Università di Pisa

Active Learning for Building a Corpus of Questions for Parsing

Jordi Atserias, Yahoo! Research, Barcelona
Giuseppe Attardi, Università di Pisa
Maria Simi, Università di Pisa
Hugo Zaragoza, Yahoo! Research, Barcelona

LREC 2010, Malta – 19-21/05/2010

Summary

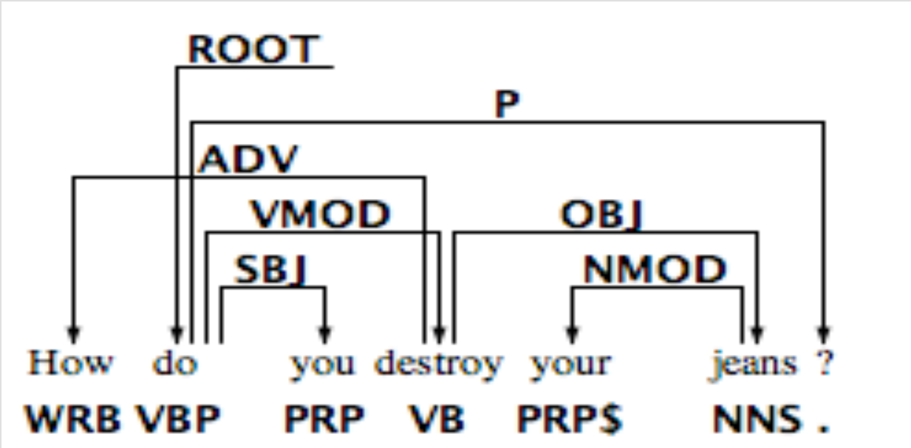
- Introduction and Goals
- Construction of a question corpus
- Experiments
  - Parsing questions / non questions
  - Smartest ways of building the corpus
    - Different criteria, batch size
    - “exploring” active learning
- Conclusions and Further

Motivations

- Accuracy in parsing questions is important
  - question answering, FAQ retrieval, dialogue systems ...
- Parsers have poor accuracy on questions
- No suitable question specific training resources are available

Need for an specific corpus

- **CoNLL 2007**
  - **only 0.75% are questions, not very representative**
  - **Annotations are sometimes inconsistent**
- **Questions have a specific structure**



Specific Motivation: Yahoo! Answers

- Several millions of questions collected from users, in several languages
- Yahoo! Answers Collection (Webscope)
  - 4,483,032 questions (and answers)
- Motivation: building a service for question retrieval (Yahoo! Quest available at http://quest.sandbox.yahoo.net)

English Question corpus

- **800 yahoo ! answers questions [relatively clean]**
- **500 questions from TREC QA**
- **PosTagged, revised and Parsed with DeSR, revised**

	Number of sentences	Average sentence length	Number of tokens
Yahoo! Answers Corpus	800	11.35	9,080
TREC QA Corpus	500	7.5	3,750
Question Corpus	1300	9.50	12,830

Active Learning for questions

- Research questions
  - Q1: how big a corpus of questions should be in order to achieve adequate accuracy?
  - Q2: Is a single corpus adequate to analyze both questions and non-questions?
  - Q3: Can we mimimize the cost of annotating the corpus?
- Active learning
  - supervised machine learning technique in which the learner is allowed to select the data
- Size of data samples
  - The smaller the set, the less efficient the process
  - Adding training data all at once, no benefit from AL

Experiment Set up

- **Question Corpus (12,830 tokens)**
  - **Divided into a *base train* and *base test* corpus**
- **Base corpus (250,805 tokens)**
  - **A sample of CoNLL 2007, without questions**
  - **Divided into a *base train* corpus and *base test* corpus**

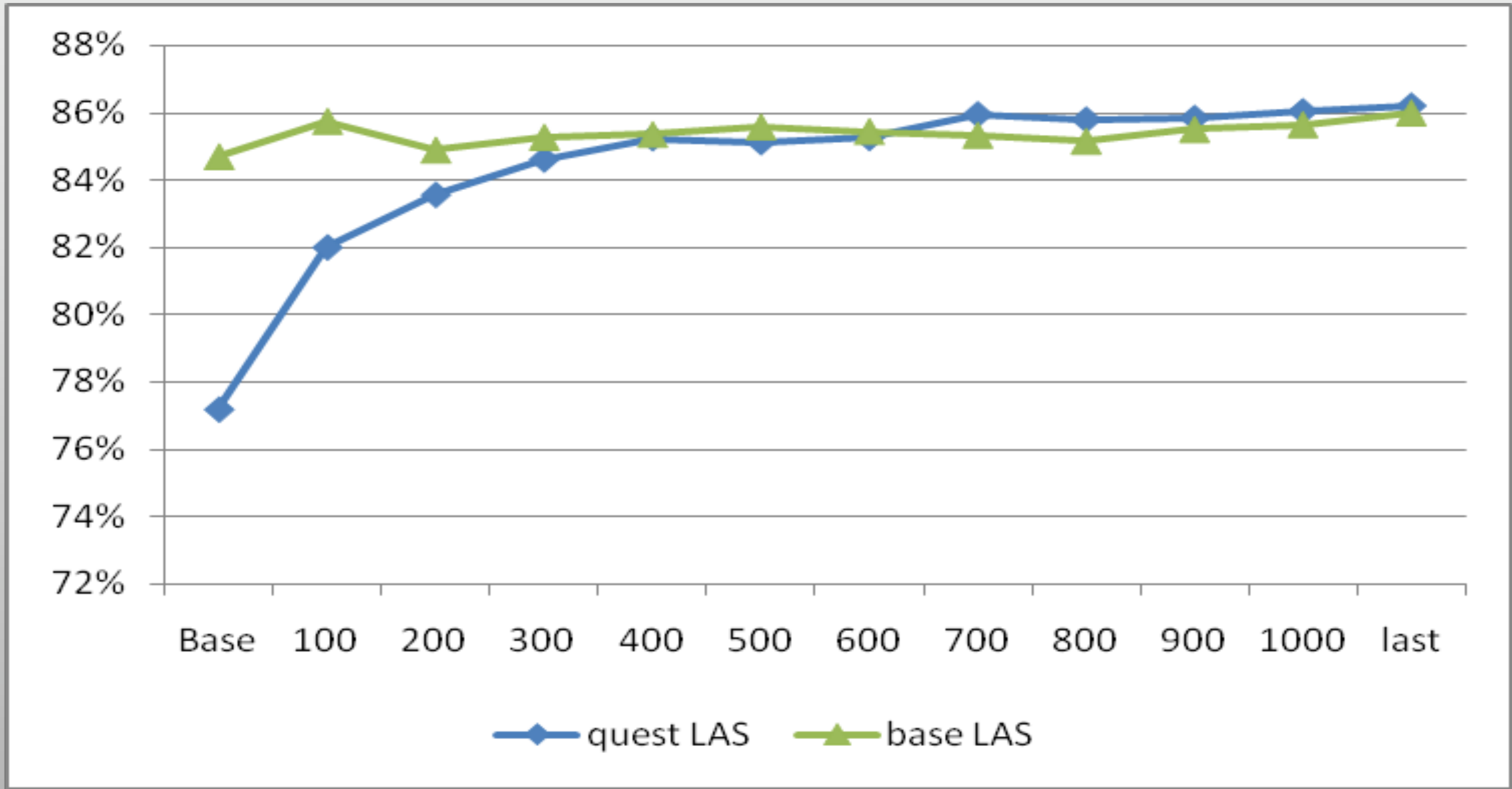
- **Baseline**
  - **Train on a corpus composed of the *base train* corpus plus random samples of questions of increasing size (0, 100, 200, 300 … 1000) extracted from the *question train* corpus**
  - **For each training corpus:**

    - evaluate on the *question test* (LAS score)

    - evaluate on the *base test* (LAS score)
  - **Repeat with different seeds (5 times), take the average LAS**

Q1 (size) & Q2 (helps and no harm) Random selection

	base	100	200	300	400	500	600	700	800	900	1000
quest LAS	**77.20%**	81.99%	**83.54%**	84.59%	85.22%	85.10%	85.23%	85.92%	85.77%	85.81%	**86.01%**
base LAS	**84.69**%	85.73%	**84.88%**	85.26%	85.34%	85.56%	85.43%	85.32%	85.15%	85.49%	**85.63%**



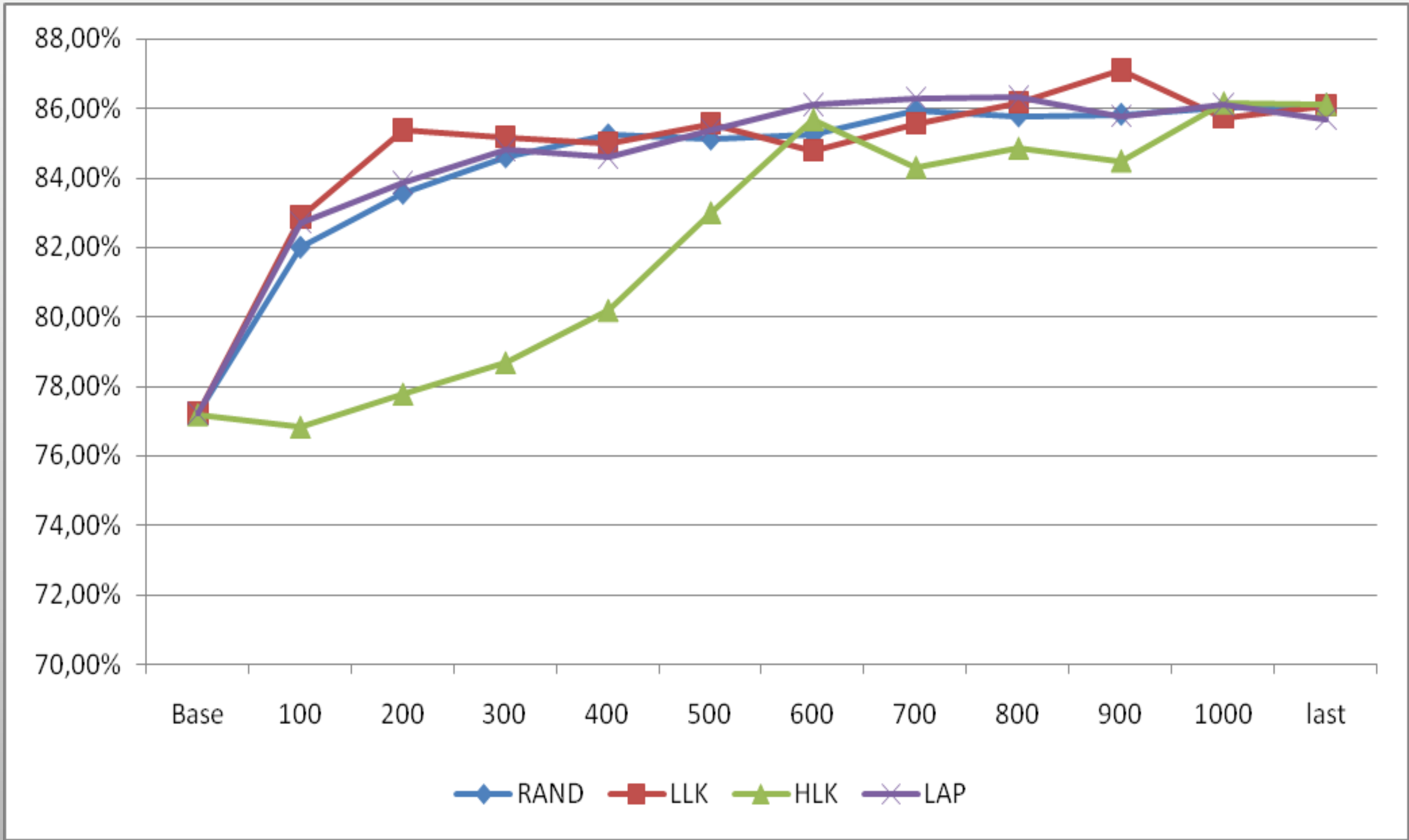
Q3: Can minimize annotation effort? Exploring Active learning

- Active learning is an iterative process
- At each step:
  - A learner is trained using the previous model
  - Using a "selection criterion" chooses "interesting" examples from a non-annotated collection (reparse the unannotated data)
  - Manually annotated and added to the training corpus
- If the selection criterion is effective, a much smaller number of examples is needed

Q3: Can minimize annotation effort? Testing selection criteria

- Selection criteria based on the output of the DeSR transition based parser
- *Likelihood* of sentence parse tree can be computed as the product of the probabilities of all parsing steps
  - *LLK: Lowest likelihood* of sentence parse tree
  - *HLK: Highest likelihood* of sentence parse tree
  - *LAP: Lowest average probability*
  - *LNL: Lowest normalized likelihood* (*likelihood/log*(#tokens)
- The sentences in the question training corpus were parsed and then ordered a priori with these criteria.
- Samples of increasing size were tested (as before)

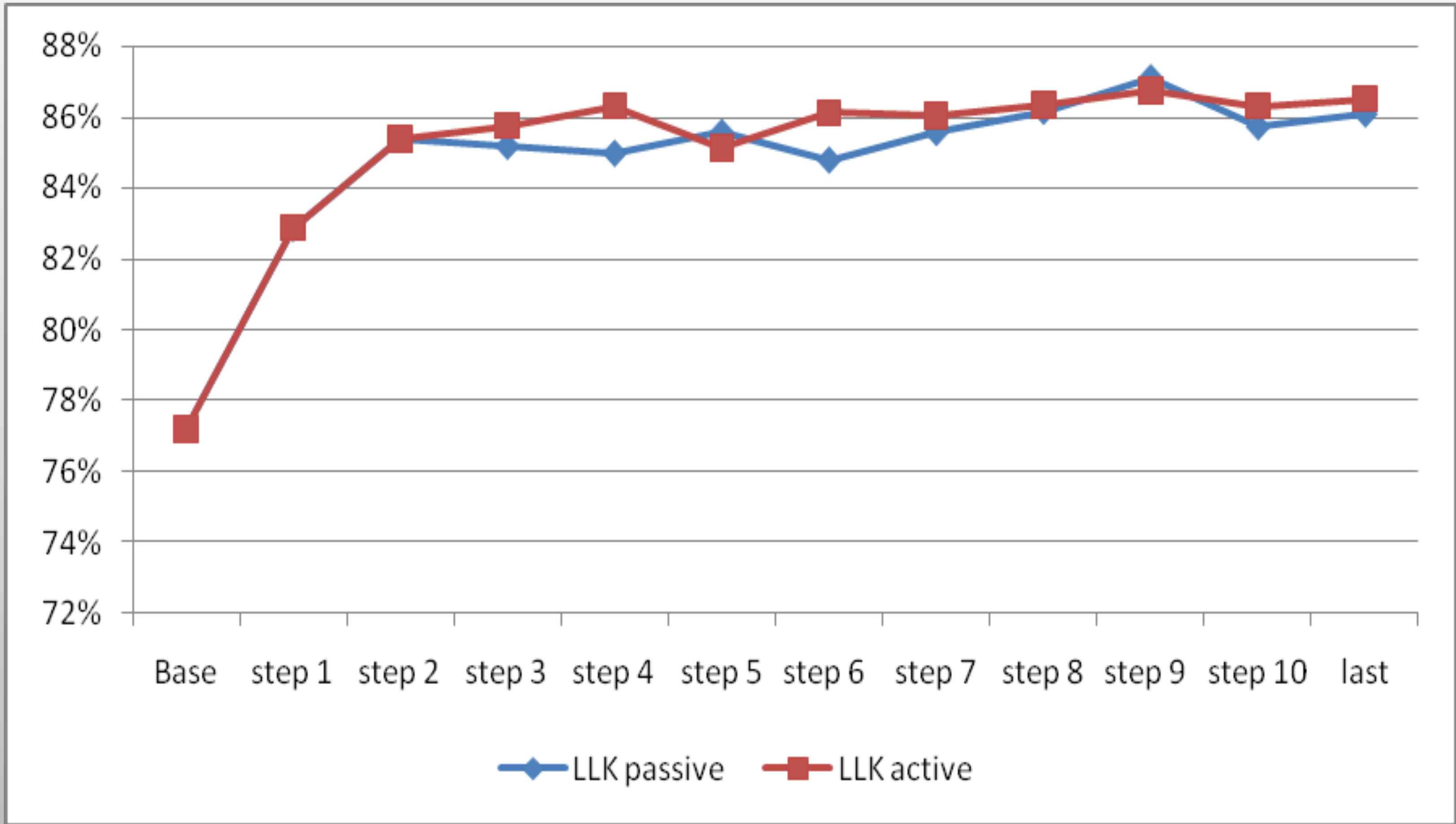
Random vs other criteria



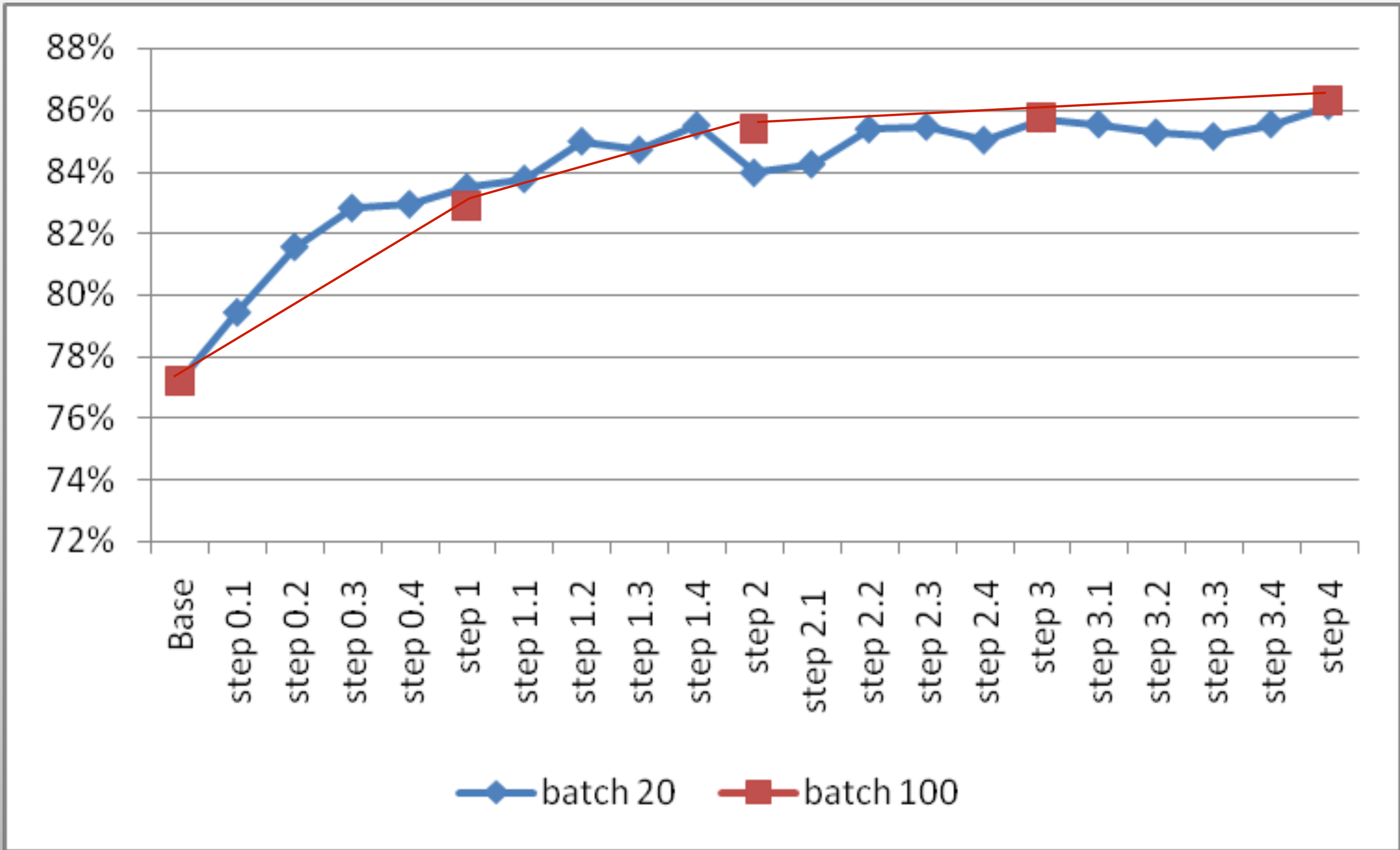
Evaluation of selection criteria

	base	100	200	300	400	500	600	700	800	900	1000
RAND	**77.20%**	81.99%	83.54%	84.59%	85.22%	85.10%	85.23%	85.92%	85.77%	85.81%	86.01%
LLK	**77.20%**	82.87%	**85.39%**	85.19%	84.99%	85.58%	84.80%	85.58%	86.18%	**87.12%**	85.74%
HLK	**77.20%**	76.84%	77.79%	78.69%	80.19%	82.99%	85.66%	84.29%	84.84%	84.48%	86.14%
LAP	**77.20%**	82.71%	83.85%	84.80%	84.60%	86.10%	86.29%	86.33%	85.78%	86.10%	85.70%
LNL	**77.20%**	82.20%	**85.47%**	85.35%	84.17%	85.66%	86.14%	85.19%	85.66%	85.98%	**86.92%**

Active vs passive



Smaller steps



Conclusions

- The corpus we have built can be useful for improving the accuracy of parsers in analysing questions
- With a relatively small corpus (about 1000 questions) quite good accuracy can be obtained in parsing questions without hurting the performance on non question sentences
- By using active learning we can further reduce the cost of building a question corpus

Future Work

- Building a larger corpus
- Try this approach on other languages
- Explore ML techniques that use unannotated data

Any Question?

Questions and feedback are highly welcome

Thanks for your attention