From Bilingual Dictionaries to Interlingual Document Representations

Jagadeesh Jagarlamudi
University of Maryland (College Park)

Hal Daumé III
University of Maryland (College Park)

Raghavendra Udupa
Microsoft Research India
Problem statement
Problem statement
Problem statement
Problem statement

From Bilingual Dictionaries to Interlingual Document Projections
Problem statement

From Bilingual Dictionaries to Interlingual Document Projections
Problem statement

- Aim is to identify the **appropriate** interlingual space
Problem statement

Aim is to identify the appropriate interlingual space

- Cross-lingual Information Retrieval
- Multilingual Web Search
- Pair-wise document similarities
- Filtering stage
- Parallel phrase extraction
- Translation Mining

From Bilingual Dictionaries to Interlingual Document Projections
Agenda

- Problem Statement
  
  Learn the Interlingual document representations for cross-lingual documents

- Related Work

- Our approach
  
  - Symmetric Noisy Alignments
  
  - Supervised Kernelized Sorting

- Experiments
Related Work

• Dictionary based approaches

• Supervised Approaches
Related Work

• Dictionary based approaches
  • Use bilingual dictionaries to compute pair-wise similarities
    • Word-by-word translation [Ballesteros and Croft, 1996, Pirkola et al., 2001]
    • Generative Models [Boyd-Graber and Blei 2009, Jagarlamudi and Daumé III 2010]

• Supervised Approaches
Related Work

- Dictionary based approaches
  - Use bilingual dictionaries to compute pair-wise similarities
    - Word-by-word translation [Ballesteros and Croft, 1996, Pirkola et al., 2001]

- Supervised Approaches
  - Use a training data of aligned document corpus
  - Discriminative approaches
    - LSI, CCA and OPCA [Littman 1996, Vinakourov et al., 2003, Platt et al., 2010]
  - Generative Models
    - PTM, CPLSA, JPLSA [Mimno et al., 2009, Platt et al., 2010, Gao et al., 2011]
Our approach

pueblo : people 0.5
vocear : people 1e-3
Our approach

pueblo : people 0.5
vocear : people 1e-3
Our approach

From Bilingual Dictionaries to Interlingual Document Projections
Our approach

Noisy Aligner

pueblo : people 0.5
vocear : people 1e-3

Noisy document pairs
Our approach

- **Noisy Aligner**
  - *pueblo*: people 0.5
  - *vocear*: people $1 \times 10^{-3}$

- **Noisy document pairs**

- **Discriminative Learner**
Our approach

Noisy Aligner

Noisy document pairs

Discriminative Learner

Projection Directions (U, V)

pueblo : people 0.5
vocear : people 1e-3
Our approach

Noisy Aligner

pueblo : people 0.5
vocear : people 1e-3

Discriminative Learner

Projection Directions (U, V)
Our approach

1. Word-by-Word translation (Flow Formulation)

Noisy Aligner

Noisy document pairs

Discriminative Learner

Projection Directions $(U, V)$

pueblo : people 0.5
vocear : people 1e-3
Our approach

Noisy Aligner

pueblo : people 0.5
vocear : people 1e-3

Noisy document pairs

Discriminative Learner

Projection Directions
(U, V)

Word-by-Word translation
(Flow Formulation)

1.

2. Supervised Kernelized Sorting
1. Noisy Document Alignments

- Word-by-word translation
  - Different results based on the direction of translation

- Transform each document into bilingual word-pairs
  - Replace every word with all the bilingual word-pairs
  - “great river” → \{ great:gran, great:grandes, river:río, river:fluvial \ldots\}
  - “río que fluye” → \{ … río:river, río:stream, fluye:flow, \ldots\}

- Compute pair wise distances ($W$)
- Solve a soft bipartite matching ($\hat{\Pi}$)
2. Supervised Kernelized Sorting

- Kernelized Sorting (unsupervised) [Quadrianto et al. 2009]
  - Uses intra-language similarities to find an alignment
  - Formally $\hat{\Pi} = \arg \max_{\Pi} \text{tr}(K\Pi L\Pi^T)$
2. Supervised Kernelized Sorting

- Kernelized Sorting (unsupervised) [Quadrianto et al. 2009]
  - Uses intra-language similarities to find an alignment
  - Formally $\hat{\Pi} = \arg\max_{\Pi} \text{tr}(K\Pi L\Pi^T)$

- Supervised Kernelized Sorting
  - Fix an alignment say $\hat{\Pi}$, and change $K$, $L$
    $$K \hat{\Pi} L \hat{\Pi}^T = (X^TX) \hat{\Pi} (Y^TY) \hat{\Pi}^T$$
    $$(X^TU^TX)\hat{\Pi} (Y^TV^TY) \hat{\Pi}^T$$

$$\arg\max_{\{U, V\}} \text{tr} \left( (X^TU^TX) \hat{\Pi} (Y^TV^TY) \hat{\Pi}^T \right) \quad s.t \quad U^TU = I \quad \& \quad V^TV = I$$

Linear dot product kernel
Low-rank projection
2. Supervised Kernelized Sorting

\[
\text{arg max}_{\{U,V\}} \text{tr } \left( (X^TUU^TX) \tilde{\Pi} (Y^TVV^TY) \tilde{\Pi}^T \right) \quad \text{s.t. } U^TU = I & V^TV = I
\]

- Use alternative optimization
- Let \( C_{xy} = X\tilde{\Pi}Y^T \)
- Fix \( V \) to \( V_0 \) and solve for \( U \) as: \( (C_{xy}V_0V_0^TC_{xy}^T) U = \lambda_u U \)
- Fix \( U \) to \( U_0 \) and solve for \( V \) as: \( (C_{xy}^TU_0U_0^TC_{xy}) V = \lambda_v V \)
2. Supervised Kernelized Sorting

\[
\arg\max_{\{U,V\}} \text{tr} \left( (X^TUU^TX) \Pi (Y^TVVTY) \Pi^T \right) \quad \text{s.t. } U^TU = I \text{ and } V^TV = I
\]

- Use alternative optimization
- Let \( C_{xy} = X\Pi Y^T \)
- Fix \( V \) to \( V_0 \) and solve for \( U \) as: \( (C_{xy}V_0V_0^TC_{xy}^T) U = \lambda_u U \)
- Fix \( U \) to \( U_0 \) and solve for \( V \) as: \( (C_{xy}^TU_0U_0^TC_{xy}) V = \lambda_v V \)
2. Supervised Kernelized Sorting

\[
\text{arg max } \{U,V\} \quad tr \left( (X^T U U^T X) \tilde{\Pi} (Y^T V V^T Y) \tilde{\Pi}^T \right) \quad \text{s.t } U^T U = I \text{ & } V^T V = I
\]

- Use alternative optimization
- Let \( C_{xy} = X\tilde{\Pi}Y^T \)
- Fix \( V \) to \( V_0 \) and solve for \( U \) as: \( (C_{xy} V_0 V_0^T C_{xy}^T) U = \lambda_u U \)
- Fix \( U \) to \( U_0 \) and solve for \( V \) as: \( (C_{xy}^T U_0 U_0^T C_{xy}) V = \lambda_v V \)
2. Supervised Kernelized Sorting

\[
\arg \max_{\{U,V\}} \text{tr} \left( (X^T U U^T X) \tilde{\Pi} (Y^T V V^T Y) \tilde{\Pi}^T \right) \quad \text{s.t.} \ U^T U = I \ & \ V^T V = I
\]

- Use alternative optimization
- Let \( C_{xy} = X \tilde{\Pi} Y^T \)
  - Fix \( V \) to \( V_0 \) and solve for \( U \) as : \( (C_{xy} V_0 V_0^T C_{xy}^T) U = \lambda_u U \)
  - Fix \( U \) to \( U_0 \) and solve for \( V \) as : \( (C_{xy}^T U_0 U_0^T C_{xy}) V = \lambda_v V \)
- For the initial iteration
2. Supervised Kernelized Sorting

\[
\arg \max_{\{U,V\}} \text{tr} \left( (X^T U U^T X) \tilde{\Pi} (Y^T V V^T Y) \tilde{\Pi}^T \right) \quad \text{s.t} \quad U^T U = I \; & \; V^T V = I
\]

- Use alternative optimization
- Let \( C_{xy} = X \tilde{\Pi} Y^T \)
- Fix \( V \) to \( V_0 \) and solve for \( U \) as: \( (C_{xy} V_0 V_0^T C_{xy}^T) U = \lambda_u U \)
- Fix \( U \) to \( U_0 \) and solve for \( V \) as: \( (C_{xy}^T U_0 U_0^T C_{xy}) V = \lambda_v V \)
- For the initial iteration

\[
USV^T = C_{xy}
\]
Our approach

1. \( \hat{\Pi} \leftarrow \text{Noisy doc. aligner} \)

2. \( USV^T = C_{xy} \)
   \( \langle C_{xy} \leftarrow X\hat{\Pi}Y^T \rangle \)

Noisy Aligner

Noisy document pairs

Discriminative Learner

Projection Directions

\( \langle U, V \rangle \)

\text{pueblo} : \text{people} \ 0.5
\text{vocear} : \text{people} \ 1e-3
Experiments

- 2500 pairs of documents from Wikipedia
  - **Task**: Align these pairs
  - **Metric**: accuracy of top ranked document
- Bilingual dictionary is learned from Europarl

<table>
<thead>
<tr>
<th></th>
<th>En-Es</th>
<th>En-De</th>
</tr>
</thead>
<tbody>
<tr>
<td>Word-by-Word</td>
<td>0.597</td>
<td>0.564</td>
</tr>
</tbody>
</table>

Use dictionary
Experiments

- 2500 pairs of documents from Wikipedia
  - **Task**: Align these pairs
  - **Metric**: accuracy of top ranked document

- Bilingual dictionary is learned from Europarl

<table>
<thead>
<tr>
<th>Method</th>
<th>En-Es</th>
<th>En-De</th>
</tr>
</thead>
<tbody>
<tr>
<td>Word-by-Word</td>
<td>0.597</td>
<td>0.564</td>
</tr>
<tr>
<td>CCA</td>
<td>0.637</td>
<td>0.487</td>
</tr>
</tbody>
</table>
Experiments

- 2500 pairs of documents from Wikipedia
  - **Task**: Align these pairs
  - **Metric**: accuracy of top ranked document
- Bilingual dictionary is learned from Europarl

<table>
<thead>
<tr>
<th></th>
<th>En-Es</th>
<th>En-De</th>
</tr>
</thead>
<tbody>
<tr>
<td>Word-by-Word</td>
<td>0.597</td>
<td>0.564</td>
</tr>
<tr>
<td>CCA</td>
<td>0.637</td>
<td>0.487</td>
</tr>
<tr>
<td>OPCA</td>
<td>0.688</td>
<td>0.530</td>
</tr>
</tbody>
</table>

From Bilingual Dictionaries to Interlingual Document Projections
Experiments

- 2500 pairs of documents from Wikipedia
  - **Task**: Align these pairs
  - **Metric**: accuracy of top ranked document
- Bilingual dictionary is learned from Europarl

<table>
<thead>
<tr>
<th>Method</th>
<th>En-Es</th>
<th>En-De</th>
</tr>
</thead>
<tbody>
<tr>
<td>Word-by-Word</td>
<td>0.597</td>
<td>0.564</td>
</tr>
<tr>
<td>CCA</td>
<td>0.637</td>
<td>0.487</td>
</tr>
<tr>
<td>OPCA</td>
<td>0.688</td>
<td>0.530</td>
</tr>
</tbody>
</table>

Use training data | Use dictionary
Experiments

- 2500 pairs of documents from Wikipedia
  - **Task**: Align these pairs
  - **Metric**: accuracy of top ranked document
  - Bilingual dictionary is learned from Europarl

<table>
<thead>
<tr>
<th>Method</th>
<th>En-Es</th>
<th>En-De</th>
</tr>
</thead>
<tbody>
<tr>
<td>Word-by-Word</td>
<td>0.597</td>
<td>0.564</td>
</tr>
<tr>
<td>CCA</td>
<td>0.637</td>
<td>0.487</td>
</tr>
<tr>
<td>CCA (trained Wiki)</td>
<td>0.75</td>
<td>0.62</td>
</tr>
<tr>
<td>OPCA</td>
<td><strong>0.688</strong></td>
<td>0.530</td>
</tr>
</tbody>
</table>

From Bilingual Dictionaries to Interlingual Document Projections

Use training data

Use dictionary
Experiments

- 2500 pairs of documents from Wikipedia
  - **Task**: Align these pairs
  - **Metric**: accuracy of top ranked document
- Bilingual dictionary is learned from Europarl

<table>
<thead>
<tr>
<th>Method</th>
<th>En-Es</th>
<th>En-De</th>
</tr>
</thead>
<tbody>
<tr>
<td>Word-by-Word</td>
<td>0.597</td>
<td>0.564</td>
</tr>
<tr>
<td>CCA</td>
<td>0.637</td>
<td>0.487</td>
</tr>
<tr>
<td>CCA (trained Wiki)</td>
<td>0.75</td>
<td>0.62</td>
</tr>
<tr>
<td>OPCA</td>
<td>0.688</td>
<td>0.530</td>
</tr>
<tr>
<td>Ours (soft-match)</td>
<td>0.67</td>
<td>0.604</td>
</tr>
<tr>
<td>Ours (one-one)</td>
<td>0.658</td>
<td>0.590</td>
</tr>
</tbody>
</table>

From Bilingual Dictionaries to Interlingual Document Projections
Experiments

- 2500 pairs of documents from Wikipedia
  - **Task**: Align these pairs
  - **Metric**: accuracy of top ranked document
- Bilingual dictionary is learned from Europarl

<table>
<thead>
<tr>
<th>Method</th>
<th>En-Es</th>
<th>En-De</th>
</tr>
</thead>
<tbody>
<tr>
<td>Word-by-Word</td>
<td>0.597</td>
<td>0.564</td>
</tr>
<tr>
<td>CCA</td>
<td>0.637</td>
<td>0.487</td>
</tr>
<tr>
<td>CCA (trained Wiki)</td>
<td>0.75</td>
<td>0.62</td>
</tr>
<tr>
<td>OPCA</td>
<td>0.688</td>
<td>0.530</td>
</tr>
<tr>
<td>Ours (soft-match)</td>
<td>0.67</td>
<td>0.604</td>
</tr>
<tr>
<td>Ours (one-one)</td>
<td>0.658</td>
<td>0.590</td>
</tr>
</tbody>
</table>
Conclusions

- Use dictionaries to compute Interlingual document projections
  - Dictionaries generalize better

- Supervised kernelized sorting
  - Modify the kernel matrices for a given alignment
  - Similar to LSI
    - $USV^T = T$ (term x document matrix)
  - SVD can be computed efficiently and accurately
    - Readily available in many programming languages
    - Thus, scalable to bigger data sets
Thank You

Gracias

Preguntas y comentarios

Questions and Comments

English