classification

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 [5] plus 3 types specific to QAst corpora (i.e., `system/method`, `shape`, and `material`). 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 [5]. Additionally, the QP component extracts and ranks relevant keywords from the question

For scenario T5, he have developed an Spanish question classifier using human translated questions from the corpus of [5] following the same machine learning approach. This classifier obtains an accuracy of 74%.

### 2.2 Passage Retrieval

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 [7]). In each iteration a Document Retrieval application (Lucene 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, you have to bear in mind that the state of the art in ASR technology is far from perfect. For example, the word error rate (WER) of the meetings automatic transcripts (T1B) is around 38% and the WER of the lectures (T2B) is over 20%, and from 10.6% to 24.1% for the T4B transcripts. Figure 2 shows three real examples of common errors when generating automatic transcripts. From the point of view of passage retrieval, imperfect

1M: *"The pattern frequency relevance rate indicates the ratio of relevant documents..."*
1A: *"the putt and frequency illustrating the case the ratio of relevant documents..."*
2M: *"The host system it is a UNIX Sun workstation"*
2A: *"that of system it is a unique set some workstation"*
3M: *"Documents must be separated into relevant documents and irrelevant documents by a manual process, which is very time consuming."*
3A: *"documents must be separated into relevant documents and in relevant document by a manual process witches' of very time consuming"*

Figure 2: Examples of manual (M) and automatic (A) transcripts.

transcripts create a new problem of incorrectly transcribed words that yield false positives and false negative for traditional search methods.

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. A detailed description of PHAST can be found in [3].

### 2.3 Answer Extraction

Identifies the exact answer to the given question within the retrieved passages. First, answer candidates are identified as the set of 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[6]. 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 English and Spanish scenarios.

## 3 Named Entity Recognition and Classification

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.

**NERC for English Manual Transcripts.**

We have used a modified version of the NERC presented in [8]. One change from the previous system is that it uses multi-class Perceptron instead of the existing SVM classifiers. As training data we annotated the NEs that occur in the QAst development corpus with their types (i.e., person, organization, location, language, measure, system/method and time) and used an 80–20% corpus split for training and testing for both lectures and meetings corpora. This experiment indicated that the development data is sufficient for good generalization for meetings (a $F_1$ score of +75 points in the development test partition) but it is insufficient in lectures: 33 points. This is most likely caused by the small size of the development corpus and the large number of topics addressed. To compensate for the insufficient training data we perform a combination of several NERC models for this task. We merged the outputs of: (a) a rule-based NERC developed previously [7], (b) the NERC trained on the existing development data, and (c) the NERC trained on the CoNLL English corpus.[1] We used the above priority ordering for conflict resolution in case of overlapping assignments (e.g., lectures model has higher priority than the CoNLL model). After model combination the NERC $F_1$ score in the development test partition did not improve but the recall did increase, so we decided to use this combination strategy in the testing since recall is paramount for QA

[1] http://cnts.ua.ac.be/conll2002/ner

**NERC for English Automatic Transcripts.** We have used a similar framework for the processing of automatic transcripts: we annotated the development corpora and trained specific NERC models for lectures and meetings. The significant difference is that here we expand the classifiers' feature sets with phonetic attributes. These features are motivated by the fact that even when the ASR incorrectly transcribes NEs the phonetic structure is by and large maintained in the transcript (e.g. in Figure 2 the name *"Sun"* is recognized as *"some"*). We used an unsupervised hierarchical clustering algorithm that groups tokens based on the similarity of their phonetic sequences. The stop condition of the algorithm is set to reach a local maximum of the Calinski criterion [1]. Then the cluster of each token is added as a feature (e.g. *"Sun"* and *"some"* share the same cluster), which helps the NERC model generalize from the correct to the incorrect transcript. We also added phonetic features that model prefix and suffix similarity.

**NERC for Spanish.** For the Spanish track T5 we have used a previously developed NERC. It uses a machine learning approach and it has been trained with the CoNLL Spanish corpus. See details in [2]. Unfortunately, this NERC can recognize only person, location and organization NE types. Thus only this types can be used as answer candidates. It supposes a serious shortcoming for QA performance as the results show in Section 4.

# 4 Experimental Results

UPC participated in 4 of the 5 scenarios, all but the French one (T3). We submitted two runs for the tasks on automatic transcripts, one using run using the standard QA system for written text ($QA_m$) and another run using the system tailored for automatic transcripts ($QA_a$). See section 2 for the differences between both. Each scenario included 100 test questions, from which 10 does not have an answer in the corpora (these are *nil* questions). Around 75% of the questions are of factual types and around 25% are definitional. 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) TOP$k$, 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 []. 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.

Table 1 summarizes our overall results for factual question only. The cost of moving from manual transcripts to automatic transcripts (i.e., the difference between TXA and TXB) is a loss in TOP1 score of at last 10% for T1, 43% for T2, 50% for T4 and 42% for T5. The performance of $QA_a$ is very similar to $QA_m$. As shown in QAst 2008 Overview paper [9], UPC has ranked among the top teams in tasks T1, T2 and T4. Our team got the best TOP1 score in T1B, T2B and TA4 tracks, although the differences were not significant. For task T5 our results were far beyond other participants.

Table 2 shows the distribution of correct answers for all tasks according to the answer type. In scenario T4, a design error prevented our NERC from recognizing entity types `Sha`, `Mat` and `Col`. Therefor there are 20 unanswerable questions from the 78 factual ones. Our system for the Spanish scenario (T5) is limited to answer types `Org`, `Per`, and `Loc`, so the real upper bound for factual questions is 36 instead of 75.

Finally, Table 3 summarizes the error analysis of QP, PR, and AE parts. 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. C.NE: number of questions where the retrieved passages contain the correct answer tagged as a NE of the right type. U.NE: number of questions where the retrieved passages contain the correct answer but it remains undetected by the NERC. Er.NE: number of questions where the retrieved passages contain the correct answer tagged as a NE with an incorrect type.

| Task, System | #Q | MRR | TOP1 | TOP5 | Task, System | #Q | MRR | TOP1 | TOP5 |
|---|---|---|---|---|---|---|---|---|---|
| T1A, $QA_m$ | 78 | 0.44 | 30 | 39 | T2A, $QA_m$ | 74 | 0.35 | 23 | 29 |
| T1B, $QA_m$ | 78 | 0.39 | 27 | 35 | T2B, $QA_m$ | 74 | 0.20 | 13 | 19 |
| T1B, $QA_a$ | 78 | 0.37 | 26 | 35 | T2B, $QA_a$ | 74 | 0.16 | 8 | 16 |
| T4A, $QA_m$ | 75 | 0.44 | 30 | 38 | T5A, $QA_m$ | 75 | 0.11 | 7 | 9 |
| T4B A, $QA_m$ | 75 | 0.22 | 15 | 18 | T5B A, $QA_m$ | 75 | 0.05 | 3 | 5 |
| T4B B, $QA_m$ | 75 | 0.18 | 12 | 15 | T5B B, $QA_m$ | 75 | 0.06 | 4 | 5 |
| T4B C, $QA_m$ | 75 | 0.11 | 7 | 11 | T5B C, $QA_m$ | 75 | 0.03 | 2 | 2 |
| T4B A, $QA_a$ | 75 | 0.16 | 10 | 16 | T5B A, $QA_a$ | 75 | 0.06 | 4 | 5 |
| T4B B, $QA_a$ | 75 | 0.16 | 10 | 14 | T5B B, $QA_a$ | 75 | 0.06 | 4 | 5 |
| T4B C, $QA_a$ | 75 | 0.11 | 6 | 11 | T5B C, $QA_a$ | 75 | 0.03 | 2 | 3 |

Table 1: Overall results for our twenty QAst runs.

| Task, System | Org | Per | Loc | Tim | Mea | Sys | Lan | Sha | Mat | Col | Def |
|---|---|---|---|---|---|---|---|---|---|---|---|
| T1A, $QA_m$ | 4/8 | 8/9 | 1/2 | 3/5 | 13/19 | 4/5 | 6/10 | 0/8 | 0/3 | 0/9 | 4/22 |
| T1B, $QA_m$ | 3/8 | 5/9 | 1/2 | 3/5 | 13/19 | 4/5 | 6/10 | 0/8 | 0/3 | 0/9 | 4/22 |
| T1B, $QA_a$ | 4/8 | 4/9 | 2/2 | 2/5 | 14/19 | 3/5 | 6/10 | 0/8 | 0/3 | 0/9 | 4/22 |
| T2A, $QA_m$ | 1/8 | 2/8 | 7/10 | 1/8 | 4/10 | 3/6 | 2/8 | 1/4 | 4/6 | 4/6 | 3/26 |
| T2B, $QA_m$ | 3/8 | 2/8 | 1/10 | 0/8 | 3/10 | 1/8 | 1/8 | 1/4 | 4/6 | 3/6 | 5/26 |
| T2B, $QA_a$ | 1/8 | 3/8 | 2/10 | 0/8 | 1/10 | 1/8 | 1/8 | 1/4 | 4/6 | 2/6 | 6/26 |
| T4A, $QA_m$ | 7/14 | 9/14 | 6/15 | 9/15 | 7/15 | 0/2 | - | - | - | - | 4/25 |
| T4B-A, $QA_m$ | 1/14 | 0/14 | 3/15 | 8/15 | 6/15 | 0/2 | - | - | - | - | 4/25 |
| T4B-B, $QA_m$ | 1/14 | 0/14 | 2/15 | 8/15 | 4/15 | 0/2 | - | - | - | - | 5/25 |
| T4B-C, $QA_m$ | 0/14 | 1/14 | 2/15 | 1/15 | 6/15 | 1/2 | - | - | - | - | 4/25 |
| T4B-A, $QA_a$ | 1/14 | 0/14 | 2/15 | 7/15 | 6/15 | 0/2 | - | - | - | - | 4/25 |
| T4B-B, $QA_a$ | 1/14 | 0/14 | 1/15 | 8/15 | 4/15 | 0/2 | - | - | - | - | 5/25 |
| T4B-C, $QA_a$ | 0/14 | 1/14 | 2/15 | 1/15 | 6/15 | 1/2 | - | - | - | - | 4/25 |
| T5A, $QA_m$ | 1/10 | 8/21 | 0/5 | 0/25 | 0/14 | - | - | - | - | - | 3/25 |
| T5B-A, $QA_m$ | 1/10 | 3/21 | 1/5 | 0/25 | 0/14 | - | - | - | - | - | 0/25 |
| T5B-B, $QA_m$ | 2/10 | 2/21 | 0/5 | 0/25 | 0/14 | - | - | - | - | - | 0/25 |
| T5B-C, $QA_m$ | 0/10 | 3/21 | 0/5 | 0/25 | 0/14 | - | - | - | - | - | 2/25 |
| T5B-A, $QA_a$ | 1/10 | 3/21 | 1/5 | 0/25 | 0/14 | - | - | - | - | - | 0/25 |
| T5B-B, $QA_a$ | 2/10 | 3/21 | 0/5 | 0/25 | 0/14 | - | - | - | - | - | 2/25 |
| T5B-C, $QA_a$ | 0/10 | 2/21 | 0/5 | 0/25 | 0/14 | - | - | - | - | - | 1/25 |

Table 2: Distribution of correct answers (TOP5) according to answer type. Org = organization, Per = person, Tim = time, Mea = measure, Met/Sys = method/system, Mat = material, Col = color, Def = definitional.

QC&PR: number of questions with correct answer type and correct passage retrieval. QC&NE: number of questions with correct answer type and correctly tagged answer in the passages. TOP5 non-nil: number of question with non-nil answer correctly answered by our system in the TOP5 candidates. Due to technical reasons this analysis has not been performed on task T2B.

We can draw several important observations from this error analysis: Question classification performs better for T1 question set than T2 and T4 question sets. This suggests that in this evaluation T1 questions were more domain specific than the others. In T5, results are really disappointing and this suggests that our Spanish classifier may be too domain dependant since it achieves 74% accuracy in our test data. "PR" is specially degraded in task T4B-C, where we processed automatic transcripts with the highest WER (24.1%). This proves that passage retrieval is indeed affected by a high WER but is robust enough to be used with a *good* ASR. Passage retrieval using PHAST performed better than the passage retrieval with classical retrieval for tasks in T5 and worse for tasks in T4. Since both scenarios have similar domain, we think this difference is due to the nature of Spanish and English phonology. Further experiments in [3] show consistently that passage retrieval in Spanish is improved by using PHAST. As the table shows, the bad performance of NERC is the critical problem of our QA system. The difference between "C.NE" and "PR" values is much bigger than between "PR" and "Q", thus the theoretical upper

| Track | System | Q | QC | PR | C. NE | U. NE | Er. NE | QC& PR | QC& NE | TOP5 non-Null |
|---|---|---|---|---|---|---|---|---|---|---|
| T1A | $QA_m$ | 78 | 70 | 69 | 42 | 21 | 6 | 62 | 38 | 37 |
| T1B | $QA_m$ | 78 | 70 | 61 | 39 | 20 | 2 | 55 | 34 | 33 |
| | $QA_a$ | 78 | 70 | 59 | 36 | 22 | 1 | 53 | 33 | 33 |
| T2A | $QA_m$ | 74 | 61 | 46 | 31 | 10 | 5 | 41 | 28 | 29 |
| T4A | $QA_m$ | 75 | 62 | 60 | 41 | 6 | 13 | 60 | 41 | 37 |
| T4B-A | $QA_m$ | 75 | 62 | 56 | 24 | 24 | 8 | 46 | 18 | 17 |
| | $QA_a$ | 75 | 62 | 56 | 24 | 24 | 8 | 44 | 16 | 15 |
| T4B-B | $QA_m$ | 75 | 62 | 55 | 21 | 26 | 8 | 45 | 16 | 14 |
| | $QA_a$ | 75 | 62 | 57 | 21 | 28 | 8 | 46 | 15 | 14 |
| T4B-C | $QA_m$ | 75 | 62 | 52 | 9 | 26 | 17 | 43 | 9 | 7 |
| | $QA_a$ | 75 | 62 | 48 | 10 | 26 | 12 | 36 | 7 | 7 |
| T5A | $QA_m$ | 75 | 18 | 55 | 21 | 31 | 3 | 21 | 8 | 9 |
| T5B-A | $QA_m$ | 75 | 18 | 54 | 14 | 36 | 5 | 5 | 3 | 5 |
| | $QA_a$ | 75 | 18 | 58 | 15 | 38 | 4 | 5 | 3 | 5 |
| T5B-B | $QA_m$ | 75 | 18 | 58 | 14 | 40 | 4 | 6 | 2 | 5 |
| | $QA_a$ | 75 | 18 | 60 | 15 | 39 | 6 | 6 | 2 | 5 |
| T5B-C | $QA_m$ | 75 | 18 | 55 | 12 | 34 | 9 | 3 | 0 | 2 |
| | $QA_a$ | 75 | 18 | 60 | 15 | 41 | 4 | 5 | 0 | 2 |

Table 3: Error analysis of the QA system components.

bound for answer extraction is limited specially by NERC performance. The average number of factual questions in all runs is 75.3, the average value for PR is 56.61 and the average for "C.NE" is 22.44, so in less than 40% of the passages the answer is correctly tagged allowing its correct extraction in the answer extraction step. "QC&NE" is a theoretical upper bound of the total score of each task. We can see that the performance of our answer extraction process is very good since "TOP5" score is very near this upper bound in all tasks. As a remark, all of the scores in T5 are above the upper bound. This is due to the combination of two factors: first, a fall-back mechanism in our answer extraction process to help overcome the `PER/ORG` ambiguity[2] in question classification, this mechanism allows to answer misclassified questions. Second, a double–error situation when the question is misclassified and the answer is erroneously tagged but matches the question type.

The impact of transcription errors in QA can be analyzed in detail thanks to the three different automatic transcripts for task T4B (WERs of T5B have very close values and our overall performance is far too poor for this analysis). Figure 3 shows graphically the values in table 3 for T4, $QA_m$. The yellow bars show the WER percentage for each transcript (0% for manual reference) and the lines show the evolution of variables "PR", "C.NE", "U.NE", "QC&NE" and "TOP5". The performance of passage retrieval decreases linearly with WER increase. The linear regression

$$\mathrm{PR} = 59.78 - 0.33 \cdot \mathrm{WER}$$

fits the data with a Pearson coefficient $r = 0.99$. Other measures such as "C.NE", "QC&NE" and "TOP5" are also strongly related to WER and its diminishment is more pronounced. All this measures decrease the same amount when going from 0% WER to 10.6% WER than from 10.6%% to 24.1%. In fact "C.NE" values fit the non-linear regression curve

$$\mathrm{C.NE} = 43.9 \cdot 0.94^{\mathrm{WER}}$$

with a coefficient $r = 0.97$. Therefore we can conclude that the passage retrieval performance decreases linearly with WER while NERC performance decreases exponentially with WER.

[2] In questions such as "*Who helped solving the packet loss problem?*" is impossible to know if the correct answer is a person name or an organization name. For this question, the answer is the name of a university.

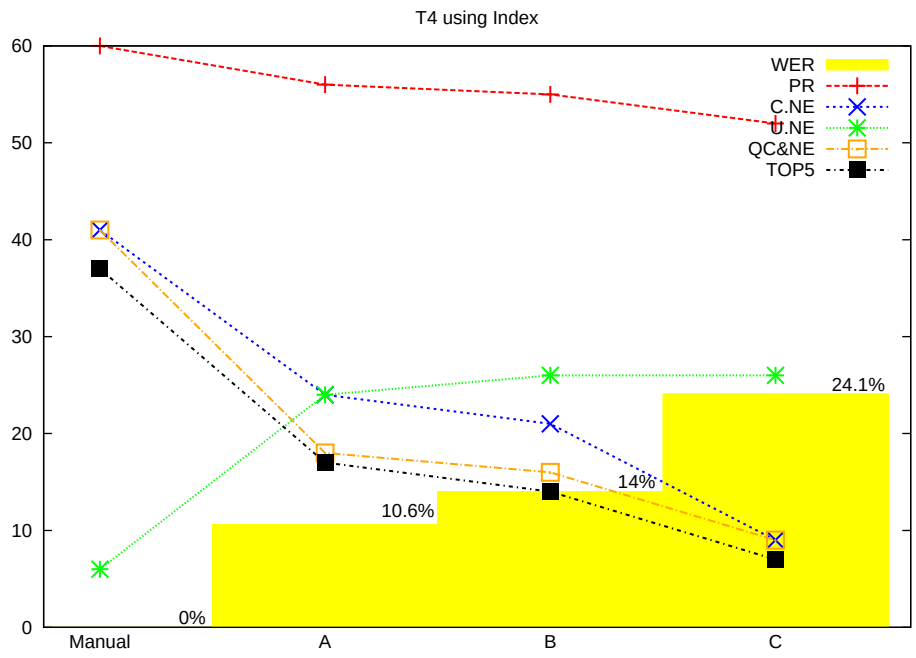


Figure 3: Impact of ASR errors

# 5 Conclusions

This paper describes UPC's participation in the CLEF 2008 Question Answering on Speech Transcripts track. We submitted runs for all English and Spanish scenarios, obtaining the best results in some tasks. In this evaluation we analyzed the behavior of two systems differing in that one is tailored for manual transcripts while the other is tailored for automatic transcripts (uses approximate keyword search based on phonetic distances and a NERC enhanced with phonetic features).

Our approximated keyword search algorithm used for passage retrieval obtains mixed results. It can improve standard search for Spanish but makes little difference for English. We think this because in some document collections it may generated too many false-positive, introducing noise in sets of candidate passages and answers. Nevertheless, we believe that this approach is a good long-term research direction because it can truly address the phenomena specific to automatic transcripts.

Finally, our results show that automatic speech recognition has critical impact on the performance of NERC but its affect on passage retrieval is much less severe.

# Acknowledgements

This work has been partially funded by the European Commission (CHIL, IST-2004-506909) and the Spanish Ministry of Science (TEXTMESS project).

# References

[1] T. Calinski and J. Harabasz. A dendrite method for cluster analysis. *Communications in Statistics*, 3, 1974.

[2] X. Carreras, Ll. Màrquez, and Ll. Padró. Named entity extraction using adaboost. *COLING-02: proceedings of the 6th conference on Natural language learning*, 2002.

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

[4] P.R. Comas, J. Turmo, and M. Surdeanu. Robust question answering for speech transcripts using minimal syntactic analysis. *Proceedings of the CLEF 2007 Workshop on Cross-Language Information Retrieval and Evaluation*, 2007.

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

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

[7] 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.

[8] M. Surdeanu, J. Turmo, and E. Comelles. Named entity recognition from spontaneous open-domain speech. *Proceedings of the International Conference on Spoken Language Processing (INTERSPEECH 2005)*, 2005.

[9] J. Turmo, P.R. Comas, S. Rosset, L. Lamel, N. Moureau, and D. Mostefa and. Overview of QAST 2008. *Proceedings of the CLEF 2008 Workshop on Cross-Language Information Retrieval and Evaluation*, 2008.