A UIMA-Based QA System (D3)

Chris Curtis, Nigel Kilmer, David McHugh

Approach

- UIMA text processing pipeline
  - DKPro suite of NLP modules
  - Custom query and answer processing modules
- MALLET for classification
- Indri for indexing and passage retrieval

UIMA

- **U**nstructured **I**nformation **M**anagement **A**rchitecture
- Provides structure for a pipeline of text processing components

- CAS - **C**ommon **A**nalysis **S**tructure
  - Stores original text and annotations (feature structures) produced by components
  - Annotations exist in the context of a type hierarchy

Question Classification

- Created UIMA wrapper for MALLET MaxEnt classifier
- Nine classes based on Li and Roth's taxonomy:
  - ABBR
  - ENTY
  - DESC
  - NUM
  - LOC
  - HUM
  - **DATE** **(DATE was a subclass of NUM)**
  - **MANNER** **(MANNER was a subclass of DESC)**
  - **REASON** **(REASON was a subclass of DESC)**

Question Classification

- On initialization, the system checks for a serialized classifier file to load
  - If it doesn't exist, the classifier is retrained on 5500 labeled questions from Li and Roth (2004)

- The only features used are unigrams, but results were fairly good

- The default MALLET stopword removal component removes question words, which led to low accuracy
  - Other than fixing this component, no other modifications to default MALLET preprocessing components were made

Question Classification

- Performed experiments with different classifiers and sets of classes

- Naive Bayes
  - Fine set: 0.520
  - Coarse (6 class) set: 0.748

- MaxEnt
  - Fine set: 0.766
  - Coarse (6 class) set: 0.844
  - Coarse (9 class) set: 0.862

Question Sets

- Anaphor resolution using context within question sets
  - Created a list of contextually relevant entities and their types
  - The list is cleared for each new question set

- Stanford NER DKPro module used to extract entities from question text
- Question classification combined with extracted answers

- For each anaphor in the question text, the most recently added entity that matches the anaphor's type is determined to be the referent
  - 'he' and 'she' are assumed to refer to human entities
  - An annotation on the anaphor is added to the CAS for use in later modules

Query Reformulation

- Question text is parsed using the Stanford factored parser DKPro module
  - Adds constituent and dependency annotations to the CAS

- Reformulates query string using techniques from MULDER (Kwok et al., 2001)
  - Subject-Aux movement
    - “was the Crip gang started” > “the Crip gang was started”
  - Subject-Verb movement
    - strip wh-words from queries
  - Verb Conversion
    - change aux + infinitive to conjugated form

Misc. Improvements

- New DKPro components integrated:
  - Chunker
  - Parser
  - NER

- Modified Indri queries to put more weight on named entities and chunks rather than unigrams

Issues

- DKPro Text Classifier module is still in development
  - Doesn’t support using a trained classifier to label new instances…
- Needed to write a fair amount of wrapper code to use MALLET as a UIMA component
  - Some duplicated components, since MALLET uses its own preprocessing pipeline
- Manipulating the parse annotations in the CAS ended up being tricky, so the query reformulation module is not yet functional
- Answer extraction is still being improved to take advantage of better techniques and information from the question classification and anaphor resolution modules

Results

- MRR scores computed on TREC 2006 test questions:
  - Strict: 0.0438 (D2: 0.0176)
  - Lenient: 0.0927 (D2: 0.0510)

- Main next step is to get improved answer extraction modules functioning and replace the baseline modules

References

Dan Klein and Christopher D. Manning. 2003. Fast exact inference with a factored model for natural language parsing. In *In Advances in Neural Information Processing Systems 15 (NIPS)*, pages 3–10. MIT Press.

Cody C. T. Kwok, Oren Etzioni, and Daniel S. Weld. 2001. Scaling question answering to the web. In *Proceedings of the 10th International Conference on World Wide Web*, WWW ’01, pages 150–161, New York, NY, USA. ACM.

Xin Li and Dan Roth. 2004. Learning question classifiers: The role of semantic information. In *In Natural Language Engineering*, pages 1–7, Taipei, Taiwan.

Andrew Kachites McCallum. 2002. Mallet: A machine learning for language toolkit. http://mallet.cs.umass.edu.