The LIMSI participation to the QAst track : experimentating on answer scoring

Guillaume Bernard, Sophie Rosset, Olivier Galibert

Spoken Language Processing Group
LIMSI-CNRS
France

The QAst tasks

Monolingual Question Answering on manual and automatic speech transcriptions. 3 tasks, 24 sub-tasks:

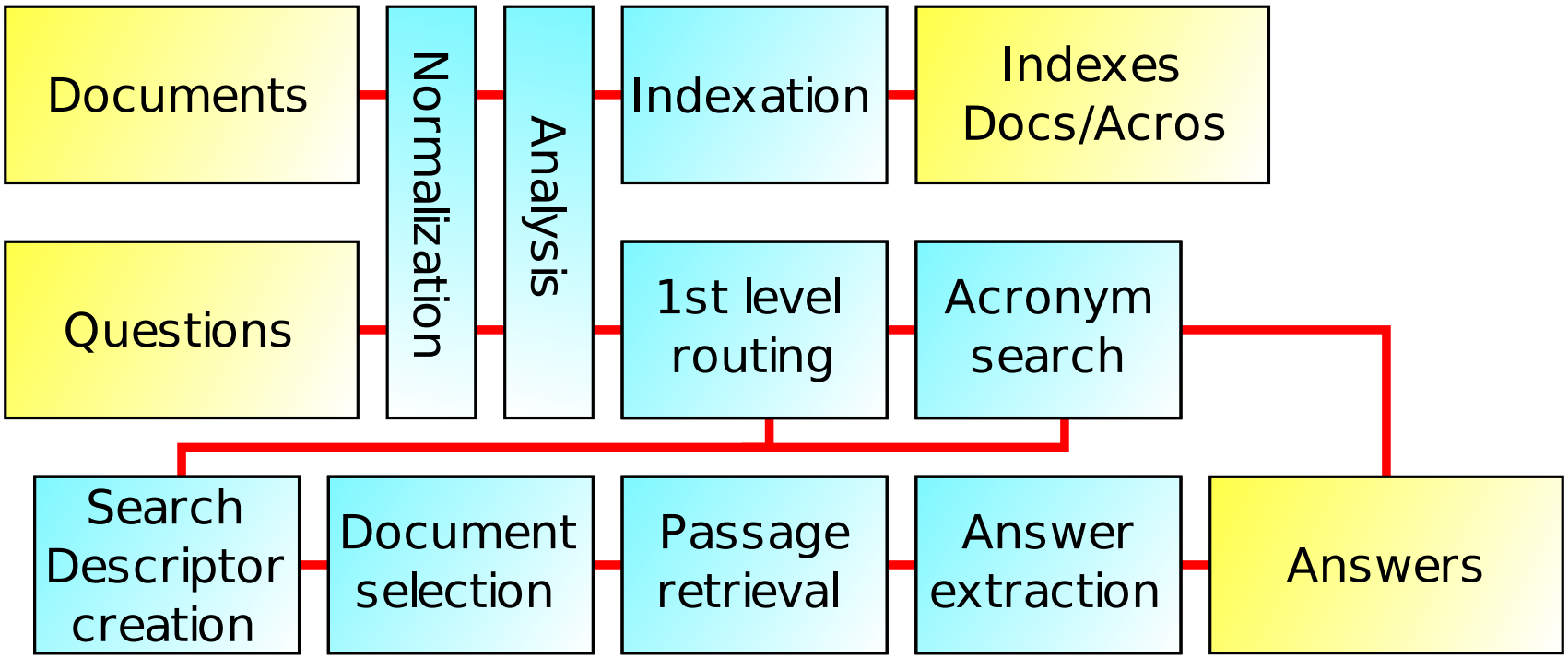
- Written and Spoken questions on:
  1. English european parliament (TCSTAR/EPPS): manual + 3 ASR
  2. Spanish european parliament (TCSTAR/EPPS): manual + 3 ASR
  3. French broadcast news (ESTER): manual + 3 ASR

LIMSI objectives

Test three different methods for answer scoring.

- Distance-based answer scoring: all tasks and sub-tasks (primary method)
- Answer scoring through bayesian modeling: English and Spanish, manually transcribed data collection
- Tree transformation-based answer re-ranking: French, manually transcribed data collection

System structure



Answer extraction and scoring

Distance-based answer scoring

- All elements of the appropriate types are candidates
- The candidates are scored using their distances to the SD elements, the snippet scores, their occurence counts
- Uses a set of tuning constants optimized by trials (dev data)

$$S(r) = \frac{\sum_{a \in A_r} (w(a) \max_{E_a} \sum_{(e,l) \in E_a} \frac{w(l)}{(1+d(e,a))^{\alpha}})^{1-\gamma} S_p(a)^{\gamma}}{C_d(r)^{\beta} C_p(r)^{\delta}}$$

$w(l)$ = line weight $\quad$ $w(a)$ = answer weight
$d(e, a)$ = element-answer distance
E_a = set of SD elements for instance a
A_r = set of instances of the answer candidate r
$S_p(a)$ = score of the snippet including a
$C_d(r)$ = instance count of r in the documents
$C_p(r)$ = instance count of r in the snippets
$\alpha, \beta, \gamma, \delta$ = tuning variables

Answer extraction and scoring

Answer scoring through bayesian modeling

- Compute the correctness probability of each candidate answer vs. all the other ones.
- Ends up as a mix of multiple sub-models:
  - Element presence probability in the presence of the correct answer or not.
  - Element co-occurence probability.
  - Out-of-context intrinsic answer probability.
- Some of the sub-models are very incorrect at that point, more studies need to be done.

Answer extraction and scoring

Tree transformation-based answer re-ranking

Baseline method based on redundancy, frequency, and distance. No use of structural information.

$\longrightarrow$ use trees produced by the multi-level analysis (and more).

$\longrightarrow$ rerank the candidate answers given a tree transformation cost

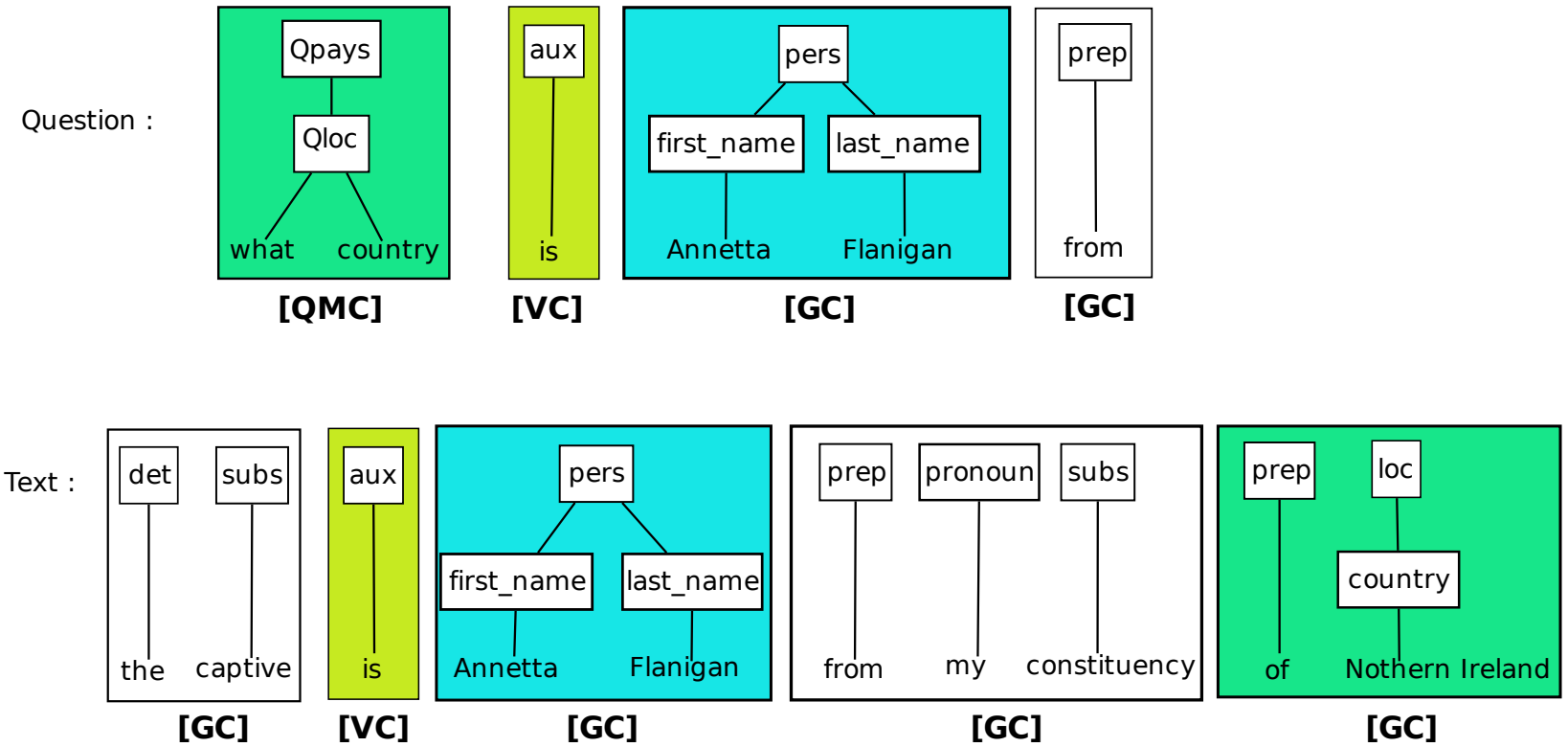
Three modules:

- Segmentation and annotation
- Relation labelling
- Text transformation cost estimation

Tree transformation-based answer re-ranking

Text transformation cost estimation

- Main objective: transformation of the snippet tree into the question tree+answer
- 3 operations: insertion, deletion and substitution



Overall results

		T1		T2		T3	
Sub-Task	**Question**	**Acc.**	**Best**	**Acc.**	**Best**	**Acc.**	**Best**
Manual	Written	27.0%	28.0%	36.0%	-	28.0%	-
	Spoken	23.0%	26.0%	36.0%	-	28.0%	-
ASR_A	Written	26.0%	-	27.0%	-	29.0%	-
	Spoken	25.0%	-	26.0%	-	29.0%	-
ASR_B	Written	21.0%	-	25.0%	-	27.0%	-
	Spoken	21.0%	-	25.0%	-	25.0%	-
ASR_C	Written	21.0%	25.0%	23.0%	-	23.0%	-
	Spoken	20.0%	25.0%	24.0%	-	22.0%	-

Comparison between different answer scoring

System	Questions	English			Spanish		
		MRR	Acc	Recall	MRR	Acc	Recall
Distance	Written	0.36	27%	53%	0.45	36.0%	61%
	Spoken	0.33	23%	45%	0.45	36.0%	62%
Bayesian	Written	0.32	23%	45%	0.34	24.0%	49%
	Spoken	0.27	19%	41%	0.34	24.0%	49%

System	Questions	French		
		MRR	Acc	Recall
Distance	Written	0.39	28.0%	60%
	Spoken	0.39	28.0%	59%
Tree	Written	0.38	27.0%	60%
	Spoken	0.39	28.0%	59%

Task difficulty evolution

Measure on French data

	Words		Nodes		Chunks	
	Mean	SD	Mean	SD	Mean	SD
Dev09 written	27	52	47	17	10	20
Dev09 spoken	28	52	47	22	10	20
Dev08	14	20	13	23	5	7

French system on dev08 obtains an accuracy of 60%, and 40% on dev09

Conclusions and perspectives

Conclusions

- Presentation of three differents methods
  - Distance-based method obtains the best results
- Significant loss between 2009 and 2008 evaluations
- More realistic and interesting task

Perspectives

- Work on tree-based transformation method
  - Work on relations costs
- Work on bayesian method
  - Work on sub-models