Knowledge-Based Systems 108 (2016) 15–24

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

# Bilingual recursive neural network based data selection for statistical machine translation

Derek F. Wong, Yi Lu, Lidia S. Chao*

*Natural Language Processing & Portuguese-Chinese Machine Translation Laboratory, Department of Computer and Information Science, University of Macau, Macau, China*

ARTICLE INFO

*Article history:*
Received 26 October 2015
Revised 28 April 2016
Accepted 6 May 2016
Available online 9 May 2016

*Keywords:*
Data selection
Machine translation
Domain adaptation
Recursive neural network
Autoencoder

ABSTRACT

Data selection is a widely used and effective solution to domain adaptation in statistical machine translation (SMT). The dominant methods are perplexity-based ones, which do not consider the mutual translations of sentence pairs and tend to select short sentences. In this paper, to address these problems, we propose bilingual semi-supervised recursive neural network data selection methods to differentiate domain-relevant data from out-domain data. The proposed methods are evaluated in the task of building domain-adapted SMT systems. We present extensive comparisons and show that the proposed methods outperform the state-of-the-art data selection approaches.



## 1. Introduction

As SMT systems acquire translation rules from training data, their performance relies heavily on the data. In general, the larger the training data, the better the translation systems can be. This is true for general-purpose SMT systems. However, experiments [1,10,36] showed that smaller but more relevant training data yields better translation quality when it comes to domain-specific translation tasks. An ideal domain-specific SMT system should be trained on a well maintained corpus which is from the domain of interest. This leads to the scenario of using in-domain data to filter out redundant and irrelevant data with the objective of regularizing the distributions of phrase pairs for domain translation [39]. On the other hand, in practice, a domain-specific corpus is difficult to obtain and usually limited in size, while general-domain data is easier to harvest and construct. In this second application scenario, general-domain data can be utilized to back up the in-domain data. The intention of this application is to broaden the content of data for training a better model [24,29].

Data selection is a complementary solution to these problems. Instead of using the large general-domain corpus, a sub-sample which is more relevant to the target domain is preferable for training a domain-specific system. Most data selection approaches use a model that is trained from a small domain-specific corpus to estimate the relevance of sentences (sentence pairs) $S_i$ in a general monolingual (bilingual) corpus $G$. The relevance is represented as a score and can be stated as follows [37]:

$$\text{Score}(S_i) \rightarrow \text{Sim}(S_i, R) \tag{1}$$

where $R$ is the abstract model to represent the target domain; meaning that we could score the relevance by measuring the similarity, $Sim(\cdot, \cdot)$, between $S_i$ and $R$. Sentences (sentence pairs) that give better scores are extracted to compose the pseudo in-domain sub-corpus $G_{sub}$.

Therefore, the main problem in data selection is finding an appropriate scoring function. Current dominant approaches are perplexity-based models [1,26]. They use language models (LMs) trained on an in-domain corpus to measure the perplexity of sentences in the general-domain corpus. Those sentences or sentence pairs which are assigned lower perplexity by the LMs are considered to be more domain relevant. However, these approaches rely solely on the surface forms and the word occurrences (word collocations) of sentences, which may be insufficient to represent domain-specific data without considering the linguistic properties and mutual translation features of a sentence pair. In addition, perplexity models tend to favor sentences containing fewer words, resulting in long but relevant sentences being filtered out.

In this work, we seek to address these issues by proposing a bilingual recursive neural network (biRNN) for the problem of data selection. The proposed model aims to learn higher level abstraction (vector representation) of sentence pairs by considering the syntactic and semantic information of sentences in a

* Corresponding author.
*E-mail addresses:* derekfw@umac.mo (D.F. Wong), takamachi660@gmail.com (Y. Lu), lidiasc@umac.mo (L.S. Chao).

http://dx.doi.org/10.1016/j.knosys.2016.05.003
0950-7051/

bilingual context [32]. The model is designed to be integrated with two recursive auto-encoders (RAE) in a bilingual context, one for each source and target sentence. To experience different neural network architectures, for the single-layer model, the representations of the source and target sentences are fed directly to a softmax layer. For the multi-layer model, we introduce a hidden layer between the representation of source and target sentences and the output layer by leveraging the recursive merging mechanism. To evaluate the proposed approaches in the task of domain adaptation, we employ our methods as well as the previous methods to build SMT systems trained on selected sub-corpora and examine their performances. Experimental results show that the proposed models yield better BLEU scores [28] than perplexity-based data selection models.

The remainder of this paper is organized as follows. We firstly review the related work in Section 2. Section 3 describes the proposed methods. Section 4 details the setup of experiments and reports the end-to-end SMT evaluation results and analysis. Finally, we draw a conclusion in Section 5.

## 2. Related works

Domain adaptation is an active field of research in both machine learning (ML) [8,9] and natural language processing (NLP) [4,38]. In particular, it has attracted much attention in field of machine translation (MT) [20,31,36]. In the big data environment, training data for SMT has increased significantly over the past decades. The data however is coming in a wider spectrum of texts encompassing different topics and genres. This is the reasons why data selection (and cleaning) has been the essential step in building a quality domain specific MT system.

The most commonly used selection criterion is the perplexity-based method [15,22]. It uses in-domain LM to measure the general-domain text. The sentences that are assigned lower perplexity are considered as domain data and are selected. The Moore-Lewis (ML) model [14,26,39] is an augmented method which considers not only the perplexity with respect to the in-domain LM but also the perplexity with respect to the general-domain LM, in order to differentiate domain specific sentences. [1] further extended the ML model to bilingual application. In general, the modified Moore-Lewis model performs better than early Information Retrieval methods [12,13,17] and other perplexity-based variants. It has been commonly used in the task of SMT domain adaptation.

Formally, given a language model $q$, the perplexity of a string $s$ with empirical $n$-gram distribution $p$ is defined by:

$$2^{H(p,q)} = 2^{-\sum_x p(x)\log q(x)}, \tag{2}$$

in which $x$ is the $n$-gram of $s$, $H(p, q)$ is the cross-entropy between $p$ and $q$. The relevance between $s$ and the target domain is calculated by bilingual cross-entropy difference [1]:

$$[H_{I-src}(s) - H_{O-src}(s)] + [H_{I-tgt}(t) - H_{O-tgt}(t)], \tag{3}$$

where $H_{I-*}(x)$ and $H_{O-*}(x)$ are the cross-entropy between the source or target side string of sentence pair $(s, t)$ and an in-domain language model $LM_{I-*}$ and an out-domain language model $LM_{O-*}$ respectively, which are trained on in-domain and out-domain data on the source and target sides. In line with this direction, there have been a number of studies. [10] employed a more robust language model based on recurrent neural networks to resolve the problem of unknown words by replacing the traditional $n$-gram LM. As stated in [3], the neural-based language model performs well in providing smooth probability estimates of unseen but relevant context. To account for linguistic features, [35] modeled various information such as lemmas, Part-of-Speech, and named entities of sentences instead of the surface form. Sentences with these additional representations that were previously limited by the naïve LM can be selected. However, these methods do not model the bilingual data at sentence level, which makes it difficult to capture the parallel sentences where their translations are in line with the target domain. Besides, the models also suffer from the problem of bias towards short sentences [1,10]. In this study, the proposed models are designed to address these issues. The models do not rely on the perplexity of domain data, but try to learn the deep representation of sentences to differentiate domain-specific from general contexts. Recently, [23] presented the utilization of the translation probabilities of bilingual phrases as additional scores in their data selection model. Their selection function is simply the sum of perplexities (of source and target sentences) and the bidirectional translation probabilities. The sentence pairs which receive higher scores are presumed to be domain-relevant data and are selected. However, the model favors the sentence pairs that are similar to the training data. Sentence pairs of unseen context are assigned lower scores by the model. In contrast, our model based on RNNs learns the abstract representation of sentences, which is able to better smooth the probabilities of the new context.

Although perplexity and neural network based data selection methods are the focus of this study, it is worth mentioning other approaches which have been proposed for SMT domain adaptation. The first attempt in domain adaptation via training data selection is the work of [12]. They adapted the method from information retrieval (IR) realm, TF-IDF, and cosine distance similarity measure to select sentences to adapt LMs in SMT systems. More followed similar technique include [17] and [13]. The standard IR approach considers bag-of-words. The alternative selection criterion is the edit-distance based similarity measure [21]. It was recently applied to domain adaptation and the empirical results showed that edit-distance criterion works well when the general corpus contains sentences that are very close to the in-domain data [37].

## 3. Bilingual recursive neural network based data selection model

In this section, the proposed data selection model is introduced. It is based on a bilingual recursive neural network, namely the RAE [32], which enables domain-specific context to be predicted to exploit syntactic and semantic information from a neural language modeling point of view [3]. An RAE is a kind of RNN which aims to find vector representations for variably sized phrases or sentences and, to some degree, is able to capture the linguistic meaning of sentences. The framework for the induction of vector space representation for sentences is first described, followed by an introduction of the bilingual setting of the network structure, objective function, and parameter inference.

### 3.1. Recursive auto-encoders

To generate phrase or sentence embedding using composition, the word is represented as a real-value vector [2,7], which serves as the basis for input to the neural network. These vectors are stacked into a word-embedding matrix $L \in \mathbb{R}^{n\times|V|}$, where $|V|$ is the size of the vocabulary. This word-embedding matrix is a parameter to be learned and subsequently modified to capture the domain information. Given a sentence as an ordered list of $m$ words, each word $w_i$ has an associated vocabulary index $k$, and retrieving the word-embedding vector from matrix $L$ can be seen as a projection layer:

$$x_i = L \cdot r_k \in \mathbb{R}^n, \tag{4}$$

where $r_k$ is a binary vector which is zero in all positions except at the $k$th index. The word vectors can be either pre-trained using an unsupervised neural language model [3,25,30], or randomly