Available online at www.sciencedirect.com

ScienceDirect

Computer Speech and Language 34 (2015) 1–17

COMPUTER SPEECH AND LANGUAGE

www.elsevier.com/locate/csl

Improving translation quality stability using Bayesian predictive adaptation☆

Germán Sanchis-Trilles*,[1], Francisco Casacuberta

Pattern Recognition and Human Language Technologies Center, Universitat Politècnica de València, 46022 Valencia, Spain

Received 27 June 2014; received in revised form 17 December 2014; accepted 9 March 2015
Available online 20 March 2015

Abstract

We introduce a Bayesian approach for the adaptation of the log-linear weights present in state-of-the-art statistical machine translation systems. Typically, these weights are estimated by optimising a given translation quality criterion, taking only into account a certain set of development data (e.g., the adaptation data). In this article, we show that the Bayesian framework provides appropriate estimates of such weights in conditions where adaptation data is scarce. The theoretical framework is presented, alongside with a thorough experimentation and comparison with other weight estimation methods. We provide a comparison of different sampling strategies, including an effective heuristic strategy and a theoretically sound Markov chain Monte-Carlo algorithm. Experimental results show that Bayesian predictive adaptation (BPA) outperforms the re-estimation from scratch in conditions where adaptation data is scarce. Further analysis reveals that the improvements obtained are due to the greater stability of the estimation procedure. In addition, the proposed BPA framework has a much lower computational cost than raw re-estimation.

Keywords: Bayesian methods; Adaptation; Natural language processing; Machine translation

1. Introduction

Adaptation has become a very popular issue in natural language processing (Kuhn and De Mori, 1990; Huo et al., 1995; Koehn and Schroeder, 2007), and more specifically in *statistical machine translation* (SMT) (Koehn, 2010). Typically, the adaptation problem arises when two very different sets of training data are available, yielding two different sets of model parameters. The first set of data, the training data $\mathcal{T}$ (e.g., obtained from the European Parliament or the United Nations) is often very large and rather generic in domain. The second set of data, the adaptation data $\mathcal{A}$, belongs to the specific task of interest, such as printer manuals or medical diagnoses, and is usually overwhelmingly smaller than $\mathcal{T}$. Then, the challenge is to modify the SMT system appropriately by taking into consideration both $\mathcal{T}$ and $\mathcal{A}$: on the one hand, $\mathcal{T}$ is should provide robustness in the estimation of the model parameters $\boldsymbol{\theta}$, and on the other hand $\mathcal{A}$ should introduce a certain bias towards the specific task.

☆ This paper has been recommended for acceptance by E. Briscoe.

* Corresponding author. Tel.: +34 963878170.
E-mail addresses: gersantr@upv.es (G. Sanchis-Trilles), fcn@prhlt.upv.es (F. Casacuberta).
[1] Currently at Sciling S.L.

http://dx.doi.org/10.1016/j.csl.2015.03.001
0885-2308/

This definition of adaptation is specially appropriate for the Bayesian learning paradigm, where the model parameters $\boldsymbol{\theta}$ are treated as (hidden) random variables governed by some kind of a priori distribution $p(\boldsymbol{\theta})$. This distribution represents our prior knowledge about what values for $\boldsymbol{\theta}$ should be good estimates. Estimating $p(\boldsymbol{\theta})$ by using a sufficiently large collection of data $\mathcal{T}$ allows us to obtain a canonical model with parameters $\boldsymbol{\theta}_{\mathcal{T}}$, and it can be assumed that such estimation is a robust estimation. As further evidence arrives in form of adaptation data $\mathcal{A}$, that such estimations are revised so that they reflect the newly arrived data. Considering $\mathcal{A}$ within the Bayesian predictive distribution leads precisely to a scenario in which the decision regarding the output sentence includes a bias towards $\mathcal{A}$, but is still guided by $p(\boldsymbol{\theta}_{\mathcal{T}})$ (i.e., the prior distribution given $\mathcal{T}$). Hence, under the *Bayesian predictive adaptation* (BPA) framework, the final translation is not computed by considering only the topic-specific data (i.e., $\mathcal{A}$), which could lead to over-trained estimations of $\boldsymbol{\theta}$: if the amount of data available is small, the parameter prior $p(\boldsymbol{\theta})$ will compensate this, providing robustness (Duda et al., 2001). However, the effect of this prior knowledge fades when incorporating further evidence, until a point in which the contribution of the parameter prior towards the complete model distribution is negligible. In addition, the Bayesian learning paradigm does not attempt to obtain a single best point estimate of $\boldsymbol{\theta}$, but rather relies on considering all possible parameter values, allowing uncertainty regarding what the best estimations of such parameters might be. In this paper, we focus on the Bayesian adaptation of the weights of the log-linear combination of features present in state-of-the-art SMT systems. Even though these weights are not very numerous (generally in the range of 10 or 20), providing the system with appropriate estimates for these weights is critical (Clark et al., 2011).

The rest of this paper is structured as follows: the related literature is reviewed in Section 2. The formal derivation of Bayesian predictive adaptation for SMT is presented in Section 3. Since the equation obtained is very costly to apply in practise, different sampling strategies are presented in Section 4. The experiments performed are detailed together with their results and the related analysis in Section 5. Finally, conclusions are presented in Section 6.

2. Related work

Adaptation in SMT is a research field that has been receiving increasing attention. Following the ideas in Kuhn and De Mori (1990), one of the first works was performed in Nepveu et al. (2004), where the authors added cache language and translation models to an interactive machine translation system. In Koehn and Schroeder (2007), different ways to combine the available data belonging to two different sources were studied. The work in Civera and Juan (2007) explores alignment model mixtures as a way of performing topic adaptation. Other authors (Zhao et al., 2004; Sanchis-Trilles et al., 2009), have proposed the use of information retrieval and clustering techniques in order to extract the sub-domains of a large corpus, and Gascó et al. (2010) and Axelrod et al. (2011) proposed to select as training data only those sentences which can be considered topic-specific. Corpus weighting strategies were analysed in Matsoukas et al. (2009), and instance weighting techniques were applied in Foster et al. (2010) in order to weight out-of-domain phrase pairs. Recently, sequential Bayesian methods were applied with the purpose of adapting the word alignments present in most state-of-the-art SMT systems (Duh et al., 2011). In such work, the authors confront the problem of adapting the probabilities of the single-word models that are used for phrase extraction. In contrast, in this work we attempt to adapt the final translation model directly. Note that none of these works confront the problem of adapting the log-linear weights $\boldsymbol{\lambda}$ of the SMT system, but rather attempt to adapt either the underlying word alignments or the final translation model features $\mathbf{h}$, and comparison with such strategies is not suitable. Hence, re-estimating $\boldsymbol{\lambda}$ from scratch is, to the best of our knowledge, the most common approach when adapting $\boldsymbol{\lambda}$. This work intends to fill this gap, and can be seen as complementary to the adaptation approaches cited above.

Although only recently applied to SMT, Bayesian adaptation has been successfully applied in other natural language processing areas, such as speech recognition (Huo et al., 1995). In fact, work done in this direction is very broad, covering both batch (Yu and Gales, 2005) and online adaptation (Yu and Gales, 2006). Variational Bayes approaches have also been studied (Valente and Wellekens, 2005), which attempt to find a lower bound to approximate the intractable marginal likelihood, yielding point estimates of the model parameters. Alternatively, BPA attempts to approximate the marginal likelihood directly by sampling from the posterior distribution, and usually leads to more robust estimates (Yu and Gales, 2005).

With respect to BPA in SMT, to our knowledge the only work published as of yet in this direction is Sanchis-Trilles and Casacuberta (2010). In that article, only the idea was introduced, together with preliminary experiments. Here, such preliminary work is widely extended both in depth and in range: