Combining Coregularization and Consensus-based Self-Training for Multilingual Text Categorization

Massih-Reza Amini[†] Cyril Goutte[†] Nicolas Usunier[‡]

[†]National Research Council Canada
Institute for Information Technology
283, boulevard Alexandre-Taché
Gatineau, J8X 3X7, Canada
First.Last@nrc-cnrc.gc.ca

[‡]Université Pierre et Marie Curie
Laboratoire d'Informatique de Paris 6
104, avenue du Président Kennedy
75016 Paris, France
Nicolas.Usunier@lip6.fr

ABSTRACT

We investigate the problem of learning document classifiers in a multilingual setting, from collections where labels are only partially available. We address this problem in the framework of multiview learning, where different languages correspond to different views of the same document, combined with semi-supervised learning in order to benefit from unlabeled documents. We rely on two techniques, coregularization and consensus-based self-training, that combine multiview and semi-supervised learning in different ways. Our approach trains different monolingual classifiers on each of the views, such that the classifiers' decisions over a set of unlabeled examples are in agreement as much as possible, and iteratively labels new examples from another unlabeled training set based on a consensus across language-specific classifiers. We derive a boosting-based training algorithm for this task, and analyze the impact of the number of views on the semi-supervised learning results on a multilingual extension of the Reuters RCV1/RCV2 corpus using five different languages. Our experiments show that coregularization and consensus-based self-training are complementary and that their combination is especially effective in the interesting and very common situation where there are few views (languages) and few labeled documents available.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: Information Storage - Record classification; I.2 [**Artificial Intelligence**]: Learning

General Terms

Algorithms, Experimentation, Theory

Keywords

Multilingual Document Classification, Learning from Multiple Views, Semi-supervised Learning

SIGIR'10, July 19–23, 2010, Geneva, Switzerland.
ACM 978-1-60558-896-4/10/07.

1. INTRODUCTION

In this paper, we address the problem of semi-supervised learning of document classifiers in a multilingual setting where documents are available as a *parallel corpus* with two or more languages for which labels are only partially available.

Our motivation is that multilingual collections are becoming more and more common in national and supranational contexts. However, the bulk of document classification and organization techniques and research is developed in the monolingual setting, most often for English. In addition, labeling text documents may require cost- and time-intensive human annotation, hence the widespread interest for semi-supervised text classification approaches that leverage unlabeled documents to speed-up the learning process.

Our work addresses the two issues of limited annotation and multilingual setting. Using the different languages as different views on a document, we develop a multiview, semi-supervised approach that learns from collection of multilingual documents.

We formalize the problem as follows. Given a collection of partially-labeled documents written in different languages and belonging to a set of classes that is fixed across languages, we wish to learn a number of monolingual classifiers for this common set of classes. Note that this problem is different from *cross-language text categorization* [5], where a document written in one language must be classified in a category system learned in another language.

In our setting, we assume that each document is available in several languages and we are interested in learning improved monolingual classifiers. We also emphasize that we wish to develop inter-dependent *monolingual* classifiers, rather than a single multilingual classifier, as we wish to be able to classify an incoming document in whatever language it is made available, without having to translate it beforehand.

There have been at least two approaches to multiview semi-supervised learning. One can use *coregularization* [19] to improve the view-specific classifiers by constraining them to agree on some unlabeled data, leveraging unlabeled data in a multiview learning framework. A more recent proposal [3], by contrast, leverages the multiple views in a semi-supervised learning framework by using the consensus between the different views in a self-training framework. Our solution is to combine those two components into a single boosting-based algorithm. View-specific classifiers are trained using coregularization, and a consensus-based self-

training process iteratively labels unlabeled examples on which the view-specific classifiers agree.

Using a large publicly available corpus of multilingual documents extracted from the Reuters RCV1 and RCV2 corpora, we show that our approach consistently improves over both coregularization and self-training taken in isolation. We also analyze the conditions in which the combination is most profitable. It turns out that adding coregularization to consensus-based self-training helps most when there are few languages and few documents available. This is a particularly interesting setting when resources are limited, and corresponds in particular to the common situation of bilingual data.

In the next section, we position our work with respect to the state of the art. In Section 3, we then present the problem of multiview semi-supervised learning for multilingual text classification. Section 4 describes the boosting-based algorithm we developed to obtain the language-specific classifiers. In Section 5, we present experimental results obtained with our approach on a subcollection of the Reuters RCV1/RCV2 corpus. Finally, in Section 6 we discuss the outcomes of this study and give some pointers to further research.

2. RELATION TO STATE-OF-THE-ART

Document classification has been a very popular application domain for Machine Learning algorithms, and in particular for multiview [7] and semi-supervised learning [16, 12]. The setting of *multilingual* document classification, however, has been much less studied so far [1, 2].

Interestingly, the original work on *co-training* [7] introduced both multiview *and* semi-supervised learning on a document classification task. Since then, both fields have developed greatly but mostly independently. Semi-supervised learning approaches include generative approaches, density-based or graph-based approaches (cf. [9] for an overview). Multiview learning techniques include *multiple kernel learning* [4] and techniques relying on kernel Canonical Correlation Analysis [11].

Some recent work more in line with the original co-training approach have introduced *coregularization* [19, 8], where classifiers are learnt in each view using a multiview regularizer that constrains predictions made in each view to be as similar as possible.

When this multiview regularizer is computed on unlabeled data, this provides a way to perform semi-supervised learning in a multiview setting. More recently, a semi-supervised multiview approach has been developed [3] where classifiers are learned on each view using standard single view training, but unlabeled examples are iteratively labeled in a *self-training* manner using the consensus across the views. The multiview consensus ensures higher confidence in the labeling, which yields improved semi-supervised learning rates.

Our work analyses and illustrates the combination of these two techniques. We use a coregularization component similar to [19, 8], with the key difference that instead of the coregularized least squares, we penalize disagreement using a Kullback-Leibler divergence which has a more natural interpretation in the context of probabilistic classifier outputs. In addition, it allows us to develop a novel boosting-based algorithm for solving the coregularized multilingual classification problem.

We combine this coregularized learning with a consensus-based self-training framework similar to [3] where unlabeled documents are iteratively labeled using the consensus prediction across the multiple views.

As both coregularization and consensus-based self-training use multiview information and unlabeled data for training, the key question we address is to see whether the two techniques can be complementary and improve on each other, as opposed to being completely redundant. We also investigate in which conditions such a complementarity may be exploited. We are particularly interested in the effects of coregularization in the common situation where the number of views is small (eg bilingual documents) and few labeled data are available.

3. FRAMEWORK

We consider V input spaces $\mathcal{X}_v \subset \mathbb{R}^{d_v}; \forall v \in \{1, .., V\}$, and an output space $\mathcal{Y}$. We take $\mathcal{Y} = \{-1, +1\}$ since we restrict our presentation to binary classification. Each *multiview document* $\mathbf{x} \in \mathcal{X}_1 \times ... \times \mathcal{X}_V$ is a sequence

$$\mathbf{x} \stackrel{def}{=} (x^1, ..., x^V)$$

where each *view* x^v provides a representation of the same document in a different vector space $\mathcal{X}_v$. In the seminal work on *co-training* [7], web pages are represented by either their textual content (first view) or anchor text pointing to them (second view). In our setting of multilingual classification, each view is the textual representation in a different language. Although typically one of the views is the original version of the document and the others are its translations, we never rely on this information and treat all views equally. Note that in this framework all views of each document are present simultaneously, hence we deal with multilingual text classification in a *parallel corpus*.

We further assume that we have a labeled training set $\mathcal{Z}_\ell = \{(\mathbf{x}_i, y_i) | i \in \{1, .., l\}\}$ and a possibly much larger set of unlabeled training data that we split into two parts denoted respectively by $X^1_\mathcal{U} = \{\mathbf{x}_{l+i} | i \in \{1, .., m_1\}\}$ and $X^2_\mathcal{U} = \{\mathbf{x}_{l+m_1+i} | i \in \{1, .., m_2\}\}$. Our goal is to obtain V binary classifiers $\{h_v : \mathcal{X}_v \to \{-1, 1\} | v \in \{1, .., V\}\}$, working each on one view, such that the predictive performance as estimated for example from a test set is optimized. Note that by construction, the label for a given document is the same for all views.

4. MODEL

We iteratively learn each classifier $h_v, \forall v \in \mathcal{V}$, while keeping fixed the classifiers for the other views, $h_u, u \in \mathcal{V} \wedge u \neq v$, by optimizing the loss

$$\mathcal{L}(h_v, \mathcal{Z}_\ell, X^1_\mathcal{U}, \lambda) = \mathcal{C}(h_v, \mathcal{Z}_\ell) + \frac{\lambda}{V-1} \sum_{u=1, u \neq v}^{V} d(h_v, h_u, X^1_\mathcal{U}), \quad (1)$$

where $\mathcal{C}(h_v, \mathcal{Z}_\ell)$ is the (monolingual) cost of h_v on the labeled training set $\mathcal{Z}_\ell$, $d(h_v, h_u, X^1_\mathcal{U})$ measures the divergence between the two classifiers h_v and h_u on the unlabelled documents in $X^1_\mathcal{U}$, and λ is a discount factor which modulates the influence of the disagreement cost on the optimization.

For the monolingual cost, we consider the standard misclassification error:

$$\mathcal{C}(h_v, \mathcal{Z}_\ell) = \frac{1}{l}\sum_{i=1}^{l} [\![y_i h_v(x_i^v) \leq 0]\!],$$

where $[\![\pi]\!]$ is equal to 1 if the predicate π is true, and 0 otherwise. As this cost is non-continuous and non-differentiable, it is typically replaced by an appropriate convex and differentiable proxy. Following standard practice in Machine Learning algorithms, we replace $[\![z \leq 0]\!]$ by the upper bound $a\log(1+e^{-z})$, with $a = (\log 2)^{-1}$. The monolingual misclassification cost becomes:

$$\mathcal{C}(h_v, \mathcal{Z}_\ell) = \frac{1}{l}\sum_{i=1}^{l} a\log(1+\exp(-y_i h_v(x_i^v))),$$

Assuming that each classifier output may be turned into a posterior class probability, we measure the disagreement between the output distributions for each view using the Kullback-Leibler (KL) divergence. Using the sigmoid function $\sigma(z) = (1+e^{-z})^{-1}$ to map the real-valued outputs of the functions h_v and h_u into a probability, and assuming that the *reference* distribution is the output of the classifier learned on the other views, $h_u, u \in \{1,...,V\} \wedge u \neq v$, the disagreement $d(h_v, h_u, X^1_{\mathcal{U}})$ becomes

$$d(h_v, h_u, X^1_{\mathcal{U}}) = \frac{1}{m_1}\sum_{i=1}^{m_1} kl(\sigma(h_u(x^u_{l+i}))||\sigma(h_v(x^v_{l+i}))),$$

where for two binary probabilities p and q, the KL divergence is defined as:

$$kl(p||q) = p\log\left(\frac{p}{q}\right) + (1-p)\log\left(\frac{1-p}{1-q}\right)$$

There are two reasons for choosing the KL divergence: first, it is the natural equivalent in the classification context of the l_2 norm used for regression in previous work on coregularization [19, 8, 18]; second, it allows the derivation of a boosting approach for minimizing the local objective function (1), as described in the following section.

4.1 A view-specific boosting-like algorithm

In order to learn the classifier h_v for view v, we need to minimize

$$\mathcal{L}(h_v, \mathcal{Z}_\ell, X^1_{\mathcal{U}}, \lambda) = \frac{1}{l}\sum_{i=1}^{l} a\log(1+\exp(-y_i h_v(x_i^v)))$$
$$+\frac{\lambda}{(V-1)m_1}\sum_{i=1}^{m_1}\sum_{u=1,u\neq v}^{V} kl(\sigma(h_u(x^u_{l+i}))||\sigma(h_v(x^v_{l+i}))) \quad (2)$$

We show how the loss-minimization of (2) is equivalent to the minimization of a Bregman distance. This equivalence will allow us to employ the boosting-like parallel-update optimization algorithm proposed by [10] to learn a linear classifier $h_v : x^v \mapsto \langle \beta_v, x^v\rangle$ minimizing (2).

A Bregman distance B_F of a convex, continuously differentiable function $F : \Omega \to \mathbb{R}$ on a set of closed convex set Ω is defined as

$$\forall p, q \in \Omega, B_F(p||q) \stackrel{def}{=} F(p) - F(q) - \langle \nabla F(q), (p-q)\rangle .$$

One optimization problem arising from a Bregman distance is to find a vector $p^* \in \Omega$, closest to a given vector $q_0 \in \Omega$ with respect to B_F, under the set of linear constraints $\{p \in \Omega | p^t M_v = \tilde{p}^t M_v\}$, where $\tilde{p} \in \Omega$ is a specified vector and M_v is a $n \times d$ matrix, with n the number of examples in the training set and d the dimension of the problem.[1]

Defining the Legendre transform as

$$L_f(q, M_v\beta_v) \stackrel{def}{=} \operatorname*{argmin}_{p\in\Omega}(B_F(p||q) + \langle M_v\beta_v, p\rangle),$$

the dual optimization problem can be stated as finding a vector q in the closure $\bar{Q}$ of the set $Q = \{L_F(q, M_v\beta_v)|\beta \in \mathbb{R}^p\}$, for which $B_F(\tilde{p}||q)$ is the lowest, under the set of linear constraints $\{q \in \Omega | q^t M_v = \tilde{p}^t M_v\}$.

It has been shown that both of these optimization problems have the same unique solution [14]. Moreover, [10] have proposed a single parallel-update optimization algorithm to find this solution in the dual form.

They have further shown that their algorithm is a general procedure for solving problems which aim to minimize the exponential loss, like in Adaboost, or a log-likelihood loss, like in logistic regression. Indeed, they showed the equivalence of these two loss minimization problems in terms of Bregman distance optimization.

In order to apply the boosting algorithm proposed by [10], we have to define a continuously differentiable function F such that by properly setting Ω, $\tilde{p}$, q_0 and M_v, the Bregman distance $B_F(0||L_F(q_0, M_v\beta_v))$ is equal to Eq. (2). Following [10], we choose:

$$\forall p \in \Omega = [0,1]^n, F(p) = \sum_{i=1}^{n} \alpha_i^v \left(p_i \log p_i + (1-p_i)\log(1-p_i)\right),$$

where α_i^v are non-negative real-valued weights associated to examples x_i^v.

This definition yields that $\forall p, q \in \Omega \times \Omega$:

$$B_F(p||q) = \sum_{i=1}^{n} \alpha_i^v \left(p_i \log\left(\frac{p_i}{q_i}\right) + (1-p_i)\log\left(\frac{1-p_i}{1-q_i}\right)\right) \quad (3)$$

$$\text{and, } \forall i, L_F(q,z)_i = \frac{q_i e^{-\frac{z_i}{\alpha_i^v}}}{1-q_i+q_i e^{-\frac{z_i}{\alpha_i^v}}} \quad (4)$$

Using Equations (3) and (4), and setting $q_0 = \frac{1}{2}\mathbf{1}$, the vector with all components set to $\frac{1}{2}$, and M_v the matrix such that $\forall i, j, (M_v)_{ij} = \alpha_i^v y_i x_{ij}^v$,[2] the Bregman distance in Equation (3) writes:

$$B_F(0||L_F(q_0, M_v\beta_v)) = \sum_{i=1}^{n} \alpha_i^v \log(1+e^{-y_i\langle\beta_v, x_i^v\rangle}). \quad (5)$$

[1] We have deliberately set the number of examples to n as in our equivalent rewriting of the minimization problem the latter is not exactly m_1.

[2] All vectors $\forall i \in \{1,..,n\}, \alpha_i y_i x_i^v$ should be normalized in order to respect the constraint $M_v \in [-1,1]^{n\times d}$.

Algorithm 1: Parallel-update optimization algorithm

Input : Matrix $\forall v, M_v \in [-1,1]^{n\times d}$.
Initialize: Let $\forall v, \beta_v \leftarrow 0$
for $v = 1, ..., V$ **do**
  for $t = 1, 2, ...$ **do**
    Let $q^{(t)}$ be the solution of $L_F(q_0, M_v\beta_v^{(t)})$;
    for $j = 1, ..., d$ **do**
      $W_{v,j}^{(t)^+} \leftarrow \sum_{i:sign((M_v)_{ij})=+1} q_i^{(t)} |(M_v)_{ij}|$;
      $W_{v,j}^{(t)^-} \leftarrow \sum_{i:sign((M_v)_{ij})=-1} q_i^{(t)} |(M_v)_{ij}|$;
      $\delta_{v,j}^{(t)} \leftarrow \frac{1}{2}\log\left(\frac{W_{v,j}^{(t)^+}}{W_{v,j}^{(t)^-}}\right)$;
    end
    $\beta_v^{(t+1)} \leftarrow \beta_v^{(t)} + \delta_v^{(t)}$;
  end
end
Output : $\forall v$, the sequence $\beta_v^{(1)}, \beta_v^{(2)}, ...$ verifying

$$\lim_{t\to\infty} B_F(0||L_F(q_0, M_v\beta_v^{(t)})) = \inf_{\beta_v\in\mathbb{R}^d} B_F(0||L_F(q_0, M_v\beta_v))$$

By developing Eq. (2), we get:

$$\mathcal{L}(h_v, \mathcal{Z}_\ell, X_\mathcal{U}^1, \lambda) = K + \frac{1}{l}\sum_{i=1}^{l} a\log(1+\exp(-y_i h_v(x_i^v))) +$$
$$\frac{\lambda}{(V-1)m_1}\sum_{i=1}^{m_1}\sum_{u=1,u\neq v}^{V}\sigma(h_u(x_{l+i}^u))\log(1+e^{-h_v(x_{l+i}^v)}) +$$
$$\frac{\lambda}{(V-1)m_1}\sum_{i=1}^{m_1}\sum_{u=1,u\neq v}^{V}(1-\sigma(h_u(x_{l+i}^u)))\log(1+e^{h_v(x_{l+i}^v)}) \quad (6)$$

where K is a constant which does not depend on h_v.

In order to make Eq. (6) identical to Eq. (5) (up to a constant), we create, for each unlabeled document $x_i^v \in X_\mathcal{U}^1$, two examples $(x_i^v, +1)$ and $(x_i^v, -1)$ (which makes $n = l + 2m_1$), and set the weights as follows:

$$\alpha_i^v = \begin{cases} \frac{a}{l} & \text{if } \mathbf{x}_i \in \mathcal{Z}_\ell, \\ \frac{\lambda}{(V-1)m_1}\sum_{u=1,u\neq v}^{V} [\![y_i = -1]\!] + y_i\sigma(h_u(x_i^u)) & \text{else.} \end{cases} \quad (7)$$

As a consequence, minimizing Eq. (2) is equivalent to minimizing $B_F(0||q)$ over $q \in \bar{Q}$, where

$$Q = \{q \in [0,1]^{l+2m_1} \mid q_i = \sigma(y_i \langle \beta_v, x_i^v\rangle), \beta_v \in \mathbb{R}^{d_v}\}.$$

This equivalence allows us to adapt the parallel-update optimization algorithm described in [10] to learn each specific-view classifier, as described in Algorithm 1.

4.2 Coregularized semi-supervised algorithm

We embed the boosting-based coregularized classifier learning inside a self-training framework (cf. [22], Section 3) which relies on consensus across views in order to automatically label documents from an unlabeled document pool $X_\mathcal{U}^2$.

Each monolingual classifier $h_v, v \in \mathcal{V}$ is first initialized on the supervised monolingual cost alone, then we iteratively optimize each of the h_v classifiers while keeping the classifiers for the other views fixed, until the global objective

$$\Delta(\otimes_{v=1}^V h_v, \mathcal{Z}_\ell \cup \mathcal{Z}_\mathcal{U}, \lambda) = \sum_{v=1}^{V}\mathcal{L}(h_v, \mathcal{Z}_\ell \cup \mathcal{Z}_\mathcal{U}, X_\mathcal{U}^1, \lambda) \quad (8)$$

has reached a (possibly local) minimum.

This alternating optimization of partial cost functions bears similarity with the block-coordinate descent technique [6]. At each iteration, block coordinate descent splits variables into different subsets, the set of the active variables and the sets of inactive ones, then minimizes the objective function along active dimensions while inactive variables are fixed at current values.

Once all language-specific classifiers have been trained we assign class labels to unlabeled examples in $X_\mathcal{U}^2$ for which all mono-lingual classifiers predict the same class label. These newly labeled examples are added to the labeled training set. We then go back to the boosting-based coregularized classifier training using the combined labeled data, and so on until either no remaining unlabeled example can be labeled by consensus, or all unlabeled examples have been labeled. As shown by [3], focusing on functions which agree across several views reduces the complexity of the function class and therefore improves the prediction ability of the resulting model.

Algorithm 2 summarizes this coregularized self-training strategy.

Algorithm 2: Coregularized semi-supervised Learning

Input : A set of labeled training examples $\mathcal{Z}_\ell$;
Two sets of unlabeled training data $X_\mathcal{U}^1$ and $X_\mathcal{U}^2$;
Initialize: Set $\mathcal{Z}_\mathcal{U} \leftarrow \emptyset$;
$\forall v, h_v^{(0)} \stackrel{def}{=} \text{argmin}_h \mathcal{C}(h, \mathcal{Z}_\ell)$;
repeat
  $t \leftarrow 1$;
  repeat
    for $v = 1, .., V$ **do**
      Learn $h_v^{(t)} = \text{argmin}_h \mathcal{L}(h, \mathcal{Z}_\ell \cup \mathcal{Z}_\mathcal{U}, X_\mathcal{U}^1, \lambda)$;
    end
    $t \leftarrow t + 1$;
  until *Convergence of* $\Delta(\otimes_{v=1}^V h_v^{(t)}, \mathcal{Z}_\ell \cup \mathcal{Z}_\mathcal{U}, \lambda)$;
  – Let $X_\mathcal{U}$ be the set of unlabeled examples in $X_\mathcal{U}^2$ on which all classifiers agree over the class label of examples ;
  $-X_\mathcal{U}^2 \leftarrow X_\mathcal{U}^2 \setminus X_\mathcal{U}$;
  $-\mathcal{Z}_\mathcal{U} \leftarrow \mathcal{Z}_\mathcal{U} \cup X_\mathcal{U}$;
until $X_\mathcal{U}^2 = \emptyset$ *or* $X_\mathcal{U} = \emptyset$;
Output : Classifiers $h_v, \forall v \in \{1, ..., V\}$

5. EXPERIMENTS

We conducted a number of experiments aimed at evaluating how the combination of coregularization and consensus-based self-training can help to take advantage of multilingual unlabeled documents in order to learn efficient classification functions.

5.1 Data set

We perform experiments on a publicly available multilingual multiview text categorization corpus extracted from

Language	# docs
English	18,758
French	26,648
German	29,953
Italian	24,039
Spanish	12,342
Total	111,740

Class	# docs
C15	18,816
CCAT	21,426
E21	13,701
ECAT	19,198
GCAT	19,178
M11	19,421

Table 1: Number of documents per language (left) and per class (right) in Reuters RCV1/RCV2 subcollection used in our experiments.

the Reuters RCV1/RCV2 corpus [3].[3] This corpus contains more than 110K documents from 5 different languages, (English, German, French, Italian, Spanish), distributed over 6 classes (Table 1). Documents that originally had more than one of these 6 labels were assigned to the smallest class. We reserved a test split containing 25% of the documents, respecting class and language proportions. Within the training set containing the remaining 75% of documents, we randomly sampled labeled documents ($\mathcal{Z}_\ell$), and split the remaining unlabeled data into two subsets: one for evaluating the coregularization term ($X^1_\mathcal{U}$), and one for the self-training process ($X^2_\mathcal{U}$). The motivation for that split is to avoid bias: as coregularization enforces agreement between classifiers, it may yield artificially high consensus for the examples used in the coregularization term.

This corpus of multilingual documents is originally a comparable corpus as it covers the same subset of topics in all languages. In order to produce multiple views for each documents, each original document extracted from the Reuters corpus was translated in all other languages using a phrase-based statistical machine translation system [20]. The indexed translations are part of the corpus distribution.

More precisely, each document is indexed by the text appearing in its title (*headline* tag) and body (*body* tag). As preprocessing, all text is lowercased, digits are mapped to a single digit token, and tokens containing non-alphanumeric characters are removed. For each language, words in a stoplist as well as tokens occurring in less than 5 documents were also filtered out. Documents were then represented as a bag of words, using a TFIDF weighting scheme based on BM25 [17].

Results are evaluated over the test set using the accuracy and the standard F_1 measure [21], which is the harmonic average of precision and recall. The reported performance is averaged over the resulting five language-specific classifiers. In addition, we also averaged over 10 random (train/unlabeled/test) sets of the initial collection.

5.2 Experimental setup

To validate the coregularized consensus-based self-training approach described in the previous section, we test the following six classification methods. The first method is a purely supervised technique which does not make use of any unlabeled examples in the training stage. The following methods make use of the multiview and semi-supervised learning approaches in different ways, using coregularization and/or consensus-based self-training separately or in combination, over different subsets of the unlabeled training documents.

- **Baseline method** [Boost]: This baseline corresponds to a supervised monolingual boosting model optimizing Eq. 2 for $\lambda = 0$.

- **Coregularized boosting** [reg-Boost]: Boosting using coregularization on $X^1_\mathcal{U}$, optimizing Eq. 2 for $\lambda \neq 0$. This constrains the supervised monolingual boosting models to achieve high agreement among their predictions on $X^1_\mathcal{U}$.

- **Boosting with self-training** [Boost-cst]: Boosting using consensus-based self-training, but no coregularization. This is similar in spirit to the iterative co-training algorithm [7]. Given the language-specific classifiers trained on an initial set of labeled examples, we iteratively assign pseudo-labels to the unlabeled examples in $X^2_\mathcal{U}$ for which all classifier predictions agree.

- **SVM with self-training** [SVM-cst]: This is similar to the previous method except that we use the SVM-Perf package [13] to learn each language-specific classifiers instead of boosting. For tuning the hyperparameter C, we first tried the leave-one-out cross-validation strategy. However, with small training sets we found out that the default $(\frac{1}{l}\sum_{i=1}^{l}||x_i||)^{-1}$ gave similar, and in some cases, better results. We therefore used that default C in all of our experiments.

- **Coregularization+self-training** [reg-Boost-cst]: Coregularized boosting using the consensus-based self-training: The coregularization term is computed over $X^1_\mathcal{U}$ and self-training iteratively labels documents from $X^2_\mathcal{U}$.

- **Boosting with full self-training** [Boost-cst*]: In order to determine when the combination of coregularization and self-training is the most useful, we also trained algorithm Boost-cst using all the unlabeled training examples $X_\mathcal{U} = X^1_\mathcal{U} \cup X^2_\mathcal{U}$ rather than just those in $X^2_\mathcal{U}$.

Our aim is to show the gradual effect of each of the multiview and semi-supervised learning approaches on the boosting algorithm, progressing from Boost to reg-Boost and Boost-cst, to reg-Boost-cst. Note that the reg-Boost and Boost-cst algorithms use the two separate unlabeled training subsets in different manners. SVM-cst is the same as Boost-cst using a SVM algorithm instead of Boosting. This will allow us to benchmark the boosting-based algorithm against the state of the art SVM model in a similar framework. Note that adding co-regularization in a SVM implementation requires some significant changes to the underlying code, which is why we do not provide reg-SVM variants. Finally, using all the unlabeled training examples in Boost-cst* and comparing the results to reg-Boost-cst will allow us to uncover the situations in which it is beneficial to combine coregularization and self-training rather than use the latter alone on the combined unlabeled data. This gives an idea of the true benefit brought by coregularization.

[3] http://multilingreuters.iit.nrc.ca/

Table 2: Test classification accuracy and F_1 of different learning algorithms on the six classes, averaged over 10 random sets of 50 labeled examples per training set. For each class, the best result is in bold, and a $\downarrow$ indicates a result that is statistically significantly worse than the best, according to a Wilcoxon rank sum test with $p < .01$.

Strategy	C15		CCAT		E21		ECAT		GCAT		M11	
	Acc.	F_1	Acc.	F_1	Acc.	F_1	Acc.	F_1	Acc.	F_1	Acc.	F_1
Boost	$0.771^{\downarrow}$	$0.506^{\downarrow}$	$0.662^{\downarrow}$	$0.398^{\downarrow}$	$0.765^{\downarrow}$	$0.323^{\downarrow}$	$0.505^{\downarrow}$	$0.347^{\downarrow}$	$0.781^{\downarrow}$	$0.587^{\downarrow}$	$0.793^{\downarrow}$	$0.586^{\downarrow}$
reg-Boost	$0.793^{\downarrow}$	$0.532^{\downarrow}$	$0.689^{\downarrow}$	$0.419^{\downarrow}$	$0.783^{\downarrow}$	$0.342^{\downarrow}$	$0.513^{\downarrow}$	$0.372^{\downarrow}$	$0.803^{\downarrow}$	$0.608^{\downarrow}$	$0.815^{\downarrow}$	$0.611^{\downarrow}$
Boost-cst	$0.804^{\downarrow}$	$0.572^{\downarrow}$	$0.708^{\downarrow}$	$0.421^{\downarrow}$	$0.794^{\downarrow}$	$0.365^{\downarrow}$	$0.511^{\downarrow}$	$0.384^{\downarrow}$	$0.866^{\downarrow}$	$0.655^{\downarrow}$	$0.848^{\downarrow}$	$0.668^{\downarrow}$
SVM-cst	0.815	0.583	$0.720^{\downarrow}$	0.438	$0.800^{\downarrow}$	$0.378^{\downarrow}$	$0.522^{\downarrow}$	$0.395^{\downarrow}$	$0.873^{\downarrow}$	$0.662^{\downarrow}$	$0.861^{\downarrow}$	$0.676^{\downarrow}$
reg-Boost-cst	**0.823**	**0.595**	**0.748**	**0.449**	**0.815**	**0.394**	**0.542**	**0.408**	**0.895**	**0.687**	**0.883**	**0.693**

5.3 Experimental Results

We start our evaluation by analyzing the gains provided by coregularization, the consensus-based self-training and the combination of both, over the baseline boosting algorithm. We measure the classification accuracy and F_1 for a fixed number of labeled and unlabeled examples in the training set. In order to study the role of unlabeled data on the learning behavior we begin our experiments with very few labeled training examples. The size of the labeled training sets in these first experiments is fixed to 50 (an average of 10 per language), with an equal sampling of 25 positive and 25 negative examples in $\mathcal{Z}_\ell$. For coregularization, results are reported for the best discount factor $\lambda = 1$, although as illustrated in Section 5.3.1, results are fairly stable across a wide range of values. We will later investigate the impact on the test performance of the number of labeled examples and the number of views (cf Sections 5.3.3 and 5.3.2).

Table 2 summarizes results obtained by `Boost`, `reg-Boost`, `Boost-cst`, `SVM-cst` and `reg-Boost-cst` averaged over five languages and 10 random splits of tests sets for our six main categories. We use bold face to indicate the highest performance rates, and the symbol $\downarrow$ indicates that performance is significantly worse than the best result, according to a Wilcoxon rank sum test used at a p-value threshold of 0.01 [15]. From these results it becomes clear that:

1. Using the first part of the unlabeled training examples (X_U^1) to coregularize the boosting algorithm, algorithm `reg-Boost` always improves over `Boost` by an average of 2-3 points in F_1.

2. The consensus-based self-training framework implemented in `Boost-cst` and `SVM-cst` also improves over the baseline. In addition, it always seems to outperform coregularization (`reg-Boost`) alone. In this self-training framework, the `SVM` classifiers `SVM-cst` tend to outperform the boosting-based classifiers `Boost-cst`.

3. Finally, the combination of coregularization and self-training (`reg-Boost-cst`) produces a further improvement of around 1-2 points in F_1 over the best semi-supervised result (`SVM-cst`). The improvement is statistically significant in four classes out of six.

Our analysis of these results is that both coregularization and the consensus-based self-training provide consistent improvements over training independent monolingual classifiers. Both are instances of multiview learning, and both rely in some way on the consensus between classifiers trained on the different views. The question therefore arises as to how redundant these two techniques are? Our experimental results suggest that these techniques are in fact complementary.

The gains provided by adding coregularization to the self-training boosting-based model is in fact similar to the gain provided by coregularization in the supervised setting, which suggest that the two effects are essentially independent and additive. In order to analyze more finely the situations in which the combination of coregularization and consensus-based self-training is more advantageous, we compared all the algorithms, including `Boost-cst`*, for different numbers of languages and different amounts of labeled documents. These results are reported in Section 5.3.2 and 5.3.3, right after we address the issue of the discount factor λ.

5.3.1 *The effect of the coregularization factor λ*

We analyze the influence of the discount factor λ on the performance of `reg-Boost-cst` for varying amounts of labeled training data.[4] The results obtained on class `E21` are presented in Figure 1. Note that λ controls the relative importance of the unlabeled data in the coregularization (with $\lambda = 0$ corresponding to no regularization). Figure 1 shows that unlabeled examples become relatively less important as more labeled data is available: as the amount of labeled training data increases from 50 to 300, the optimal discount factor λ moves away from 1.

We recall that for $\lambda = 1$, unlabeled data plays the same role in the training procedure as labeled data.

Note also that in all cases, the performance of the resulting classifiers seems relatively stable for a wide range of values of λ. This suggests that the results are not overly sensitive to a precise choice of discount factor λ.

5.3.2 *The value of labeled data*

We also analyze the behavior of the various algorithms for growing initial amounts of labeled data in the training set. Figure 2, illustrates this by showing the F_1 measures on classes `CCAT` and `ECAT` with respect to the number of labeled documents in the initial labeled training set $\mathcal{Z}_\ell$. For all labeled data sizes, the proportion of negative/positive examples is maintained at 50%. As expected, all performance curves increase monotonically with respect to the additional

[4]We always maintain the proportion of positive/negative documents in the labeled training set to 50%/50%.

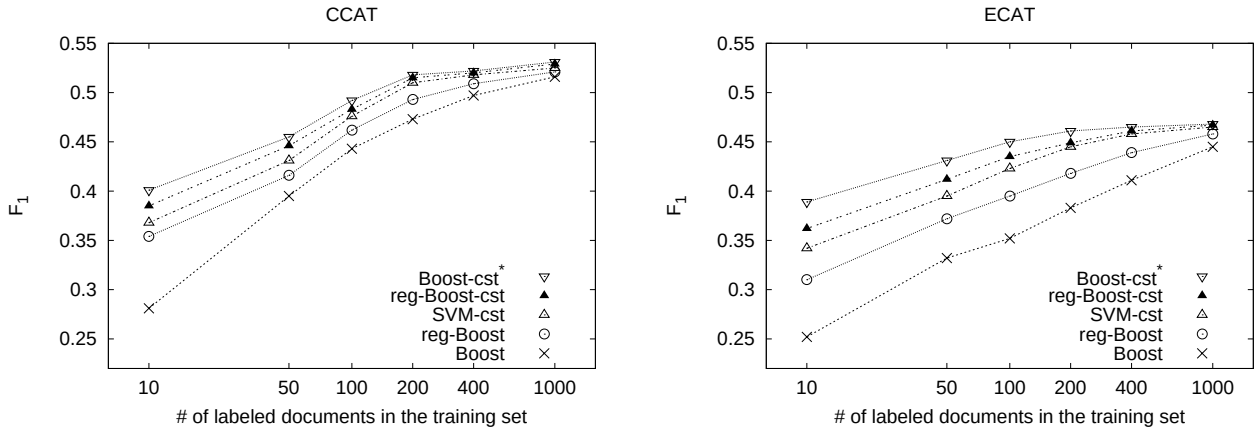

Figure 2: F_1 on classes CCAT and ECAT with respect to the number of labeled documents in the initial labeled training set $\mathcal{Z}_\ell$.

labeled data. When there are sufficient labeled examples, all algorithms actually converge to the same F_1 value, suggesting that the labeled data carries sufficient information and that no *additional* information could be extracted from unlabeled examples. For a low number of labeled training data, the contribution of each of the algorithms that use unlabeled data is clearly shown. Note that these curves are obtained using five languages, such that the highest performance is achieved by Boost-cst*, which is consistent with the findings of the previous section. When fewer views are available, the relative positions of the top algorithms are different, but the effect is similar in that the gains are more important when fewer initial labeled documents are available.

5.3.3 The effect of the number of languages

In our experiments, the unlabeled training set was split in two parts, one for coregularization and one for self-training. Our motivation was to examine the effect of each of the techniques individually without introducing any bias by performing coregularization and self-training on the same unlabeled data. The previous results suggest that the performance gain is higher when unlabeled examples are iteratively labeled in the self-training framework than when they are used in coregularization to enforce agreement between the language-specific classifiers. The question therefore arises as to what the performance would be if all the unlabeled examples were used in consensus-based self-training rather than being split between coregularization and self-training? In addition, the consensus is expected to be more reliable when there are many views than when there are few, in which case the language-specific classifiers could agree by chance but erroneously. We therefore investigate the effect of the number of views on the performance of the reg-Boost-cst and Boost-cst* algorithms. Figure 3 depicts these results by comparing both algorithms for varying numbers of languages on two classes, E21 and C15. All re-

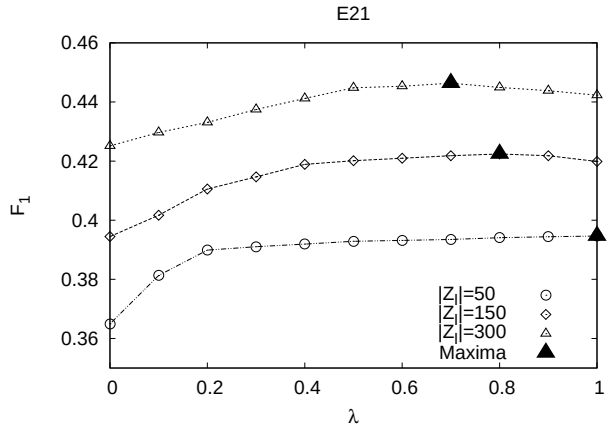

Figure 1: F_1 with respect to the coregularization factor λ for different labeled training sizes on class E21.

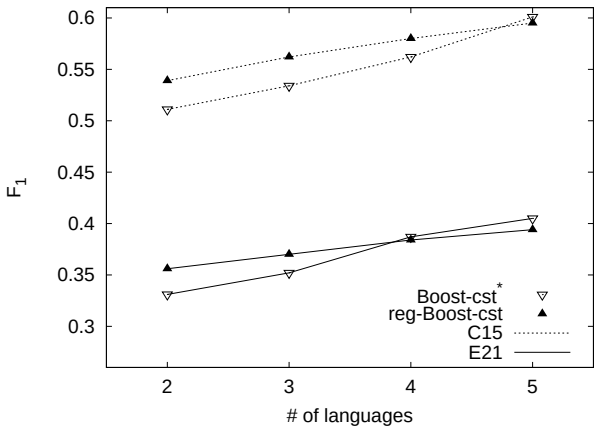

Figure 3: F_1 with respect to the number of languages used for coregularization and self-training on classes E21 (solid) and C15 (dash). Comparisons involve reg-Boost-cst (▲) and the boosting algorithm using the unlabeled examples ($X_{\mathcal{U}}^1 \cup X_{\mathcal{U}}^2$) for self-training Boost-cst* (▽).

sults obtained for less than five languages are averaged over all possible such combinations of languages.

These results show that for five languages, using all the unlabeled data for self-training is slightly more efficient than reserving part of it for coregularization. However, when the number of views is smaller, the combination of both coregularization and consensus-based self-training is more advantageous. Note that this is a common situations, for example when only bilingual documents are available.

This result suggests that in the situation where we have few views, reducing the disagreement between language specific classifiers through coregularization may lead to a more effective use of consensus-based labeling, decreasing the number of noisy examples added to the training set during self-training. On the other hand, when the number of views is large, the consensus is usually reliable enough without the need for coregularization.

6. CONCLUSION

In this paper we proposed a multiview semi-supervised boosting algorithm for multilingual document classification. We have shown how to embed a disagreement-based coregularization term into a classification objective function using a Bregman distance. This embedding allowed us to adapt an existing boosting algorithm to learn language-specific classifiers while enforcing consistency in prediction across languages. We then proposed a self-training algorithm which assigns class labels to unlabeled data based on the consensus of the classifier predictions across the different views.

Our results show clearly that the consensus based self-training allows to reach high performance in the situation where few initial labeled training documents are available. We also showed that when there are fewer languages, combining coregularization with the consensus-based self-training approach provides a better leverage of the unlabeled data by improving the quality of the consensus.

Acknowlegdements

This work was supported in part by the IST Program of the European Community, under the PASCAL2 Network of Excellence, IST-2002-506778.

7. REFERENCES

[1] J. J. G. Adeva, R. A. Calvo, and D. L. de Ipiña. Multilingual Approaches to Text Categorisation. *UPGRADE: The European Journal for the Informatics Professional*, VI(3):43–51, 2005.

[2] M.-R. Amini and C. Goutte. A Co-classification Approach to Learning from Multilingual Corpora. *Machine Learning*, 79(1-2):105–121, 2010.

[3] M.-R. Amini, N. Usunier, and C. Goutte. Learning from Multiple Partially Observed Views - an Application to Multilingual Text Categorization. In *Advances in Neural Information Processing Systems 22 (NIPS 2009)*, pages 28–36, 2009.

[4] F. R. Bach, G. R. G. Lanckriet, and M. I. Jordan. Multiple Kernel Learning, Conic Duality, and the SMO Algorithm. In *Proc. 21st International Conference on Machine Learning (ICML 2004)*, 2004.

[5] N. Bel, C. H. Koster, and M. Villegas. Cross-lingual Text Categorization. In *ECDL-2003*, pages 126–139, 2003.

[6] D. P. Bertsekas. *Nonlinear Programming*. Athena Scientific, 1999.

[7] A. Blum and T. M. Mitchell. Combining Labeled and Unlabeled Data with Co-Training. In *Proc. 11th Annual Conference on Learning Theory (COLT 1998)*, pages 92–100, 1998.

[8] U. Brefeld, T. Gärtner, T. Scheffer, and S. Wrobel. Efficient Co-regularised Least Squares Regression. In *Proc. 23rd International Conference on Machine Learning (ICML 2006)*, pages 137–144, 2006.

[9] O. Chapelle, B. Schölkopf, and A. Zien, editors. *Semi-Supervised Learning*. MIT Press, 2006.

[10] M. Collins, R. E. Schapire, and Y. Singer. Logistic regression, Adaboost and Bregman Distances. *Machine Learning*, 48(1-3):253–285, 2002.

[11] J. D. Farquhar, D. R. Hardoon, H. Meng, J. Shawe-Taylor, and S. Szedmak. Two View Learning: SVM-2k, Theory and Practice. In *Advances in Neural Information Processing 18 (NIPS 2005)*, pages 355–362, 2005.

[12] T. Joachims. Transductive Inference for Text Classification using Support Vector Machines. In *Proc. of the Sixteenth International Conference on Machine Learning (ICML 1999)*, pages 200–209, 1999.

[13] T. Joachims. Training Linear SVMs in Linear Time. In *Proceedings of the ACM Conference on Knowledge Discovery and Data Mining (KDD)*, pages 217–226, 2006.

[14] J. D. Lafferty, S. D. Pietra, and V. D. Pietra. Statistical Learning Algorithms Based on Bregman Distances. In *Canadian Workshop on Information Theory*, 1997.

[15] E. Lehmann. *Nonparametric Statistical Methods Based on Ranks*. McGraw-Hill, New York, 1975.

[16] K. Nigam, A. McCallum, S. Thrun, and T. M. Mitchell. Learning to Classify Text from Labeled and Unlabeled Documents. In *Proc. of the 15th National Conference on Artificial intelligence (AAAI/IAAI 1998*, pages 792–799, 1998.

[17] S. E. Robertson, S. Walker, S. Jones, M. Hancock-Beaulieu, and M. Gatford. Okapi at TREC-3. In *Proc. 3rd Text Retrieval Conference (TREC)*, pages 109–126, 1994.

[18] D. S. Rosenberg and P. L. Bartlett. The Rademacher Complexity of Co-regularized Kernel Classes. In *Proc. of the 11th International Conference on Artificial Intelligence and Statistics (AISTATS 2007)*, pages 396–403, 2007.

[19] V. Sindhwani, P. Niyogi, and M. Belkin. A Co-regularization Approach to Semi-supervised Learning with Multiple Views. In *Proceedings of the ICML-05 Workshop on Learning with Multiple Views*, pages 74–79, 2005.

[20] N. Ueffing, M. Simard, S. Larkin, and J. H. Johnson. NRC's PORTAGE system for WMT 2007. In *ACL-2007 Second Workshop on SMT*, 2007.

[21] C. J. Van Rijsbergen. *Information Retrieval*. Butterworth-Heinemann, Newton, MA, 1979.

[22] X. Zhu. Semi-supervised Learning Literature Survey. Technical report, University of Wisconsin Madison, 2008.