National Research Council Canada
Conseil national de recherches Canada

NRC·CNRC Institute for Information Technology

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
  $\longrightarrow$ bipartite ranking
- Multilingual documents
  $\longrightarrow$ multiview learning
- Incomplete ranking
  $\longrightarrow$ 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

Multiview ranking framework

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

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

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

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

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

Eg: document $\mathbf{x}$ available in V languages: x_1, x_2, $\dots x_V$.

Goal: learn ranking functions $h_v : \mathcal{X}_v \to \mathbb{R}$, $v \in \{1, \dots V\}$.

Ranking Risk(s)

Ranking = minimize ranking risk:[1]

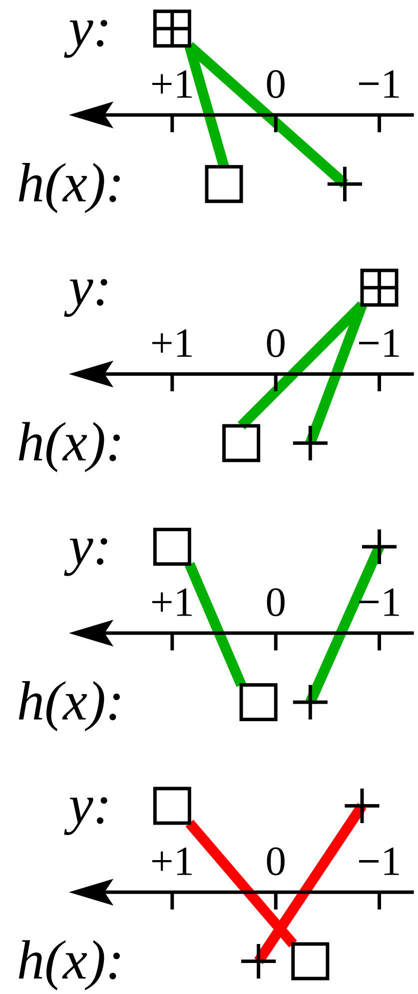
$$L(h) = \mathbb{P}\big((Y - Y')sgn(h(X) - h(X')) < 0\big)$$

which may be estimated by the empirical estimate:

$$\hat{L}_Z(h) = \frac{1}{n(n-1)} \sum_{i,j} \mathbb{I}_{\{y^i > y^j\}} \mathbb{I}_{\{h(\mathbf{x}^i) \leq h(\mathbf{x}^j)\}}$$

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

Plus: want rankers to agree on all views.



[1] Clémençon, Lugosi, Vayatis (2005) Ranking and scoring using empirical risk minimization, *COLT*.

(Dis)Agreement Constraint

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

Constraining view-specific predictors to agree $\Rightarrow$ Reduce function space $\Rightarrow$ Regularization $\Rightarrow$ Better generalization.

(Dis)agreement estimated without labels $\Rightarrow$ semisupervised learning.

Using Rademacher complexity argument,[2] given disagreement threshold t:

$$\forall (h_1, ..., h_V) \in \mathcal{H}(t)\,, \underbrace{\frac{1}{V}\sum_{v=1}^{V} L(h_v)}_{\text{true risk}} \leq \underbrace{\frac{1}{V}\sum_{v=1}^{V} \hat{L}_Z(h_v)}_{\text{emp. risk}} + \underbrace{\mathcal{R}_n(\mathcal{H}(t)\,, \delta)}_{\text{complexity penalty}}\,.$$

$\rightarrow$ Principle of semisupervised multiview ranking:

- small empirical risk on labeled data.
- small empirical disagreement on unlabeled data.

[2] Usunier, Amini, Gallinari (2005) A data-dependent generalization error bound for the AUC, *ICML workshop*.

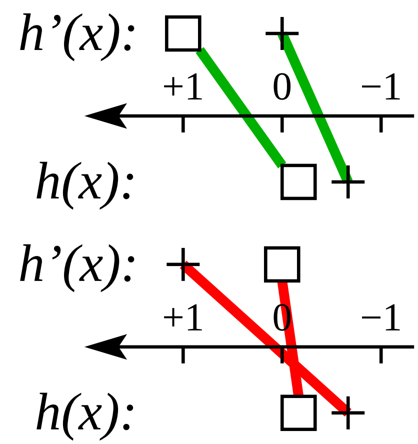
Disagreement for Bipartite Ranking

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

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

May be estimated on unlabeled data:

$$\widehat{D}_U(h_v, h_{v'}) \propto \sum_{i \neq j} \mathtt{I}_{\left\{ \left(h_v(x_v^{n+i}) - h_v(x_v^{n+j})\right)\left(h_{v'}(x_v^{n+i}) - h_{v'}(x_v^{n+j})\right) < 0 \right\}}$$



Same as Kendall's tau statistic.

To extend to any number of views:

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

Index

Algorithm

Iterative pseudolabeling, relying on efficient supervised bipartite ranking algo: label examples on which all view-specific models agree.
$\rightarrow$ 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 $Z = (\mathbf{x}^i, y^i)_{i=1}^n$ and $U = (\mathbf{x}^{n+j})_{j=1}^m$;
Supervised bipartite ranking algorithm $\mathcal{A}$; sampling size S.

Initialize: $t \leftarrow 0$

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

Repeat: $t \leftarrow t + 1$;

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

Until $\hat{D}_U\left(h_1^{(t)}, ..., h_V^{(t)}\right) \geq \hat{D}_U\left(h_1^{(t-1)}, ..., h_V^{(t-1)}\right)$

Output: $\forall v \in \{1, \dots V\}, h_v^{(t)}$

Index

Experiments: Data

Publicly available: `http://multilingreuters.iit.nrc.ca/` $\leftarrow$ Ad

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

	# docs
En	18,758
Fr	26,648
Ge	29,953
It	24,039
Sp	12,342
$\Sigma =$	111,740

cat	# docs	(%)
C15	18,816	16.84
CCAT	21,426	19.17
ECAT	13,701	12.26
E21	19,198	17.18
GCAT	19,178	17.16
M11	19,412	17.39

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.

[3] Amini, Truong, Goutte (2008) A boosting algorithm for learning bipartite ranking functions. . . , *SIGIR*.
[4] Amini, Usunier, Goutte (2009) Learning from multiple partially observed views. . . , *NIPS-22*.

Experiments: Performance (AUC)

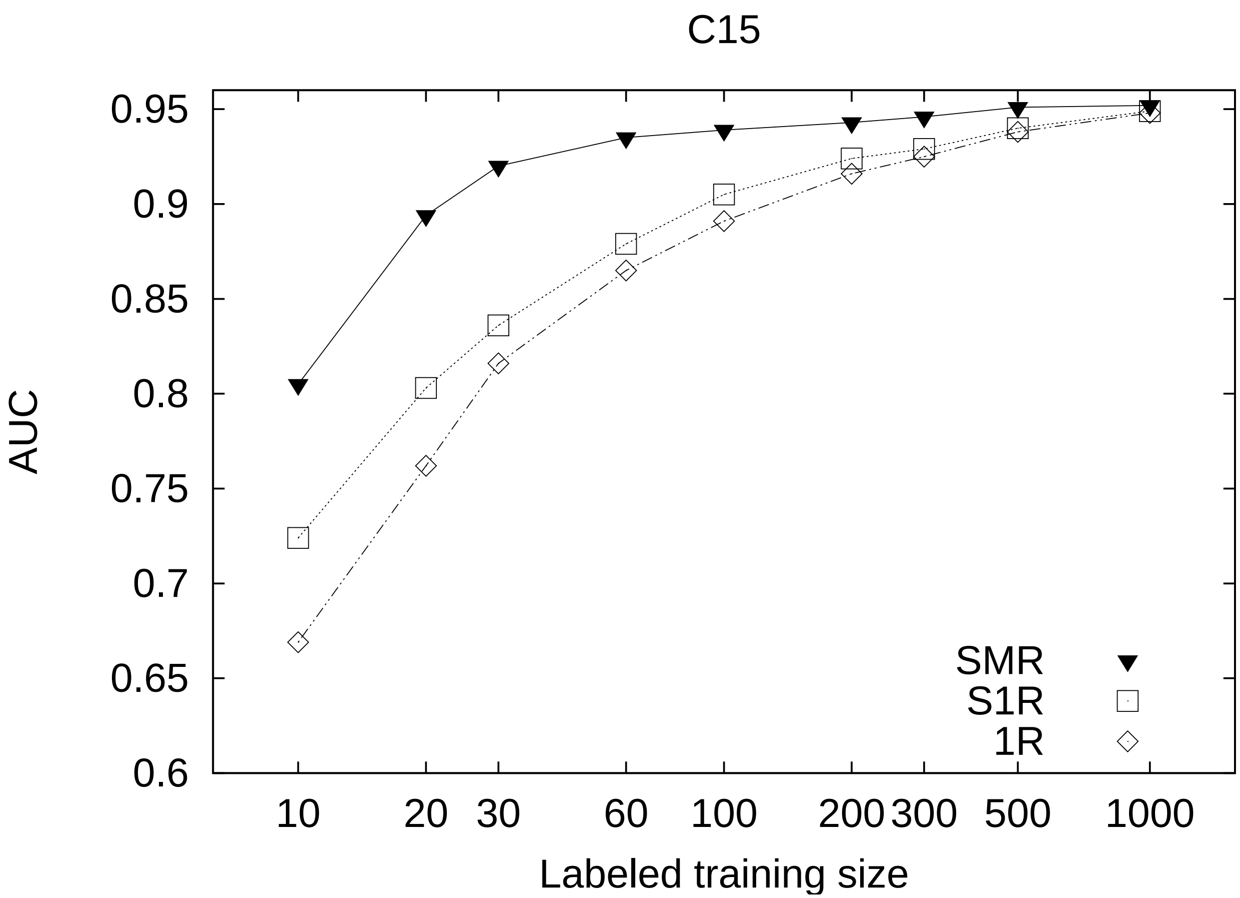
Model	C15	CCAT	E21	ECAT	GCAT	M11
1R	.669$^{\downarrow}$	.624$^{\downarrow}$	.621$^{\downarrow}$	.638$^{\downarrow}$	.755$^{\downarrow}$	.811$^{\downarrow}$
SMC	.698$^{\downarrow}$	.645$^{\downarrow}$	.652$^{\downarrow}$	.649$^{\downarrow}$	.773$^{\downarrow}$	.821$^{\downarrow}$
S1R	.724$^{\downarrow}$	.658$^{\downarrow}$	.665$^{\downarrow}$	.662$^{\downarrow}$	.802$^{\downarrow}$	.836$^{\downarrow}$
SCR	.752$^{\downarrow}$	.679$^{\downarrow}$	.672$^{\downarrow}$	.671$^{\downarrow}$	.839$^{\downarrow}$	.875$^{\downarrow}$
SMR	**.805**	**.727**	**.681**	**.694**	**.866**	**.901**

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

Our method (semisupervised multiview ranking, SMR) improves over

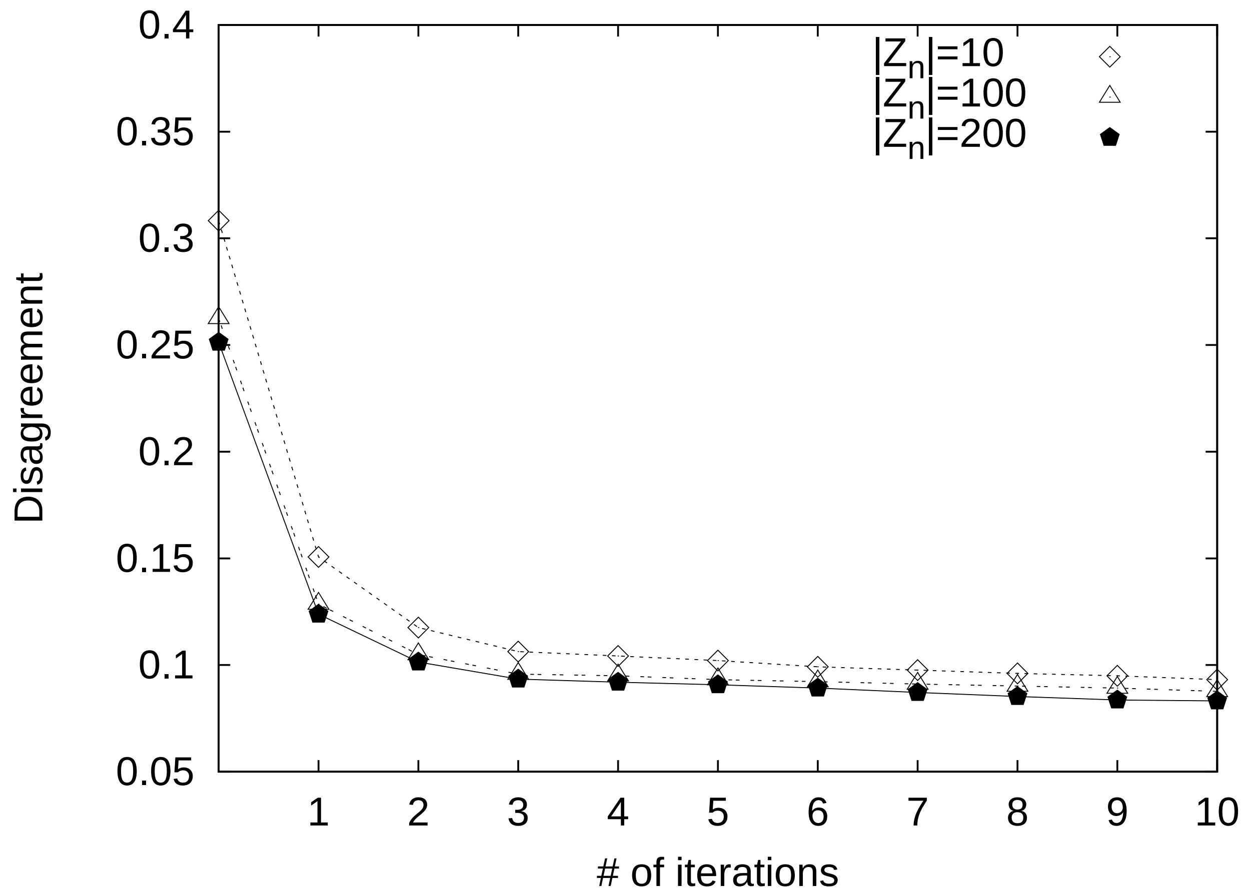
- (semi-supervised) single view ranking,
- (semi-supervised) multiview classification,
- (semi-supervised) ranking on concatenated views.

Performance vs. training set size



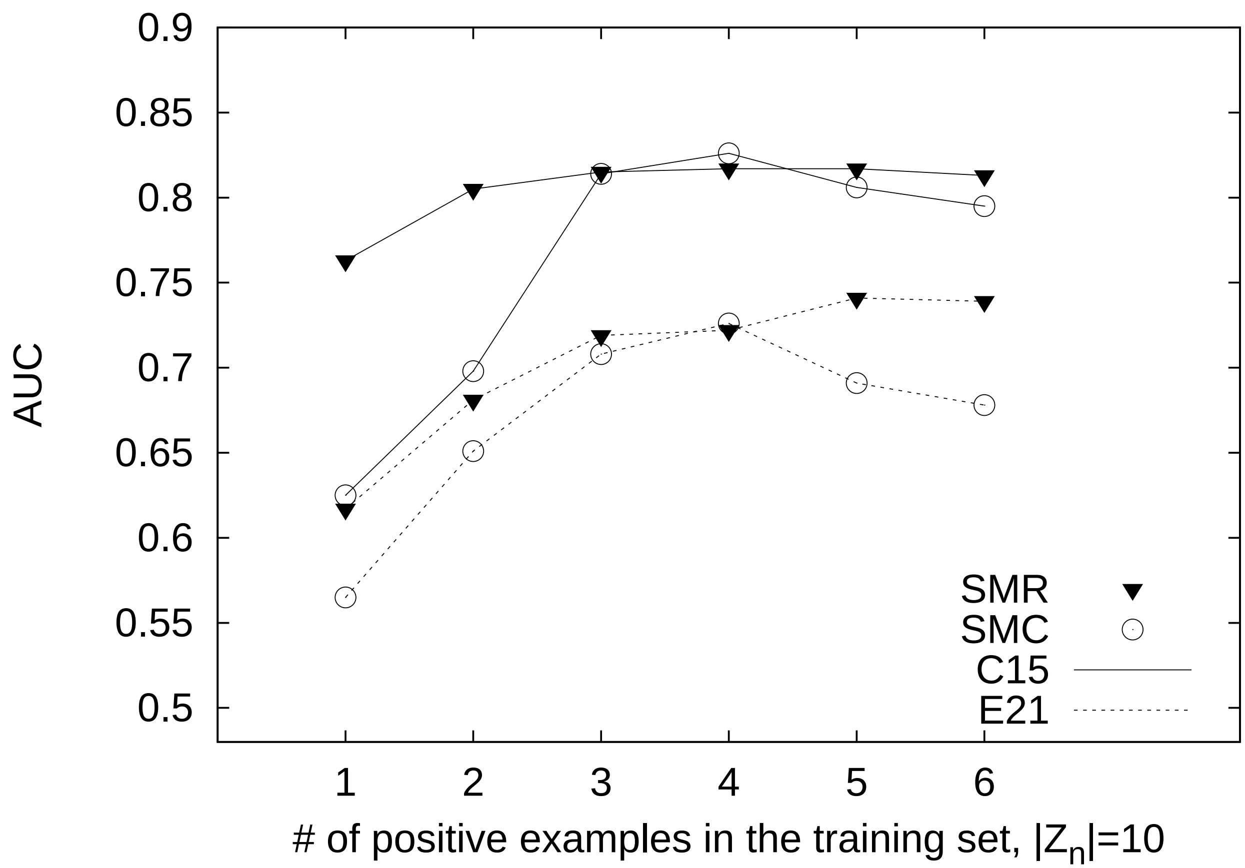
Performance improves with more labeling (duh!) and difference decreases.

Disagreement during learning



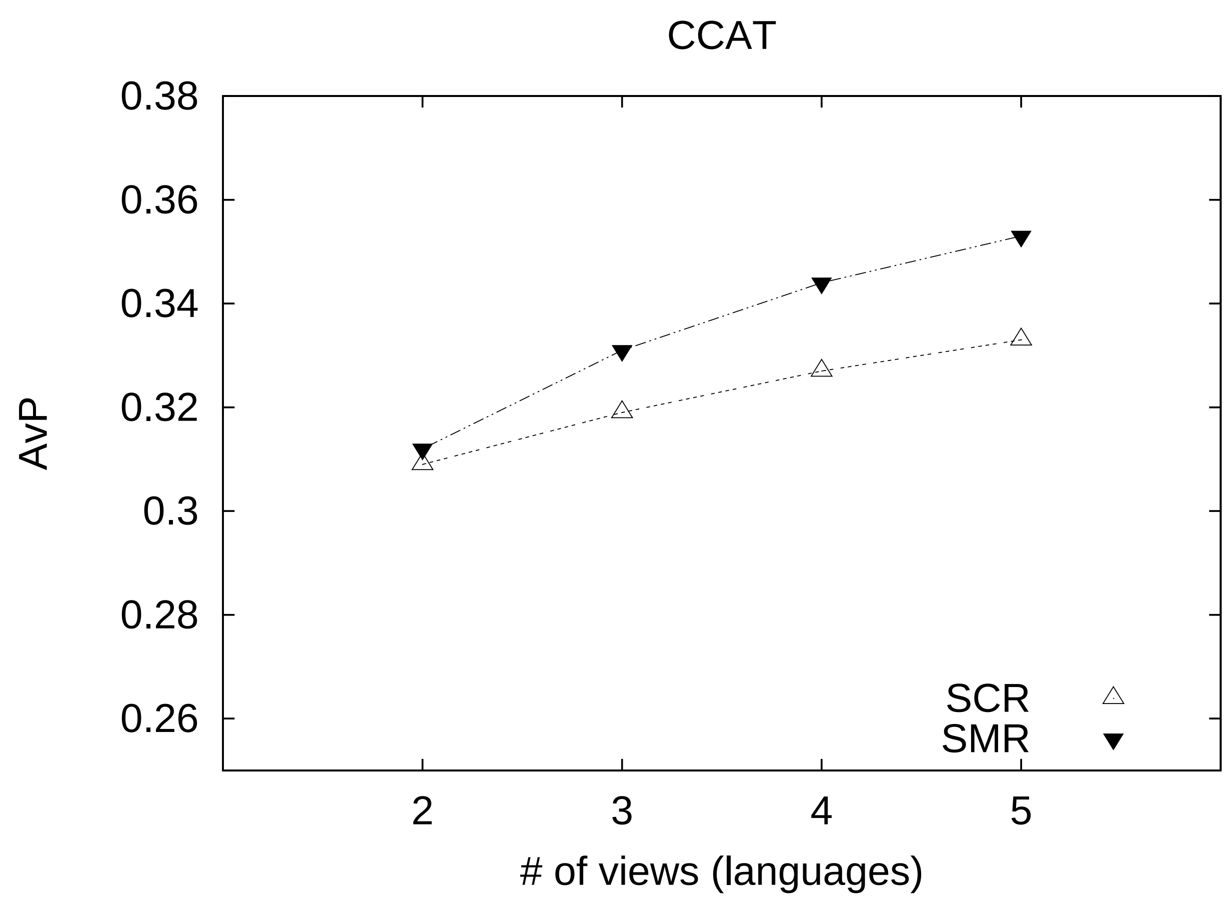
Algorithm effectively enforces agreement $\Rightarrow$ better generalization.
One iteration with 10 examples yields better agreement than 200 at start.

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

Conclusion

- Consider learning from multilingual document as a *multiview* problem.
- Learn multiview (bipartite) ranking from partially annotated data.

- Outperform independant 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