What is Semantic Matching?
Some Well-known NLP Systems

- Information Retrieval and Question Answering
  - Google Search
  - Sogo Question Answering
  - IBM Watson
  - Apple Siri
  - Walframe Alpha
  - Microsoft SharePoint Search

- Machine Translation
  - Google Translate
  - Baidu Translation
  - Translation Memory: Trados
  - Xunfei Voice Assistant

- Text Mining
  - SAS TextMiner
  - Microsoft SQL Server Text Mining
  - Autonomy
It is really hot
[Intent: please open the window]

Mary is loved by John
[act: love; agt: John; obj: Mary]

I ate icecream [with a spoon]
I ate icecream [with chocolate]

[南京市] [长江大桥]
[南京] [市长] [江大桥]
Accuracies of Natural Language Analysis

- Lexical Analysis (word segmentation and part-of-speech tagging): practically usable
- Syntactic Analysis: almost usable
- Semantic Analysis: still difficult
- Programmatic Analysis: ?

<table>
<thead>
<tr>
<th></th>
<th>English</th>
<th>Chinese</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pragmatic Analysis</td>
<td>?</td>
<td>?</td>
</tr>
<tr>
<td>Semantic Role Labeling</td>
<td>&gt;=87%</td>
<td>&gt;=75%</td>
</tr>
<tr>
<td>Syntactic Analysis</td>
<td>&gt;=90%</td>
<td>&gt;=80%</td>
</tr>
<tr>
<td>Part of Speech Tagging</td>
<td>&gt;=97%</td>
<td>&gt;=93%</td>
</tr>
<tr>
<td>Word Segmentation</td>
<td>NA</td>
<td>&gt;=95%</td>
</tr>
</tbody>
</table>
Current Approach: Avoid Understanding and Conduct *Matching*
Matching between Two Languages

- Machine Translation: A language = source language, B language = target language
- Question Answering: A language = question, B language = answer
- Information Retrieval: A language = query, B language = document
State of the Art Translation Models

Mining word translation, phrase translation, and syntactic translation rules from parallel corpora, generate new translation for given sentence based on the rules

• **Word based Model**
  – Traditional model: IBM Model

• **Phrase based Model**
  – State-of-the-art
  – Google Translate

• **Syntax based Model**
  – Main focus of current research

Alignment is made on parallel corpus
State of the Art Information Retrieval Models

Calculating relevance of query with respect to document according to matching degree between query and document

• Term based Model
  – BM25, LM for IR
• Topic Model
  – LSI, RLSI

Intuition: if query words occur many times in document, then document is viewed relevant.
<table>
<thead>
<tr>
<th>task</th>
<th>types of texts</th>
<th>relation between texts</th>
</tr>
</thead>
<tbody>
<tr>
<td>search</td>
<td>A=query, B=document</td>
<td>relevance</td>
</tr>
<tr>
<td>question answering</td>
<td>A=question, B=answer</td>
<td>answer to question</td>
</tr>
<tr>
<td>cross-language IR</td>
<td>A=query, B=document (in diff. lang.)</td>
<td>relevance</td>
</tr>
<tr>
<td>short text conversation</td>
<td>A=text, B=text</td>
<td>message and comment</td>
</tr>
<tr>
<td>similar document detection</td>
<td>A=text, B=text</td>
<td>similarity</td>
</tr>
<tr>
<td>online advertising</td>
<td>A=query, B=ads.</td>
<td>relevance as ads.</td>
</tr>
<tr>
<td>paraphrasing</td>
<td>A=sentence, B=sentence</td>
<td>equivalence</td>
</tr>
<tr>
<td>textual entailment</td>
<td>A=sentence, B=sentence</td>
<td>entailment</td>
</tr>
</tbody>
</table>
### Biggest Challenge: Mismatch between Query and Document

<table>
<thead>
<tr>
<th>Query</th>
<th>Document</th>
<th>Term Match</th>
<th>Semantic Match</th>
</tr>
</thead>
<tbody>
<tr>
<td>seattle best hotel</td>
<td>seattle best hotels</td>
<td>partial</td>
<td>yes</td>
</tr>
<tr>
<td>pool schedule</td>
<td>swimmingpool schedule</td>
<td>partial</td>
<td>yes</td>
</tr>
<tr>
<td>natural logarithm transformation</td>
<td>logarithm transformation</td>
<td>partial</td>
<td>yes</td>
</tr>
<tr>
<td>china kong</td>
<td>china hong kong</td>
<td>partial</td>
<td>no</td>
</tr>
<tr>
<td>why are windows so expensive</td>
<td>why are macs so expensive</td>
<td>partial</td>
<td>no</td>
</tr>
</tbody>
</table>
Next Opportunity: Semantic Matching?

Avoid doing understanding, conducting matching instead.
Semantic Matching in Search

Hang Li  
Huawei Technologies, Hong Kong  
hangli.hl@huawei.com

Jun Xu  
Huawei Technologies, Hong Kong  
nkxujun@gmail.com
Semantic Matching in Search

Semantic Matching
- Form
- Phrase
- Sense
- Topic
- Structure

Term Matching
Matching by Query Reformulation

• Key Idea: reformulating query and matching reformulated query and document

• Issues: query reformulation, search result blending, similar query mining

• Query reformulation includes spelling error correction, query segmentation, query expansion, etc
Matching with Term Dependency Model

• Key Idea: considering dependency between terms and conducting matching between dependent terms in query and document
• MRF based models, extension of IR models, proximity
Matching with Translation Model

• Key Idea: learning translation probabilities of words from click-through data
• Model: IBM model one
• Self-translation probability
• E.g., aircraft $\rightarrow$ aircraft 0.70, aircraft $\rightarrow$ airplane 0.20
Matching with Topic Model

- Key Idea: mining topics from document collection, representing query and document with topics, and matching query topics and document topics
- Probabilistic models: PLA, LDA
- Non-probabilistic models: LSI, NMF, RLSI
Matching with Latent Space Model

- Key Idea: matching query and document into latent space and conducting matching in latent space
- Linear models: RMLS, SSI, BLTM
- Non-linear models: DSSM
- Contain traditional IR models
Open Problems and Future Directions

• Accuracy: how to deal with topic drift
• Scalability: how to handle big data
• Natural Language: how to leverage natural language processing techniques
• Knowledge: how to incorporate existing knowledge
• Evaluation: how to develop new evaluation schemes
Thank you!

hangli.hl@huawei.com