Information Extraction & Domain Adaptation

A Tiered Approach
Overview

Numbers & Meaning:
- Building semantic financial statements
- XBRL (now required by SEC in many cases)

Information Retrieval (IR)
- Passage Extraction
- IE

Domain Adaptation
- Domains & Taxonomies
- Evaluation

Project Status
Problem

The analysis & comparison of financial statements is hindered by the unstructured, non-standard nature of the filings.

Given a filing, an analyst desires to compare that filing to:

● the same company and document type in other years
● other companies in the same industry and same document type in the same year
● companies in other industries and the same document type in the same year
Solution

XBRL

eXtensible Business Reporting Language


- **taxonomy:**
  - `xml-schema`: data definition
  - `linkbases`: data relationships
  - defines concepts (element definitions manifest)

- **instance:**
  - discoverable taxonomy set
  - `context`: facts, units, measures, periods, entities, scenarios
  - `fact`: tuples, items
  - `data`
<?xml version="1.0" encoding="utf-8"?>
<schema
xmlns="http://www.w3.org/2001/XMLSchema"
xmlns:xbrli="http://www.xbrl.org/2003/instance"
xmlns:link="http://www.xbrl.org/2003/linkbase"
xmlns:xlink="http://www.w3.org/1999/xlink"
xmlns:samp="http://www.iqinfo.com/xbrl/taxonomy"
targetNamespace="http://www.iqinfo.com/xbrl/taxonomy"
elementFormDefault="qualified"
attributeFormDefault="unqualified">
  <annotation>
    <appinfo>
      <link:linkbaseRef xlink:type='simple'
xlink:role='http://www.xbrl.org/2003/role/presentationLinkbaseRef'
xlink:arcrole='http://www.w3.org/1999/xlink/properties/linkbase' />
      <link:linkbaseRef xlink:type='simple'
xlink:role='http://www.xbrl.org/2003/role/presentationLinkbaseRef'
xlink:arcrole='http://www.w3.org/1999/xlink/properties/linkbase' />
      <link:linkbaseRef xlink:type='simple'
xlink:role='http://www.xbrl.org/2003/role/calculationLinkbaseRef'
xlink:arcrole='http://www.w3.org/1999/xlink/properties/linkbase' />
      <link:linkbaseRef xlink:type='simple'
xlink:role='http://www.xbrl.org/2003/role/calculationLinkbaseRef'
xlink:arcrole='http://www.w3.org/1999/xlink/properties/linkbase' />
    </appinfo>
  </annotation>
  <import namespace="http://www.xbrl.org/2003/instance"
</schema>
Taxonomy : Domain

- USGAAP stub
- table version IFRS
- GAAP schema

There are several USGAAP taxonomies based on industry type and or year. Each taxonomy type represents a domain for the purposes of this project.

Given a specific taxonomy and labelled instances adapt to a new taxonomy and minimal labelled instances such that precision and recall are improved.
<?xml version="1.0" encoding="UTF-8"?>
<xbrli:xbrl
xmlns:iso4217="http://www.xbrl.org/2003/iso4217"
xmlns:xbrli="http://www.xbrl.org/2003/instance"
xmlns:xbrll="http://www.xbrl.org/2003/linkbase"
xmlns:xlink="http://www.w3.org/1999/xlink">
  <ifrs-gp:OtherOperatingIncomeTotalFinancialInstitutions contextRef="J2004" decimals="0" unitRef="EUR">38679000000</ifrs-gp:OtherOperatingIncomeTotalFinancialInstitutions>
  <ifrs-gp:OtherAdministrativeExpenses contextRef="J2004" decimals="0" unitRef="EUR">35996000000</ifrs-gp:OtherAdministrativeExpenses>
  <ifrs-gp:OtherOperatingExpenses contextRef="J2004" decimals="0" unitRef="EUR">870000000</ifrs-gp:OtherOperatingExpenses>
  ...
  <ifrs-gp:OtherOperatingIncomeTotalByNature contextRef="J2004" decimals="0" unitRef="EUR">870000000</ifrs-gp:OtherOperatingIncomeTotalByNature>
  <xbrli:context id="BJ2004">
    <xbrli:entity>
      <xbrli:identifier scheme="www.iqinfo.com/xbrl">ACME</xbrli:identifier>
    </xbrli:entity>
    <xbrli:period><xbrli:instant>2004-01-01</xbrli:instant></xbrli:period>
  </xbrli:context>
  <xbrli:context id="EJ2004">
    <xbrli:entity>
      <xbrli:identifier scheme="www.iqinfo.com/xbrl">ACME</xbrli:identifier>
    </xbrli:entity>
    <xbrli:period><xbrli:instant>2004-12-31</xbrli:instant></xbrli:period>
  </xbrli:context>
  <xbrli:context id="J2004">
    <xbrli:entity>
      <xbrli:identifier scheme="www.iqinfo.com/xbrl">ACME</xbrli:identifier>
    </xbrli:entity>
    <xbrli:period>
      <xbrli:startDate>2004-01-01</xbrli:startDate>
      <xbrli:endDate>2004-12-31</xbrli:endDate>
    </xbrli:period>
  </xbrli:context>
  <xbrli:unit id="EUR">
    <xbrli:measure>iso4217:EUR</xbrli:measure>
  </xbrli:unit>
</xbrli:xbrl>
Instance Problems

- XBRL filings contain many, many customized attributes.
- Filings prior to June 15, 2011 may not contain XBRL data.
- Private companies (Getty, Dell) may not contain XBRL data.
- XBRL data may not align with ascii/html.

Given the official financial filings, extract generic XBRL based on the taxonomies, so that comparability is maximized across domains.
Problem

Taxonomies and html files are huge. It's possible to waste time attempting to extract from an html file that has no bearing on the slot. If documents are huge, hardware limitations may cause issues, as well (memory, cpu usage).

Given an asci/html file section it according to high level sections of its DTS.
Solution

- Passage Extraction ("chunking", IR)
  Given a taxonomy infer rules for a given ascii/html text so that the text is "chunked" into (labelled) passages.

- Information Extraction (IE)
  Given a set of high relevance passages extract facts so that the taxonomy schema slots are filled.
Passage Extraction

Extraction of Coherent Relevant Passages Using Hidden Markov Models

HMM used to detect accurate boundaries between coherent relevant passages in a document.

*Given a document section it into passages that are topic related and query-specific such that a high level taxonomic slot can be filled.*
Passage Extraction: Challenges

- Address Length Variation
  - passages vary in length between documents

- Exploit Coherence
  - query dependent (topic)
  - document dependent (boundaries)
Passage Extraction : HMM

Set of observable output symbols
Set of hidden states
An initial state probability distribution
A state transition probability distribution for each state
An output probability for each state

A sequence of output symbols generated from a sequence of hidden states generated by a stochastic process that produces state transitions
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

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

Estimate Relevance Model

\( (q, wd, cd) \)

\( p = \) short relevant passage generated by \( q \)

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

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

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

Train transition probabilities (Baum-Welch)

Find most likely state sequence (Viterbi)
Passage Extraction: Results

Evaluation:

\[ N_o : \text{Length of overlapping passage} \]

\[ N_e : \text{Length of extracted passage} \]

\[ N_t : \text{True length of passage} \]

Methods:

- **BL-s**: first-last occurrence of any query term
- **BL-win**: fixed sized windows max occurrence of any query term
- **TD-IF**
  - **BL-cos**: cosine similarity
  - **BL-pivoted**

\[
P = \frac{N_o}{N_e}, \quad R = \frac{N_o}{N_t}, \quad F1 = \frac{2 \times P \times R}{P + R}
\]

<table>
<thead>
<tr>
<th>Collection</th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>DOE</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BL-s</td>
<td>0.869</td>
<td>0.591</td>
<td>0.632</td>
</tr>
<tr>
<td>BL-win</td>
<td>0.779</td>
<td>0.777</td>
<td>0.730</td>
</tr>
<tr>
<td>BL-cos</td>
<td>0.764</td>
<td>0.763</td>
<td>0.717</td>
</tr>
<tr>
<td>BL-pivoted</td>
<td>0.749</td>
<td>0.745</td>
<td>0.701</td>
</tr>
<tr>
<td>HMM-q</td>
<td>0.940</td>
<td>0.500</td>
<td>0.561</td>
</tr>
<tr>
<td>HMM-ud</td>
<td>0.932</td>
<td>0.630</td>
<td>0.659</td>
</tr>
<tr>
<td>HMM-cd</td>
<td>0.941*</td>
<td>0.858*</td>
<td>0.862*</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Collection</th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>HARD04</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BL-s</td>
<td>0.670</td>
<td>0.909</td>
<td>0.666</td>
</tr>
<tr>
<td>BL-win</td>
<td>0.668</td>
<td>0.759</td>
<td>0.621</td>
</tr>
<tr>
<td>BL-cos</td>
<td>0.671</td>
<td>0.781</td>
<td>0.628</td>
</tr>
<tr>
<td>BL-pivoted</td>
<td>0.672</td>
<td>0.783</td>
<td>0.629</td>
</tr>
<tr>
<td>HMM-q</td>
<td>0.709*</td>
<td>0.726</td>
<td>0.585</td>
</tr>
<tr>
<td>HMM-ud</td>
<td>0.686</td>
<td>0.877</td>
<td>0.656</td>
</tr>
<tr>
<td>HMM-cd</td>
<td>0.671</td>
<td>0.969*</td>
<td>0.706*</td>
</tr>
</tbody>
</table>

Stars indicate the best performance among all methods on the same dataset. *HMM-cd* outperformed all other methods on both datasets (in terms of F1 measure).
## Passage Extraction: Feedback

### Table IV. Effect of Pseudo-Feedback on DOE Data—F1

<table>
<thead>
<tr>
<th></th>
<th>BL-s</th>
<th>BL-win (with different k)</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>50</td>
<td>100</td>
<td>150</td>
<td>200</td>
<td>250</td>
<td>300</td>
<td>350</td>
<td>400</td>
</tr>
<tr>
<td>BL</td>
<td>0.632</td>
<td>0.344</td>
<td>0.529</td>
<td>0.636</td>
<td>0.696</td>
<td>0.750</td>
<td>0.749</td>
<td>0.756</td>
</tr>
<tr>
<td>HMM-ud</td>
<td>0.705</td>
<td>0.712</td>
<td>0.749</td>
<td>0.771</td>
<td>0.772</td>
<td>0.775</td>
<td>0.774</td>
<td>0.768</td>
</tr>
<tr>
<td>Improv.</td>
<td>+12%</td>
<td>+107%</td>
<td>+42% +21% -11% +6% -4% -1% -0%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HMM-cd</td>
<td>0.874</td>
<td>0.845</td>
<td>0.872</td>
<td>0.876</td>
<td>0.860</td>
<td>0.842</td>
<td>0.825</td>
<td>0.809</td>
</tr>
<tr>
<td>Improv.</td>
<td>+38%</td>
<td>+146%</td>
<td>+65% +38% +24% +15% +10% +7% +5%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Table V. Effect of Pseudo-Feedback on DOE Data—Precision

<table>
<thead>
<tr>
<th></th>
<th>BL-s</th>
<th>BL-win (with different k)</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>50</td>
<td>100</td>
<td>150</td>
<td>200</td>
<td>250</td>
<td>300</td>
<td>350</td>
<td>400</td>
</tr>
<tr>
<td>BL</td>
<td>0.591</td>
<td>0.234</td>
<td>0.425</td>
<td>0.573</td>
<td>0.647</td>
<td>0.777</td>
<td>0.847</td>
<td>0.899</td>
</tr>
<tr>
<td>HMM-ud</td>
<td>0.703</td>
<td>0.646</td>
<td>0.721</td>
<td>0.789</td>
<td>0.841</td>
<td>0.880</td>
<td>0.926</td>
<td>0.950</td>
</tr>
<tr>
<td>Improv.</td>
<td>+19%</td>
<td>+176%</td>
<td>+76% +38% +22% +15% +9% +6% +3%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HMM-cd</td>
<td>0.893</td>
<td>0.824</td>
<td>0.891</td>
<td>0.924</td>
<td>0.943</td>
<td>0.957</td>
<td>0.968</td>
<td>0.978</td>
</tr>
<tr>
<td>Improv.</td>
<td>+51%</td>
<td>+252%</td>
<td>+110% +61% +37% +23% +14% +9% +5%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Table VI. Effect of Pseudo-Feedback on DOE Data—Recall

<table>
<thead>
<tr>
<th></th>
<th>BL-s</th>
<th>BL-win (with different k)</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>50</td>
<td>100</td>
<td>150</td>
<td>200</td>
<td>250</td>
<td>300</td>
<td>350</td>
<td>400</td>
</tr>
<tr>
<td>BL</td>
<td>0.666</td>
<td>0.280</td>
<td>0.411</td>
<td>0.547</td>
<td>0.621</td>
<td>0.658</td>
<td>0.679</td>
<td>0.689</td>
</tr>
<tr>
<td>HMM-ud</td>
<td>0.683</td>
<td>0.665</td>
<td>0.688</td>
<td>0.698</td>
<td>0.705</td>
<td>0.703</td>
<td>0.703</td>
<td>0.703</td>
</tr>
<tr>
<td>Improv.</td>
<td>+3%</td>
<td>+138%</td>
<td>+67% +28% +13% +7% +4%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HMM-cd</td>
<td>0.709</td>
<td>0.694</td>
<td>0.701</td>
<td>0.707</td>
<td>0.707</td>
<td>0.707</td>
<td>0.707</td>
<td>0.707</td>
</tr>
<tr>
<td>Improv.</td>
<td>+6%</td>
<td>+146%</td>
<td>+71% +29% +14% +7% +4%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Table VII. Effect of Pseudo-Feedback on HARD04 Data—F1

<table>
<thead>
<tr>
<th></th>
<th>BL-s</th>
<th>BL-win (with different k)</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>50</td>
<td>100</td>
<td>150</td>
<td>200</td>
<td>250</td>
<td>300</td>
<td>350</td>
<td>400</td>
</tr>
<tr>
<td>BL</td>
<td>0.670</td>
<td>0.769</td>
<td>0.696</td>
<td>0.674</td>
<td>0.668</td>
<td>0.662</td>
<td>0.657</td>
<td>0.657</td>
</tr>
<tr>
<td>HMM-ud</td>
<td>0.688</td>
<td>0.671</td>
<td>0.687</td>
<td>0.681</td>
<td>0.689</td>
<td>0.687</td>
<td>0.687</td>
<td>0.687</td>
</tr>
<tr>
<td>Improv.</td>
<td>-1%</td>
<td>-5%</td>
<td>-4% -2% -1% -1% +0%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HMM-cd</td>
<td>0.664</td>
<td>0.670</td>
<td>0.686</td>
<td>0.664</td>
<td>0.660</td>
<td>0.658</td>
<td>0.657</td>
<td>0.657</td>
</tr>
<tr>
<td>Improv.</td>
<td>-1%</td>
<td>-6%</td>
<td>-4% -1% -1% -1% -0%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Table VIII. Effect of Pseudo Feedback on HARD04 Data—Precision

<table>
<thead>
<tr>
<th></th>
<th>BL-s</th>
<th>BL-win (with different k)</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>50</td>
<td>100</td>
<td>150</td>
<td>200</td>
<td>250</td>
<td>300</td>
<td>350</td>
<td>400</td>
</tr>
<tr>
<td>BL</td>
<td>0.909</td>
<td>0.219</td>
<td>0.374</td>
<td>0.606</td>
<td>0.759</td>
<td>0.846</td>
<td>0.846</td>
<td>0.907</td>
</tr>
<tr>
<td>HMM-ud</td>
<td>0.946</td>
<td>0.859</td>
<td>0.910</td>
<td>0.947</td>
<td>0.967</td>
<td>0.970</td>
<td>0.970</td>
<td>0.966</td>
</tr>
<tr>
<td>Improv.</td>
<td>+4%</td>
<td>+292%</td>
<td>+143% +56% +27% +15% +7%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HMM-cd</td>
<td>0.984</td>
<td>0.949</td>
<td>0.961</td>
<td>0.981</td>
<td>0.986</td>
<td>0.987</td>
<td>0.987</td>
<td>0.991</td>
</tr>
<tr>
<td>Improv.</td>
<td>+8%</td>
<td>+333%</td>
<td>+157% +62% +30% +17% +9%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Passage Extraction: Conclusion

Significant gains were made with the inclusion of documents (cd).

Would there be more gains should the documents stem from different domains?
IE: Overview

Given an input $x$ predict the label $y$
Examples include:
Sentiment Extraction: $y$ is binary (bag of words)
Relationship Extraction: triple with a closed class of relationship types (though later we see that this could be done unsupervised)
Entity Extraction: label an entity in a sentence (also assumed here to be supervised)
IE: Current Techniques

- **Select Relevant Examples:**
  - higher weights for closer similarity
    - unknown probability distribution in each domain
    - estimation is difficult
  - Mean Matching
  - Classifier to Estimate the Probability

- **Remove Irrelevant Features**
  - penalize a feature which has a large difference in expected src value and expected dest value

- **Add New Features**
  - Anchor features with strong correlation to labels in both domains
IE: DA for Record Extraction

"Regularity-rich Aspect Property"
Instance-label pairs where $y$ is probabilistically modelled
Set of properties that generally take only one value in a domain
Clique potential: in a graph a clique is a sub set where for each vertex pair there is an edge connecting the two vertices.
IE: Properties

Decomposable properties:

\[ p(x, y) \triangleq \begin{cases} 
\emptyset & \text{if } \forall c : p(x, y_c, c) = \emptyset \\
v & \text{if } \forall c : p(x, y_c, c) \in \{v, \emptyset\} \\
\bot & \text{otherwise.} 
\end{cases} \]
Markov Random Fields are used to model 2D/3D markov processes. A markov chain in HMM is 1D (time axis e.g.).

This is not useful at the current time for the current tasks. However, poses interesting possibilities for later...changes over time and another factor.
IE: Subsequent

Cliques, Potts Potential, negative Gini et. al. This is probably less relevant to this class.

I was fascinated by a 2D Markov model and the idea of folding in properties to models.
IE: Results with Properties

And the results are pretty solid at varying training levels

<table>
<thead>
<tr>
<th>Train %</th>
<th>Title</th>
<th></th>
<th>Venue</th>
<th></th>
<th>Author</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Base</td>
<td>CI</td>
<td>Base</td>
<td>CI</td>
<td>Base</td>
</tr>
<tr>
<td>5</td>
<td>70.7</td>
<td>74.8</td>
<td>58.8</td>
<td>62.5</td>
<td>74.1</td>
</tr>
<tr>
<td>10</td>
<td>78.0</td>
<td>82.1</td>
<td>69.2</td>
<td>72.2</td>
<td>75.6</td>
</tr>
<tr>
<td>20</td>
<td>85.8</td>
<td>88.6</td>
<td>76.7</td>
<td>78.9</td>
<td>80.7</td>
</tr>
<tr>
<td>30</td>
<td>91.7</td>
<td>93.0</td>
<td>81.5</td>
<td>82.6</td>
<td>87.7</td>
</tr>
<tr>
<td>50</td>
<td>92.3</td>
<td>94.2</td>
<td>83.5</td>
<td>84.5</td>
<td>89.4</td>
</tr>
</tbody>
</table>
Open IE: Overview

- Unsupervised extraction of entities and their relationships from text cues
- POS & NP chunking to identify entities
- CRF to glean relationships (HMM undirected graphs)
- Additional processing necessary for domain relations
Open IE: Class Recognizer

- (regex)
- argument values form domain relation
- trigger words
- winIndicator/LossIndicator
Open IE: Domain Relation Mapping

Rule is set of constraints on tuple arguments and on context to left or right of tuple

If constraints met: Rule specifies domain relation type, argument types, location of each argument value

Long distance constraints handled by left context/ right context
Covering Algorithm

Base rule applied to each tuple
if extraction & correct (c)
else error (e)
If not (subphrase to extract from tuple) remove constrint

Confidence smoothing \( c/(c + e + 1) \)
Open IE: Active Learning

UI presented to user
k random tuples from a rule, which the user tags as correct

Rules refined by
● Confidence
● Saturation \((\beta - \alpha - 1)/(\beta - \alpha)\)
● Constraint Penalty \(1/(\gamma) \times \#\text{constraints}\)
Open IE: Results
Open IE: Conclusions

Intelligence results not as good (open classes, multiple meanings, NER coarse grained)

Complex NP phrases not handled, especially when combined with long distance dependencies
Sources

Jing Jiang, Chengxiang Zhai *Extraction of Coherent Relevant Passages Using Hidden Markov Models*