



# Generating Annotated Corpora for Reading Comprehension and Question Answering Evaluation

Tiphaine Dalmas    Jochen L. Leidner    Bonnie Webber

Claire Grover    Johan Bos

Institute for Communicating and Collaborative Systems  
School of Informatics, University of Edinburgh

Workshop Natural Language Processing for Question Answering  
Held at EACL 2003, Budapest, Hungary

# Initial Summary



1. We have developed a richly annotated corpus resource for Reading Comprehension (RC).
2. We show that the approach taken in the MITRE Deep Read system can be carried out simply on this corpus and produces similar results.
3. We are making this corpus available through the MITRE Corporation.



# Value of RC

- small corpora: controlled experiments using “deep” NLP
  - have an established “difficulty” correlated to human performance
- Charniak’s Ph.D. thesis (Charniak, 1972)
- **Deep Read** baseline (Hirschman et al., 1999)
- ANLP-NAACL workshop (2000)
- Summer workshop at the Johns Hopkins University (2000)

# The CBC Corpus / 1



- Original corpus developed by MITRE corporation based on news stories from [www.cbc4kids.ca](http://www.cbc4kids.ca)
- News stories (around 500 words each) on Canadian themes made accessible for 9-12 year-old children
- MITRE added to each of 125 stories:
  - set of 6-10 questions
  - human answers ranked for difficulty

# The CBC Corpus / 2



Tragedy Strikes a Northern Village  
January 4, 1998

( 1\_1 )The six hundred mostly Inuit residents of the northern Quebec village of Kangiqsuajuq had planned to bury the bodies of nine of their friends and children in a funeral this afternoon. ( 1\_2 )But the bad weather that resulted in their deaths has also delayed the funeral until Tuesday.

*What delayed the funeral of the those who were killed?*

Level 2

Human answer bad weather

Level 2

Human answer a blizzard

( 2\_1 )Kangiqsuajuq\* is about 1,500 kilometres north of Montreal, at the mouth of the George River on Ungava Bay.( 2\_2 )This region is known as Nunavik.

*How far is Kangiqsuajuq from Montreal?*

Level 1

Human answer 1,500 kilometres

# Types of Annotation



- **linguistic annotation** generated by off-the-shelf tools and integrated into layers
- **annotation for evaluation**
- **system results** available for comparison



# Linguistic Annotation

Stories, questions and answers all annotated

Sentence Boundaries

MXTERMINATOR

Tokenization

Penn tokenizer.sed, Tree-Tagger

Part-of Speech

MXPOST, Tree-Tagger

Lemmatizer

CASS “stemmer”, Tree-Tagger

Stemmer

Porter stemmer

Stop-Word

Deep Read

Syntactic Trees

Apple Pie Parser, Collins

Chunks

CASS chunk trees

Dependencies

Minipar relations, CASS tuples

Semantic Type

Wordnet

(Named Entity

LTG MUC-7)



# Evaluation Annotation

>> 1999-WF08-4-2

*What do the Kurdish separatists want?*

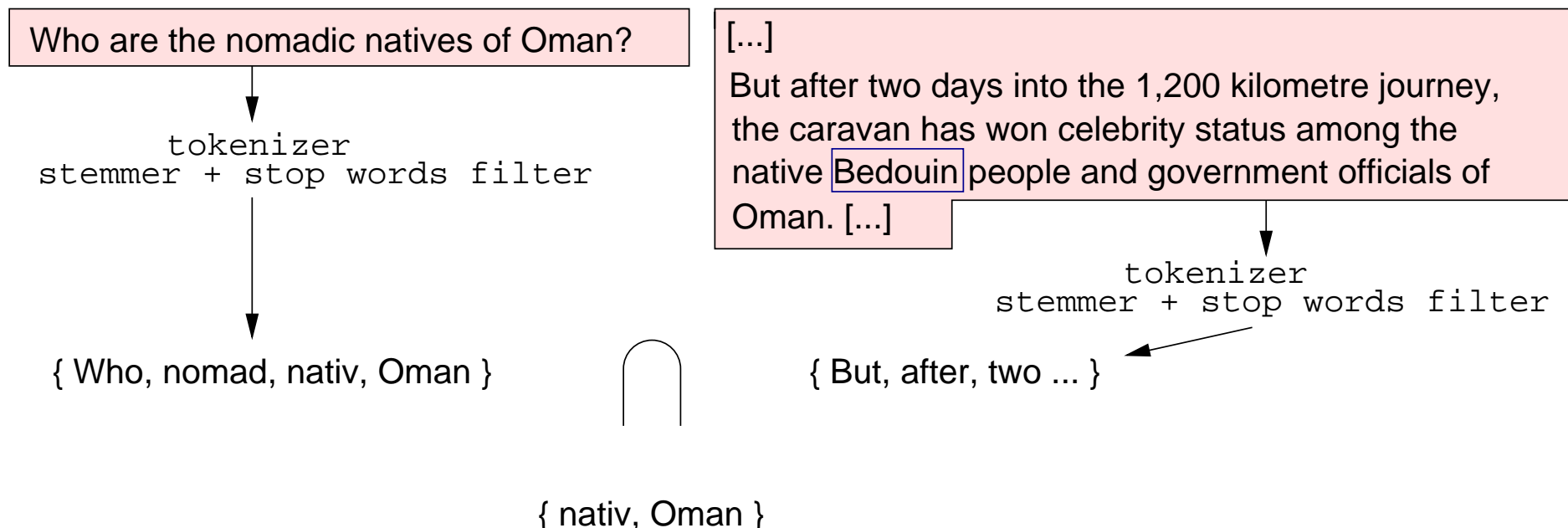
|              |   |
|--------------|---|
| Level        | 2   |
| Human answer | an independent Kurdistan  |
| AutSent      | Ocalan leads a separatist group that has fought for the creation of an independent Kurdistan. (recall: 1.0, id-ref:2_1 )                  |
| HumSent      | Ocalan leads a separatist group that has fought for the creation of an independent Kurdistan. (id-ref:2_1 )                               |
| Level        | 2   |
| Human answer | a separate Kurdish homeland   |
| AutSent      | Turkey considers Ocalan an enemy because he's been fighting an ongoing battle for a separate Kurdish homeland. (recall: 1.0, id-ref:7_1 ) |
| HumSent      | Turkey considers Ocalan an enemy because he's been fighting an ongoing battle for a separate Kurdish homeland. (id-ref:7_1 )              |





# Deep Read Approach / 1

- QA system based on stem overlap



- carried out on Remedia corpus
- component evaluation (stems, Wordnet ...)

# Deep Read Approach / 2



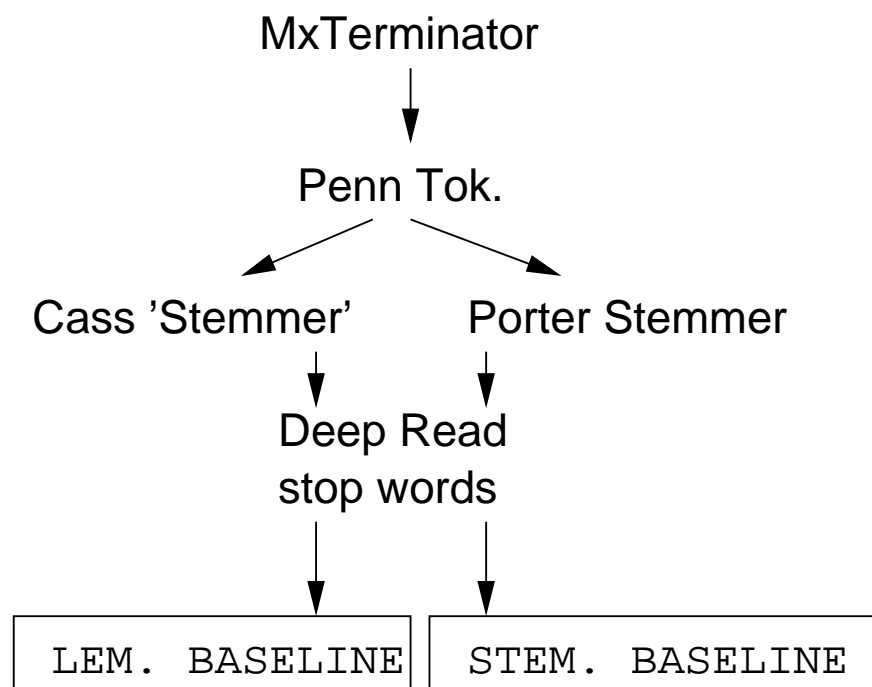
s : {But, after, ..., nativ, Bedouin, ...}  
ha : “the Bedouin” → {Bedouin}

- recall:  $|e_s \cap e_{ha}| / |e_{ha}|$
- precision:  $|e_s \cap e_{ha}| / |e_s|$
- humSent: *sentences considered as answers by a human annotator, either 1 or 0*
- autSent: *sentences for which recall against human answer is  $> 0$ , either 1 or 0*



# Experimental Results/1

A Deep Read approach on CBC corpus using stem and lemma overlap



# Experimental Results / 2



| Difficulty | QC  | R    | P    | AutSent | HumSent |
|------------|-----|------|------|---------|---------|
| Easy       | 237 | 0.74 | 0.18 | 0.75    | 0.74    |
| Moderate   | 177 | 0.57 | 0.22 | 0.55    | 0.57    |
| Difficult  | 67  | 0.49 | 0.19 | 0.43    | 0.43    |
| Average    | 481 | 0.63 | 0.19 | 0.62    | 0.63    |

- higher performance on CBC than Deep Read on Remedia (HumSent: 0.29, R: 0.29)
- within the theoretical bounds established for this corpus (Light et al., 2001)



# Overlap-Based QA

## ■ QA Process

1. Layer selection using XPath

```
/DOC/QA-SET/QA/Q/TOKENS/TOKEN[@process='LEMMA2_CLEMMMA2']  
/DOC/TEXT/P/S/TOKENS/TOKEN[@process='LEMMA2_CLEMMMA2']
```

2. Sets Intersection (overlap) and ranking

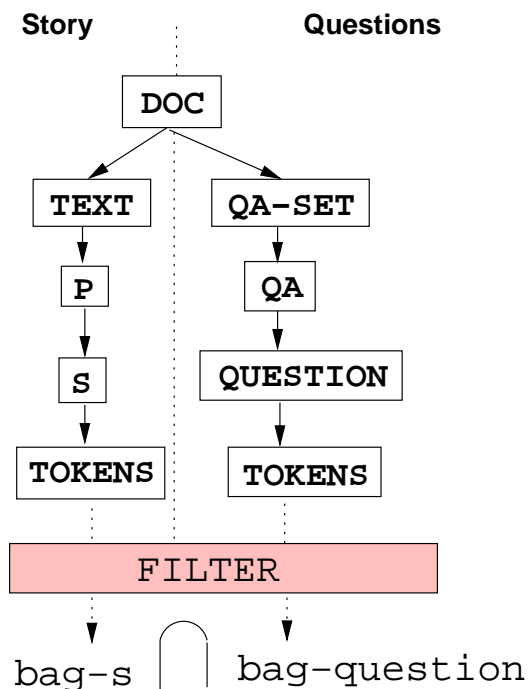
## ■ QA Evaluation

1. Evaluation layer selection

```
./ANSWERS/HUMAN-ANSWER/TOKENS/TOKEN[@process='LEMMA2_CLEMMMA2']  
./ANSWERS/AUT-SENTS/AUT-SENT  
./ANSWERS/HUM-SENTS/HUM-SENT
```

2. Metrics computation

3. Result integration



# Future Work



- Facilitated by existing annotation layers
  - overlap of chunks
- Requiring additional annotation layers
  - predicate-argument structure
  - answer comparison



# Conclusion

- reusable resource in XML
- allows for reformatting
- linguistic annotation and evaluation metrics
- result to be distributed to the community for free (for non-commercial purpose) by MITTRE (contact [Lisa Ferro, lferro@mitre.org](mailto:Lisa.Ferro@mitre.org))

## Acknowledgements:

Thanks to Lynette Hirschman and Lisa Ferro at MITRE  
Financial support of the German Academic Exchange Service (DAAD)  
under grant D/02/01831, of Linguist GmbH (research contract UK-2002/2),  
and of the School of Informatics, University of Edinburgh.

## References

- E. Charniak. 1972. *Toward a Model of Children's Story Comprehension*. Ph.D. thesis, Massachusetts Institute of Technology.
- L. Hirschman, M. Light, E. Breck, and J. D. Burger. 1999. Deep Read: A reading comprehension system. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*.
- M. Light, G. Mann, E. Riloff, and E. Breck. 2001. Analyses for elucidating current question answering technology. *Journal of Natural Language Engineering*, 7(4):325–342.