Milestone 1 Presentation

Team 3
Hector Liu
Da Teng
Guoqing Zheng
Xiawen Chu
Ryan Carlson

Pipeline Overview

1. Pre-process training data
2. Linking to external resources
3. Indexing and searching
4. Semantic and dependency parsing
5. Question Types
6. Answer Classifiers
7. Combining!

Pre-Processing Training Data

- Raw text is kind of a mess

- There’s useful information in there
  - but also a lot of not-so-useful stuff.

- Regexs, POS tagging to clean up data

- Tag with “relevant info”, “author”, “citation”, etc

Linking to External Resources

- External knowledge can be useful
  - If we know the background information of a term, we can know its type, its description and many other details
  - For example, we would like to know that “alzheimer” is a kind of disease
- We use DBpedia as an external resource
  - We might be more interested in Bio related resources in the future

Indexing & Searching

- Built our own Solr server;

- Annotations from previous phases can serve to construct better queries (new fields, new queries, etc.);

- Beyond the current implementation, we may also try to use higher order interactions from terms and alternative retrieval models to help retrieve better results from the index.

Semantic & Dependency Parsing

- Stanford CoreNLP:
    - StanfordCorenlpSentence: Sentences in docs
    - StanfordCorenlpToken: Tokens in docs
    - StanfordDependencyNode
    - StanfordDependencyRelation: dependency relation in sentence
    - StanfordEntityMention: different type of name entities
- FanseNLP:
    - FanseDependencyRelation:label the dependency relations between lexical items
    - FanseSemanticRelation: annotate basic semantic relations for each items
    - FanseTokenAnnotation: Annotate token with its depedency and semantic relations.

Question Types

- 5 question types
  - factoid, causal, method, purpose, true/false

- Train a model to identify these types given the text (and possibly the training data)
  - manually annotate (some of?) the ~300 examples
  - cross-validate to get a measure of generalizability
  - probably also need an i_have_no_idea tag

- Also annotate with NOT where appropriate

Answer Classifiers

- Given a question type, use the training data and our data type representation to select an answer

(very open question right now)

Putting it all together

- We might have several methods for answering any given question, need some mechanism to combine them

- Linearly interpolate system values (based on confidences?)

No PMI Baseline

- We focus more on clean up the current annotations
- We set up the original baseline (voter) provided using the cleaned text
  - c@1 score:0.22
  - c@1 score:0.11000000000000001
  - c@1 score:0.11000000000000001
  - c@1 score:0.3
  - Avg : 0.1925
- We are not sure about the performance, probably due to the mixture of our annotation with the old one

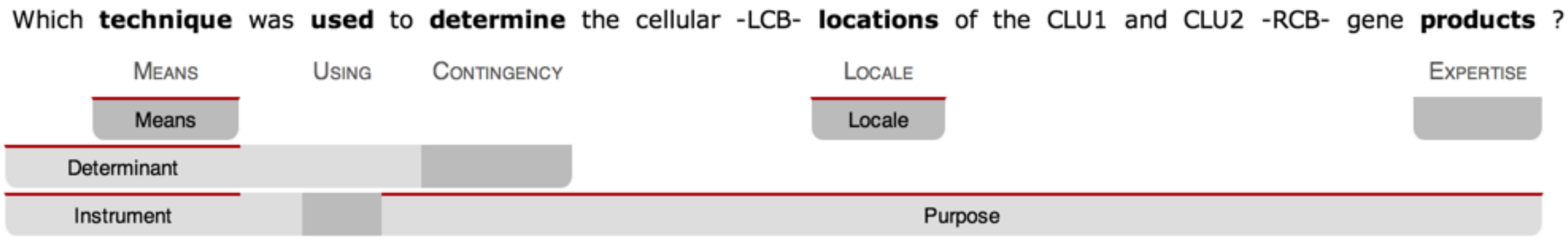
PMI Baseline

- We also tried a baseline using the PMI, this time we achieve something higher than random
  - c@1 score:0.22
  - c@1 score:0.55
  - c@1 score:0.33
  - c@1 score:0.2
  - Avg : 0.325
- We will use this as the baseline to beat and build our system on top of it.

Future Plan

- Semantic based re-ranking
  - Question: Which technique was used to determine the cellular

    CLU1 and CLU2 gene products?



Which **technique** was **used** to **determine** the cellular -LCB- **locations** of the CLU1 and CLU2 -RCB- gene **products** ?

  - Answer sentence: {immunofluorescence and Western blot studies:Answer Phrase:Arg0} {indicate:Answer Head} that {CLU1 and CLU2:Arg0-clause} both {produce:clause Head} secreted proteins that are similar to those detected {in the human brain:ArgM-Loc}
- Two step approach
  - Ranking first, semantic for reranking

Thanks!

Questions?

Comments?

Suggestions?

Concerns?

Gripes?