

CoNLL-2012 Shared Task: Modeling Multilingual Unrestricted Coreference in OntoNotes

Sameer Pradhan
Raytheon BBN Technologies,
Cambridge, MA 02138
USA

pradhan@bbn.com

Alessandro Moschitti
University of Trento,
38123 Povo (TN)
Italy

moschitti@disi.unitn.it

Nianwen Xue
Brandeis University,
Waltham, MA 02453
USA

xuen@cs.brandeis.edu

Olga Uryupina
University of Trento,
38123 Povo (TN)
Italy

uryupina@gmail.com

Yuchen Zhang
Brandeis University,
Waltham, MA 02453
USA

yuchenz@brandeis.edu

Abstract

The CoNLL-2012 shared task involved predicting coreference in three languages – English, Chinese and Arabic – using OntoNotes data. It was a follow-on to the English-only task organized in 2011. Until the creation of the OntoNotes corpus, resources in this sub-field of language processing have tended to be limited to noun phrase coreference, often on a restricted set of entities, such as ACE entities. OntoNotes provides a large-scale corpus of general anaphoric coreference not restricted to noun phrases or to a specified set of entity types and covering multiple languages. OntoNotes also provides additional layers of integrated annotation, capturing additional shallow semantic structure. This paper briefly describes the OntoNotes annotation (coreference and other layers) and then describes the parameters of the shared task including the format, pre-processing information, evaluation criteria, and presents and discusses the results achieved by the participating systems. Being a task that has a complex evaluation history, and multiple evaluation conditions, it has, in the past, been difficult to judge the improvement in new algorithms over previously reported results. Having a standard test set and evaluation parameters, all based on a resource that provides multiple integrated annotation layers (pares, semantic roles, word senses, named entities and coreference) that could support joint models, should help to energize ongoing research in the task of entity and event coreference.

