Spoken Question Answering

Defensa del DEA i Projecte de Tesi

Pere R. Comas i Umbert

Directors

Jordi Turmo i Lluís Màrquez
Outline

1. Introduction
2. State of the Art
3. Evaluation Framework QAst
4. Our Proposal
5. QAst 2007 results
6. PhD. Thesis Project
7. Publications
Part I

Introduction
Outline of Part I

1. Question Answering
2. Motivation
3. Our Proposal
Question Answering consists in providing short answers to natural language questions.
Question Answering (QA)

- QA is a Text Mining application
  - Information Extraction
  - Automatic summarization
- NLP
- More complex tasks than Information Retrieval
Motivation

- QA technology is focused on written text
- Most human interaction occurs through spontaneous speech
- Speech media could be mined by QA systems
- Spontaneous speech sources brings QA closer to real world
Differences

Lexical/Gramatical differences

- False starts and interrupted sentences
- Repetition of words
  “uh, i don’t know where where the people will be”
- The use of onomatopoeias
  “Rufford, hum, Sanatorium, that’s right”
- Ungrammatical
Differences

Automatic Speech Recognizer (ASR)
- Lack of punctuation marks
- Lack of capitalization
- Recognition errors

“Barcelona” → bars alone
“a study at CMU” → a study it seems you
Example

Q: Who wrote about dynamic programming?

Audio: “We have one paper by Hermann Ney about like uhm yeah dynamic programming ”

ASR: we have one paper {fw} by him and they about like {fw} yeah dynamic programming

What is the correct answer?

- The text pointing to the correct answer in the audio stream
- Even better with timestamps
Q: Who wrote about dynamic programming?

Audio: “We have one paper by Hermann Ney about like uhm yeah dynamic programming ”

ASR: we have one paper {fw} by him and they about like {fw} yeah dynamic programming

What is the correct answer?

The text pointing to the correct answer in the audio stream

Even better with timestamps
Example

Q: Who wrote about dynamic programming?

- Audio: “We have one paper by [Hermann Ney] about like uhm yeah dynamic programming”
- ASR: we have one paper {fw} by him and they about like {fw} yeah dynamic programming

What is the correct answer?

- The text pointing to the correct answer in the audio stream
- Even better with timestamps
Our Proposal

- Robust tools and algorithms for QA on speech transcripts
- Evaluation framework for oral QA
Part II

State of the Art
Outline of Part II

4 Classification of QA systems

5 Question Answering
Several classifications are possible:

- Domain application
- Knowledge used
- Information access method
- Types of question can answer

International QA evaluations such as TREC, CLEF and NTCIR classify systems according to the types of questions.
Question Types

**Factual questions:** Names of entities like persons, organizations, places, dates.

**List questions:** A list of entities such as from factual questions.

**Definition questions:** Definition of some entity or concept. As in “What is ELDA?” or “Who was Mervyn Peake?”

**Interactive questions:** Human/machine interact through a dialog system.
The scope of this thesis project is the factual question answering using spoken documents obtained by automatic transcription of spontaneous speech.
Commonly-used pipeline schema splits the process into three sequential phases:

1. Question Processing (QP)
2. Passage Retrieval (PR)
3. Answer Extraction (AE)

Some kind of data process is carried previously.

- Named Entity Recognition and Classification (NERC)
Spoken Question Answering

- **Question Processing**
  - Question

- **Passage Retrieval**
  - Keywords
  - Question type
  - Documents
  - Passages

- **Answer Extraction**
  - Answers
Classification of QA systems

Spoken Question Answering

- Question Processing
  - Written Question
  - ASR

- Keywords
- Passage Retrieval
  - Question type
  - Transcripts

- Passages
- Answer Extraction
  - Answers
  - Synthesizer
Automatic Speech Recognition

- Search for the most likely word sequence that could produce the signal
- *Acoustic model*: relates signal and phones
- *Language model*: word sequence probability
- Given input signal $A$,

$$\hat{W} = \arg\max_i P(A|W_i)P(W_i)$$

- Recognition errors
- Limited vocabulary
Automatic Speech Recognition

- Search for the most likely word sequence that could produce the signal
- *Acoustic model*: relates signal and phones
- *Language model*: word sequence probability
- Given input signal $A$,

$$\tilde{W} = \arg \max_i P(A|W_i)P(W_i)$$

- Recognition errors
- Limited vocabulary
Oral Question

Voice-Activated Question Answering
[Harabagiu 2002, Tür 2006, Schooten 2007]

- Answering spoken questions
  - Integration of QA and ASR
- Interactive questions
- Iterative refinement of voice recognition
  - Ask again
  - Clarify some parts
- Focusing in time performance

Spoken documents can not be iteratively refined!
Oral Question

Voice-Activated Question Answering
[Harabagiu 2002, Tür 2006, Schooten 2007]

- Answering spoken questions
  - Integration of QA and ASR
- Interactive questions
- Iterative refinement of voice recognition
  - Ask again
  - Clarify some parts
- Focusing in time performance

Spoken documents can not be iteratively refined!
Passage Retrieval

Classically: integration of ASR with IR

- Use of the one-best output [Gorafolo 2000]
- Use of traditional IR ranking algorithms
  - Okapi BM25 [Robertson 1999]
  - Divergence From Randomness [Amati 2002]
  - Vector Space Models [Salton 1988]
- Diverse term detection methods [Inkpen 2005]
  - $n$-gram of words
  - $n$-gram of phones
- Other standard IR techniques (query expansion, relevance feedback...) [Inkpen 2006, Jones 2006]
Although there is a huge state-of-the-art in written QA, there is little bibliography about QA on audio transcripts.
CLT adapted previous system for written text:

- Doesn’t use syntactic or semantic information
- NERC: specific for automatic speech transcripts
  - Hand-crafted regular expressions
  - Name gazetteers
  - Machine learning part
  - It can assign multiple tags to one word
- PR: Retrieval by question/passage word overlap
- AE: NERC confidence score
DFKI adapted previous system for written text:

- Previous automatic sentence splitting
- NERC:
  - Combination of 3 NERCs
  - Statistical, trained on written text
- IR: Question fills IR query pattern
- AE: NE frequency
LIMSI. Designed for transcripts.
- Automatic reconstruction of case and punctuation.
- POS, chunks, NEs
- NERC: Hand-crafted rule-based
- PR: hand-crafted “query descriptor” patterns
- AE: Empirical ranking formula
Tokyo Institute of Technology.

- Non-linguistic QA, generative model
- Based on noisy-channel model
- Parameters specially tuned to corpus
- NERC: none
- IR: $P(Q|D)$
- AE: $P(A|Q) = P(A|W, X)$
Results?

- [Mollá 2007] and [Neumann 2007] adapted systems for written text
- [Rosset 2007] outperforms all the others
Part III

Evaluation Framework QAst
Outline of Part III

6. QAst Evaluation

7. QAst Setting
What is QAst?

- First evaluation of QA over speech transcripts
- In the Cross-Language Evaluation Forum (CLEF)
- Testbed for research in speech transcripts
- Pilot edition in 2007, it will last at least until CLEF 2009
- Organized by UPC, ELDA and LIMSI
- Coordinated by J. Turmo and P. Comas
- [http://www.lsi.upc.edu/~qast](http://www.lsi.upc.edu/~qast)
QAst Objectives

- Evaluate the performance of oral QA
- Motivating the design of novel QA architectures for automatic speech transcripts
- Comparing the performances of manual vs. automatic transcript
- Measuring the effect of ASR inaccuracies
**Question:** A text written question.

**Answer:** The minimal sequence of words that includes the correct exact answer to the question in the audio stream.
Q: Which organization has worked with the University of Karlsruhe on the meeting transcription system?

Manual: uhm this is joint work between the University of Karlsruhe and Carnegie Mellon, so also here in these files you find uh my colleagues and uh Tanja Schultz. Also...

Automatic: {breath} {fw} and this is joint work between University of Karlsruhe and coming around so {fw} all sessions once you find {fw} like only stringent custom film canals communicates on on
Q: Which organization has worked with the University of Karlsruhe on the meeting transcription system?

Manual: uhm this is joint work between the University of Karlsruhe and Carnegie Mellon, so also here in these files you find uh my colleagues and uh Tanja Schultz. Also...

Automatic: {breath} {fw} and this is joint work between University of Karlsruhe and coming around so {fw} all sessions once you find {fw} like only stringent custom film canals communicates on on
Two different resources

- The CHIL corpus\(^1\)
  - 25 lectures of one hour (59kw)
  - WER of 20%

- The AMI corpus\(^2\)
  - 100 hours of meetings (764kw)
  - WER of 38%

\(^1\)http://chil.server.de
\(^2\)http://www.amiproject.org
Four tasks have been defined in QAst

- T1: QA in manual transcriptions of lectures
- T2: QA in automatic transcriptions of lectures
- T3: QA in manual transcripts of meetings
- T4: QA in automatic transcriptions of meetings
Factual questions, answer is a Named Entity

<table>
<thead>
<tr>
<th>Type</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>person</td>
<td>Stefan Kantak, Maria Danninger, Phil</td>
</tr>
<tr>
<td>organization</td>
<td>AT&amp;T, Carnegie Mellon</td>
</tr>
<tr>
<td>location</td>
<td>Dallas, Texas, USA, Geneva</td>
</tr>
<tr>
<td>time</td>
<td>January, thursday twenty-four</td>
</tr>
<tr>
<td>measure</td>
<td>twenty-five point zero percent, ten grams</td>
</tr>
<tr>
<td>method</td>
<td>flexible three clustering,</td>
</tr>
<tr>
<td>system</td>
<td>Emacs, Java</td>
</tr>
<tr>
<td>language</td>
<td>English, French</td>
</tr>
<tr>
<td>color</td>
<td>red, green, blue</td>
</tr>
<tr>
<td>shape</td>
<td>triangle, banana</td>
</tr>
<tr>
<td>material</td>
<td>silver, plastic, fiberglass</td>
</tr>
</tbody>
</table>
Question Sets

- **Development set**
  - T1/T2: 10 seminars and 50 questions
  - T3/T4: 50 meetings and 50 questions

- **Evaluation set**
  - T1/T2: 15 seminars and 100 questions
  - T3/T4: 118 meetings and 100 questions
Evaluation

- Manually judged by human assessors from ELDA (Evaluation and Language resources Distribution Agency)
  - Correct
  - Incorrect
  - Non-exact
  - Unsupported

- Measures
  - Accuracy: answer in first position
  - Mean Reciprocal Rank (MRR): $\sum 1/k$
Part IV

Our proposal
Outline of Part IV

8. General Overview of the System Architecture
9. Named Entity Recognition and Classification
10. Question Processing
11. Passage Retrieval
12. Answer Extraction
Architecture Overview

Question Processing -> Keywords

Passage Retrieval

Passages -> Answer Extraction

Question type

Documents

Answers
Machine learning: multi-class perceptron
Hand-tagging of NEs in development corpus
One NERC model for each task
Combination of NERC models
<table>
<thead>
<tr>
<th>Architecture</th>
<th>NERC</th>
<th>Question Processing</th>
<th>Passage Retrieval</th>
<th>Answer Extraction</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**NERC**

<table>
<thead>
<tr>
<th>Recall in dev.</th>
<th>CHIL REF</th>
<th>CHIL AUTO</th>
<th>AMI REF</th>
<th>AMI AUTO</th>
<th>CoNLL 2002</th>
</tr>
</thead>
<tbody>
<tr>
<td>T1 45.98%</td>
<td>√</td>
<td></td>
<td></td>
<td>√</td>
<td>√</td>
</tr>
<tr>
<td>T2 28.87%</td>
<td>√</td>
<td>√</td>
<td></td>
<td></td>
<td>√</td>
</tr>
<tr>
<td>T3 79.53%</td>
<td></td>
<td></td>
<td>√</td>
<td></td>
<td>√</td>
</tr>
<tr>
<td>T4 47.76%</td>
<td></td>
<td></td>
<td>√</td>
<td>√</td>
<td>√</td>
</tr>
</tbody>
</table>
Question Processing

- Multi-class perceptron classifier
- Lexical, semantic and syntactic features
- 56 open domain answer types
- Mapping from answer types to NE types
- 69% precision in T1/T2
- 90% precision in T3/T4
Passage Retrieval

- **Keyword ranking**
- Passage retrieval based on query relaxation and keyword proximity [Pașca 2001]
- Two systems with search engines
  - $QAm$: Term based search
  - $QAa$: Phonetic similarity
Phonetic similarity example

Reference:
“The host system it is a UNIX Sun workstation”

Automatic:
“That of system it is a unique set some workstation”

<table>
<thead>
<tr>
<th>∫æt æβ sistəm it iz æ</th>
<th>junik sɛt səm</th>
<th>wəərkstɛfən</th>
</tr>
</thead>
<tbody>
<tr>
<td>junik s sən</td>
<td>← UNIX Sun</td>
<td></td>
</tr>
</tbody>
</table>
Answer Extraction

- Candidates matching Question Type
- Robust heuristic scoring function

\[ \text{score} = H_1 + H_2 + 2 \cdot H_3 + H_4 + H_5 - \frac{1}{4} \sqrt{H_6 - H_7} \]
Part V

QAst 2007 Results
Outline of Part V

13 Participants

14 Results

15 Analysis

16 QAst Conclusions
Participants

- CLT, Center for Language Technology, Australia
- DFKI, Deutsches Forschungszentrum für Künstliche Intelligenz, Germany
- LIMSI, Laboratoire d’Informatique et de Mécanique des Sciences de l’Ingénieur, France
- TOKYO, Tokyo Institute of Technology, Japan
- UPC

Five groups participated in both T1 and T2 tasks
Three groups in both T3 and T4 tasks
## Results T1: manual transcripts of lectures

<table>
<thead>
<tr>
<th>System</th>
<th>Questions</th>
<th>C. answers</th>
<th>MRR</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>clt1</td>
<td>98</td>
<td>16</td>
<td>0.09</td>
<td>0.06</td>
</tr>
<tr>
<td>clt2</td>
<td>98</td>
<td>16</td>
<td>0.09</td>
<td>0.05</td>
</tr>
<tr>
<td>dfki1</td>
<td>98</td>
<td>19</td>
<td>0.17</td>
<td>0.15</td>
</tr>
<tr>
<td>limsi1</td>
<td>98</td>
<td>43</td>
<td>0.37</td>
<td>0.32</td>
</tr>
<tr>
<td>limsi2</td>
<td>98</td>
<td>56</td>
<td>0.46</td>
<td>0.39</td>
</tr>
<tr>
<td>tokyo1</td>
<td>98</td>
<td>32</td>
<td>0.19</td>
<td>0.14</td>
</tr>
<tr>
<td>tokyo2</td>
<td>98</td>
<td>34</td>
<td>0.20</td>
<td>0.14</td>
</tr>
<tr>
<td>upc QA&lt;sub&gt;m&lt;/sub&gt;</td>
<td>98</td>
<td>54</td>
<td>0.53</td>
<td>0.51</td>
</tr>
</tbody>
</table>
Results T2: automatic transcripts of lectures

<table>
<thead>
<tr>
<th>System</th>
<th>Questions</th>
<th>C. answers</th>
<th>MRR</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>clt1</td>
<td>98</td>
<td>13</td>
<td>0.06</td>
<td>0.03</td>
</tr>
<tr>
<td>clt2</td>
<td>98</td>
<td>12</td>
<td>0.05</td>
<td>0.02</td>
</tr>
<tr>
<td>dfki1</td>
<td>98</td>
<td>9</td>
<td>0.09</td>
<td>0.09</td>
</tr>
<tr>
<td>limsi1</td>
<td>98</td>
<td>28</td>
<td>0.23</td>
<td>0.20</td>
</tr>
<tr>
<td>limsi2</td>
<td>98</td>
<td>28</td>
<td>0.24</td>
<td>0.21</td>
</tr>
<tr>
<td>tokyo1</td>
<td>98</td>
<td>17</td>
<td>0.12</td>
<td>0.08</td>
</tr>
<tr>
<td>tokyo2</td>
<td>98</td>
<td>18</td>
<td>0.12</td>
<td>0.08</td>
</tr>
<tr>
<td>upc QA&lt;sub&gt;m&lt;/sub&gt;</td>
<td>98</td>
<td>36</td>
<td>0.36</td>
<td>0.36</td>
</tr>
<tr>
<td>upc QA&lt;sub&gt;a&lt;/sub&gt;</td>
<td>97</td>
<td>29</td>
<td>0.25</td>
<td>0.24</td>
</tr>
<tr>
<td>upc QA&lt;sub&gt;a&lt;/sub&gt;*</td>
<td>98</td>
<td>36</td>
<td>0.34</td>
<td>0.34</td>
</tr>
</tbody>
</table>
## Results T3:
manual transcripts of meetings

<table>
<thead>
<tr>
<th>System</th>
<th>Questions</th>
<th>C. answers</th>
<th>MRR</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>clt1</td>
<td>96</td>
<td>31</td>
<td>0.23</td>
<td>0.16</td>
</tr>
<tr>
<td>clt2</td>
<td>96</td>
<td>29</td>
<td>0.25</td>
<td>0.20</td>
</tr>
<tr>
<td>limsi1</td>
<td>96</td>
<td>31</td>
<td>0.28</td>
<td>0.25</td>
</tr>
<tr>
<td>limsi2</td>
<td>96</td>
<td>40</td>
<td>0.31</td>
<td>0.25</td>
</tr>
<tr>
<td>upc QA_m</td>
<td>95</td>
<td>27</td>
<td>0.26</td>
<td>0.25</td>
</tr>
<tr>
<td>upc QA_m*</td>
<td>96</td>
<td>42</td>
<td>0.37</td>
<td>0.34</td>
</tr>
</tbody>
</table>
## Results T4: automatic transcrips of meetings

<table>
<thead>
<tr>
<th>System</th>
<th>Questions</th>
<th>C. answers</th>
<th>MRR</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>clt1</td>
<td>93</td>
<td>17</td>
<td>0.10</td>
<td>0.06</td>
</tr>
<tr>
<td>clt2</td>
<td>93</td>
<td>19</td>
<td>0.13</td>
<td>0.08</td>
</tr>
<tr>
<td>limsi1</td>
<td>93</td>
<td>21</td>
<td>0.19</td>
<td>0.18</td>
</tr>
<tr>
<td>limsi2</td>
<td>93</td>
<td>21</td>
<td>0.19</td>
<td>0.17</td>
</tr>
<tr>
<td>upc QA\textsubscript{m}</td>
<td>93</td>
<td>22</td>
<td>0.22</td>
<td>0.21</td>
</tr>
<tr>
<td>upc QA\textsubscript{a}</td>
<td>92</td>
<td>17</td>
<td>0.15</td>
<td>0.13</td>
</tr>
<tr>
<td>upc QA\textsubscript{a*}</td>
<td>93</td>
<td>19</td>
<td>0.17</td>
<td>0.17</td>
</tr>
</tbody>
</table>
## Overall

<table>
<thead>
<tr>
<th>Task</th>
<th>#Q</th>
<th>MRR</th>
<th>Accuracy</th>
<th>TOP1</th>
<th>TOP5</th>
</tr>
</thead>
<tbody>
<tr>
<td>T1, QA &lt;sub&gt;m&lt;/sub&gt;</td>
<td>98</td>
<td>0.53</td>
<td>0.51</td>
<td>50</td>
<td>54</td>
</tr>
<tr>
<td>T2, QA &lt;sub&gt;a&lt;/sub&gt;</td>
<td>98</td>
<td>0.34</td>
<td>0.34</td>
<td>33</td>
<td>36</td>
</tr>
<tr>
<td>T2, QA &lt;sub&gt;m&lt;/sub&gt;</td>
<td>98</td>
<td>0.36</td>
<td>0.36</td>
<td>35</td>
<td>36</td>
</tr>
<tr>
<td>T3, QA &lt;sub&gt;m&lt;/sub&gt;</td>
<td>96</td>
<td>0.37</td>
<td>0.34</td>
<td>34</td>
<td>42</td>
</tr>
<tr>
<td>T4, QA &lt;sub&gt;a&lt;/sub&gt;</td>
<td>93</td>
<td>0.17</td>
<td>0.17</td>
<td>16</td>
<td>19</td>
</tr>
<tr>
<td>T4, QA &lt;sub&gt;m&lt;/sub&gt;</td>
<td>93</td>
<td>0.22</td>
<td>0.21</td>
<td>20</td>
<td>22</td>
</tr>
</tbody>
</table>
## Overall

<table>
<thead>
<tr>
<th>Task</th>
<th>#Q</th>
<th>MRR</th>
<th>Accuracy</th>
<th>TOP1</th>
<th>TOP5</th>
</tr>
</thead>
<tbody>
<tr>
<td>T1, $QA_m$</td>
<td>98</td>
<td>0.53</td>
<td>0.51</td>
<td>50</td>
<td>54</td>
</tr>
<tr>
<td>T2, $QA_a$</td>
<td>98</td>
<td>0.34</td>
<td>0.34</td>
<td>33</td>
<td>36</td>
</tr>
<tr>
<td>T2, $QA_m$</td>
<td>98</td>
<td>0.36</td>
<td>0.36</td>
<td>35</td>
<td>36</td>
</tr>
<tr>
<td>T3, $QA_m$</td>
<td>96</td>
<td>0.37</td>
<td>0.34</td>
<td>34</td>
<td>42</td>
</tr>
<tr>
<td>T4, $QA_a$</td>
<td>93</td>
<td>0.17</td>
<td>0.17</td>
<td>16</td>
<td>19</td>
</tr>
<tr>
<td>T4, $QA_m$</td>
<td>93</td>
<td>0.22</td>
<td>0.21</td>
<td>20</td>
<td>22</td>
</tr>
</tbody>
</table>

- QC is worse for CHIL than AMI
Overall

<table>
<thead>
<tr>
<th>Task</th>
<th>#Q</th>
<th>MRR</th>
<th>Accuracy</th>
<th>TOP1</th>
<th>TOP5</th>
</tr>
</thead>
<tbody>
<tr>
<td>T1, QA&lt;sub&gt;m&lt;/sub&gt;</td>
<td>98</td>
<td>0.53</td>
<td>0.51</td>
<td>50</td>
<td>54</td>
</tr>
<tr>
<td>T2, QA&lt;sub&gt;a&lt;/sub&gt;</td>
<td>98</td>
<td>0.34</td>
<td>0.34</td>
<td>33</td>
<td>36</td>
</tr>
<tr>
<td>T2, QA&lt;sub&gt;m&lt;/sub&gt;</td>
<td>98</td>
<td>0.36</td>
<td>0.36</td>
<td>35</td>
<td>36</td>
</tr>
<tr>
<td>T3, QA&lt;sub&gt;m&lt;/sub&gt;</td>
<td>96</td>
<td>0.37</td>
<td>0.34</td>
<td>34</td>
<td>42</td>
</tr>
<tr>
<td>T4, QA&lt;sub&gt;a&lt;/sub&gt;</td>
<td>93</td>
<td>0.17</td>
<td>0.17</td>
<td>16</td>
<td>19</td>
</tr>
<tr>
<td>T4, QA&lt;sub&gt;m&lt;/sub&gt;</td>
<td>93</td>
<td>0.22</td>
<td>0.21</td>
<td>20</td>
<td>22</td>
</tr>
</tbody>
</table>

- QA<sub>a</sub>: similar for T2, worse for T4
## Overall

<table>
<thead>
<tr>
<th>Task</th>
<th>#Q</th>
<th>MRR</th>
<th>Accuracy</th>
<th>TOP1</th>
<th>TOP5</th>
</tr>
</thead>
<tbody>
<tr>
<td>T1, QA&lt;sub&gt;m&lt;/sub&gt;</td>
<td>98</td>
<td>0.53</td>
<td>0.51</td>
<td>50</td>
<td>54</td>
</tr>
<tr>
<td>T2, QA&lt;sub&gt;a&lt;/sub&gt;</td>
<td>98</td>
<td>0.34</td>
<td>0.34</td>
<td>33</td>
<td>36</td>
</tr>
<tr>
<td>T2, QA&lt;sub&gt;m&lt;/sub&gt;</td>
<td>98</td>
<td>0.36</td>
<td>0.36</td>
<td>35</td>
<td>36</td>
</tr>
<tr>
<td>T3, QA&lt;sub&gt;m&lt;/sub&gt;</td>
<td>96</td>
<td>0.37</td>
<td>0.34</td>
<td>34</td>
<td>42</td>
</tr>
<tr>
<td>T4, QA&lt;sub&gt;a&lt;/sub&gt;</td>
<td>93</td>
<td>0.17</td>
<td>0.17</td>
<td>16</td>
<td>19</td>
</tr>
<tr>
<td>T4, QA&lt;sub&gt;m&lt;/sub&gt;</td>
<td>93</td>
<td>0.22</td>
<td>0.21</td>
<td>20</td>
<td>22</td>
</tr>
</tbody>
</table>

- AE performs well in T1/T2
## Overall

<table>
<thead>
<tr>
<th>Task</th>
<th>#Q</th>
<th>MRR</th>
<th>Accuracy</th>
<th>TOP1</th>
<th>TOP5</th>
</tr>
</thead>
<tbody>
<tr>
<td>T1, QA&lt;sub&gt;m&lt;/sub&gt;</td>
<td>98</td>
<td>0.53</td>
<td>0.51</td>
<td>50</td>
<td>54</td>
</tr>
<tr>
<td>T2, QA&lt;sub&gt;a&lt;/sub&gt;</td>
<td>98</td>
<td>0.34</td>
<td>0.34</td>
<td>33</td>
<td>36</td>
</tr>
<tr>
<td>T2, QA&lt;sub&gt;m&lt;/sub&gt;</td>
<td>98</td>
<td>0.36</td>
<td>0.36</td>
<td>35</td>
<td>36</td>
</tr>
<tr>
<td>T3, QA&lt;sub&gt;m&lt;/sub&gt;</td>
<td>96</td>
<td>0.37</td>
<td>0.34</td>
<td>34</td>
<td>42</td>
</tr>
<tr>
<td>T4, QA&lt;sub&gt;a&lt;/sub&gt;</td>
<td>93</td>
<td>0.17</td>
<td>0.17</td>
<td>16</td>
<td>19</td>
</tr>
<tr>
<td>T4, QA&lt;sub&gt;m&lt;/sub&gt;</td>
<td>93</td>
<td>0.22</td>
<td>0.21</td>
<td>20</td>
<td>22</td>
</tr>
</tbody>
</table>

- AE drops in T3/T4 due to NERC
Remarks

- Moving from manual to automatic
  - 35% drop in CHIL
  - 29% drop in AMI
- Drop in AMI is smaller than CHIL
  - AMI task is more difficult
  - AMI corpus contains more redundancy
  - CHIL task is easier
  - CHIL corpus is smaller
Analysis

- \( QA_a \) performs worse than \( QA_m \)
  - Phonetic similarity triggered too many false keyword matches noisy passages
  - NERC training data is insufficient to learn phonetic generalizations. No improvement for automatic transcripts
  - Good results in IR (tbp)

- Worst results in types: Org, Loc, Tim, or Mea for T3/T4
  - Types with high variation
  - Small training data

- Similar drop-off in all types of NE
Conclusions

- QA technology can be successfully used in speech scenarios
- $Q A_m$ is robust
- PR based on query relaxation adapts to automatic transcripts
- QA performance drop follows WER in small corpora but is smaller than WER in larger corpora
- PR is affected by a high WER (T4)
- NERC enhanced with phonetic features deserves further research
Part VI

PhD. Thesis Project
Outline of Part VI

17  Short Term Work

18  Mid Term Work

19  Work Plan
Short Term

Organization of QAst 2008
- English, Spanish, French
- ASR output with different WER
- Definitional questions
- Factual NP questions
- From February 11th to June 30th
Short Term

- Research in IR
  - Error detection mechanisms
  - Use of semantic information
- Improve NERC module
  - Add phonetic information
  - Error detection mechanisms
- Participate in QAst 2008
  - Adapt to Spanish
Mid Term

- Future QAst 2009
- Research in Answer Extraction
  - Add Semantic Roles
    - Agent, Patient, Instrument, Temporal, etc.
    - SR are key for answering “Who”, “When”, “What”, “Where”...
Mid Term

- Future QAst 2009
- Research in Answer Extraction
  - Add Semantic Roles
  - Agent, Patient, Instrument, Temporal, etc.

Short Term Work

Mid Term Work

Work Plan
Work Plan

Year 2008

QAst 2008

IR

NERC

Year 2009

QAst 2009

AE

Writing
Part VII

Publications
Outline of Part VII

20 Question Answering

21 Spoken Document Retrieval

22 Semantic Role Labeling
J. Turmo, P.R. Comas, C. Ayache, D. Mostefa, S. Rosset and L. Lamel
*Overview of QAST 2007*
CLEF 2007

P.R. Comas, J. Turmo and M. Surdeanu
*Robust Question Answering for Speech Transcripts Using Minimal Syntactic Analysis*
CLEF 2007

M. Surdeanu, D. Dominguez-Sal and P.R. Comas
*Design and Performance Analysis of a Factoid Question Answering System for Spontaneous Speech Transcriptions*
INTERSPEECH 2006
P.R. Comas and J. Turmo

*PHAST: Spoken Document Retrieval Based on Sequence Alignment*

Report de Recerca del LSI: LSI-08-2-R
Semantic Role Labeling

- M. Surdeanu, L. Màrquez, X. Carreras and P.R. Comas
  *Combination Strategies for Semantic Role Labeling*
  JAIR 2007

- L. Màrquez, M. Surdeanu, P.R. Comas and J. Turmo
  *A Robust Combination Strategy for Semantic Role Labeling*
  EMNLP 2005

- L. Màrquez, P.R. Comas, J. Giménez and N. Català
  *Semantic Role Labeling as Sequential Tagging*
  CoNLL 2005
Acknowledgements

- CHIL project, IP-506909
- TEXT-MESS project, TIN2006-15265-C06
Thank you!

Question Answering?
S. Altschul, W. Gish, W. Miller, E.W. Meyers and D.J. Lipman
Basic local alignment search tool
*Journal of Molecular Biology, 1990*

G. Kondrak.
Algorithms for Language Reconstruction
*PhD thesis, University of Toronto, 2002*
S. Rosset, O. Galibert, G. Adda and E. Bilinski.

D. Mollá, S. Cassidy and M. van Zaanen.
Bibliography III

CLEF2007 question answering experiments at tokyo institute of technology.
*Clef, 2007.*

S. Harabagiu, D. Moldovan, and J. Picone.
Open-domain voice-activated question answering.


G. Amati and C.J. Van Rijsbergen. Probabilistic models of information retrieval based on measuring the divergence from randomness. 


G. Salton, editor. 

*Automatic text processing.* 

M. Alzghool and D. Inkpen. 
University of Ottawa’s participation in the CL-SR task at CLEF 2006. 
*CLEF, 2006.*

D. Inkpen, M. Alzghool, and A. Islam. 
Using various indexing schemes and multiple translations in the CL-SR task at CLEF 2005. 
*CLEF, 2006.*
Dublin city university at CLEF 2006: Cross-language speech retrieval (CL-SR) experiments.

*CLEF, 2006.*
# TREC 2007 results factual

<table>
<thead>
<tr>
<th>Submitter</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lymba Corporation</td>
<td>0.706</td>
</tr>
<tr>
<td>Language Computer Corporation</td>
<td>0.494</td>
</tr>
<tr>
<td>Saarland University</td>
<td>0.289</td>
</tr>
<tr>
<td>University of Lethbridge</td>
<td>0.258</td>
</tr>
<tr>
<td>Concordia University</td>
<td>0.256</td>
</tr>
<tr>
<td>Fudan University</td>
<td>0.236</td>
</tr>
<tr>
<td>Universita di Roma</td>
<td>0.222</td>
</tr>
<tr>
<td>State University of New York</td>
<td>0.222</td>
</tr>
<tr>
<td>CMU and University of Karlsruhe</td>
<td>0.208</td>
</tr>
<tr>
<td>Tsinghua University</td>
<td>0.206</td>
</tr>
</tbody>
</table>