R\_10 7

### ERCIM NEWS

www.ercim.eu

08 R 17

R\_05

Special theme:

# Digital Humanities

Also in this issue:

**Research and Innovation:** 

**Trend Analysis of Underground Marketplaces** 

ERCIM News is the magazine of ERCIM. Published quarterly, it reports on joint actions of the ERCIM partners, and aims to reflect the contribution made by ERCIM to the European Community in Information Technology and Applied Mathematics. Through short articles and news items, it provides a forum for the exchange of information between the institutes and also with the wider scientific community. This issue has a circulation of about 6,000 printed copies and is also available online.

ERCIM News is published by ERCIM EEIG BP 93, F-06902 Sophia Antipolis Cedex, France Tel: +33 4 9238 5010, E-mail: contact@ercim.eu Director: Philipp Hoschka, ISSN 0926-4981

### Contributions

Contributions should be submitted to the local editor of your country

### Copyright notice

All authors, as identified in each article, retain copyright of their work. ERCIM News is licensed under a Creative Commons Attribution 4.0 International License (CC-BY).

### Advertising

For current advertising rates and conditions, see http://ercim-news.ercim.eu/ or contact peter.kunz@ercim.eu

### **ERCIM** News online edition

http://ercim-news.ercim.eu/

### Next issue

October 2017, Special theme: Digital Humanities

Subscribe to ERCIM News by sending an email to en-subscriptions@ercim.eu or by filling out the form at the ERCIM News website: http://ercim-news.ercim.eu/

### Editorial Board:

Central editor:

Peter Kunz, ERCIM office (peter.kunz@ercim.eu) Local Editors:

Austria: Erwin Schoitsch (erwin.schoitsch@ait.ac.at) Belgium:Benoît Michel (benoit.michel@uclouvain.be)

Cyprus: Ioannis Krikidis (krikidis.ioannis@ucy.ac.cy)

France: Steve Kremer (steve.kremer@inria.fr)

Germany: Michael Krapp (michael.krapp@scai.fraunhofer.de)

Greece: Lida Harami (lida@ics.forth.gr),

Artemios Voyiatzis (bogart@isi.gr)

Hungary: Andras Benczur (benczur@info.ilab.sztaki.hu) Italy: Maurice ter Beek (maurice.terbeek@isti.cnr.it)

Luxembourg: Thomas Tamisier (thomas.tamisier@list.lu) Norway: Poul Heegaard (poul.heegaard@item.ntnu.no)

Poland: Hung Son Nguyen (son@mimuw.edu.pl)

Portugal: José Borbinha, Technical University of Lisbon

Spain: Silvia Abrahão (sabrahao@dsic.upv.es)

Sweden: Kersti Hedman-Hammarström (kersti@sics.se)

Switzerland: Harry Rudin (hrudin@smile.ch) The Netherlands: Annette Kik (Annette.Kik@cwi.nl)

W3C: Marie-Claire Forgue (mcf@w3.org)

### **JOINT ERCIM ACTIONS**

- 4 ERCIM Membership
- ERCIM "Alain Bensoussan" Fellowship Programme
- **HORIZON 2020 Project** Management
- HRADIO A New Project to Tackle Future of Internet Radio

### **SPECIAL THEME**

The special theme section "Digital Humanities" has been coordinated by George Bruseker (ICS-FORTH), László Kovács (MTA-SZTAKI), and Franco Niccolucci (University of Florence)

### Digital Source Indexing and Analysis

- In Codice Ratio: Scalable **Transcription of Vatican Registers** by Donatella Firmani, Paolo Merialdo (Roma Tre University), Marco Maiorino (Vatican Secret Archives)
- **Teaching Archaeology to Machines: Extracting Semantic Knowledge** from Free Text Excavation Reports by Achille Felicetti (Università degli Studi di Firenze)
- 11 MTAS Extending Solr into a **Scalable Search Solution and Analysis Tool on Multi-Tier Annotated Text** by Matthijs Brouwer (Meertens Institute)
- 12 Phonetic Search in Audio and Video Recordings

by Ioannis Dologlou and Stelios Bakamidis (RC ATHENA)

13 KA3: Speech Analytics for Oral History and the Language Sciences by Joachim Köhler (Fraunhofer IAIS), Nikolaus P. Himmelmann (Universität zu Köln) and Almut Leh (FernUniversität in Hagen)

- 14 Graph-Based Entity-Oriented Search: Imitating the Human **Process of Seeking and Cross Referencing Information** by José Devezas and Sérgio Nunes (INESC TEC and FEUP)
- 16 ContentCheck: Content **Management Techniques and Tools** for Fact-checking

by Ioana Manolescu (Inria Saclay and Ecole Polytechnique, France)

17 Argumentation Analysis of **Engineering Research** by Mihály Héder (MTA-SZTAKI)

### Information Management Strategy

- 18 A Back Bone Thesaurus for Digital Humanities by Maria Daskalaki and Lida Charami (ICS-FORTH)
- 19 Meeting the Challenges to Reap the **Benefits of Semantic Data in Digital** Humanities

by George Bruseker, Martin Doerr and Chryssoula Bekiari (ICS-FORTH)

- 21 HumaReC: Continuous Data **Publishing in the Humanities** by Claire Clivaz, Sara Schulthess and Anastasia Chasapi (SIB)
- 22 A Data Management Plan for Digital Humanities: the PARTHENOS Model by Sheena Bassett (PIN Scrl), Sara Di Giorgio (MIBACT-ICCU) and Paola Ronzino (PIN Scrl)
- 23 Trusting Computation in Digital **Humanities Research** by Jacco van Ossenbruggen (CWI)

### Research Infrastructure Development

- 25 The DARIAH ERIC: Redefining **Research Infrastructure for the Arts** and Humanities in the Digital Age by Jennifer Edmond (Trinity College Dublin), Frank Fischer (National Research University Higher School of Economics, Moscow), Michael Mertens (DARIAH EU) and Laurent Romary (Inria)
- 26 DIGILAB: A New Infrastructure for Heritage Science by Luca Pezzati and Achille Felicetti (INO-CNR)
- 28 Building a Federation of Digital **Humanities Infrastructures** by Alessia Bardi and Luca Frosini (ISTI-CNR)
- 29 Knowledge Complexity and the Digital Humanities: Introducing the KPLEX Project.

by Jennifer Edmond and Georgina Nugent Folan (Trinity College Dublin)

### 3D Analysis Techniques

- 30 Restoration of Ancient Documents
  Using Sparse Image Representation
  by Muhammad Hanif and Anna
  Tonazzini (ISTI-CNR)
- 32 St Paul's Cathedral Rises From the Dust News From the Virtual St Paul's Cathedral Project

by John N. Wall (NC State University), John Schofield (St Paul's Cathedral, London), David Hill (NC State University and Yun Jing (NC State University)

33 Immersive Point Cloud Manipulation for Cultural Heritage Documentation

by Jean-Baptiste Barreau (CNRS/CReAAH UMR 6566), Ronan Gaugne (Université de Rennes 1/ IRISA-Inria) and Valérie Gouranton (INSA Rennes/IRISA-Inria)

35 Culture 3D Cloud: A Cloud Computing Platform for 3D Scanning, Documentation, Preservation and Dissemination of Cultural Heritage

by Pierre Alliez (Inria), François Forge (Reciproque), Livio de Luca (CNRS MAP), Marc Pierrot-Deseilligny (IGN) and Marius Preda (Telecom SudParis)

36 Physical Digital Access Inside Archaeological Material

> by Théophane Nicolas (Inrap/UMR 8215 Trajectoires), Ronan Gaugne (Université de Rennes 1/IRISA-Inria), Valérie Gouranton (INSA de Rennes/ IRISA-Inria) and Jean-Baptiste Barreau (CNRS/CReAAH UMR 6566)

### Information Visualization and Communication

37 Reinterpreting European History Through Technology: The CrossCult Project

by Susana Reboreda Morillo (Universidad de Vigo), Maddalena Bassani (Università degli Studi di Padova) and Ioanna Lykourentzou (LIST)

39 Cultural Opposition in former European Socialist Countries: Building the COURAGE Registry by András Micsik , Tamás Felker and Balázs Nász (MTA SZTAKI) 40 Locale, an Environment-Aware Storytelling Framework Relying on Augmented Reality

by Thomas Tamisier, Irene Gironacci and Roderick McCall (Luxembourg Institute of Science and Technology)

42 The Biennale 4D Project

by Kathrin Koebel, Doris Agotai, Stefan Arisona (FHNW) and Matthias Oberli (SIK-ISEA)

43 The Clavius Correspondence: From Digitization to Visual Exploration of Knowledge

> by Matteo Abrate, Angelica Lo Duca, Andrea Marchetti (IIT-CNR)

45 Service-oriented Mobile Social Networking

by Stefano Chessa and Michele Girolami (ISTI-CNR)

### RESEARCH AND INNOVATION

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

46 Collaboration Spotting: A Visual Analytics Platform to Assist Knowledge Discovery

by Adam Agocs, Dimitris Dardanis, Richard Forster, Jean-Marie Le Goff, Xavier Ouvrard and André Rattinger (CERN)

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

by Alejandro Moreo Fernández, Andrea Esuli and Fabrizio Sebastiani

49 Measuring Bias in Online Information

by Evaggelia Pitoura (University of Ioannina), Irini Fundulaki (FORTH) and Serge Abiteboul (Inria & ENS Cachan)

50 The Approximate Average Common Submatrix for Computing the Image Similarity

by Alessia Amelio (Univ. of Calabria)

52 My-AHA: Stay Active, Detect Risks Early, Prevent Frailty!

by Nadine Sturm, Doris M. Bleier and Gerhard Chroust (Johanniter Österreich) 53 A New Vision of the Cruise Ship Cabin

by Erina Ferro (ISTI-CNR)

54 Trend Analysis of Underground Marketplaces

by Klaus Kieseberg, Peter Kieseberg and Edgar Weippl: (SBA Research)

55 RDF Visualiser: A Tool for Displaying and Browsing High Density of RDF Data

by Nikos Minadakis, Kostas Petrakis, Korina Doerr and Martin Doerr (FORTH)

57 Services for Large Scale Semantic Integration of Data

by Michalis Mountantonakis and Yannis Tzitzikas (ICS-FORTH)

58 On the Interplay between Physical and Digital World Accessibility

by Christophe Ponsard, Jean Vanderdonckt and Lyse Saintjean (CETIC)

60 Highly Sensitive Biomarker Analysis Using On-Chip Electrokinetics

> by Xander F. van Kooten, Federico Paratore (Israel Institute of Technology and IBM Research – Zurich), Moran Bercovici (Israel Institute of Technology) and Govind V. Kaigala IBM Research – Zurich)

### **EVENTS, IN BRIEF**

Call for Proposals

60 Dagstuhl Seminars and Perspectives Workshops

Report

60 A Conference for Advanced Students: IFIP TC6's AIMS

62 W3C Workshop on WebVR Authoring: Opportunities and Challenges

In Brief

63 SICS becomes RISE SICS

63 W3C Brings Web Developer Skills to the European Workforce

3

63 Awards for Breaking SHA-1 Security Standard in Practice

ERCIM NEWS 111 October 2017

## 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, opinionladen 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 deter-

schimmsten (worst) is (tiresome)
erbarmich (stitful) v unrealistic

fehlt (lacks)

fehlt (lacks)

erschrecked (terrifvip)
nett (nesh leck) (ferrifv)

nett (nesh leck) (ferrifv)

entäuschend (disapointed)
entäuschend (disapointed)
unglaublich (unbelievable)
unglaublich (unbelievable)
verdienen (earn)
bipte (please)
beschreiben (describe)

wirklich (truly)
besser (better)
expecting

erwartet (expecting)

mined by its distribution in text, and by the terms it tends to cooccur 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 crosslingual 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

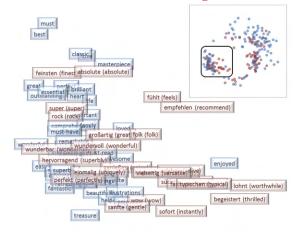


Figure 1: Cross-lingual (English-German) and cross-domain (book-music) alignment of word embeddings through DCI. The left part is a cluster of negative-polarity words, the right part of positive-polarity ones.