MACAQ : A Multi Annotated Corpus to study how we adapt Answers to various Questions

Anne Garcia-Fernandez, Sophie Rosset, Anne Vilnat
LIMSI-CNRS and University Paris Orsay

21/05/2010

LREC 2010, Valletta, Malta

Overview

1. **Why a corpus of human answers?**
2. **Corpus constitution**
   - List of questions
   - Corpus of answers
3. **Corpus annotation**
   - Automatic non-specific annotations
   - Manual specific annotations
4. **Conclusion**

Provide a *natural-language* answer

What is QA?

Q: *Where is the Mona Lisa?*
A1: *Louvre Museum*
A2: The Mona Lisa is in the Louvre Museum in Paris.

Our Goal

Provide a corpus of answers in natural language

From *QA systems* answer to *natural-language* answer

Evaluation campaigns answer

in the Louvre + an extract of the document where the answer was found

Multiple natural-language answer forms

The Mona Lisa *is exhibited* *in the Louvre Museum*.
It *is exhibited* *in the Louvre Musuem*.
The Mona Lisa *is* *in the Louvre Museum*.
It is *in the Louvre Museum* *the Mona Lisa* *is exhibited*.
The Mona Lisa *is a work of Léonard de Vinci* *exhibited*
in the Louvre Museum. ...

Which surface form?

What to say? How to say it?

The Mona Lisa *is exhibited in the Louvre Museum.*

It *is exhibited in the Louvre Musuem.*

?

Which surface form?

What to say? How to say it?

The Mona Lisa is exhibited *in the Louvre Museum* *.*

It is *in the Louvre Museum* *the Mona Lisa is exhibited.*

?

Which surface form?

What to say? How to say it?

The Mona Lisa is exhibited in the Louvre Museum.

The Mona Lisa | *is a work of Léonard de Vinci* | *exhibited in the Louvre Museum.*

?

Which surface form?

What to say? How to say it?

The Mona Lisa is exhibited in the Louvre Museum.
The Mona Lisa | *is a work of Léonard de Vinci* | *exhibited in the Louvre Museum.*

?

Hypothesis

Answer surface form depends on question surface form

Which surface form?

What to say? How to say it?

The Mona Lisa is exhibited in the Louvre Museum.
The Mona Lisa | *is a work of Léonard de Vinci* | *exhibited in the Louvre Museum.*

?

Hypothesis

Answer surface form depends on question surface form

The corpus

Human answers to various questions

Using existing corpora

Existing answers: QA campaign answers

in the Louvre Museum, Paris, France, ...

Too short

Using existing corpora

Existing answers: collaborative QA website

Q: What was Lewis Carroll's first job?
A: Lewis Carroll (Charles Lutwidge Dodgson) was born in Daresbury Parsonage, Daresbury, Cheshire, on 27 January 1832, the third child and eldest son of Rev. Charles Dodgson and his wife Frances. Altogether, there were eleven Dodgson children, and all of them survived; quite unusual for those days!
Too long, complex

Using existing corpora

~~Existing answers~~

Existing questions: QA campaign answers

How many chickens are available for adoption at the Camden County Animal Shelter?

Topic too complex, not available in different syntactic forms

Using existing corpora

~~Existing answers~~

Existing questions: collaborative QA website

Under what condition the average speed is equal to the magnitude of the average velocity?

Too long, topic complex

Using existing corpora

~~**Existing answers**~~

~~**Existing questions**~~

Modality of interaction

QA systems are available for written and speech interaction...
Corpora are not.

Using existing corpora

~~Existing answers~~

~~Existing questions~~

Modality of interaction

QA systems are available for written and speech interaction...
Corpora are not.

We need to build a new corpus

Corpus acquisition methodology

We ask a question, the user answers it.

Corpus of questions:

- controlled variations of the same question
- easy questions: to minimize "I don't know"-type answers

Users:

- French native speakers

Modalities:

- oral interaction over the phone
- written interaction on a website

Protocol:

- 18 to 24 questions per session

Corpus of questions

Factoid and simple

question markers | principal verb | nominal phrase (focus) [other]

Example

Combien | pèse | un bébé | à la naissance ?

How much | does | a baby | weight | at birth ?

Corpus of questions

Factoid and simple

question markers | principal verb | nominal phrase (focus) [other]

Variation:

Où est la Joconde ? (*Where is the Mona Lisa?*)

Quand sont les JO ? (*When are the Olympic Games?*)

Combien mesure la Tour Eiffel ? (*How tall is the Eiffel Tower?*)

Corpus of questions

Factoid and simple

question markers | principal verb | nominal phrase (focus) [other]

Variation:

Où est la Joconde ? (*Where is the Mona Lisa?*)

Dans quel musée est la Joconde ? (*In which museum is the Mona Lisa?*)

La Joconde est-elle au Louvre ? (*Is the Mona Lisa in the Louvre Museum?*)

Corpus of questions

Factoid and simple

question markers | principal verb | nominal phrase (focus) [other]

Variation:

Où | est | la Joconde ? (*Where is the Mona Lisa?*)

La Joconde | est | où ? (*The Mona Lisa is where?*)

Je voudrais savoir où | est | la Joconde ? (*I would like to know where is the Mona Lisa?*)

Corpus of questions

Factoid and simple

question markers | principal verb | nominal phrase (focus) [other]

Variation:

Où est la Joconde ? (*Where is the Mona Lisa?*)

Où est exposée la Joconde ? (*Where is exhibited the Mona Lisa?*)

Corpus of questions

Factoid and simple

question markers | principal verb | nominal phrase (focus) [other]

Variation:

Dans quel **musée** est la Joconde? — *NE museum*

Dans quel **pays** est la Joconde? — *NE country*

La Joconde **est-elle** au Louvre ? — *Yes-No*

Corpus of questions

Factoid and simple

question markers | principal verb | nominal phrase (focus) [other]

Variation:

Où est la Joconde ? (*Where is the Mona Lisa?*)

Où est le Rhin ? (*Where is the Rhin River?*)

General desription of the corpus

More than 3,000 answers and 6 answers per question

	Written	Speech	Total
# answers	2,088	1,044	**3,132**
# different questions	507	*493*	**507**
# subjects	99	53	**152**
# subjects/question	4.12	2.12	**6.17**
# words	17,976	7,128	25,104
# different words	3,363	1,634	4,574

Table: General characteristics of the corpus

Annotations

Automatic annotations

Non specific to the QA context

Manual annotations

Specific to the QA context

Automatic non-specific annotations

Lemmatisation

Raw	La	Joconde	est	actuellement	au	Louvre
Lemmatised	le	Joconde	être	actuellement	au	Louvre

Part-of-speech tagging

Syntactic parsing

Automatic non-specific annotations

Lemmatisation

Part-of-speech tagging

Raw	La	Joconde	est	actuellement	au	Louvre
POS	DET	NAM	VER	ADV	PRP:det	NAM

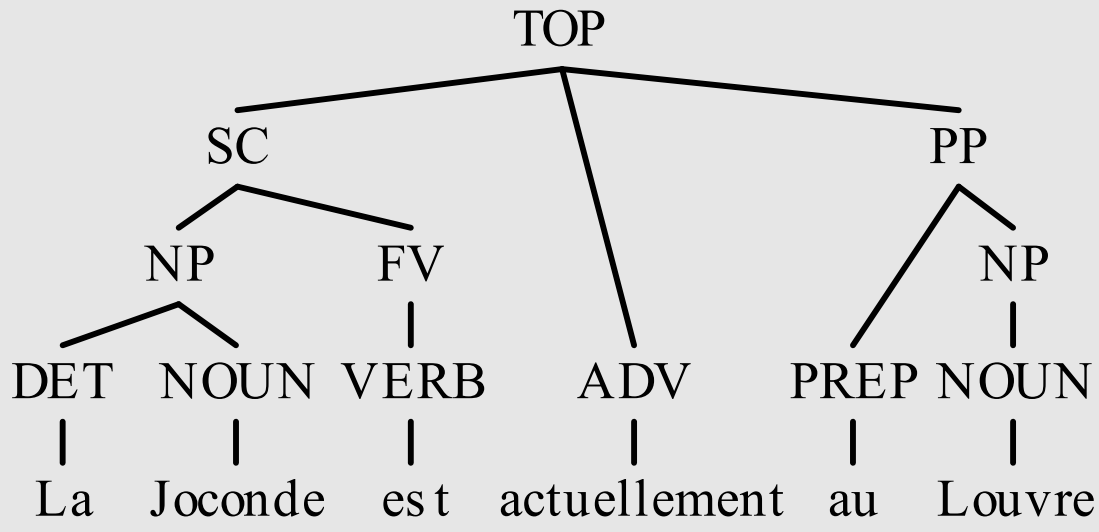
Syntactic parsing

Automatic non-specific annotations

Lemmatisation

Part-of-speech tagging

Syntactic parsing

Raw	La Joconde est actuellement au Louvre
Syntactic Parsing	

Manual specific annotations

Linguistic link between Q and A

Words reused from the question in the answer:

question markers | principal verb | nominal phrase (focus) [other]

Question	Où est la Joconde ?
Answer	La Joconde est actuellement au Louvre

The answer itself

Additionnal elements

Manual specific annotations

Linguistic link between Q and A

The answer itself

Definition of the Information-answer as the shortest part of the answer which consists either:
(1) of a new information which corresponds to the question expected type (EN, Yes-No)
(2) or of an admission of incompetence

Raw	La Joconde est actuellement au Louvre
Annotation	La Joconde est actuellement au Louvre

Additionnal elements

Manual specific annotations

Linguistic link between Q and A

The answer itself

Additionnal elements

Completion, suggestion and irrelevance were manually annotated

Completion	Le 11 novembre 1918 à Rethondes
	November 11th 1918 in Rethondes
Suggestion	Je ne suis pas sûr, il faut chercher dans un dictionnaire.
	I am not sure, you should look in a dictionary.
Irrelevance	vas dans ta chambre :P [sic]
	Go to your room :P

Manual specific annotations

Linguistic link between Q and A

The answer itself

Additionnal elements

Examples

La Joconde est actuellement au Louvre à Paris.

Manual specific annotations

Linguistic link between Q and A

The answer itself

Additionnal elements

Examples

La Joconde | est | actuellement | au Louvre | à Paris .

Conclusion

A corpus of natural-language answers

- answering questions with controlled linguistic features
- speech and written modalities
- annotated automatically and manually

Useful for...

- comparison speech/written
- answer generation
- interactive question-answering
- ... and what **you** want!

Thanks!