Cross-Lingual Word Representations: Induction and Evaluation

EMNLP 2017 Tutorial

Manaal Faruqui[1], **Anders Søgaard**[2], **Ivan Vulić**[3]
[1] Google Research, New York
[2] Dpt. of Computer Science, University of Copenhagen
[3] Language Technology Lab, DTAL, University of Cambridge
mfaruqui@google.com soegaard@hum.ku.dk iv250@cam.ac.uk

1 Motivation and Objectives

In recent past, NLP as a field has seen tremendous utility of distributional word vector representations as features in downstream tasks. The fact that these word vectors can be trained on unlabeled monolingual corpora of a language makes them an inexpensive resource in NLP. With the increasing use of monolingual word vectors, there is a need for word vectors that can be used as efficiently across multiple languages as monolingually. Therefore, learning bilingual and multilingual word embeddings/vectors is currently an important research topic. These vectors offer an elegant and language-pair independent way to represent content across different languages.

This tutorial aims to bring NLP researchers up to speed with the current techniques in cross-lingual word representation learning. We will first discuss how to induce cross-lingual word representations (covering both bilingual and multilingual ones) from various data types and resources (e.g., parallel data, comparable data, non-aligned monolingual data in different languages, dictionaries and theasuri, or, even, images, eye-tracking data). We will then discuss how to evaluate such representations, intrinsically and extrinsically. We will introduce researchers to state-of-the-art methods for constructing cross-lingual word representations and discuss their applicability in a broad range of downstream NLP applications.

We will deliver a detailed survey of the current methods, discuss best training and evaluation practices and use-cases, and provide links to publicly available implementations, datasets, and pre-trained models.

2 Tutorial Overview

2.1 Introduction

An overview of the cross-lingual NLP landscape, situating the current work on cross-lingual representation learning and motivating the need for multilingual training and cross-lingual transfer for resource-poor languages. A discussion on various types of bilingual resources available: e.g., dictionaries, parallel vs. comparable vs. non-aligned monolingual data.

2.2 Part I: Learning from Word Alignments and Dictionaries

In the first part of the tutorial, after important preliminaries (i.e., standard learning techniques in monolingual settings which lend themselves to cross-lingual scenarios: dimensionality reduction, learning from context, etc.), we will present a typology of cross-lingual models roughly clustered according to the cross-lingual signal needed for training (e.g., translation pairs, document-aligned data, images), as well as their ability to exploit both multilingual and more abundant monolingual data in training. We will then zoom in the group of models which learn directly from available dictionaries and word alignment information, drawing comparisons with older baseline work on bilingual/multilingual clustering and traditional distributional cross-lingual spaces based on one-to-one translation lexicons, and analyzing their modeling assumptions and protocols. Throughout the tutorial (Part I - Part III), we will demonstrate how to extend the current cross-lingual representation models from bilingual to true multilingual settings (three or more languages), as such extensions are not straightforward for plenty of modeling frameworks.

2.3 Part II: Learning from Sentence and Document Alignments

We will focus on collections of cross-lingual representation models that learn from sentence-aligned and document-aligned bilingual data. We will again draw links to older cross-lingual learning frameworks from the same data types: word alignment algorithms and translation tables, multilingual topic modeling, cross-lingual LSA and ESA. We will then analyze several representative cross-lingual word vector models from sentence-/document-aligned and outline more recent developments.

2.4 Part III: Learning from Other Resources

We will first analyze cross-lingual models that can leverage (typically more abundant) monolingual data together with multilingual data for training. We will then show how to combine available linguistic information (e.g., WordNet, BabelNet) with corpora-driven representations. Finally, we will discuss alternative sources of bilingual information for learning cross-lingual word representations: image data, eye-tracking data, etc.

2.5 Part IV: Evaluation and Application

In the final part, we will focus on evaluation and application of cross-lingual word representations. First, we will discuss the differences between intrinsic and extrinsic evaluations, and current evaluation protocols. We will demonstrate how cross-lingual representations may boost monolingual NLP tasks, and how such representations can support fundamental cross-lingual tasks such as word alignment induction, bilingual lexicon learning, or machine translation. Following that, we will demonstrate the importance of cross-lingual transfer in NLP, and discuss what kind of knowledge (semantic vs. syntactic information) can actually transfer across languages. We will then show how to use cross-lingual word vectors to accomplish such transfers for a (didactically chosen) selection of NLP downstream tasks.

2.6 Discussion and Final Remarks

We will conclude by listing publicly available software packages and implementations, available training datasets and evaluation protocols, and sketching future research avenues in this domain.

3 Structure

- **Introduction**: Motivating and grounding cross-lingual representation learning (*15 minutes*)
- **Part I:** Cross-lingual representation learning from word alignments and dictionaries (*40 minutes*)
- **Part II:** Cross-lingual representation learning from sentence and document alignments (*35 minutes*)
- **Coffee Break** (*30 minutes*)
- **Part III:** Cross-lingual representation learning from other resources (*30 minutes*)
- **Part IV:** Evaluation and Application (*40 minutes*)
- **Discussion and Final Remarks** (*20 minutes*)

4 About the Speakers

Manaal Faruqui is a research scientist at Google NYC currently working on industrial-scale NLP problems. Manaal received his PhD in the Language Technologies Institute at Carnegie Mellon University. He has worked on problems in the areas of representation learning, distributional semantics and multilingual learning. He has won one of the best paper awards at NAACL 2015. He organized the workshop on cross-lingual and multilingual models in NLP at NAACL 2016.
`http://www.manaalfaruqui.com/`

Anders Søgaard is a full professor of Computer Science (NLP and Machine Learning) at the University of Copenhagen. Anders is interested in transfer learning and has worked on semi-supervised learning, domain adaptation, and cross-language adaptation of NLP models. He is particularly interested in transferring models to very low-resource languages. He holds an ERC Starting Grant, as well as several grants from national research councils and private research foundations. He has won three best paper awards at major ACL conferences. He gave a tutorial on domain adaptation at COLING 2014.
`http://cst.dk/anders/`

Ivan Vulić is a research associate at the University of Cambridge. He received his PhD *summa cum laude* at KU Leuven in 2014. Ivan is interested in representation learning, distributional and multi-modal semantics in monolingual and multilingual contexts, and transfer learning for enabling cross-lingual NLP applications. His work has been published in top-tier *ACL and *IR conferences. He gave a tutorial on topic models at ECIR 2013 and WSDM 2014, and co-organised a Vision & Language workshop at EMNLP 2015. https://sites.google.com/site/ivanvulic/