Multiview Semi-Supervised Learning for Ranking Multilingual Documents

Nicolas Usunier*, Massih Amini*† and Cyril Goutte†

*LIP6, University of Paris 6, and
†Interactive Language Technologies, National Research Council Canada
Ranking Multilingual Documents

Ranking documents for

- Relevance (eg search),
- Importance (eg summarization),
- Recommendation...
Ranking Multilingual Documents

Ranking documents for

- Relevance (eg search),
- Importance (eg summarization),
- Recommendation...

Many countries and organizations handle multiple languages:

- Canada: English and French;
- European Union: 23 official languages and more...
- United Nations: 6 official languages;
- PAHO: Spanish, English, Portuguese, French.

Yet most document processing is monolingual (often English).
Semisupervised Ranking of Multilingual Documents

- **Ranking** documents
  - bipartite ranking

- **Multilingual** documents
  - multiview learning

- **Incomplete** ranking
  - semisupervised learning

We propose

1. Efficient multilingual ranking;
2. Multiview learning from partially observed labels;
3. Improvement over single-view semisupervised ranking;
4. Improvement over semisupervised multiview classification.
Index

1 Motivation ▷ 1

○ 2 Framework ▷ 4

3 Algorithm ▷ 9

4 Experimental results ▷ 12

5 Conclusion ▷ 20
Multiview ranking framework

Bipartite ranking labeled data $Z = (x^i, y^i)_{i=1}^n$:

- Observations $x^i$, sampled i.i.d. from fixed but unknown distribution,
- $y^i \in \{-1, +1\}$ the relevance of observation $x^i$.

Unlabeled data $U = (x^{n+j})_{j=1}^m$ i.i.d. from same distribution.

Goal: ranking observations $x$ so that relevant ($y = +1$) observations are above non relevant ($y = -1$) observations.

Multiview observations $x = (x_1, \ldots, x_V)$, $x_v \in \mathcal{X}_v, v \in \{1 \ldots V\}$.

Eg: document $x$ available in $V$ languages: $x_1, x_2, \ldots x_V$.

Goal: learn ranking functions $h_v : \mathcal{X}_v \rightarrow \mathbb{R}$, $v \in \{1, \ldots V\}$.
Ranking Risk(s)

Ranking = minimize ranking risk:  

\[
L(h) = \mathbb{P}( (Y - Y') \text{sgn}(h(X) - h(X')) < 0 )
\]

which may be estimated by the empirical estimate:

\[
\hat{L}_Z(h) = \frac{1}{n(n-1)} \sum_{i,j} \mathbb{I}\{y_i > y_j\} \mathbb{I}\{h(x^i) \leq h(x^j)\}
\]

**Multiview** learning: minimize average risk of view-specific scoring functions \(h_v\).

Plus: want rankers to agree on all views.

---


Cyril Goutte
(Dis)Agreement Constraint

Joint learning of view-specific rankers = reduce risk + constrain to agree.

Constraining view-specific predictors to agree ⇒ Reduce function space ⇒ Regularization ⇒ Better generalization.

(Dis)agreement estimated without labels ⇒ semisupervised learning.

Using Rademacher complexity argument, given disagreement threshold $t$:

$$\forall (h_1, \ldots, h_V) \in \mathcal{H}(t), \frac{1}{V} \sum_{v=1}^{V} L(h_v) \leq \frac{1}{V} \sum_{v=1}^{V} \hat{L}_Z(h_v) + R_n(\mathcal{H}(t), \delta).$$

→ Principle of semisupervised multiview ranking:

▶ small empirical risk on labeled data.

▶ small empirical disagreement on unlabeled data.

---

Disagreement for Bipartite Ranking

Natural measure: probability that $h_v$ and $h_{v'}$ disagree over two observations:

$$D(h_v, h_{v'}) = \mathbb{P}(\text{sgn}(h_v(X) - h_v(X')) \neq \text{sgn}(h_{v'}(X) - h_{v'}(X')))$$

May be estimated on unlabeled data:

$$\hat{D}_U(h_v, h_{v'}) \propto \sum_{i \neq j} \mathbb{I}\{(h_v(x^{n+i}_v) - h_v(x^{n+j}_v))(h_{v'}(x^{n+i}_v) - h_{v'}(x^{n+j}_v)) < 0\}$$

Same as Kendall’s tau statistic.

To extend to any number of views:

$$D(h_1, \ldots, h_V) = \frac{2 \sum_{v < v'} D(h_v, h_{v'})}{V(V - 1)} \quad \text{and} \quad \hat{D}_U(h_1, \ldots, h_V) = \frac{2 \sum_{v < v'} \hat{D}_U(h_v, h_{v'})}{V(V - 1)}$$
Index

1 Motivation ▷ 1

2 Framework ▷ 4

3 Algorithm ▷ 9

4 Experimental results ▷ 12

5 Conclusion ▷ 20
Algorithm

Iterative pseudolabeling, relying on efficient supervised bipartite ranking algo: label examples on which all view-specific models agree. → a natural way to get low disagreement.

In classification, checking consensus and labeling examples is straightforward.

Could do the same in ranking by labeling pairs of examples, but:
  ▶ labeling arbitrary pairs may be inconsistent with bipartite ranking,
  ▶ needs a pass over pairs of examples ($O(\ell^2)$), and
  ▶ need algorithm that learns from arbitrary pairs ($O(\ell^2)$).

Solve this by
  ▶ Subsampling pairs of example for pseudolabeling;
  ▶ Weighted pseudolabeling: examples may be included several times;
  ▶ Relying on efficient ($O(\ell)$) algorithms for bipartite ranking (linear SVM).
Semisupervised Multiview Ranking Algorithm

**Input:** Labeled and unlabeled sets $Z = (x^i, y^i)_{i=1}^n$ and $U = (x^{n+j})_{j=1}^m$; supervised bipartite ranking algorithm $A$; sampling size $S$.

**Initialize:** $t \leftarrow 0$

- Train $h_v^{(0)}$ on $Z$ with $A$, $\forall v = 1 \ldots V$.

**Repeat:** $t \leftarrow t + 1$

- For $s = 1 \ldots S$
  - Sample $(i, j) = \text{from } \{(k, \ell) \in \{1, \ldots, m\}^2, k \neq \ell\}$,
  - If $\forall v, h_v^{(t)}(x_v^{n+i}) > h_v^{(t)}(x_v^{n+j})$ then $Z \leftarrow Z \cup \{(x_v^{n+i}, +1), (x_v^{n+j}, -1)\}$
  - If $\forall v, h_v^{(t)}(x_v^{n+i}) < h_v^{(t)}(x_v^{n+j})$ then $Z \leftarrow Z \cup \{(x_v^{n+i}, +1), (x_v^{n+j}, -1)\}$
  - Train $h_v^{(t)}$ on $Z$ with $A$, $\forall v = 1 \ldots V$.

**Until** $\hat{D}_U(h_1^{(t)}, \ldots, h_V^{(t)}) \geq \hat{D}_U(h_1^{(t-1)}, \ldots, h_V^{(t-1)})$

**Output:** $\forall v \in \{1, \ldots V\}, h_v^{(t)}$
# Index

1. Motivation ▶ 1

2. Framework ▶ 4

3. Algorithm ▶ 9

   4. *Experimental results* ▶ 12

5. Conclusion ▶ 20
Experiments: Data

Publicly available: http://multilingreuters.iit.nrc.ca/

- Extracted from RCV1/RCV2;
- 6 categories;
- 5 languages / views;
- All docs translated to all languages;
- ⇒ 111k docs, 5 views.

<table>
<thead>
<tr>
<th>Language</th>
<th># docs</th>
<th>Cat</th>
<th># docs</th>
<th>(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>En</td>
<td>18,758</td>
<td>C15</td>
<td>18,816</td>
<td>16.84</td>
</tr>
<tr>
<td>Fr</td>
<td>26,648</td>
<td>CCAT</td>
<td>21,426</td>
<td>19.17</td>
</tr>
<tr>
<td>Ge</td>
<td>29,953</td>
<td>ECAT</td>
<td>13,701</td>
<td>12.26</td>
</tr>
<tr>
<td>It</td>
<td>24,039</td>
<td>E21</td>
<td>19,198</td>
<td>17.18</td>
</tr>
<tr>
<td>Sp</td>
<td>12,342</td>
<td>GCAT</td>
<td>19,178</td>
<td>17.16</td>
</tr>
<tr>
<td>Σ</td>
<td>111,740</td>
<td>M11</td>
<td>19,412</td>
<td>17.39</td>
</tr>
</tbody>
</table>

Documents indexed using title+body, lowercased, filtering stopwords, non words and low frequency tokens, digit-mapped, tf-idf weighting.

Split 75-25% for training-testing.

10 random labeled/unlabeled/test splits.

Evaluation in Average Precision (AvP) and Area Under the ROC Curve (AUC).
Experiments: Models

**1R:** fully supervised, single view ranking. (step 0 in algo) → absolute baseline in ranking.

**S1R:** semisupervised single view ranking.\(^3\) → adds semisupervised learning,
   → checks performance of single view vs. multiview.

**SMC:** semisupervised multiview classification.\(^4\) → classification counterpart to our approach,
   → checks performance of classification vs. ranking.

**SCR:** semisupervised ranking on concatenated views.
   → alternate, “baseline” semisup multiview ranking,
   — requires having all views available at test time!

**SMR:** semi-supervised multi-view ranking.
   → our approach.


Experiments: Performance (AUC)

<table>
<thead>
<tr>
<th>Model</th>
<th>C15</th>
<th>CCAT</th>
<th>E21</th>
<th>ECAT</th>
<th>GCAT</th>
<th>M11</th>
</tr>
</thead>
<tbody>
<tr>
<td>1R</td>
<td>.669↓</td>
<td>.624↓</td>
<td>.621↓</td>
<td>.638↓</td>
<td>.755↓</td>
<td>.811↓</td>
</tr>
<tr>
<td>SMC</td>
<td>.698↓</td>
<td>.645↓</td>
<td>.652↓</td>
<td>.649↓</td>
<td>.773↓</td>
<td>.821↓</td>
</tr>
<tr>
<td>S1R</td>
<td>.724↓</td>
<td>.658↓</td>
<td>.665↓</td>
<td>.662↓</td>
<td>.802↓</td>
<td>.836↓</td>
</tr>
<tr>
<td>SCR</td>
<td>.752↓</td>
<td>.679↓</td>
<td>.672↓</td>
<td>.671↓</td>
<td>.839↓</td>
<td>.875↓</td>
</tr>
<tr>
<td>SMR</td>
<td><strong>.805</strong></td>
<td><strong>.727</strong></td>
<td><strong>.681</strong></td>
<td><strong>.694</strong></td>
<td><strong>.866</strong></td>
<td><strong>.901</strong></td>
</tr>
</tbody>
</table>

AUC averaged over 10 random splits (10 labeled examples) and 5 languages.

Our method (semisupervised multiview ranking, SMR) improves over

- (semi-supervised) single view ranking,
- (semi-supervised) multiview classification,
- (semi-supervised) ranking on concatenated views.
Performance improves with more labeling (duh!) and difference decreases.
Disagreement during learning

Algorithm effectively *enforces agreement* ⇒ *better generalization*. One iteration with 10 examples yields better agreement than 200 at start.
Effect of class imbalance

Ranking outperforms classification when classes are imbalanced.
Comparison with concatenated views

Better than concatenation (SCR) especially when many views are available.
Index

1 Motivation ▶ 1

2 Framework ▶ 4

3 Algorithm ▶ 9

4 Experimental results ▶ 12

5 Conclusion ▶ 20
Conclusion

- Consider learning from multilingual document as a multiview problem.
- Learn multiview (bipartite) ranking from partially annotated data.
- Outperform independent single-view ranking;
- Outperform multiview classification;
- Outperform simple view concatenation.
- Better performance when 1) few annotated examples, 2) unbalanced data and 3) many views.
- Importance of optimizing a ranking (vs. binary classification) criterion.
- May generalize to arbitrary ranking (with complexity hit?).
The end

Thank you.

Questions?
## Index

1. Motivation ▶ 1
2. Framework ▶ 4
3. Algorithm ▶ 9
4. Experimental results ▶ 12
5. Conclusion ▶ 20