Full Machine Translation for Factoid Question Answering

Cristina España-Bonet and **Pere R. Comas**

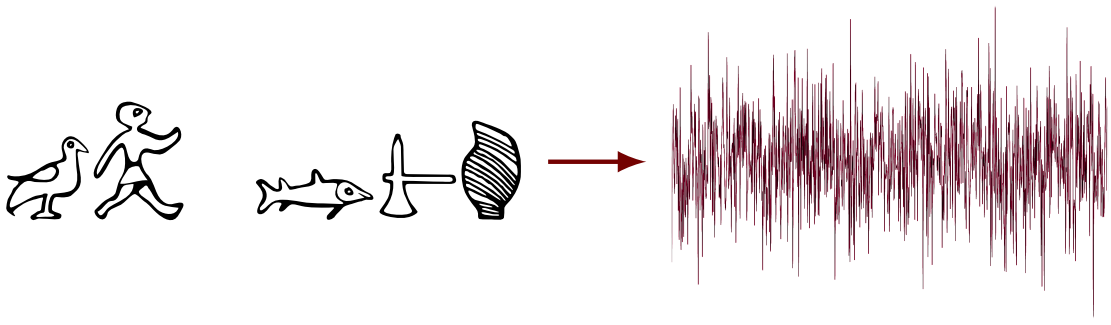
Universitat Politècnica de Catalunya
TALP Research Center

Joint ESIRMT and HyTra Workshop

Avignon, April 23rd, 2012

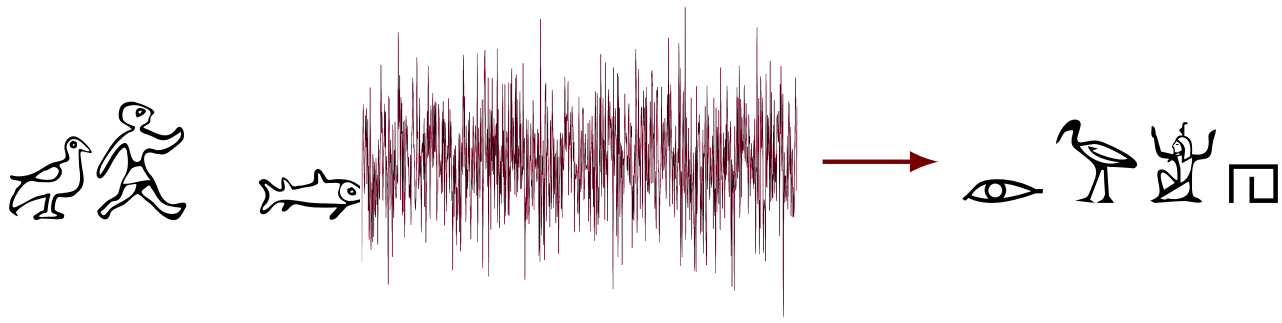
Full MT for factoid QA

The Noisy Channel



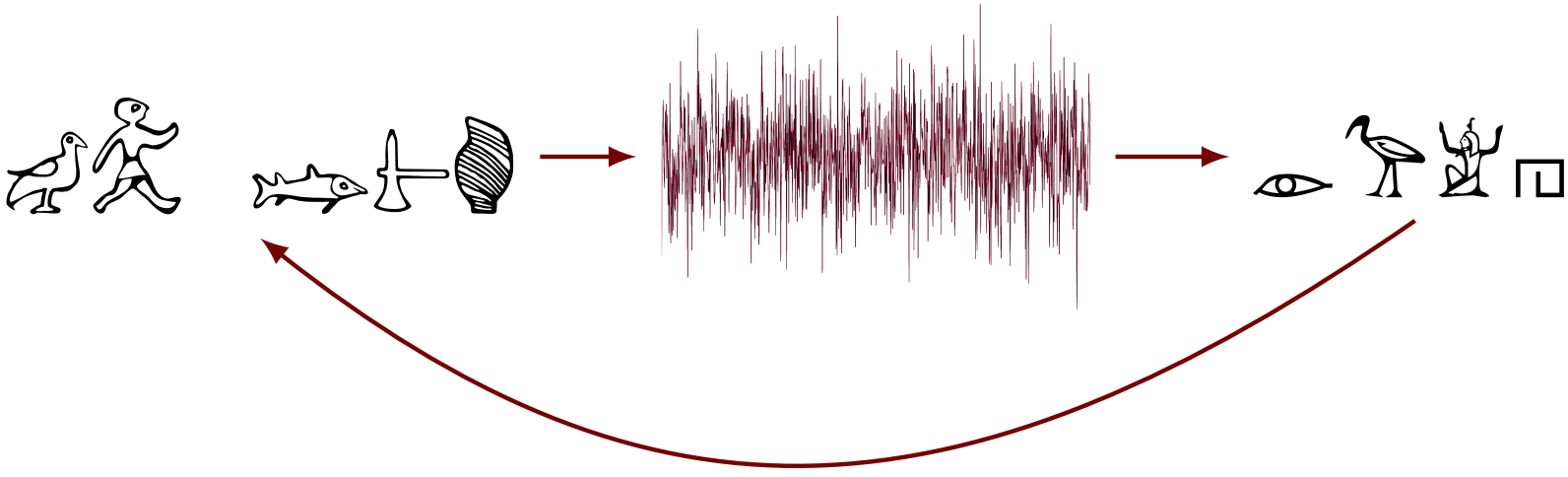
Full MT for factoid QA

The Noisy Channel



Full MT for factoid QA

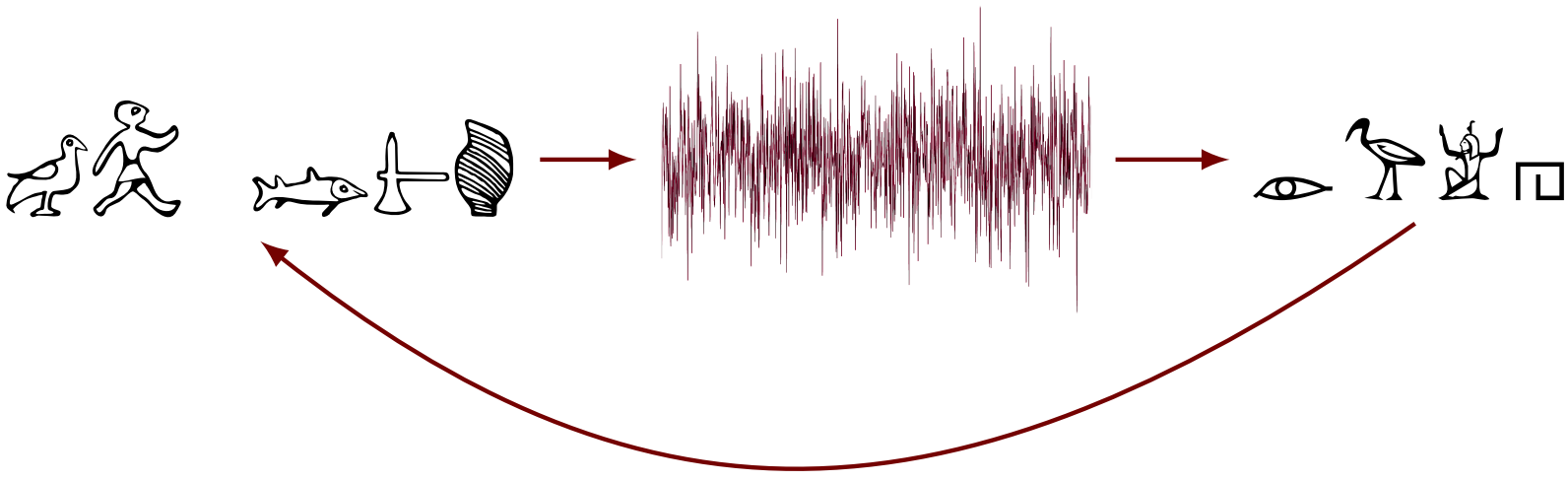
The Noisy Channel



Translation

Full MT for factoid QA

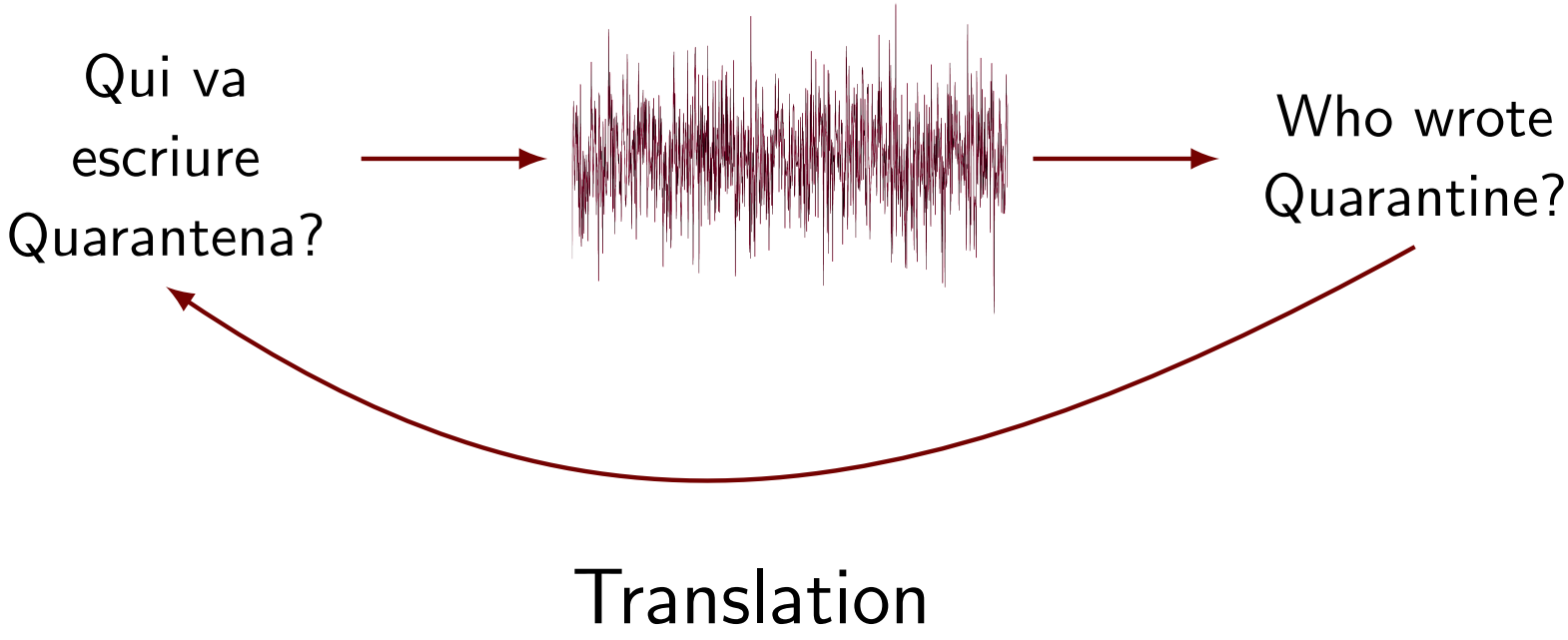
The Noisy Channel



Question Answering!

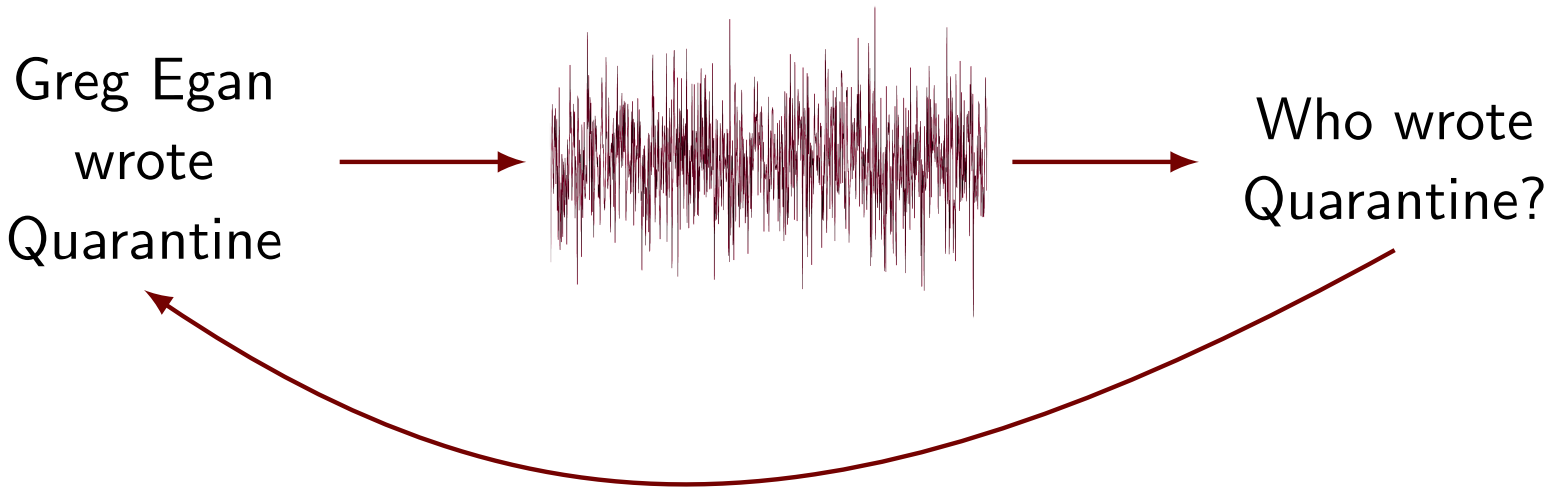
Full MT for factoid QA

The Noisy Channel



Full MT for factoid QA

The Noisy Channel



Question Answering

Full MT for factoid QA

Is it the same?

Mathematically,

$$P(O|I) = \frac{P(O)\,P(I|O)}{P(I)}$$

SMT:

$$\mathcal{T}(f) = \text{argmax}_{\text{e}}\ P(e|f) = \text{argmax}_{\text{e}}\ P(f|e)\,P(e)$$

QA:

$$\mathcal{A}(Q) = \text{argmax}_{A}\ P(A|Q) = \text{argmax}_{A}\ P(Q|A)\,P(A)$$

Full MT for factoid QA

Is it the same?

How old was Greg Egan when he wrote Quarantine?

SMT: Divide and conquer

```
How old ||| Quina edat ||| prob1
How old ||| Quants ||| prob2
How old ||| Quants anys ||| prob3
...
old ||| Quina edat ||| prob4
old ||| vell ||| prob5
old ||| gran ||| prob6
...
```

$\Longrightarrow$

Quina edat tenia en Greg Egan quan va escriure Quarantena?

Full MT for factoid QA

Is it the same?

How old was Greg Egan when he wrote Quarantine?

QA: Divide and conquer?

How old is Johnny Depp? Johnny Depp is 49

When was Quarantine writen? Quarantine was writen in 1992

When did he write his first novel? He published his first work in 1983

How old was Greg Egan when he wrote Permutation City? 33 years old

Full MT for factoid QA

Is it the same?

How old was Greg Egan when he wrote Quarantine?

QA: Divide and conquer?

How old is Johnny Depp? Johnny Depp is 49

When was Quarantine writen? Quarantine was writen in 1992

When did he write his first novel? He published his first work in 1983

How old was Greg Egan when he wrote Permutation City? 33 years old

Alignments depend on concrete questions

Full MT for factoid QA

Overview

1. The QA system
2. Experiments
3. Final thoughts

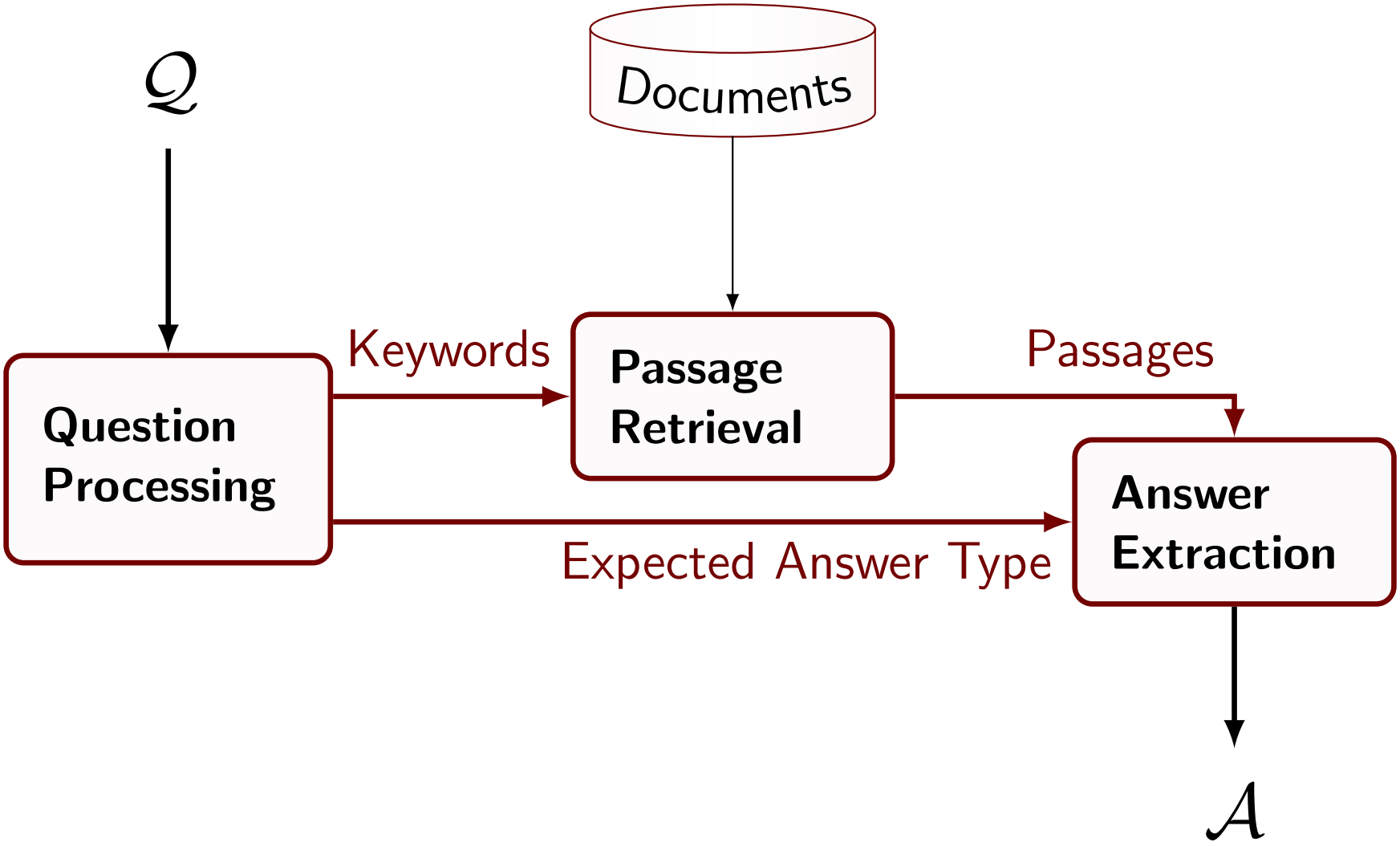
The QA system

Definition

Question Answering /ˈkwestʃən ˈɑːnsərɪŋ/ n. Task of extracting short, relevant textual answers from a given document collection in response to natural language questions.

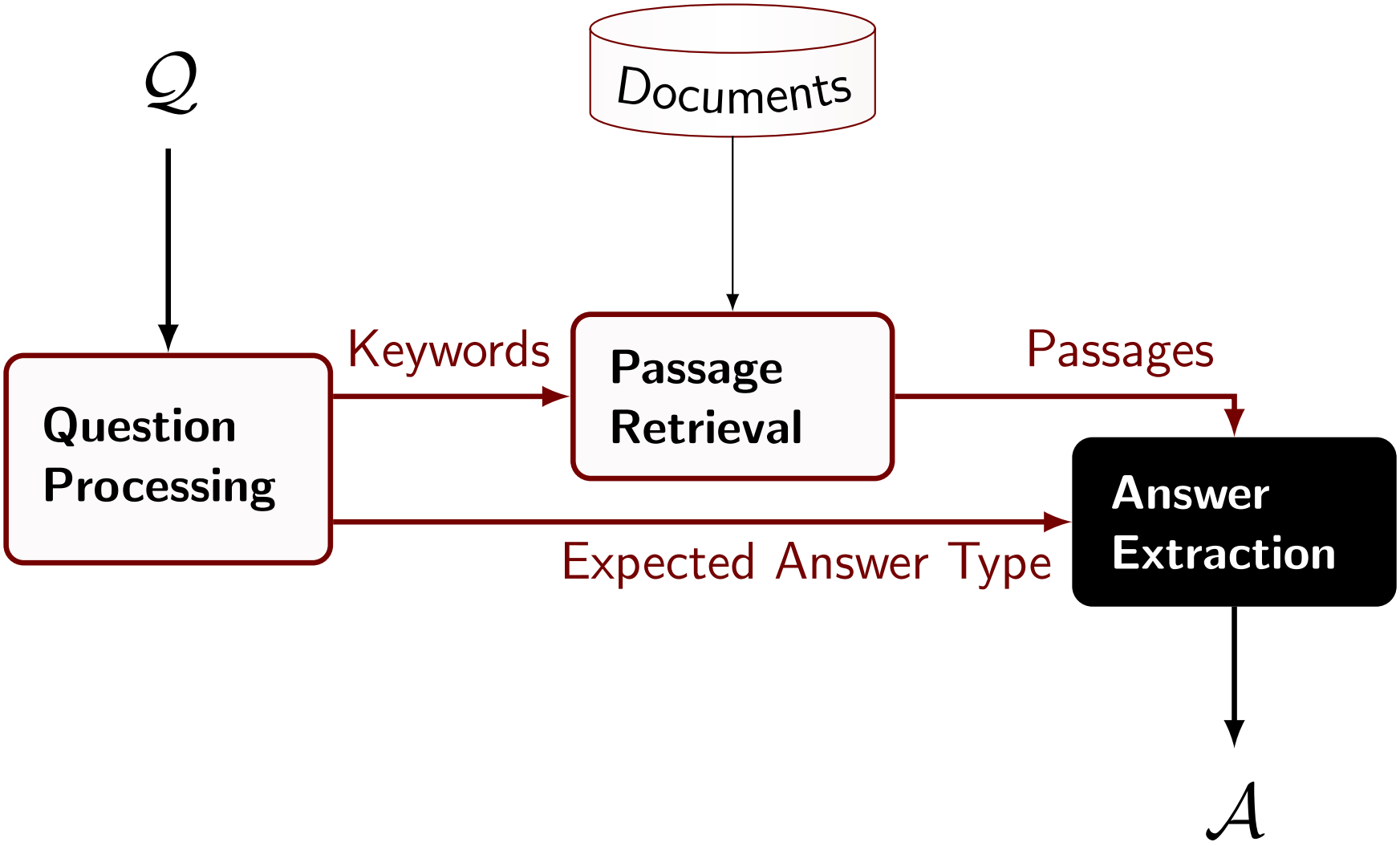
The QA system

Architecture



The QA system

SMT within the architecture

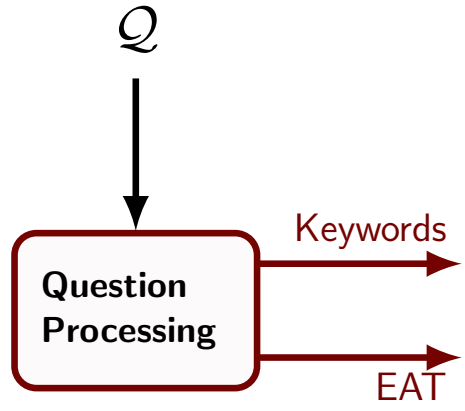


The QA system

The Question Processing Module

Question processing

Annotation with
PoS, chunks, NERC,
most frequent WordNet sense

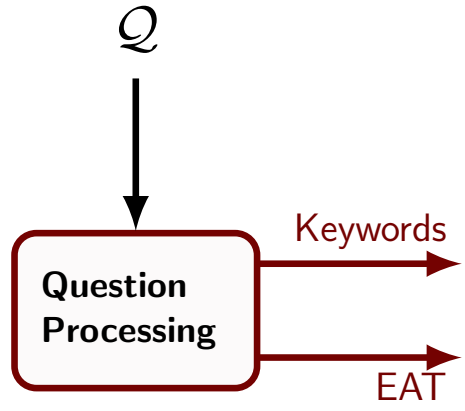


The QA system

The Question Processing Module

Question processing

Annotation with
PoS, chunks, NERC,
most frequent WordNet sense



Keywords extraction

According to the **salience** of words

Expected Answer Type

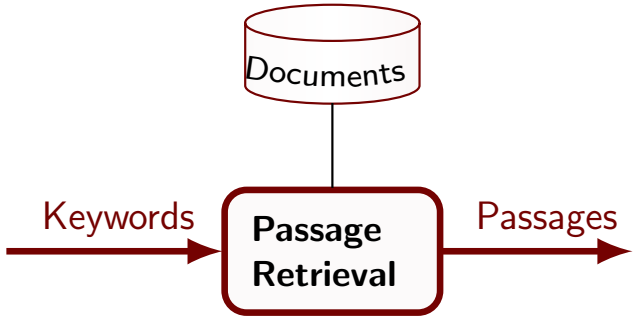
ME classifier for Li and Roth (2005) answer types

The QA system

The Passage Retrieval Module

Document retrieval

Keywords query with `Lucene`
IR engine



Passage building

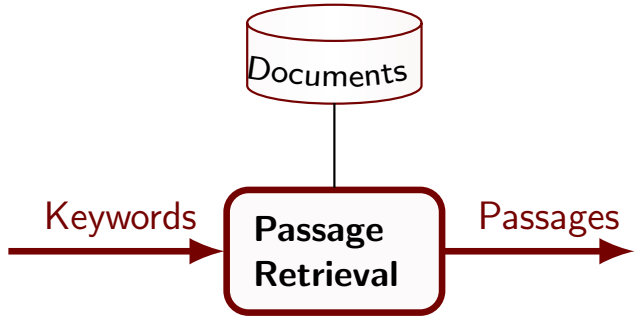
Segments with two keywords separated less than t words

The QA system

The Passage Retrieval Module

Document retrieval

Keywords query with `Lucene`
IR engine



Passage building

Segments with two keywords separated less than t words

Passages processing

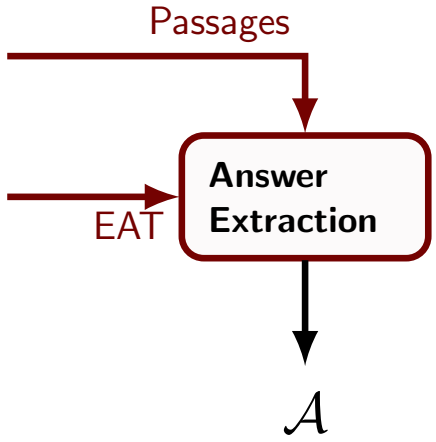
Split into sentences; annotate with PoS, chunks, NERC and most frequent WordNet sense

The QA system

The Answer Extraction Module

Answer candidates

NEs and phrases with a noun within the passages

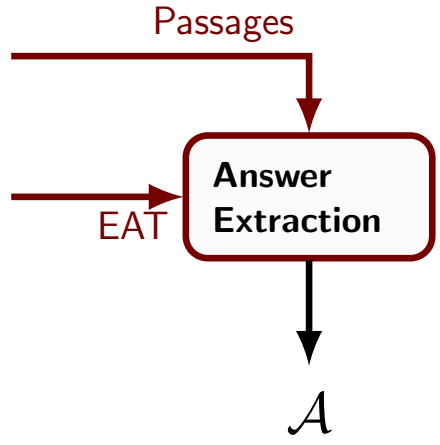


The QA system

The Answer Extraction Module

Answer candidates

NEs and phrases with a noun within the passages



Answer Ranking

Candidate answer sentences are ranked according to their similarity to SMT **Question-to-Answer translations**

Scoring

MT related scores; EAT score

The QA system

Question-to-Answer translation

Log-linear model (generalisation of the Noisy Channel)

$$\mathcal{A}(Q) = \text{argmax}_A \sum_m \lambda_m h_m(Q, A)$$

Level 1 Q: What is Karl Malone's nickname ?

Level 1 A: Malone , whose overall consistency has earned him the nickname `ANSWER` , missed both of them with nine seconds remaining .

The QA system

Question-to-Answer translation

Log-linear model (generalisation of the Noisy Channel)

$$\mathcal{A}(Q) = \text{argmax}_A \sum_m \lambda_m h_m(Q, A)$$

Level 1 Q: What is Karl Malone's nickname ?

Level 1 A: Malone , whose overall consistency has earned him the nickname ANSWER , missed both of them with nine seconds remaining .

The QA system

Question-to-Answer translation

Level 1 Q: What is Karl Malone's nickname ?

Level 1 A: Malone , whose overall consistency has earned him the nickname ANSWER , missed both of them with nine seconds remaining .

Context generalisation, patterns

Level 2 Q: What STATIVE B-PERSON 's COMMUNICATION ?

Level 2 A: B-PERSON , whose overall ATTRIBUTE POSSESSION POSSESSION him the COMMUNICATION ANSWER , PERCEPTION both of them with B-NUM TIME CHANGE

The QA system

Question-to-Answer translation

Level 1 Q: What is Karl Malone's nickname ?

Level 1 A: Malone , whose overall consistency has earned him the nickname ANSWER , missed both of them with nine seconds remaining .

Context generalisation, patterns for NEs

Level 2 Q: What STATIVE B-PERSON 's COMMUNICATION ?

Level 2 A: B-PERSON , whose overall ATTRIBUTE POSSESSION POSSESSION him the COMMUNICATION ANSWER , PERCEPTION both of them with B-NUM TIME CHANGE

The QA system

Question-to-Answer translation

Level 1 Q: What is Karl Malone's nickname ?

Level 1 A: Malone , whose overall consistency has earned him the nickname ANSWER , missed both of them with nine seconds remaining .

Context generalisation, patterns for verbs & nouns

Level 2 Q: What STATIVE B-PERSON 's COMMUNICATION ?

Level 2 A: B-PERSON , whose overall ATTRIBUTE POSSESSION POSSESSION him the COMMUNICATION ANSWER , PERCEPTION both of them with B-NUM TIME CHANGE

The QA system

Question-to-Answer translation

Level 1 Q: What is Karl Malone's nickname ?

Level 1 A: Malone , whose overall consistency has earned him the nickname ANSWER , missed both of them with nine seconds remaining .

Context generalisation, patterns for remaining words

Level 2 Q: What STATIVE B-PERSON 's COMMUNICATION ?

Level 2 A: B-PERSON , whose overall ATTRIBUTE POSSESSION POSSESSION him the COMMUNICATION ANSWER , PERCEPTION both of them with B-NUM TIME CHANGE

The QA system

Question-to-Answer translation

Answering... What is Karl Malone's nickname ?

The QA system

Question-to-Answer translation

Answering What is Karl Malone's nickname ?

Abstraction... What STATIVE B-PERSON 's COMMUNICATION ?

The QA system

Question-to-Answer translation

Answering What is Karl Malone's nickname ?

Abstraction What STATIVE B-PERSON 's COMMUNICATION ?

Translating... The B-ORGANIZATION B-LOCATION , B-DATE (B-ORGANIZATION) - B-PERSON , whose COMMUNICATION STATIVE " ANSWER . "

The QA system

Question-to-Answer translation

Answering What is Karl Malone's nickname ?

Abstraction What STATIVE B-PERSON 's COMMUNICATION ?

Translating...

1st best: The B-ORGANIZATION B-LOCATION , B-DATE (B-ORGANIZATION) - B-PERSON , whose COMMUNICATION STATIVE " ANSWER . "

...

50th best: The ANSWER ANSWER , B-DATE (B-ORGANIZATION) - B-PERSON , the PERSON of ANSWER , the most popular ARTIFACT , serenely COGNITION COMMUNICATION .

The QA system

Question-to-Answer translation

Translation/Answer

The B-ORGANIZATION B-LOCATION , B-DATE (B-ORGANIZATION) - B-PERSON , whose COMMUNICATION STATIVE " ANSWER . "

Not a real answer!

The QA system

Question-to-Answer translation

Translation/Answer

The B-ORGANIZATION B-LOCATION , B-DATE (B-ORGANIZATION) - B-PERSON , whose COMMUNICATION STATIVE " ANSWER . "

Not a real answer!

The ANSWER is found in the document collection:

- Search for the **most similar** candidate sentence obtained with the Passage Retrieval Module

The QA system

Answer scoring

Ranking of candidate answer sentences done by a **combination of scores**.

Context scores (B, R)

- *n*-gram matching metrics: BLEU & ROUGE
- Scores the similarity between translations and candidates in L2 representation

Are they enough?

The QA system

Answer scoring

Level 1 Qa: Where was C.S. Lewis born ?

Level 1 Qb: Where did Hans Christian Anderson die ?

Level 2 Qx: Where STATIVE PERSON STATIVE ?

The QA system

Answer scoring

Level 1 Qa: Where was C.S. Lewis born ?

Level 1 Qb: Where did Hans Christian Anderson die ?

Level 2 Qx: Where STATIVE PERSON STATIVE ?

Language scores (L_b, L_r, L_f)

- Similarity between translations and candidates in L1 representation (L_b, L_r)
- Scores candidate's words according to their frequency in the translations (L_f)

The QA system

Answer scoring

Level 2 Aa: The B-ORG B-LOCATION , B-DATE (B-ORGANIZATION) - B-PERSON , whose COMMUNICATION STATIVE ANSWER .

Level 2 Ab: The ANSWER B-LOCATION , B-DATE (B-ORGANIZATION) - B-PERSON , whose COMMUNICATION STATIVE B-PERSON .

The QA system

Answer scoring

Level 2 Aa: The B-ORG B-LOCATION , B-DATE (B-ORGANIZATION) - B-PERSON , whose COMMUNICATION STATIVE ANSWER .

Level 2 Ab: The ANSWER B-LOCATION , B-DATE (B-ORGANIZATION) - B-PERSON , whose COMMUNICATION STATIVE B-PERSON .

Expected Answer Type score (E)

- EAT mapped to NE or WN supersenses
- Candidates scored with the normalised probability of the ME classifier

Experiments

Overview

Experiments

Corpora

TREC evaluation campaigns [TREC9,TREC16]

Document collection. Newspapers
(Tipster, Acquaint, Acquaint2)

Question sets. Questions and answer keys

Experiments

Corpora

TREC evaluation campaigns [TREC9,TREC16]

Document collection. Newspapers
(Tipster, Acquaint, Acquaint2)

Question sets. Questions and answer keys

	Q	A	TRECs
Train	2264	12116	9,10,12,13,14,15,16
Dev	219	219	9,10,12
Test	500	2551	11

Experiments

SMT system

- **Language model**: 5-gram interpolated Kneser-Ney discounting, SRILM Toolkit
- **Alignments**: GIZA++ Toolkit
- **Translation model**: Moses package
- **Weights optimization**: MERT against BLEU
- **Decoder**: Moses

Experiments

SMT model

Characteristics

(experiments detailed in the paper)

- 8 standard **features**: $P(A)$, $lex(Q|A)$ and $lex(A|Q)$, $P_t(Q|A)$ and $P_t(A|Q)$, $P_d(A, Q)$, $ph(A)$ and $w(A)$
- 5-gram **language model**
- **100-best list** of translations

Experiments

QA system

- Applied to **Factoid questions**
- The **Question Analysis** module has been adapted from the QA system SIBYL (Comas, 2012)
- **Passage Retrieval** module (SIBYL)

  500 questions Q
  373,323 candidate answer sentences (747 per Q)
  2,866,098 candidate answers (5,732 per Q)

  Upper bound: 66,7%

Experiments

QA & SR results

	QA			SR		
Metric	T1	T50	MRR	T1	T50	MRR
B	0.018	0.292	0.049	0.084	0.540	0.164
R	0.018	0.283	0.045	0.119	0.608	0.209
B+R	0.022	0.294	0.053	0.097	0.573	0.180
BR	0.027	0.294	0.057	0.137	0.591	0.211
L_f	0.016	0.286	0.046	0.137	0.605	0.236
L_b	0.022	0.304	0.054	0.100	0.581	0.192
L_r	0.018	0.326	0.060	0.131	0.627	0.225
L_{brf}	0.038	0.330	0.079	0.147	0.622	0.238
E	0.044	0.373	0.096	0.058	0.579	0.142
EL_{brf}	0.018	0.293	0.048	0.118	0.623	0.214
BL_{brf}	0.051	0.337	0.091	0.184	0.616	0.271
RL_{brf}	0.033	0.346	0.069	0.191	0.618	0.279
BRL_{brf}	0.042	0.350	0.082	0.182	0.616	0.273
$(B+R)L_{brf}$	0.044	0.346	0.085	0.187	0.618	0.273
BE	0.035	0.384	0.084	0.086	0.579	0.179
RE	0.035	0.377	0.086	0.131	**0.630**	0.228
BRE	0.049	0.377	0.098	0.135	0.608	0.220
(B+R)E	0.040	**0.386**	0.091	0.102	0.596	0.196
BEL_{brf}	0.093	0.379	0.137	0.200	0.621	0.283
REL_{brf}	0.071	0.377	0.123	**0.208**	0.619	**0.294**
$BREL_{brf}$	0.091	0.379	0.132	0.200	0.622	0.287
$(B+R)EL_{brf}$	**0.100**	0.377	**0.141**	0.204	0.621	0.286

Experiments

QA & SR results

Buf!

	QA			SR		
Metric	T1	T50	MRR	T1	T50	MRR
B	0.018	0.292	0.049	0.084	0.540	0.164
R	0.018	0.283	0.045	0.119	0.608	0.209
B+R	0.022	0.294	0.053	0.097	0.573	0.180
BR	0.027	0.294	0.057	0.137	0.591	0.211
L_f	0.016	0.286	0.046	0.137	0.605	0.236
L_b	0.022	0.304	0.054	0.100	0.581	0.192
L_r	0.018	0.326	0.060	0.131	0.627	0.225
L_{brf}	0.038	0.330	0.079	0.147	0.622	0.238
E	0.044	0.373	0.096	0.058	0.579	0.142
EL_{brf}	0.018	0.293	0.048	0.118	0.623	0.214
BL_{brf}	0.051	0.337	0.091	0.184	0.616	0.271
RL_{brf}	0.033	0.346	0.069	0.191	0.618	0.279
BRL_{brf}	0.042	0.350	0.082	0.182	0.616	0.273
$(B+R)L_{brf}$	0.044	0.346	0.085	0.187	0.618	0.273
BE	0.035	0.384	0.084	0.086	0.579	0.179
RE	0.035	0.377	0.086	0.131	**0.630**	0.228
BRE	0.049	0.377	0.098	0.135	0.608	0.220
(B+R)E	0.040	**0.386**	0.091	0.102	0.596	0.196
BEL_{brf}	0.093	0.379	0.137	0.200	0.621	0.283
REL_{brf}	0.071	0.377	0.123	**0.208**	0.619	**0.294**
$BREL_{brf}$	0.091	0.379	0.132	0.200	0.622	0.287
$(B+R)EL_{brf}$	**0.100**	0.377	**0.141**	0.204	0.621	0.286

Experiments

QA & SR results

Buf!

Note that with QA, SR comes at the same price.

	QA			SR		
Metric	T1	T50	MRR	T1	T50	MRR
B	0.018	0.292	0.049	0.084	0.540	0.164
R	0.018	0.283	0.045	0.119	0.608	0.209
B+R	0.022	0.294	0.053	0.097	0.573	0.180
BR	0.027	0.294	0.057	0.137	0.591	0.211
L_f	0.016	0.286	0.046	0.137	0.605	0.236
L_b	0.022	0.304	0.054	0.100	0.581	0.192
L_r	0.018	0.326	0.060	0.131	0.627	0.225
L_{brf}	0.038	0.330	0.079	0.147	0.622	0.238
E	0.044	0.373	0.096	0.058	0.579	0.142
EL_{brf}	0.018	0.293	0.048	0.118	0.623	0.214
BL_{brf}	0.051	0.337	0.091	0.184	0.616	0.271
RL_{brf}	0.033	0.346	0.069	0.191	0.618	0.279
BRL_{brf}	0.042	0.350	0.082	0.182	0.616	0.273
$(B+R)L_{brf}$	0.044	0.346	0.085	0.187	0.618	0.273
BE	0.035	0.384	0.084	0.086	0.579	0.179
RE	0.035	0.377	0.086	0.131	**0.630**	0.228
BRE	0.049	0.377	0.098	0.135	0.608	0.220
(B+R)E	0.040	**0.386**	0.091	0.102	0.596	0.196
BEL_{brf}	0.093	0.379	0.137	0.200	0.621	0.283
REL_{brf}	0.071	0.377	0.123	**0.208**	0.619	**0.294**
$BREL_{brf}$	0.091	0.379	0.132	0.200	0.622	0.287
$(B+R)EL_{brf}$	**0.100**	0.377	**0.141**	0.204	0.621	0.286

Experiments

QA & SR results: comments

Individual metrics

Weak because:

- **B** and **R** do not take into account the lexical realisation
- $\mathbf{L}_x$ gives the same score to all candidates in the same sentence (better in SR)
- **E** gives the same score to all candidates of the same type (better in QA)

Experiments

QA & SR results: comments

Combination of metrics

	QA			SR		
Metric	**T1**	**T50**	**MRR**	**T1**	**T50**	**MRR**
BEL$_{brf}$	0.093	0.379	0.137	0.200	0.621	0.283
REL$_{brf}$	0.071	0.377	0.123	**0.208**	0.619	**0.294**
BREL$_{brf}$	0.091	0.379	0.132	0.200	0.622	0.287
(B+R)EL$_{brf}$	**0.100**	0.377	**0.141**	0.204	0.621	0.286

Experiments

QA & SR results: comments

Combination of metrics

	QA			SR		
Metric	**T1**	**T50**	**MRR**	**T1**	**T50**	**MRR**
BEL$_{brf}$	0.093	0.379	0.137	0.200	0.621	0.283
REL$_{brf}$	0.071	0.377	0.123	**0.208**	0.619	**0.294**
BREL$_{brf}$	0.091	0.379	0.132	0.200	**0.622**	0.287
(B+R)EL$_{brf}$	**0.100**	0.377	**0.141**	0.204	0.621	0.286

Final thoughts

Overview

Final thoughts

Comments, conclusions & summary

Approximation to QA as an MT problem

- T1 $\sim$ 10% is in the lowest part of TREC11 evaluation.
- Other approaches that use translation probabilities (Echihabi and Marcu, 2003) are better ranked.
- This approach is more similar to Ravichandran and Hovy (2002) who learn patterns to find answer contexts.

Final thoughts

Comments, conclusions & summary

Sentence retrieval as a complement to QA

- T50 in SR is close to the upper bound given by the Passage Retrieval module.
- **E** is not discriminative enough: T50 drops almost to a half in QA.

Final thoughts

Comments, conclusions & summary

Ranking the candidates, the key point

- Substitute **E** for an MT-based metric.
- Introduce new scoring metrics such as the score given by the decoder in translation.
- Improve the retrieval and therefore the upper bound (query expansion with SMT?).

Gràcies...

... i bon St. Jordi

Thank you!

Full Machine Translation for Factoid Question Answering

Cristina España-Bonet and **Pere R. Comas**

Universitat Politècnica de Catalunya
TALP Research Center

Joint ESIRMT and HyTra Workshop

Avignon, April 23rd, 2012

Extra slides

QA system: The Question Processing Module

Saliences

9 Words within quotes

8 Named entities

7 Sequences of nouns and adjectives

6 Sequences of nouns

5 Adjectives

4 Nouns

3 Verbs and adverbs

2 Question focus word

1 Any non-stop word

Extra slides

QA system: The Question Processing Module

Answer types, Li and Roth (2005)

ABBREVIATION:abb	ENTITY:other	LOCATION:mount
ABBREVIATION:exp	ENTITY:plant	LOCATION:other
DESCRIPTION:def	ENTITY:product	LOCATION:state
DESCRIPTION:desc	ENTITY:religion	NUMBER:code
DESCRIPTION:manner	ENTITY:sport	NUMBER:count
DESCRIPTION:reason	ENTITY:substance	NUMBER:date
ENTITY:animal	ENTITY:symbol	NUMBER:distance
ENTITY:body	ENTITY:techmeth	NUMBER:money
ENTITY:color	ENTITY:termeq	NUMBER:order
ENTITY:cremat	ENTITY:veh	NUMBER:other
ENTITY:currency	ENTITY:word	NUMBER:perc
ENTITY:dismed	HUMAN:description	NUMBER:period
ENTITY:event	HUMAN:group	NUMBER:speed
ENTITY:food	HUMAN:individual	NUMBER:temp
ENTITY:instrument	HUMAN:title	NUMBER:volsize
ENTITY:lang	LOCATION:city	NUMBER:weight
ENTITY:letter	LOCATION:country	

Extra slides

QA system: The Passage Retrieval Module

Passage building

Query Keywords: relevant, documents, process

Passage:

Extra slides

QA system: Question-to-Answer translation

The log-linear model

$$
\begin{aligned}
\mathcal{A}(Q) &= \hat{A} = \operatorname{argmax}_{\mathrm{A}} \log P(A|Q) = \\
&+ \lambda_{lm} \log P(A) + \lambda_d \log P_d(A, Q) \\
&+ \lambda_{lg} \log lex(Q|A) + \lambda_{ld} \log lex(A|Q) \\
&+ \lambda_g \log P_t(Q|A) + \lambda_d \log P_t(A|Q) \\
&+ \lambda_{ph} \log ph(A) + \lambda_w \log w(A)
\end{aligned}
$$