Report for IE Assignment

Longfei Zhao, u5976992

Q2. Named Entity Recognition baselines

1. Rule-based: use some language rule to do that. For example, words ending with 'me' in Spanish are person or words ending with 'estado' are verb.
2. mechine learning: we could use knn to deal with train set. After that, for any word, we will know its same class words. We choose the most frequent label.

Q3. HMM

- Hidden state: summary, non-summary
- obversation: we could use some features as obversations, such as,
    - Position of the sentence in the doc
    - Number of terms in the sentence
    - Likelihood of the sentence terms given the document terms

Q4. Automatic summarization

Type of summary: Multi-document summary. It's best for Multi-document summarization.
Relevance method: Hybrid. It Combines different methods together, which is more valuable.

1. Choose an compression parameter. The compression ratio for Multi-document summarization will typically be much smaller than for single document summarization. For example, for 200 documents, 1% or 0.1% may be better.
2. Segment the documents into sentences.
3. Assign to each sentence a score using relevance method.
4. Discard sentences which below threshold. Notice that threshold should be related with compression parameter. It could aviod sentences similarity metric too large.
5. Compute sentence similarity metric which contains each pair of sentence similarity.
6. Depending on the compression parameter, to select sentences. Notice we need to balance relevance score and sentence similarity. Sentences' relevance scores should be high, meanwhile their similarities should be low. We could use some math functions to combine them as one single formula then to optimize it.