Domain Adaptation for Ontology Localisation

Journal of Web Semantics, Vol 36, Jan 2015

John P. McCrae\textsuperscript{a,b}, Mihael Arcan\textsuperscript{a}, Kartik Asooja\textsuperscript{b,c}, Jorge Gracia\textsuperscript{c}, Paul Buitelaar\textsuperscript{a} and Philipp Cimiano\textsuperscript{b}

\textsuperscript{a}Insight Centre for Data Analytics, National University of Ireland, Galway
\textsuperscript{b}CIT-EC, Bielefeld University
\textsuperscript{c}Ontology Engineering Group, Universidad Politécnica de Madrid
Is the Semantic Web multilingual?

“Guidelines for multilingual linked data” Gómez-Pérez et al. 2013
Ontology Localisation

● Different languages (cultures) have different concepts
  ○ Cross-lingual ontology alignment
  ○ Not the focus of this paper

● Translation of labels
  ○ Manually is time-intensive and costly
  ○ Standard MT has poor performance for ontologies

● Domain ontologies
  ○ Small, focused ontologies
  ○ Large, general purpose ontologies (e.g., DBpedia) should be divided into more sections
Why is **ontology** translation hard?

Vessel@en
Why is **ontology** translation hard?

Vessel@en
Why is **ontology** translation hard?

Vessel@en  

船@ja
Why is **ontology** translation hard?
Why is **ontology** translation hard?

Vessel@en

Artery@en

Vein@en

船@ja

血管@ja
Why is **ontology** translation hard?
Why is ontology translation hard?
Three-pronged attack

Find new translations
(Domain Lexicon)

 Prefer in-domain translation
(Domain Features)

Choose in-domain translations
(Domain Language Model)
Phrase-based Machine Translation

- The ‘workhorse’ of machine translation is the phrase-based model (Koehn et al., 2003, 2007, 2010)
- We want to translate a foreign sentence, $f$, into a translation, $t$.
- We divide $f$ into a sequence of phrases (one or more consecutive words)
  - $f = \{f_1, ..., f_n\}$
- Generate a permuted sequence of translations from a phrase table
  - $t = \{t_{d(1)}, ..., t_{d(n)}\}$
Phrase-based Machine Translation II

● We search for a sequence of phrases that maximizes
  ○ $[\sum_i \sum_j \alpha_j \varphi_j(f_i, t_i)] + \alpha_l l(t) + \alpha_d d(f, t)$

● Where $\varphi_j(f_i, t_i)$ are the feature scores
  ○ Log probability of $f_i$ given $t_i$
  ○ Log probability of $t_i$ given $f_i$
  ○ Lexical weighting of $f_i$ given $t_i$
  ○ Lexical weighting of $t_i$ given $f_i$
  ○ Out-of-vocabulary?
  ○ Constant 1 (to bias towards using fewer phrases)

● $l(t)$ is a language model score
  ○ Prefer fluent translations
• $d(f, t)$ is the *distortion* score
  ○ Measures how much the translation has been rearranged
• A *decoder* heuristically finds the optimal division into phrases and the permutation of phrases
Phrase-based Machine Translation III

- $d(f, t)$ is the *distortion* score
  - Measures how much the translation has been rearranged
- A *decoder* heuristically finds the optimal division into phrases and the permutation of phrases.

Moses SMT
Monnet Architecture for MT

"Minimum Finance Lease Payment"

Pre-processing

Sources

Wikipedia

IATE

Linguee

Ranker

CLESA

Feature Extraction

Coder

Language Model

LM Adaptation

„Minimale Finanz-Leasing-Zahlungen“
Domain lexicon

- Parallel text may not have domain translations
- We add additional data
- Wikipedia
  - We find all articles with names matching an ontology label
  - Find all categories for these articles
  - Filter categories by threshold
  - Add all translations from DBpedia for these categories
## Domain lexicon

- Parallel text may not have domain translations
- We add additional data

### Wikipedia

- We find all articles with names matching an ontology label
- Find all categories for these articles
- Filter categories by threshold
- Add all translations from DBpedia for these categories

<table>
<thead>
<tr>
<th>Frequency</th>
<th>Wikipedia Category</th>
</tr>
</thead>
<tbody>
<tr>
<td>95</td>
<td>Economics Terminology</td>
</tr>
<tr>
<td>62</td>
<td>Generally Accepted Accounting Principles</td>
</tr>
<tr>
<td>61</td>
<td>Macroeconomics</td>
</tr>
<tr>
<td>55</td>
<td>Accounting Terminology</td>
</tr>
<tr>
<td>47</td>
<td>Finance</td>
</tr>
<tr>
<td>44</td>
<td>Economic Theories</td>
</tr>
<tr>
<td>42</td>
<td>International Trade</td>
</tr>
</tbody>
</table>
Domain lexicon II

● Linguee
  ○ Large source of parallel data from the Web
  ○ We queried ontology labels
  ○ Generates domain parallel corpus (Financial Domain: ~24,000 sentences)
  ○ Train phrase table (Moses)

● IATE (Interactive Terminology for Europe)
  ○ Translations in EU languages
  ○ Weighted by Cross-lingual Explicit Semantic Analysis (Sorg et al., 2008)
Domain features

We used Cross-lingual extension (Sorg & Cimiano, 2008) of Explicit Semantic Analysis (Gabrilovich & Markovitch, 2007)
Domain features

We used Cross-lingual extension (Sorg & Cimiano, 2008) of Explicit Semantic Analysis (Gabrilovich & Markovitch, 2007)
Domain Language Model

- The language model estimates the likelihood of the target sentence
  - \( p(w_1 \ldots w_n) = \prod p(w_i|w_{i-n} \ldots w_{i-1}) \)
  - \( p(w_i|w_{i-n} \ldots w_{i-1}) = c(w_{i-n} \ldots w_{i-1}w_i) / c(w_{i-n} \ldots w_{i-1}^*) \)
- For each document in our corpus we estimate its domain relevance (using ONETA, McCrae et al., 2014)
  - \( s_O(d) \)
- The count is weighted by document relevance
  - \( c(w_{i-n} \ldots w_i) = \sum s_O(d) c_d(w_{i-n} \ldots w_i) \)
The language model estimates the likelihood of the target sentence:

\[ p(w_1 \ldots w_n) = \prod p(w_i|w_{i-n} \ldots w_{i-1}) \]

For each document in our corpus we estimate its domain relevance (using ONETA, McCrae et al., 2014):

\[ s_{O}(d) \]

The count is weighted by document relevance:

\[ c(w_{i-n} \ldots w_i) = \sum s_{O}(d) c_{d}(w_{i-n} \ldots w_i) \]

\[ p(Welcome to Kobe) = p(Welcome) \times p(to|Welcome) \times p(Kobe|Welcome to) \]

\[ p(Kobe|Welcome to) = \frac{c(Welcome to Kobe)}{c(Welcome to *)} \]
Domain Language Model

- The language model estimates the likelihood of the target sentence
  - \( p(w_1 \ldots w_n) = \prod p(w_i|w_{i-n} \ldots w_{i-1}) \)
  - \( p(w_i|w_{i-n} \ldots w_{i-1}) = \frac{c(w_{i-n} \ldots w_{i-1}, w_i)}{c(w_{i-n} \ldots w_{i-1,*})} \)

- For each document in our corpus we estimate its domain relevance (using ONETA, McCrae et al., 2014)
  - \( s_O(d) \)

- The count is weighted by document relevance
  - \( c(w_{i-n} \ldots w_i) = \Sigma s_O(d) \ c_d(w_{i-n} \ldots w_i) \)
Evaluation

Metrics

- BLEU
- BLEU-2
- METEOR
- NIST
- PER
- WER
- TER

<table>
<thead>
<tr>
<th>Ontology</th>
<th>Size</th>
<th>Language</th>
</tr>
</thead>
<tbody>
<tr>
<td>IFRS 2009</td>
<td>2,757</td>
<td>🇪🇸 🇳🇱 🇩🇪 🇪🇸</td>
</tr>
<tr>
<td>DE-GAAP</td>
<td>2,782</td>
<td>🇪🇸 🇩🇪</td>
</tr>
<tr>
<td>LAG</td>
<td>196</td>
<td>🇪🇸 🇮-den</td>
</tr>
<tr>
<td>RB</td>
<td>1,449</td>
<td>🇪🇸 🇳🇱</td>
</tr>
<tr>
<td>HB</td>
<td>857</td>
<td>🇪🇸 🇳🇱</td>
</tr>
</tbody>
</table>

Finance Ontologies

Public Service Ontologies

BLEU is gold standard for machine translation but has been shown to perform very poorly on short texts (McCrae et al., 2011)
Results

Financial English to Spanish

- Moses (Baseline)
- Lexicon
- Features
- Language Model
- All

Metrics: BLEU, BLEU-2, METEOR, NIST, PER, TER, WER
Discussion

- Strongest improvement is provided by domain lexicon
  - Generalized for any corpus/domain (Arcan et al., 2015)
- Other methods alone did not significantly improve over the baseline
- The combination of all approaches better than domain lexicon in most settings (33/42 settings)
  - Main exception was English to Dutch for public services
- Domain adaptation was more effective for financial domain than public services domain
Conclusion

● Domain lexicon improves translation
  ○ Capable of suggesting new translations
  ○ Requires parallel text
● Domain selection weaker but still useful
● Code is open source:
  ○ https://github.com/monnetproject/translation
● Online demo (OTTO Ontology Translator)
  ○ http://server1.nlp.insight-centre.org/otto/


References


