Available online at www.sciencedirect.com
ScienceDirect
Computer Speech & Language 00 (2017) 1–19
www.elsevier.com/locate/csl

Domain adaptation using neural network joint model

Shafiq Joty, Nadir Durrani*, Hassan Sajjad, Ahmed Abdelali

Arabic Language Technologies, Qatar Computing Research Institute, HBKU, Qatar

Received 6 May 2016; received in revised form 10 November 2016; accepted 7 December 2016

Abstract

We explore neural joint models for the task of domain adaptation in machine translation in two ways: (*i*) we apply state-of-the-art domain adaptation techniques, such as mixture modelling and data selection using the recently proposed Neural Network Joint Model (NNJM) (Devlin et al., 2014); (*ii*) we propose two novel approaches to perform adaptation through instance weighting and weight readjustment in the NNJM framework. In our first approach, we propose a pair of models called Neural Domain Adaptation Models (NDAM) that minimizes the cross entropy by regularizing the loss function with respect to in-domain (and optionally to out-domain) model. In the second approach, we present a set of Neural Fusion Models (NFM) that combines the in- and the out-domain models by readjusting their parameters based on the in-domain data.

We evaluated our models on the standard task of translating English-to-German and Arabic-to-English TED talks. The NDAM models achieved better perplexities and modest BLEU improvements compared to the baseline NNJM, trained either on in-domain or on a concatenation of in- and out-domain data. On the other hand, the NFM models obtained significant improvements of up to +0.9 and +0.7 BLEU points, respectively. We also demonstrate improvements over existing adaptation methods such as instance weighting, phrasetable fill-up, linear and log-linear interpolations.

Keywords: Machine translation; Domain adaptation; Neural network joint model; Distributed representation of texts; Noise contrastive estimation

1. Introduction

Parallel data required to train Statistical Machine Translation (SMT) systems is often inadequate as it is typically collected opportunistically from wherever available (Koehn and Schroeder, 2007). The conventional wisdom is that more data improves the translation quality. Additional data however, may not be best suited for tasks such as translating TED talks (Cettolo et al., 2014), patents (Fujii et al., 2010) and educational content (Abdelali et al., 2014), that posit the challenges of dealing with word-sense ambiguities and stylistic variance of other genres. When additional data, later referred as *out-domain* data, is much larger than *in-domain* data, the resultant distribution can get biased towards out-domain, yielding a sub-optimal system. For example, an Arabic-to-English SMT system trained by simply concatenating in- and out-domain data translates the Arabic phrase "عن مشكلة الحمل الزائد للاختيار", to "about the problem of unwanted pregnancy". This translation is inaccurate in the context of the in-domain data,

* Corresponding author.
E-mail address: sjoty@qf.org.qa (S. Joty), ndurrani@qf.org.qa (N. Durrani), hsajjad@qf.org.qa (H. Sajjad), aabdelali@qf.org.qa (A. Abdelali).

http://dx.doi.org/10.1016/j.csl.2016.12.006

where it should be translated to "about the problem of choice overload". The sense of the Arabic phrase taken from out-domain data completely changes the meaning of the sentence.

Domain adaptation aims to preserve the identity of the in-domain data while exploiting the out-domain data in favor of the in-domain data and avoid possible drift towards out-domain jargon and style. This is typically done either by selecting a subset from the out-domain data, which is closer to the in-domain (Matsoukas et al., 2009; Moore and Lewis, 2010), or by reweighting the probability distribution in favor of the in-domain data (Foster and Kuhn, 2007; Sennrich, 2012).

Joint sequence ngram-based models (Mariño et al., 2006; Durrani et al., 2013) have shown to be effective in improving the quality of machine translation and have achieved state-of-the-art performance recently. Their ability to capture non-local dependencies makes them superior to the traditional models, which do not consider contextual information across phrasal boundaries. Such models however suffer from data sparsity. As the length of the sequence increases, the test sequences are likely to be different from the ones used for training the models. To overcome this problem, a transition towards continuous space modeling using neural networks has been proposed (Devlin et al., 2014; Bengio et al., 2003; Le et al., 2012). In this framework, a distributed representation is learned for each word in the process of modeling the word sequences.

We hypothesize that the distributed vector representation of neural models helps to bridge the lexical differences between the in-domain and out-domain data, and adaptation is necessary to avoid deviation and drift of the model from the in-domain, which otherwise happens because of the large out-domain data.

In this paper, we explore Neural Network Joint Model (NNJM) proposed by Devlin et al. (2014) for the task of domain adaptation in Statistical Machine Translation (SMT). Preliminarily, we customize state-of-the-art methods in domain adaptation, such as mixture modelling (Foster and Kuhn, 2007) and MML (Axelrod et al., 2011) to be used with NNJM. We train NNJM models from in- and out-domain data individually and interpolate them linearly or log-linearly to perform adaptation. We also tried NNJM-based data selection similar to MML filtering. Later, we propose two sets of novel models based on the NNJM framework: (*i*) Neural Domain Adaptation Models (NDAM), where we minimize the cross entropy by regularizing the loss function with respect to in-domain (and optionally to out-domain) model(s); (*ii*) Neural Fusion Models (NFM), where we combine in- and out-domain models and readjust their parameters to minimize the loss on the in-domain sequences.

The NDAM models use data dependent regularizations in their loss functions to perform instance weighting. In our first NDAM model (NDAM-v1), we use a regularizer based on an in-domain model to bias the resultant model towards the in-domain data. In the second model (NDAM-v2), we additionally use an out-domain model to penalize out-domain sequences that are similar to the out-domain data. The regularizers in our loss functions are inspired from the data selection methods proposed in Moore and Lewis (2010); Axelrod et al. (2011).

In the NFM models we train in- and out-domain NNJM models and fuse them by readjusting their parameters towards in-domain data. This is achieved by backpropogating errors from the output layer to the word embedding layer of each model. In a variant of the NFM model, we restrict backpropogation to only the outermost hidden layer and adjust only the final layer combination weights.

We evaluated our models against strong baselines on a standard task of translating IWSLT TED talks for English-to-German (EN-DE) and Arabic-to-English (AR-EN) language pairs using BLEU (Papineni et al., 2002). The baseline MT system uses an NNJM trained on a concatenation of in- and out-domain data ($NNJM_c$). In the adapted models we simply replace the baseline NNJM model with our adapted versions. The most relevant baseline to our work is the fine-tuning method proposed in Luong and Manning (2015). This method first learns a neural model on the concatenated data, then trains it further on the in-domain data to tune the model towards in-domain. We applied fine-tuning to the NNJM model and report it as an additional baseline.

We also compared our models against state-of-the-art model adaptation techniques including phrase-table fill-up (Bisazza et al., 2011), phrasetable interpolation and instance (phrase-level) weighting (Foster and Kuhn, 2007; Sennrich, 2012), and also against existing data selection methods like Modified-Moore-Lewis (MML) (Axelrod et al., 2011). Below is a summary of our main findings:

- The NDAM models gave an average improvement of +0.4 BLEU points over the best baseline. NDAM-v1 performed better on English-to-German and NDAM-v2 performs better on Arabic-to-English.
- The NNJM mixture models also gave average improvements of up to +0.4 BLEU points. Log-linear interpolation performed slightly worse than linear interpolation on Arabic-to-English.

Please cite this article as: S. Joty et al., Domain adaptation using neural network joint model, Computer Speech & Language (2017), http://dx.doi.org/10.1016/j.csl.2016.12.006