Passage Extraction : Overview

The document is a sequence generated by two language models: relevance and background. A hidden stochastic process determines the switch between relevance and background models.

A hidden process generates a sequence of transitions. This sequence maps the boundaries between document segments.
Domain Adaptation & Passage Extraction
INPUT, OUTPUT, ACCESS

• Inputs
  o Keywords from hierarchy of the taxonomy
    ▪ Risk Factors is an element that is a text block
    ▪ Riskfactors has many sub elements in the hierarchy
  o .html/.txt financial statement

• Output
  o Passage of text from html document to compare to xbrl

• Access
  o downloaded data to local machine (still need to upload to patas)
Previous Work

• Passage extraction is only mentioned in passing generally for Q&A and bioinformatics

My Approach

Using the xbrl schema provided by GAAP for commercial & industry and Yeti and I discovered keywords used to assess risk.

Running a subset of the documents through an hmm, I expanded the query terms across the domain.

Todo: run the model against insurance and/or banking and compare the results.
Yeti: Gathering Keywords
Passage Extraction: Overview

The document is a sequence generated by two language models: relevance and background.

A hidden stochastic process determines the switch between relevance and background models.

A hidden process generates a sequence of transitions. This sequence maps the boundaries between document segments.
Passage Extraction: HMM Structure

B* background states
R relevant state
B2 automatic tuning

Fig. 3. Final HMM structure.
Passage Extraction: Auto-smoothing

Fig. 4. Within-document pseudo-feedback.
Passage Extraction: Algorithm

Estimate Background Model

\[ p(w_i|B) = p(w_i|C) = \frac{c(w_i, C)}{\sum_{j=1}^{m} c(w_j, C)} \]

Estimate Relevance Model

(q, wd, cd)

\[ p(w_i|Q) = \frac{c(w_i, q)}{\sum_{j=1}^{m} c(w_j, q)} \]

\[ p(w_i|R) = \frac{c(w_i, p)}{\sum_{j=1}^{m} c(w_j, p)} \]

\[ p(w_i|Q') = \frac{\sum_{k=1}^{l} c(w_i, p_k)}{\sum_{k=1}^{l} \sum_{j=1}^{m} c(w_j, p_k)} \]

p = short relevant passage generated by q

Train transition probabilities (Baum-Welch)

Find most likely state sequence (Viterbi)
The Datasets and Baselines

• Datasets:
  o 2012 QTR 3 10-K financial statements
  o 2011 QTR 3 10-K financial statements

• Baselines
  o Fixed window 300 tokens
Experimental Results

TBD -- I'm still working on the baselines.

Currently – I’ve downloaded qtr 3 2012 and qtr 3 2011 10-K forms (annual report). The html has been stripped from the html and it has been tokenized by white space.

In addition to xdocument, I am extrapolating to xdomain.
Conclusion & Future Direction

• It is very difficult to create data. Most of my time has been spent scraping/cleaning data.

• Passage extraction is a useful method to create substrata used to classify a document along different dimensions (i.e., risk (text) vs. earnings(numbers) vs. performance over the years).

• Future direction is building new domains for documents based on their text (instead of extrinsic factors).