# ERCIM NEWS

M_08

R_17

R_05

R_10

R_11

R_08

R_09

R

06

**Special theme:**

# Digital Humanities

**Also in this issue:**

Research and Innovation:

**Trend Analysis of Underground Marketplaces**

# Contents

## RESEARCH AND INNOVATION

This section features news about research activities and innovative developments from European research institutes

## EVENTS, IN BRIEF

# Distributional Correspondence Indexing for Cross-Lingual and Cross-Domain Sentiment Classification

by Alejandro Moreo Fernández, Andrea Esuli and Fabrizio Sebastiani

*Researchers from ISTI-CNR, Pisa (in a joint effort with the Qatar Computing Research Institute), have developed a transfer learning method that allows cross-domain and cross-lingual sentiment classification to be performed accurately and efficiently. This means sentiment classification efforts can leverage training data originally developed for performing sentiment classification on other domains and/or in other languages.*

Sentiment Classification (SC) is the task of classifying opinion-laden documents in terms of the sentiment (positive, negative, or neutral) they express towards a given entity (e.g., a product, a policy, a political candidate). Determining the user's stance towards such an entity is of the utmost importance for market research, customer relationship management, the social sciences, and political science, and several automated methods have been proposed for this purpose. SC is usually accomplished via supervised learning, whereby a sentiment classifier is trained on manually labelled data. However, when SC needs to deal with a completely new context, it is likely that the amount of manually labelled, opinion-laden documents available to be used as training data, is scarce or even non-existent. Transfer Learning (TL) is a class of context adaptation methods that focus on alleviating this problem by allowing the learning algorithm to train a classifier for a "target" context by leveraging manually labelled examples available from a different, but related, "source" context.

In this work the authors propose the Distributional Correspondence Indexing (DCI) method for transfer learning in sentiment classification. DCI is inspired by the "Distributional Hypothesis", a well-know principle in linguistics that states that the meaning of a term is somehow determined by its distribution in text, and by the terms it tends to co-occur with (a famous motto that describes this assumption is "You shall know a word by the company it keeps"). DCI mines, from unlabelled datasets and in an unsupervised fashion, the distributional correspondences between each term and a small set of "pivot" terms, and is based on the further hypothesis that these correspondences are approximately invariant across the source context and target context for terms that play equivalent roles in the two contexts. DCI derives term representations in a vector space common to both contexts, where each dimension (or "feature") reflects its distributional correspondence to a pivot term. Term correspondence is quantified by means of a distributional correspondence function (DCF); in this work the authors propose and experiment with a number of efficient DCFs that are motivated by the distributional hypothesis. An advantage of DCI is that it projects each term into a low-dimensional space (about 100 dimensions) of highly predictive concepts (the pivot terms), while other competing methods (such as Explicit Semantic Analysis) need to work with much higher-dimensional vector spaces.

The authors have applied this technique to sentiment classification across domains (e.g., book reviews vs DVD reviews) and across languages (e.g., reviews in English vs reviews in German), either in isolation (cross-domain SC, or cross-lingual SC) or – for the first time in the literature – in combination (cross-domain cross-lingual SC, see Figure 1). Experiments run by the authors (on "Books", "DVDs", "Electronics", "Kitchen Appliances" as the domains, and on English, German, French, and Japanese as the languages) have shown that DCI obtains better sentiment classification accuracy than current state-of-the-art techniques (such as Structural Correspondence Learning, Spectral Feature Alignment, or Stacked Denoising Autoencoder) for cross-lingual and cross-domain sentiment classification. DCI also brings about a significantly reduced computational cost (it requires modest computational resources, which is an indication that it can scale well to huge collections), and requires a smaller amount of human intervention than competing approaches in order to create the pivot set.

**Link:** http://jair.org/papers/paper4762.html

**Please contact:**
Fabrizio Sebastiani, ISTI-CNR, Italy
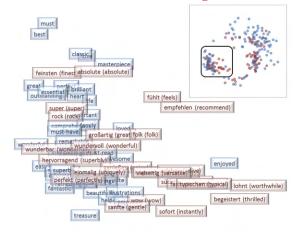+39 050 6212892, fabrizio.sebastiani@isti.cnr.it

*Figure 1: Cross-lingual (English-German) and cross-domain (book-music) alignment of word embeddings through DCI. The left part is a cluster of negative-polarity words, the right part of positive-polarity ones.*