Phoebe Mulcaire

phoebe.mulcaire@gmail.com | (707)-888-5253 | pmulcaire.github.io

PROFILE

I am interested in applying natural language processing to low-resource languages, and in developing language-universal or language-agnostic NLP methods. In particular, I am interested in the potential of continuous representations of words and subword tokens for sharing information between languages, and methods of transferring semantic information from high-resource to low-resource languages.

EDUCATION

UNIVERSITY OF CALIFORNIA, BERKELEY – 2010-2014

B.A. IN COMPUTER SCIENCE AND COGNITIVE SCIENCE

I studied machine learning, artificial intelligence and the overlap of cognitive science and computer science.

UNIVERSITY OF WASHINGTON – 2015-PRESENT

PhD student in NLP, advised by Noah Smith. My focus has been on multilingual NLP, including crosslingual transfer and multilingual representation learning.

INTERNSHIPS

FACEBOOK, SUMMER 2020

Worked with Alexis Conneau and the LATTE team on two projects: one to learn a discrete vocabulary representation of text via a vector quantization approach, and a second focusing on substituting vocabulary embeddings in a large language model to learn alignments between word representations in different languages.

PUBLICATIONS

GROUNDED COMPOSITIONAL OUTPUTS FOR ADAPTIVE LANGUAGE MODELING – EMNLP 2020

Nikolaos Pappas, Phoebe Mulcaire, and Noah A. Smith

Language models have emerged as a central component across NLP, and a great deal of progress depends on the ability to cheaply adapt them (e.g., through finetuning) to new domains and tasks. A language model's vocabulary—typically selected before training and permanently fixed later—affects its size and is part of what makes it resistant to such adaptation. Prior work has used compositional input embeddings based on surface forms to ameliorate this issue. In this work, we go one step beyond and propose a fully compositional output embedding layer for language models, which is further grounded in information from a structured lexicon (WordNet), namely semantically related words and free-text definitions. To our knowledge, the result is the first word-level language model with a size that does not de-pend on the training vocabulary. We evaluate the model on conventional language modeling as well as challenging cross-domain settings with an open vocabulary,

finding that it matches or outperforms previous state-of-the-art output embedding methods and adaptation approaches. Our analysis attributes the improvements to sample efficiency: our model is more accurate for low-frequency words.

LOW-RESOURCE PARSING WITH CROSSLINGUAL CONTEXTUALIZED REPRESENTATIONS – CONLL 2019

Phoebe Mulcaire, Jungo Kasai, and Noah A. Smith.

We assess recent approaches to multilingual contextual word representations (CWRs), and compare them for crosslingual transfer from a language with a large treebank to a language with a small or nonexistent treebank, by sharing parameters between languages in the parser itself. We experiment with a diverse selection of languages in both simulated and truly low-resource scenarios, and show that multilingual CWRs greatly facilitate low-resource dependency parsing even with-out crosslingual supervision such as dictionaries or parallel text. Furthermore, we examine the non-contextual part of the learned language models (which we call a "decontextual probe") to demonstrate that polyglot language models better encode crosslingual lexical correspondence compared to aligned monolingual language models. This analysis provides further evidence that polyglot training is an effective approach to crosslingual transfer.

POLYGLOT CONTEXTUAL REPRESENTATIONS IMPROVE CROSSLINGUAL TRANSFER – NAACL 2019

Phoebe Mulcaire, Jungo Kasai, and Noah A. Smith.

We introduce Rosita, a method to produce multilingual contextual word representations by training a single language model on text from multiple languages. Our method combines the advantages of contextual word representations with those of multilingual representation learning. We produce language models from dissimilar language pairs (English/Arabic and English/Chinese) and use them in dependency parsing, semantic role labeling, and named entity recognition, with comparisons to monolingual and non-contextual variants. Our results provide further evidence for the benefits of polyglot learn-ing, in which representations are shared across multiple languages.

POLYGLOT SEMANTIC ROLE LABELING – ACL 2018

Phoebe Mulcaire, Swabha Swayamdipta, and Noah A. Smith.

Previous approaches to multilingual semantic dependency parsing treat languages independently, without exploiting the similarities between semantic structures across languages. We experiment with a new approach where we combine resources from a pair of languages in the CoNLL 2009 shared task (Hajič et al., 2009) to build a polyglot semantic role labeler. Notwithstanding the absence of parallel data, and the dissimilarity in annotations between languages, our approach results in an improvement in SRL performance on multiple languages over a monolingual baseline. Analysis of the polyglot model shows it to be advantageous in lower-resource settings.

A NEURAL MODEL FOR LANGUAGE IDENTIFICATION IN CODE-SWITCHED TWEETS – LICS 2016

Aaron Jaech, Phoebe Mulcaire, Mari Ostendorf, and Noah A. Smith.

Language identification systems suffer when working with short texts or in domains with unconventional spelling, such as Twitter or other social media. These challenges are explored in a shared task for Language Identification in Code-Switched Data (LICS 2016).

We apply a hierarchical neural model to this task, learning character and contextualized word-level representations to make word-level language predictions. This approach performs well on both the 2014 and 2016 versions of the shared task.

HIERARCHICAL CHARACTER-WORD MODELS FOR LANGUAGE IDENTIFICATION – SOCIALNLP 2016

Aaron Jaech, Phoebe Mulcaire, Shobhit Hathi, Mari Ostendorf, and Noah A. Smith.

Social media messages' brevity and unconventional spelling pose a challenge to language identification. We introduce a hierarchical model that learns character and contextualized word-level representations for language identification. Our method performs well against strong baselines, and can also reveal code-switching.

MANY LANGUAGES, ONE PARSER – TACL, JULY 2016

Waleed Ammar, Phoebe Mulcaire, Miguel Ballesteros, Chris Dyer, and Noah A. Smith.

We train one model for dependency parsing and use it to parse competitively in several languages. The parsing model uses multilingual word clusters and multilingual word embeddings alongside learned and specified typological information, enabling generalization based on linguistic universals and typological similarities. Our model can also incorporate language-specific features (e.g., fine POS tags), enabling still letting the parser to learn language-specific behaviors. Our parser compares favorably to strong baselines in a range of data scenarios, including when the target language has a large treebank, a small treebank, or no treebank for training.

MASSIVELY MULTILINGUAL WORD EMBEDDINGS – FEB. 2016

Waleed Ammar, Phoebe Mulcaire, Yulia Tsvetkov, Guillaume Lample, Chris Dyer, and Noah A. Smith.

We introduce new methods for estimating and evaluating embeddings of words in more than fifty languages in a single shared embedding space. Our estimation methods, multiCluster and multiCCA, use dictionaries and monolingual data; they do not require parallel data.

Our new evaluation method, multiQVEC-CCA, is shown to correlate better than previous ones with two downstream tasks (text categorization and parsing). We also describe a web portal for evaluation that will facilitate further research in this area, along with open-source releases of all our methods.

TESTING THE LEARNABILITY OF WRITING SYSTEMS - BERKELEY LINGUISTICS SOCIETY, 2013

Sharon Inkelas, Keith Johnson, Charles Lee, Emil Minas, Phoebe Mulcaire, Gek Yong Keng, and Tomomi Yuasa.

The world's sound-based writing systems differ according to the size of the typical speech chunk which is mapped to a symbol: the phone, in so-called alphabetic writing systems, and the mora, demisyllable, or syllable, in so-called syllabaries. This paper reports the results of an artificial learning study designed to test whether the acoustic stability of the speech chunks mapped to symbols is a factor in subjects' ability to learn a novel writing system.