And now for something real complicated:

Latent Semantic Indexing for Multi-lingual Searches (LSIMS)

Vinayak Ganeshan
The 196 document mini-WDL

<table>
<thead>
<tr>
<th>Language</th>
<th>Number terms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chinese</td>
<td>38</td>
</tr>
<tr>
<td>(metadata in English)</td>
<td></td>
</tr>
<tr>
<td>English</td>
<td>320</td>
</tr>
<tr>
<td>French</td>
<td>181</td>
</tr>
<tr>
<td>Portuguese</td>
<td>39</td>
</tr>
<tr>
<td>Russian</td>
<td>95</td>
</tr>
<tr>
<td>Spanish</td>
<td>99</td>
</tr>
</tbody>
</table>

5 most popular terms: карта, america, carte, russia, губерния
Many terms uncommon (i.e. old English terms, rare references)
Scalability requires efficiency

Search algorithm must be

– scalable in terms of number of languages
– scalable in terms of number of documents
– able to deal with very diverse set of documents

*Buzz words: efficiency, automation, robustness*

Multi-lingual searching critical for success of WDL
Can we depend on translation?

Possible approach: translate both query and data

But,
• machine translation is not yet up to scratch
• scalability is questionable

So, let’s look for something else
Is it possible to use LSI?

LSI must be trained to recognize connections between similar terms in different languages (without intermediate translation step)

*So, how about this?*

Find a very large collection of phrases translated in many languages that can supply such connections

Words in language A are considered connected to words in language B if they occur in same phrase
Quiz

Document Phrases
- must cover large part of commonly used terms
- must be translated in many different languages

Can you think of one such document?
How about ....

Bible
Corpus of bible verses
Translated into at least 2400 languages
80 versions available on the world wide web

Europarl
Corpus of proceedings from European parliament
In English, French, Spanish and Portuguese

and various smaller corpora are available
The Bible as a training corpus

An LSI is performed on the training corpus
Documents: bible verses
Terms: words (in all languages) in verses

Documents of the WDL are “folded in”

Resulting accuracy of search depends on
• coverage provided by the Bible
• accuracy of the LSI mappings
How is mini-WDL covered by Bible?

<table>
<thead>
<tr>
<th>Language</th>
<th>Number terms</th>
<th>Bible coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chinese</td>
<td>38</td>
<td>21%</td>
</tr>
<tr>
<td>English</td>
<td>320</td>
<td>71%</td>
</tr>
<tr>
<td>French</td>
<td>181</td>
<td>62%</td>
</tr>
<tr>
<td>Portuguese</td>
<td>39</td>
<td>85%</td>
</tr>
<tr>
<td>Russian</td>
<td>95</td>
<td>40%</td>
</tr>
<tr>
<td>Spanish</td>
<td>99</td>
<td>81%</td>
</tr>
</tbody>
</table>
Putting the mini-WDL to the biblical test

DEMO
It’s expected to be only so-so

Coverage provided by the Bible insufficient
Many terms do not exist in the Bible

Mini-WDL data very sparse and very diverse

Note that previously reported results are positive
because searches performed for similar documents
(e.g. the Koran)
Idea: Add a second training corpus

<table>
<thead>
<tr>
<th>Language</th>
<th>Number terms</th>
<th>Bible coverage</th>
<th>Bible + Europarl</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chinese</td>
<td>38</td>
<td>21%</td>
<td>34%</td>
</tr>
<tr>
<td>English</td>
<td>320</td>
<td>71%</td>
<td>96%</td>
</tr>
<tr>
<td>French</td>
<td>181</td>
<td>62%</td>
<td>84%</td>
</tr>
<tr>
<td>Portuguese</td>
<td>39</td>
<td>85%</td>
<td>90%</td>
</tr>
<tr>
<td>Russian</td>
<td>95</td>
<td>40%</td>
<td>40%</td>
</tr>
<tr>
<td>Spanish</td>
<td>99</td>
<td>81%</td>
<td>100%</td>
</tr>
</tbody>
</table>

Bible + Europarl gives improved coverage
Putting the mini-WDL to the biblical and Euro test

DEMO
A completely open area of research

Existing algorithms not robust for highly diverse data

LSIMS promising, but
• must find good training corpora
• may need to specialize LSI

*Multi-lingual searching offers exciting research with potentially very large impact*