Extracting bilingual wordlists from Wikipedia

Francis M. Tyers\textsuperscript{1,2} and Jacques A. Pienaar\textsuperscript{3}

\textsuperscript{1}Departament de Llenguatges i Sistemes Informàtics, Universitat d’Alacant, E-03071 Alacant (Spain)

\textsuperscript{2}Prompsit Language Engineering, S.L., E-03690 St. Vicent del Raspeig (Spain)

\textsuperscript{3}Centre for Text Technology, Language and literature in the South African context, North-West University, Potchefstroom 2520 (South Africa)

27th May 2008
Contents

1 Introduction
   - Bilingual wordlists
   - Scarcity of resources
   - Wikipedia
   - Interwiki links
   - Ambiguity

2 Algorithm
   - Requirements
   - Description

3 Method
   - Method
   - Results
   - Analysis

4 Discussion
   - Discussion
   - Future research
   - Conclusions
### What is a bilingual wordlist

<table>
<thead>
<tr>
<th>English</th>
<th>Afrikaans</th>
</tr>
</thead>
<tbody>
<tr>
<td>group</td>
<td>groep</td>
</tr>
<tr>
<td>software</td>
<td>sagteware</td>
</tr>
<tr>
<td>programmer</td>
<td>programmeerder</td>
</tr>
<tr>
<td>chatroom</td>
<td>kletskamer</td>
</tr>
<tr>
<td>developer</td>
<td>ontwikkelaar</td>
</tr>
<tr>
<td>planet</td>
<td>planeet</td>
</tr>
<tr>
<td>solar system</td>
<td>sonnestelsel</td>
</tr>
<tr>
<td>god</td>
<td>god</td>
</tr>
<tr>
<td>universe</td>
<td>heelal</td>
</tr>
<tr>
<td>middle</td>
<td>middel</td>
</tr>
<tr>
<td>astronomer</td>
<td>sterrekundige</td>
</tr>
<tr>
<td>life</td>
<td>lewe</td>
</tr>
<tr>
<td>page</td>
<td>bladsy</td>
</tr>
</tbody>
</table>

Francis M. Tyers and Jacques A. Pienaar

Extracting bilingual wordlists from Wikipedia
Bilingual wordlists

Why do we want bilingual wordlists?

- Cross-language information retrieval (CLIR)
  - Finding search results across languages.
  - Finding aligned phrase pairs.

- Machine translation
  - RBMT: start of a bilingual part-of-speech tagged dictionary.
  - SMT: improving the quality of translation models.
Scarcity of resources

For lesser-used languages resources (e.g. bilingual wordlists) are often scarce, they may be:

- **Inexistent**
  - Catalan–Faroese,
  - Basque–Russian etc.

- **Unavailable**
  - not available online
  - available at (great) cost, or
  - in legacy encodings or formats.

- **Prohibitively licensed**
  - non-commercial
  - research and education only, or
  - fully proprietary.
What is Wikipedia? (I)

- An online, collaboratively edited encyclopaedia – content created and managed by volunteers
- Articles are available in over 250 languages – including many lesser-used and under-resourced languages
- Liberally licensed – it is freely available and freely distributable both for non-commercial and commercial use
- Requires no specific expertise other than ability to use web browser and using a simple markup language to create entries
What’s interesting about Wikipedia?

- Wikipedia encyclopaedias are freely-editable
- Content and structure amenable to linguistic research
- Breadth of language coverage make useful/appropriate for creating linguistic resources
- Inter-language (interwiki) page links
What are interwiki links?

Divisions, specialties, and subfields

The field of general linguistics traditionally attempts to characterize the nature of human language and it is an individual knows when said to know a language; and to explain how individuals learn language. All humans (setting aside extremely pathological cases) achieve competence in what linguists assume, the ability to acquire and use language is an innate, biologically-based ability to walk. There is no discernible genetic process responsible for differences between the language(s) they are exposed to as a child, regardless of parentage or ethnic origin, in nativist schools of linguistics.
What are interwiki links?

- Link between same article in different languages
- Links are periodically maintained by both human editors and bots\(^1\)
- Harvesting these links provide useful translation equivalents

\(^1\)As used by Wikipedia, a bot is a software program that makes automated changes to Wikipedia.
# Interwiki links: How links are maintained

## Manual
- Normally done when someone starts a new article
- Looks for the page in another language and copies in links

## Bot assisted
- Can be supervised or unsupervised
- Supervised mode prompts the user to resolve ambiguities
- Unsupervised mode discards ambiguous links
Ambiguity: Disambiguation pages

Sometimes articles can simply be pages with lists of other articles.
Other times, they can just *forward* the user from many article titles to one.
Requirements

- **A seed word list**
  - Recommended practice: start with the better sourced language.
  - Wikipedia titles are almost exclusively nouns, therefore a list made up of nouns and proper will have greatest success.

- **Computer**
  - Algorithm not computationally expensive $\implies$ desktop PC sufficient.

- **Internet connection**
  - Only text downloaded.
  - Low bandwidth usage and requirement.
Algorithm

Foreach \( w \in \text{Word-List} \)

Extract links

\( a \leftarrow \text{Retrieve-Page(SourceWikipedia, } w) \)

\( \ell \leftarrow \text{Extract-Links}(a) \);

Foreach \( t \in \text{Target-Languages} \)

\( t \in \ell \)

√

Add pair

\( \text{Add-Pair}(w, \ell[t]) \)

Extracted bilingual word-list

Francis M. Tyers and Jacques A. Pienaar

Extracting bilingual wordlists from Wikipedia
Algorithm

\[
\text{Foreach } w \in \text{Word-List} \\
\text{Extract links} \\
a \leftarrow \text{Retrieve-Page(SourceWikipedia, } w) \\
\ell \leftarrow \text{Extract-Links}(a); \\
\text{Foreach } t \in \text{Target-Languages} \\
t \in \ell \\
\sqrt{\text{Add pair}} \\
\text{Add-Pair}(w, \ell[t]) \\
\text{Extracted bilingual word-list}
\]
Algorithm

\[
\text{Foreach } w \in \text{Word-List} \\
\quad \text{Extract links} \\
\quad a \leftarrow \text{Retrieve-Page(SourceWikipedia, } w) \\
\quad \ell \leftarrow \text{Extract-Links}(a); \\
\text{Foreach } t \in \text{Target-Languages} \\
\quad t \in \ell \quad \sqrt{\text{Add pair}} \\
\quad \text{Add-Pair}(w, \ell[t]) \\
\text{Extracted bilingual word-list}
\]
Algorithm

Foreach \( w \in \text{Word-List} \)

\[
a \leftarrow \text{Retrieve-Page(SourceWikipedia, w)}
\]

\[
\ell \leftarrow \text{Extract-Links}(a);
\]

Foreach \( t \in \text{Target-Languages} \)

\[
t \in \ell
\]

√

Add pair

\[
\text{Add-Pair}(w, \ell[t])
\]

Extracted bilingual word-list

Francis M. Tyers and Jacques A. Pienaar

Extracting bilingual wordlists from Wikipedia
Extracted bilingual word-list

\[
\text{Foreach } w \in \text{Word-List} \\
\text{Extract links} \\
a \leftarrow \text{Retrieve-Page(SourceWikipedia, } w) \\
\ell \leftarrow \text{Extract-Links}(a); \\
\text{Foreach } t \in \text{Target-Languages} \\
t \in \ell \\
\sqrt{t \in \ell} \\
\text{Add pair} \\
\text{Add-Pair}(w, \ell[t]) \\
\text{Extracted bilingual word-list}
\]
Exhaust bilingual word-list

Foreach \( w \in \text{Word-List} \)

\[
\begin{align*}
    &\text{Extract links} \\
    &a \leftarrow \text{Retrieve-Page(SourceWikipedia, } w) \\
    &\ell \leftarrow \text{Extract-Links}(a); \\
    \end{align*}
\]

Foreach \( t \in \text{Target-Languages} \)

\[
\begin{align*}
    &t \in \ell \\
    &\sqrt{\text{Add pair}} \\
    &\text{Add-Pair}(w, \ell[t]) \\
\end{align*}
\]

Extracted bilingual word-list

Francis M. Tyers and Jacques A. Pienaar

Extracting bilingual wordlists from Wikipedia
Method

- Wordlist of 11,393 English nouns – these were all lemmas.
- Extracted from the apertium-en-ca language pair – slightly biased towards scientific and technical vocabulary.
- Rationale: A post-edited list would be immediately useful in an Apertium translation pair.
- 10,024 of the 11,393 lemmas were entries in the English Wikipedia.
- Generated word lists (in Swedish, Macedonian, Afrikaans and Iranian Persian) were given to native speaker to check.
- Positive result is when translation is judged correct:
  - has the right form,
  - right sense and
  - is in correct register.
Results

Table: Results for the language pairs

<table>
<thead>
<tr>
<th>Language</th>
<th>Pages</th>
<th>Links</th>
<th>Correct</th>
<th>Recall</th>
<th>Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>Swedish</td>
<td>273,291</td>
<td>4,913</td>
<td>3,428</td>
<td>34%</td>
<td>69%</td>
</tr>
<tr>
<td>Iranian Persian</td>
<td>32,194</td>
<td>1,605</td>
<td>1,487</td>
<td>14%</td>
<td>92%</td>
</tr>
<tr>
<td>Macedonian</td>
<td>14,887</td>
<td>779</td>
<td>631</td>
<td>6%</td>
<td>81%</td>
</tr>
<tr>
<td>Afrikaans</td>
<td>9,183</td>
<td>444</td>
<td>354</td>
<td>3%</td>
<td>79%</td>
</tr>
</tbody>
</table>
Analysis of errors (I)

- (af) Right sense, wrong surface form – *vandal* translated as *vandale* (vandals).
- (sv) Right sense, wrong register – *nephrolithiasis* translated as *njursten* (kidney stone).
- (af) Wrong sense, right domain – *sociolinguist* translated as *sosiolinguistieik* (sociolinguistics).
- (af) Wrong sense, wrong domain – *solidarity* translated as *Solidarność* (a Polish trade union)
Analysis of errors (II)

- Errors generally found to exist at approximately same frequency.
- None particularly more frequent than others.
- The errors examined seem to have been caused by redirects.
- *No full quantitative analysis done.*
Method provides basic bilingual word list.

Precision is good but recall low.

Recall will increase over time as more articles are added, and the link coverage improves.
Further and future research

- Double-check each pair – Ensuring that a retrieved link points back to the same source.
- Avoid following redirects.
- Retrieve gender information from article leads.
- Use category of Wikipedia pages.
- Creation of gazetteers.
- More thorough evaluation.
Conclusions

- Presented simple, computationally inexpensive and fast means of automatically obtaining bilingual word list.
- Good accuracy (lowest 69%) obtained.
- Poor recall.
- Method could be useful for bootstrapping more complex induction techniques for under-resourced languages.