Fast On-line Learning for Multilingual Categorization

Michelle Kovesi
Interactive Language Tech.
National Research Council
283 Alexandre-Taché
Gatineau, QC, Canada

Cyril Goutte
Interactive Language Tech.
National Research Council
283 Alexandre-Taché
Gatineau, QC, Canada
Cyril.Goutte@nrc.ca

Massih-Reza Amini
University P. & M. Curie
Lab. Info. Paris 6
4, place Jussieu
75252 Paris, France
amini@poleia.lip6.fr

ABSTRACT

Multiview learning has been shown to be a natural and efficient framework for supervised or semi-supervised learning of multilingual document categorizers. The state-of-the-art co-regularization approach relies on alternate minimizations of a combination of language-specific categorization errors and a disagreement between the outputs of the monolingual text categorizers. This is typically solved by repeatedly training categorizers on each language with the appropriate regularizer. We extend and improve this approach by introducing an on-line learning scheme, where language-specific updates are interleaved in order to iteratively optimize the global cost in one pass. Our experimental results show that this produces similar performance as the batch approach, at a fraction of the computational cost.

Categories and Subject Descriptors

I.2.7 [**Artificial Intelligence**]: Natural Lang. Processing

General Terms

Algorithms, Experimentation, Languages

Keywords

Multilingual text categorisation, on-line learning

1. INTRODUCTION

Large annotated multilingual corpora are massively produced by many national or supra-national initiatives and are now available for various purposes, e.g. machine translation [4], semantic web [6] or classification [1]. For the latter, multiview learning was shown to be a natural and efficient framework for supervised or semi-supervised learning of multilingual document categorizers [2].

We propose a new online learning algorithm for multilingual document categorization that is as efficient as, and much faster than, state-of-the-art multilingual categorization algorithms. Our approach operates by learning two language-specific categorizers, iteratively adjusting their parameters on the basis of the prediction error on each language and the disagreement between the predictions on either language. The main difference with previous co-classification work is that we leverage the structure of perceptron updates in order to learn both categorizers simultaneously instead of alternatingly. Experiments carried out on `Reuters` RCV1/RCV2 multilingual documents show that our approach performs at least as well as state-of-the art strategies, while being consistently and significantly faster.

SIGIR'12, August 12–16, 2012, Portland, Oregon, USA.
ACM 978-1-4503-1472-5/12/08.

2. FRAMEWORK

We consider the representation of a *bilingual document* as a pair of two vectors $\mathbf{x} \stackrel{def}{=} (x^1, x^2)$, where each *vector* x^v provides the representation of the same document in a given vectorial space $\mathcal{X}_v, v \in \{1, 2\}$ associated to each language. Our goal is to learn languages-specific categorizers $h^1(x^1)$ and $h^2(x^2)$ that minimize prediction error over new documents in either language. Original work [1] solved this by minimizing a global loss on a training set $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^N$:

$$\mathcal{L}(h^1, h^2, S) = \underbrace{\mathcal{C}(h^1, S) + \mathcal{C}(h^2, S)}_{\text{misclassification}} + \underbrace{D(h^1, h^2)}_{\text{disagreement}} \quad (1)$$

using either logistic regression or Boosting. These are optimized in turn on each language by minimizing the misclassification loss for that language and the disagreement, and alternating between languages until convergence[1, 2].

3. ALGORITHM FOR ONLINE LEARNING

When many examples are available, online learning has been shown to provide an efficient way to train models. Instead of optimizing the model over the entire training set, online learning randomly picks one example and adjusts the model based on that example alone. Although it doesn't directly optimize the full cost function, it can handle large datasets very efficiently[3].

We consider two linear binary categorizers with parameters $w^v \in \mathcal{X}_v, v \in \{1, 2\}$ associated to each of the two languages. Each may be trained independently by on-line learning in the co-classification framework outlined above. However, we propose to jointly train both models in the same stochastic online learning process. At each iteration t, a multilingual document $(\mathbf{x}_t, y_t)$ is picked randomly from the training set and presented to both model. If it is misclassified by classifier v (i.e. $\langle w^v, x_t^v \rangle . y_t < 0$), the categorizer

is updated using the following rule:

$$w^v \leftarrow w^v + \eta \left(y_t x_t^v + \lambda \frac{\partial \mathcal{D}}{\partial w^v}(\mathbf{x}_t, w^v, w^{\overline{v}}) \right) \quad (2)$$

where $w^{\overline{v}}$ stands for the categorizer on the other language, and η is the learning rate, while λ weighs the influence of the disagreement term (part of D in eq. 1). Following [1], we set the disagreement function $\mathcal{D}$ to the Kullback-Leibler divergence between the outputs of the categorizers on each view, h^v and $h^{\overline{v}}$, mapped to $[0;1]$ with a sigmoid transformation ($\sigma(x) = 1/(1+e^{-x})$). We then get:

$$\frac{\partial \mathcal{D}}{\partial w^v}(\mathbf{x}, w^v, w^{\overline{v}}) = x^v \left(\sigma(\langle w^v, x^v \rangle) - \sigma(\langle w^{\overline{v}}, x^{\overline{v}} \rangle) \right) \quad (3)$$

The algorithm is summarized as follows:

Algorithm: Online co-classification

repeat
  for $i = 1, ..., m$ **do**
    Pick example $(\mathbf{x}_t, y_t) \in \mathcal{Z}_l$ at random;
    Update w^1 for example $\mathbf{x}_t$ (eq. 2-3);
    Update w^2 for example $\mathbf{x}_t$ (eq. 2-3)
  end
until *Convergence of global loss* ;

4. EXPERIMENTAL RESULTS

We illustrate our method on data from a large extract of the RCV1 corpus [5], which was processed and made freely available for multiview multilingual learning experiments [2].[1]

We used the entirety of the English and French documents ($N = 111,740$ documents), comprised of both originals and machine-translated Reuters newswire stories covering 6 categories: C15, CCAT, E21, ECAT, GCAT and M21. In these experiments, we randomly sampled 20 different training sets of 10,000 documents from the full corpus, each with a corresponding (non-overlapping) test set of 90,000 documents. All results are averaged over these 20 samples. For each category, we measured the performance using the F-score, and the training time (in seconds), excluding data load.

Table 1 shows the experimental results obtained using the state-of-the-art batch algorithm as well as using online approaches with interleaved updates. As observed in the original work of [1], the performance achieved by the English and French categorizers are very similar, within one point in F-score. This is due to the fact that we minimize the disagreement between the categorizers, therefore biasing them to provide the same categorization (and therefore reach similar scores) in either languages. We see also see that the F-scores are very close for both algorithms, apart from ECAT where the online version does about 1 point better. This is expected as both algorithms learn similar, linear models minimizing the same cost using similar updates. We do not expect large differences in performance, we expect large differences in training time [3]. Here, the online approach provides a very clear and consistent speedup (rightmost column). It is about three times faster, completing training within 11 to 27 seconds depending on the category, while the original batch approach requires 34-71 seconds.

[1] `http://multilingreuters.iit.nrc.ca`

Table 1: Online vs. batch results (F-score and time).

	Perf. English		Perf. French		Time (s)		
Cat.	batch	online	batch	online	bat	onl	×
C15	79.9	79.8	78.8	78.7	51	17	3.0
CCAT	70.0	70.0	69.1	69.0	65	27	2.4
E21	72.8	72.7	71.9	72.0	56	17	3.3
ECAT	68.7	69.6	67.8	68.9	71	27	2.7
GCAT	78.1	78.1	77.1	77.0	53	17	3.2
M11	88.5	88.3	87.5	87.6	34	11	3.0

In a typical scenario, convergence is achieved by the online approach (at the chosen convergence threshold) using 7 to 10 epochs over the entire training set. By contrast, the batch algorithm performed as little as 2 (but usually more) alternate otimizations of the models in each view, each composed of several epochs on the training set. This therefore multiplies the overall training time.

We confirmed this by running experiments with increasing training set sizes, from 10k to 100k training documents. In all experiments, the online algorithm yields performance similar to batch, while requiring significantly less time. The speedup grows from around 3 (Table 1) to around 6 for 100k documents.

5. DISCUSSION AND CONCLUSIONS

We propose a novel online algorithm that builds on the co-classification work of [1], but leverages the structure of online perceptron learning in order to simultaneously update the models on both views at each example presentation, yielding a clear and consistent speedup while providing similar performance.

This provides a natural way to learn multiple categorizers in a multiview learning framework, without the need to resort to alternating optimizations of regularized view-specific losses as in the original co-classification approach. In addition, although we have not yet obtained evidence for that, we expect that it will scale up better to more than 2 languages, because all models can be updated simulatneously at each iteration, instead of alternating view-specific optimizations. We expect that further work will shed light on the scalability to larger corpora and more languages.

6. REFERENCES

[1] M.-R. Amini and C. Goutte. A co-classification approach to learning from multilingual corpora. *Machine Learning*, 79(1-2), 2010.

[2] M. R. Amini, C. Goutte, and N. Usunier. Combining coregularization and consensus-based self-training for multilingual text categorization. In *SIGIR'10*, 2010.

[3] L. Bottou and Y. LeCun. Large scale online learning. In *NIPS 16*, 2004.

[4] A. Eisele and Y. Chen. MultiUN: A multilingual corpus from united nation documents. In *LREC'10*, 2010.

[5] D. D. Lewis, Y. Yang, T. Rose, and F. Li. A new benchmark collection for text categorization research. *J. Machine Learning Research*, 5:361–397, 2004.

[6] B. Pouliquen, R. Steinberger, and C. Ignat. Automatic annotation of multilingual text collections with a conceptual thesaurus. *CoRR*, abs/cs/0609059, 2006.