Coling 2010

23rd International Conference on Computational Linguistics

Kernel Engineering for Fast and Easy Design of Natural Language Applicationsa

Alessandro Moschitti

Department of Information Engineering and Computer Science
University of Trento
August 22, 2010

To order the CD of Coling 2010 and its Workshop Proceedings, please contact:

Chinese Information Processing Society of China
No.4, Southern Fourth Street
Haidian District, Beijing, 100190
China
Tel: +86-010-62562916
Fax: +86-010-62562916
cips@iscas.ac.cn

Tutorial Instuctor

Alessandro Moschitti: *DISI, University of Trento, Italy. E-mail: moschitti@disi.unitn.it*

Alessandro is a professor at the Information Engineering and Computer Science Department of the University of Trento. In 2003, he obtained his PhD in Computer Science at the University of Rome "Tor Vergata" and between 2002 and 2004, he worked as an associate researcher in the University of Texas at Dallas for two years. His expertise concerns machine learning approaches to Natural Language Processing, Information Retrieval and Data Mining. In particular, he has recently devised innovative kernels within Support Vector and other kernel-based machines for advanced syntactic/semantic processing. He is author of more than 100 scientific articles published in the major conferences of different research communities, e.g. ACL, ICML, CIKM and ICDM. He has participated in several projects of the European Community (EC), e.g. LIVING-KNOWLEDGE 2008, PRESTOSPACE 2004, NAMIC 2000 and TREVI 1998 and in two US projects: MTBF 2008 (Con-Edison) and ARDA AQUAINT PROGRAM (IQAS 2002). He is currently project coordinator of the EC project, EternalS.

Outline

Previous work on the use of Machine Learning for Computational Linguistics has shown that most of the design effort is devoted to feature engineering. Indeed, the latter requires expertise, intuition and deep knowledge about the target problem to convert linguistic objects into attribute-value representations. Kernel Methods (KM) are powerful techniques, which can simplify data modeling by defining abstract representations and implicit feature spaces. More in particular, KM allow for: (a) directly using a similarity function between instances in learning algorithms, thus avoiding explicit feature design; and (b) implicitly defining huge feature spaces, e.g. structures can be represented in the substructure space.

In this tutorial, practical recipes to successfully use KM for target language applications will be presented: first, after an introduction to Support Vector Machines (explained from an application viewpoint), KM theory will be explained in a way that useful practical methods can be derived. Second, basic kernels, such as linear, polynomial, sequence and tree kernels will be presented, by focusing on the implementation, accuracy and efficiency perspectives. KM application to typical natural language tasks, e.g. text categorization, question and answer classification, semantic role labeling, textual entailment and so on, will be shown. The aim is to provide practical procedures for the selection and exploitation of the right kernel for the target task. Third, the SVM-Light-TK toolkit, which encodes several kernels in SVMs, will be illustrated along with the associated data structures and its practical use in NL tasks. Finally, the tutorial will illustrate how innovative and effective kernels can be engineered starting from basic kernels and using systematic data transformation. Such know-how allows for a very fast and accurate design of applications even if the underlying language phenomena and properties are still not very well understood, e.g. Arabic SRL or relation extraction between pairs of text fragments.