A question-answer distance measure to investigate QA system progress

Guillaume Bernard, Sophie Rosset, Martine Adda-Decker and Olivier Galibert

Groupe Traitement du langage parlé
LIMSI-CNRS, FRANCE
http://www.limsi.fr/tlp/

20 May 2010
Questions-answering (QA) systems

- Provide precise answers to the user questions
- Search the answer through a corpus of documents

Example

Question: *Besides France and Germany, where have we seen case of mad-cows disease?*
Answer: *In Belgium*

Importance of evaluating the evolution of such systems

- Evaluation campaigns: TREC, QA@CLEF, QAst ...
Evaluation campaigns on questions-answering systems

- Documents come from various origins: newspaper, meetings transcriptions ...
- Question corpus
  - Questions built by evaluators using the document corpus
- Measure the progress in QA domain

Issue addressed

Can we compare a QA system on successives evaluation campaigns?
→ Assessing the evolution of the evaluation criteria

Context of the work

- QAst: Questions-Answering on Speech Transcriptions
The QAst evaluation campaign

- Evaluate systems on speech transcriptions
- Three different languages: French, English and Spanish
- QAst 2009: a new building procedure for the questions
A new question corpus building procedure

- 2008: questions created from the documents
- 2009: more “spontaneous” questions provided by naive users:
  - Use of excerpts of document
  - Ask questions on information related to these excerpts

Example

Text fragment: *Jacques Chirac is the previous President of France.*

2008 question: *Who is the previous President of France?*

2009 question: *What is the age of Jacques Chirac?*

Question

- Does this questions building methodology changes the evaluating features of the QAst campaign?
- Observation of the impact on the results of the QA systems: comparison between 2008 and 2009 results
## Observations on QAst 2008 and 2009 results

### Results on LIMSI system

| System   | French |   | English |   | Spanish |   |
|----------|--------|---|---------|---|---------|---|
|          | Acc(%) | Δ | Acc(%)  | Δ | Acc(%)  | Δ |
| QAst 2008| 50     | -22| 52      | -25| 56      | -20|
| QAst 2009| 28     |   | 27      |   | 36      |   |

### Results for the other participants

| System       | English |
|--------------|---------|
|              | Acc(%)  | Δ    |
| INAOE 2008   | 33%     | -5   |
| INAOE 2009   | 28%     |      |
| UPC 2008     | 34%     | -13  |
| UPC 2009     | 21%     |      |
Observations on QAst 2008 and 2009 results

**Observations**
- Strong decrease between 2008 and 2009

**Hypothesis for the loss**
- Influence of the way the questions were built
  - Greater distance between the text fragment used to create the question and the answer

**Example**
- **Text fragment:** Jacques Chirac is the previous President of France.
- **2008 question:** Who is the previous President of France?
- **2009 question:** What is the age of Jacques Chirac?

**Idea**
- Quantifying the influence of the new building procedure
A measure for the question corpus

Aim of the measure

- Evaluation of the distance between the elements of each question of a corpus and the corresponding answers
  - Question elements considered: named entities and multi-words expressions
- Gives two values: the average distance and the standard deviation

Computing of a global distance for each question

- Distance evaluated in words
- Average of distances between the elements of the question found in the document and the answer
A measure for the question corpus

Example

Q: Which **Belgian** **organization** has been declared **criminal**?

A: **Vlaams Block**

TS: The Belgian supreme Court has upheld a previous ruling that declares Vlaams Blok a criminal organization.
### Example

| Q   | Which **Belgian** organization has been declared **criminal**? |
|-----|-------------------------------------------------------------|
| A   | Vlaams Block                                                |
| TS  | The **Belgian** supreme Court has upheld a previous ruling that declares **Vlaams Blok** a criminal organization. |
A measure for the question corpus

Example

Q: Which *Belgian* organization has been declared *criminal*?

A: Vlaams Block

TS: The *Belgian* supreme Court has upheld a previous ruling that declares Vlaams Blok a criminal *organization*. 
A measure for the question corpus

Example

Q: Which **Belgian** organization has been declared **criminal**?

A: **Vlaams Block**

TS: The **Belgian** supreme Court has upheld a previous ruling that declares **Vlaams Blok** a **criminal** organization.
A measure for the question corpus

Example

Q: Which **Belgian** organization has been declared **criminal**?

A: **Vlaams Block**

TS: The **Belgian** supreme Court has upheld a previous ruling that declares **Vlaams Blok** a **criminal organization**.

Global distance of the question: \( \text{Average}(10+2+1) = 4 \)
### Evolution of the Average Distance and Standard Deviation

|        | French |     |     |        | English |     |     |        |        | Spanish |     |     |     |
|--------|--------|-----|-----|--------|---------|-----|-----|--------|--------|---------|-----|-----|-----|
|        | AD     | SD  | Δ   | AD     | SD     | Δ   | AD  | SD     | Δ      | AD      | SD  | Δ   |
| 2008   | 45     | 100 | +98 | 97     | 284    | +39 | 381 | 851    | -359   | 22      | 73  |     |
| 2009   | 143    | 431 |     | 136    | 310    |     | 22  | 73     |        |         |     |     |

- Strong increase on French and English, but also a very strong decrease on Spanish
  - New building procedure does not always imply an increase of the distance
  - The corpora have not the same features for 2008 and 2009
- High Standard Deviation: strong distance variations in corpus
Focus on question distances

Average distance values - 2008 and 2009 test corpus

- X axis: distance classes (DC); Y axis: #questions in DC
- Evolution of the question corpus between 2008 and 2009
- Strong dispersion for the three languages
## Discussion

### Correlation with evaluation campaign results

- Segmentation of the documents by the QA systems
  - Use a window size fixed by tuning on the 2008 corpus
    - In 2009, the snippets are either too small (French, English) or too big (Spanish)

→ Potential explanation for the strong loss

### Usability for futures evaluations

- Measure based on our representation of the elements of a question
  - Can be generalized on other systems using different representations (e.g. keywords)
- Measure can be used as a control parameter criterion for building question corpus
  - Allow to evaluate the features of a campaign
Conclusions and perspectives

Conclusions

- Huge performance loss between QAst 2008 and 2009 evaluations
  - New building procedure for the question corpus of 2009
- Application of a measure based on a distance between the elements of the question and the answer
  - Strong variations between the two instances of the QAst campaign
- The strong variations can potentially explain the bad results of the QA systems
- The measure can control for variations between two instances of a campaign

Perspectives

- Complementary measures
  - Referential expressions
  - Language-specific features
Thank you for listening! Any questions?