Evaluations and Language resources Distribution Agency
55/57, rue Brillat Savarin
75 013 Paris
Tél. : 01 43 13 33 33
Fax : 01 43 13 33 30
Web : www.elda.org
Email : info@elda.org

The QAst pilot task in CLEF 2007

- Guidelines for participants -

Author(s)	Christelle Ayache
Organism	ELDA
Address	55-57, rue Brillat-Savarin 75013 Paris, France
E-mail	ayache@elda.org
Date	25th January, 2007
Version	1.1

Versions of the document

Version	**Date**	**Status**	**Notes**
1.0	01/25/07	Guidelines	
1.1	02/09/07	Guidelines	Metrics updated
1.2	05/10/07	Guidelines	Output format updated

CONTENTS

1. Tasks

The participants in QAst track will be able to deal with four different tasks, each one involving different data sets (automatic or manual transcriptions) related to two different scenarios (lectures in English about *"speech and language processing"* or meetings in English about *"design of television remote control"*).

The tasks defined in QAst track are the following:

- T1: Question Answering in manual transcriptions of lectures (CHIL corpus).
- T2: Question Answering in automatic transcriptions of lectures (CHIL corpus).
- T3: Question Answering in manual transcriptions of meetings (AMI corpus).
- T4: Question Answering in automatic transcriptions of meetings (AMI corpus).

2. Data collections

The data for the QAst pilot track consists of two different resources, one for dealing with the lecture scenario and the other for dealing with the meeting scenario:

A. Lecture scenario

The **CHIL corpus** is used. It consists of around 25 hours (around 1 hour per lecture) both manually and automatically transcribed (LIMSI has volunteered to produce the ASR transcriptions). In addition, the set of lattices and confidences for each lecture will be probably provided (also produced by LIMSI). The domain of the lectures is *"speech and language processing"*. The language is European English, mostly spoken by non native speakers. Lectures are provided with simple tags: a tag for the delimitation of the document (<DOC>), a tag indicating the document's id (<DOC_ID>) and a tag indicating the topic of the seminar (<TOPIC>). A tag has been added to identify the type of document. For this task, all documents are either "Manual Transcription" or either "Automatic Transcription" (<DOC_TYPE>). Seminars are formatted as plain text files (ISO-8859-1).

The following is an extract of the CHIL corpus:

```
<DOC>
<DOC_ID>ISL_20050420</DOC_ID>
<TOPIC>SPEECH RECOGNITION SYSTEM</TOPIC>
<DOC_TYPE>MANUAL TRANSCRIPTION</DOC_TYPE>
yeah so if somebody uhm
want to do something with signal processing I would suggest a paper from
Hermanski
perceptional linear prediction .
and then we have one paper by
Herman
about
like
uhm
yeah dynamic programming which is basically a basic
algorithm and speech .
you want to say something more or ?
so at least the people who who enter the cognitive systems lecture should have known about
dynamic programming you heard about it .
...
</DOC>
```

B. <u>Meeting scenario</u>:

The **AMI corpus** is used. It consists of around 100 hours (around 170 meetings) both manually and automatically transcribed (the AMI RT 2006 ASR will be used). The domain of this meetings is *"design of television remote control"*. The language is European English, mostly spoken by non native speakers. Meetings are provided with simple tags: a tag for the delimitation of the document (<DOC>) and a tag indicating the document's id (<DOC_ID>). A tag has been added to identify the type of document. For this task, documents are either "Manual Transcription" or either "Automatic Transcripts" (<DOC_TYPE>). A new tag has been added on these meetings: a tag for the delimitation of the speakers. Contrary to the CHIL corpus, in each AMI meeting, several speakers appear. Thus, we added a tag of delimitation for different speakers (<speaker name="A">) and a tag for the delimitation of text (<TEXT>). Meetings are formatted as plain text files (ISO-8859-1).

The following is an extract of the AMI corpus:

```
<DOC>
<DOC_ID>IB4011a</DOC_ID>
<DOC_TYPE>MANUAL TRANSCRIPTION</DOC_TYPE>
<TEXT>
<speaker name="B">
 It's just that I need uh some some <ne type"50">white</ne> <ne type"70">paper</ne>.
</speaker>
<speaker name="A">
 Are you gonna take the minutes?
</speaker>
<speaker name="B">
Uh yes.
</speaker>
<speaker name="D">
 Yep. Um
</speaker>
<speaker name="A">
Good.
</speaker>
<speaker name="D">
So hello everybody. So uh you everybody had the a agenda for today, you received it or
</speaker>
…
</TEXT>
</DOC>
```

The data will be prepared and made available by ELDA.

3. Questions (input format)

All the questions in QAst track are "Factual questions", i.e. questions whose expected answer is a Named Entity (location, date, person, organisation, etc.) such as defined in specific NE's guidelines.

<u>Questions Generation</u>

We have selected a list of 10 named entity types in the corpus:

- ⇨ **Person**: names of humans, both real and fictional, fictional or real non-human individuals (the names of animals are not tagged as "Person" but as "Misc"). For example: "Mirjam Killer", "John", "Jesus", etc.

- ⇨ **Organisation**: names of business, multinational organizations, political parties, religious groups, etc. For example: “CIA”, “IBM”, but also sometimes the “Bulls”, “Washington” when they display the characteristics of an organisation.

- ⇨ **Location**: geographical, political or astronomical entities. For example: “California”, “South of California”, “Earth”...

- ⇨ **Time**: a date or a specific moment in time, absolute and relative time expressions. For example: “March 28th”, “last week”, at “four o’clock in the morning”

- ⇨ **Measure**: measures of length, width or weight. Generally, a quantity and a unit of measurement. For example: “five kilometres”, “20 hertz”... But also ages, period of time…

- ⇨ **Method / System**: names of a specific method or a specific system. For example: “VTLN”, “Hidden Markov Models”...

- ⇨ **Language**: natural languages. For example: “Spanish”, “Chinese”, etc.

- ⇨ **Colour**: basic (green, yellow) and elaborate colour terms (yellowy orange). In general, if the expression includes a colour-denoting adjective, it should be considered as a whole (dark blue).

- ⇨ **Shape**: identifiable geometrics shapes (squarely, kidney, curved…) or an idiomatic use of the notion, in which case the fact that this is a reference to shape has to be explicit (friendly or ergonomic shape).

- ⇨ **Material**: entities referring to substances things are made of. It is only used for generic mentions of materials For example: wood, titanium, rubber, plastic, etc.

The two data collections are first tagged NE.
Then, an English native speaker created questions for each NE tagged session.
Each created question has as an answer a named entity tagged in the corpus.

NB: For each corpus, 10 percent of the questions created have no answer in the document collection, and in this case the correct answer is a blank string with docid “NIL”. A question is assumed to have no right answer when neither human assessors nor participating systems can find one.

<u>Sets of questions</u>

Two sets of questions will be provided to the participants :
- the Development set (30 January 2007)
- the Evaluation set (15 June 2007)

For each one of the scenarios, a development set and an evaluation set will be released:

- ➢ Development sets:

- Lectures: 10 seminars and 50 questions (the same 10 seminars used in the CHIL evaluation).
- Meetings: 50 meetings and 50 questions.

- Evaluation sets:

Lectures: 15 seminars (probably more if more records become available) and 100 questions.
Meetings: 120 meetings and 100 questions.

Development sets will be formatted as plain text files, with one question per line. Each line will be structured as follows:

```
<id-question> <question>
```

where:

<id-question> is the number of the question in the file of questions,
<question> contains the text of the question.

For each set, columns are separated by a single blank space:

```
...
37 <question>
38 <question>
39 <question>
40 <question>
...
```

4. Answers (output format)

An answer is basically structured as an [answer-string, document-id] pair, where the answer-string contains nothing more than a complete and exact answer (a “Named Entity”) and the document-id is the unique identifier of a document that supports the answer.
There are no particular restrictions on the length of an answer-string (which is normally very short), but unnecessary pieces of information will be penalized, since the answer will be marked as non-exact. Assessors will focus mainly on the “responsiveness” and “usefulness” of the answers.

Each submitted run is a single text file that contain one line per answer. Each line is in the following format:

```
<id-question> <id-team+id-run+id-task> <id-document> <answer> <ranking> <score>
```

where:

<id-question> is the number of the question in the file of questions,

<id-team+id-run+id-task> is the name of the participating team plus the run identifier (1 for run1 or 2 for run2) plus the task identifier (t1, t2, t3 or t4).

<*id-document*> is the name of the file containing the document in which the <answer> was found.

<*answer*> contains the answer-string (or a blank if the docid is "NIL").

<*ranking*> is the ranking of the answer within the list of 5 possible answers returned for a question.

<*score*> (or confidence score) is a mandatory score-value per answer (see below for more detail).

Examples:

```
...
38 limsi1_t1 ISL_20050420 Cambridge 1 0.76
38 limsi1_t1 NIL 2 0.68
39 limsi1_t1 ISL_20050420 VTLN 1 0.52
40 limsi1_t1 NIL 1 0.49
...
```

- **the first element** is the identifiant of the question: answers must be returned in the same ascending (increasing) order in which questions appear in the datasets (from 01 to n).

- **the second element** describes:

  - the name of the participating team ("limsi", "upc", etc.),

  + the identifier of the run (1 if it is the first one, or 2 if it is the second one),

  + the identifier of the task (_t1, _t2, _t3, or _t4).

  Clearly, the content of this field never changes within the same submission file. Each submission file must be named after this column, with a .txt extension. Therefore, the lines in the examples above will be part of the file limsi1_t1.txt.

- **the third element** contains the docid of the document that supports the given answer. Some questions may not have any known response in the document collection: in this case, the correct answer will be a blank answer-string, and the string NIL will replace the docid (see third line in the examples above).

- **the fourth element** contains the answer-string (or a blank if the docid is NIL). There can be no line breaks in the answer-string. Human assessors will judge both the "correctness" and "exactness" of the submitted answers, so unnecessary additional information will be marked as non-exact. The same judgment will be assigned to responses that lack significant bits of information.

- **the fifth element** is the ranking of the answer within the list of 5 possible answers returned for a question.

- **the sixth element** is a mandatory score-value that can range between 0.00 and 1.00 (inclusive), where 0.00 means that the system has no evidence of the correctness of the answer, and 1.00 means that the system is absolutely confident about the correctness of the answer. If the system does not produce any normalized score number, it must return "NIL". This score has no impact on the evaluation process.

5. Human judgment

The files submitted by participants (deadline: 30 June 2007) will be manually judged by native speaking assessors. Assessors will consider **correctness** (i.e. responsiveness) and **exactness** (i.e. the quantity of information) of the returned answers. They will also check that the document labelled with the returned docid supports the given response. Each [answer-string, docid] pair will be assessed and marked with one of the following judgments, that have also been used at TREC:

- **incorrect**: the answer-string does not contain a correct answer or the answer is not responsive;
- **unsupported**: the answer-string contains a correct answer but the docid does not support it;
- **non-exact**: the answer-string contains a correct answer and the docid supports it, but the string has bits of the answer missing or is longer than the required length of the answer;
- **correct**: the answer-string consists of the relevant information (exact answer), and the answer is supported by the returned document.

Individual results will be released to each participating group from the 30th of July. All runs will be judged, and assessors will attach their judgments to the beginning of each line returned by the system, as follows:

...
R 099 limsi1_t1 ISL_20050420 Cambridge 1 0.92
R 100 limsi1_t1 ISL_20050420 VTLN 1 0.89
W 101 limsi1_t1 NIL 1 0.69
...

where:

W stands for wrong,
U stands for unsupported,
X stands for inexact,
R stands for right.

6. Evaluation measures

The two following metrics used in CLEF will be used in the evaluation:
Mean Reciprocal Ranked (MRR) measures how well ranked is the right answer in the list of 5 possible answers in average.
Accuracy: The fraction of correct answers ranked in the first position in the list of 5 possible answers.

7. Important dates – QAst pilot task – CLEF 2007

1	**Development set release**	1 February 2007	CHIL corpus: 50 questions, 10 lectures AMI corpus: 50 questions, 50 meetings
2	**Evaluation set release**	15 June 2007	CHIL corpus: 100 questions, 15 lectures AMI corpus: 100 questions, 118 meetings
3	**Submission deadline**	30 June 2007	
4	**Individual results release**	30 July 2007	
5	**Paper submission**	20 August 2007	
6	**Workshop**	19-21 September 2007	