Use of NLP

• Syntactic
  – Parsing to identify phrases
  – Full syntactic structure comparison

• Semantic
  – Building an understanding of a document’s content

• Discourse
  – Exploiting document structure?
Syntactic

• Parsing to identify phrases
  – The issues.
  – Explain how it’s done (a bit).
  – Is it worth it?

• Other possibilities
  – Grammatical tagging
  – Full syntactic structure comparison
    • Explain how it’s done (a little bit).
    • Show results.
Simple phrase identification

• High frequency terms could be good candidates.
  – Why?

• Terms co-occurring more often than chance.
  – Within small number of words.
  – Surrounding simple terms.
  – Not surrounding punctuation.
Problems

• Close words that aren’t phrases.
  • “the use of computers in science & technology”

• Distant words that are phrases.
  • “preparation & evaluation of abstracts and extracts”
Parsing for phrases

• Using parsers to identify noun phrases.
• Make a phrase out of a head and the head of its modifiers.

“automatic analysis of scientific text”
Errors

• Not a perfect rule by any means.
  – Need restrictions to eliminate bogus phrases.

“automatic analysis of these four scientific texts”
Do they work?

• Fagan compared statistical with syntactic, statistics won, just

• More research has been conducted.
Check out TREC

- Overview of the Seventh Text REtrieval Conference (TREC-7), E. Voorhees, D. Harman (National Institute of Standards and Technology)
  - http://trec.nist.gov/
  - Ad hoc track
    - Fairly even between statistical phrases, syntactic phrases and no phrases.
Grammatical tagging?

- Tag document text with grammatical codes?

- Doesn’t appear to work
Syntactic structure comparison

• Has been tried...

• Method
  – Parse sentences into tree structures
  – When you get a phrase match
    • Look at linking syntactic operator.
    • Look at the residual tree structure that didn’t match

• Does not to work
Semantic

• Disambiguation
  – Given a word appearing in a certain context, disambiguators will tell you what sense it is.

• IR system
  – Index document collections by senses rather than words
  – Ask the users what senses the query words are
  – Retrieve on senses
Disambiguation

• Does it work?
  – No (well maybe)
    • M. Sanderson, Word sense disambiguation and information retrieval, in *Proceedings of the 17th ACM SIGIR Conference*, Pages 142-151, 1994
Partial conclusions

• NLP has yet to prove itself in IR
  – Agree
  – Sort of don’t agree
Mark’s idle speculation

• What people think is going on always
Mark’s idle speculation

• What’s usually actually going on

Keywords

NLP
Areas where NLP does work

• Systems with the following ingredients.
  – Collection documents cover small domain.
  – Language use is limited in some manner.
  – User queries cover tight subject area.
  – Documents/queries very short

• Image captions
  – LSI, pseudo-relevance feedback

– People willing to spend money getting NLP to work
RIME & IOTA

• From Grenoble

• Medical record retrieval system
• Some database’y parts
• Free text descriptions of cases
Indexing

• “an opacity affecting probably the lung and the trachea”
Retrieval

• How do we match a user’s query to these structures?
  – Using transformations - bit like logic.

\[
\begin{align*}
\{[\text{bears-on}], \text{SGN}\} & \quad \Rightarrow \quad \{[\text{lung}], \text{LOC}\}, t \\
\{[\text{opacity}], \text{SGN}\} & \quad \Rightarrow \quad \{[\text{opacity}], \text{SGN}\}, t
\end{align*}
\]
Tree transformation

{[has-for-value], SGN}

{[bears-on], SGN}

{[opacity], SGN}  {[lung], LOC}

{[has-for-value], SGN}

{[contour], SGN}  {[blurred], LOC}

⇒

{[has-for-value], SGN}, t

{[opacity], SGN}

{[has-for-value], SGN}

{[contour], SGN}  {[blurred], LOC}
Term transforms

- Basic medical terms stored in a hierarchy.
  - Transformations possible again with uncertainty added.

<table>
<thead>
<tr>
<th>Level 1</th>
<th>Level 2</th>
<th>Level 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>tumour</td>
<td>cancer</td>
<td>sarcoma</td>
</tr>
<tr>
<td></td>
<td>hygroma</td>
<td>polykystosis</td>
</tr>
<tr>
<td></td>
<td>kyste</td>
<td></td>
</tr>
<tr>
<td></td>
<td>pseudokyst</td>
<td></td>
</tr>
<tr>
<td></td>
<td>polyp</td>
<td>polyposis</td>
</tr>
</tbody>
</table>
Isn’t this a bit slow?

• Yes

• Optimisation
  – Scan for potential documents.
  – Process them intensively.

• Evaluation?
  – Not in that paper.
Not unique

- SCISOR

Why do they work?

• Because of the restrictions
  – Small subject domain.
  – Limited vocabulary.
  – Restricted type of question.

• Compare with large scale IR system.
  – Keywords are good enough.
  – Long time to set up.
  – Hard to adapt to new domain.
Anything else for NLP?

• Text Generation
  – IR system explaining itself?
Conclusions

• By now, you will be able to
  – Outline a range of attempts to get NLP to work with IR systems
  – Idly speculate on why they failed
  – Describe the successful use of NLP in a limited domain