Mach Translat (2017) 31:187–224
https://doi.org/10.1007/s10590-018-9216-8

A survey of domain adaptation for statistical machine translation

Hoang Cuong[1] · **Khalil Sima'an**[2]

Received: 2 June 2017 / Accepted: 24 February 2018 / Published online: 17 March 2018

Abstract Differences in domains of language use between training data and test data have often been reported to result in performance degradation for phrase-based machine translation models. Throughout the past decade or so, a large body of work aimed at exploring domain-adaptation methods to improve system performance in the face of such domain differences. This paper provides a systematic survey of domain-adaptation methods for phrase-based machine-translation systems. The survey starts out with outlining the sources of errors in various components of phrase-based models due to domain change, including lexical selection, reordering and optimization. Subsequently, it outlines the different research lines to domain adaptation in the literature, and surveys the existing work within these research lines, discussing how these approaches differ and how they relate to each other.

Keywords Statistical machine translation · Domain adaptation · Survey

1 Introduction

Machine translation (MT) systems are often applied in settings where the test data might be sampled from a distribution that differs from the training data, usually due to different domains of language use. This domain mismatch between training and test data often leads to performance degradation, usually due to lexical differences between

✉ Hoang Cuong
ch2107@hunter.cuny.edu

Khalil Sima'an
k.simaan@uva.nl

[1] The City University of New York, New York, NY, USA

[2] ILLC, University of Amsterdam, Amsterdam, The Netherlands

the domains. When a word in the test data is found in the training data, its most suitable translation in the test domain might be different from that in the training domain. For example, when translating from English to Russian, the most natural translation for the word '*code*' would be 'шифр' ('cipher'), 'закон' ('law') or '' ('program') if we consider *cryptography*, *legal* and *software development* domains, respectively. Given parallel training data originating from one of those domains, training an MT system on the data would produce a rather suboptimal translation for the other domains.

Surprisingly, degradation in translation quality is observed even when we train an MT system on large heterogeneous corpora such as EuroParl (Koehn 2005),[1] Common Crawl Corpus,[2] UN Corpus,[3] and News Commentary[4] (Shah et al. 2012; Carpuat et al. 2014; Cuong et al. 2016b). For instance, Axelrod et al. (2011) show that when it comes to a domain-specific task, a small percentage of well-selected data can outperform the full heterogeneous dataset for training MT systems (Biçici and Yuret 2011; Poncelas et al. 2017). Shah et al. (2010) show that it would benefit from training word alignment with weighting sentence pairs according to their relevance to a domain-specific task.

In this paper, we provide a comprehensive survey of domain adaptation for statistical machine translation (SMT), aimed particularly at phrase-based systems (Koehn et al. 2003). A very basic question is what constitutes a domain? There are different definitions in the literature, for example:

- The 'provenance' of the training data (Foster and Kuhn 2007; Moore and Lewis 2010; Sennrich 2012);
- The difference of words and grammars between corpora (Pecina et al. 2012);
- The thematic content in the training data, such as topic (Hasler et al. 2014; Hu et al. 2014);
- A particular combination of many factors: genres, topics, dialects and writing styles (Chen et al. 2013a).

We do not aim to find the best answer, as the concept of domain is still an open question and has not been well-defined in the literature (see Van Der Wees et al. (2015) for a discussion). We rather provide a systematic overview of previous approaches to domain adaptation, showing their advantages/disadvantages, as well as how they relate to and differ from each other.

The survey is organized as follows. We first introduce SMT in general, with a focus on aspects of SMT relevant to domain adaptation (Sects. 2, 3).[5] The survey identifies components that need to be adapted when an SMT system is applied to new domains (Sect. 4). We explain what may go wrong in translation by analyzing potential sources of translation errors and providing an explanation as to why each specific type of error may happen.

[1] See Ozdowska and Way (2009) for a clear demonstration that building MT systems with more EuroParl data does not always lead to better translation results.

[2] http://www.statmt.org/wmt13/training-parallel-commoncrawl.tgz.

[3] http://www.statmt.org/wmt13/training-parallel-un.tgz.

[4] http://www.statmt.org/wmt15/training-parallel-nc-v10.tgz.

[5] Readers may refer to Lopez (2008) or Koehn (2010) for a comprehensive survey of SMT in general.

Subsequently, we present a general picture of domain adaptation for SMT where we outline the main general approaches (Sect. 5). A major part focuses on the induction (Sect. 6) and combination (Sect. 7) of domain-focused phrase translation tables, lexical weights and reordering probabilities. The induction of domain-focused sparse features and word-alignment probabilities are discussed in Sects. 8.1 and 8.2.

Finally we also cover several practical adaptation scenarios, including adapting an existing system to multiple specific domains at the same time (Sect. 8.3). Another scenario addressed in (Sect. 8.4) is embedding an SMT system into a Cross-lingual Information Retrieval (CLIR) system (i.e. automatically translating queries into different languages, so that a search engine can return search results in the corresponding languages). We also discuss how web-based translation services such as Bing Translator[6] and Google Translate[7] can be improved when the domain of a new request is not known in advance. Specifically, we cover cache-based adaptive models (Sect. 8.5) and rewarding domain invariance for adaptation (Sect. 8.6).

2 Statistical machine translation

In SMT, we aim to translate a source (foreign) sentence $\mathbf{f}$ into a sentence in the target language $\mathbf{e}$. Among the target translation hypotheses, the translation hypothesis $\hat{\mathbf{e}}$ with the highest probability given the source sentence is selected, as in (1):

$$\hat{\mathbf{e}} = \text{argmax}_{\mathbf{e}}\{P(\mathbf{e}|\,\mathbf{f})\} = \text{argmax}_{\mathbf{e}}\{P(\mathbf{e})P(\mathbf{f}|\,\mathbf{e})\}. \tag{1}$$

This approach to modeling translation is referred to as the noisy-channel framework. The architecture of the framework includes two components: the translation model (i.e. $P(\mathbf{f}|\,\mathbf{e})$) and the language model (i.e. $P(\mathbf{e})$).

A more powerful approach exploits a log-linear formulation, more formally, where the posterior probability $P(\mathbf{e}|\,\mathbf{f})$ is modeled with a set of M feature functions $\boldsymbol{\phi}(\mathbf{e},\ \mathbf{f}) = \{\phi_1(\mathbf{e},\ \mathbf{f}),\ \ldots,\ \phi_M(\mathbf{e},\ \mathbf{f})\}$ with model parameters $\mathbf{w} = \{w_1,\ \ldots,\ w_M\}$ as in (2):

$$P(\mathbf{e}|\,\mathbf{f}) \propto \exp(\mathbf{w} \cdot \boldsymbol{\phi}(\mathbf{e},\ \mathbf{f})). \tag{2}$$

Under this framework, we obtain the decision rule in (3):

$$\hat{\mathbf{e}} = \text{argmax}_{\mathbf{e}}\mathbf{w} \cdot \boldsymbol{\phi}(\mathbf{e},\ \mathbf{f}). \tag{3}$$

The decision rule is simple as we can safely ignore the daunting normalization factor.

The model was first proposed by Och and Ney (2002), forming the basis of phrase-based SMT systems. It is straightforward to see that this framework contains the noisy-channel framework as a special case. Its advantage lies in its flexibility, relative to the noisy-channel framework, as one can extend a basic SMT system containing translation and language models by including arbitrary feature functions of the source

[6] https://www.bing.com/translator/.

[7] https://translate.google.com.

and the target sentences. There are many possibilities for defining feature functions that help the SMT system to improve translation, such as linguistic features, word and phrase penalties, reordering features, and rule counting. Simply adding feature functions from the target to source language also often improves translation.

Learning the model parameters $\mathbf{w} = \{w_1, \ldots, w_M\}$) using a held-out development set is crucial to improving translation quality. In principle, training for log-linear models can be done using maximum likelihood or related criteria (e.g. *cross-entropy*, *perplexity*). Such an objective function is convex, and global optimization is possible. The main difficulty, however, is that we need to compute the normalization factor during learning. This is intractable, as we cannot explore the full space of all translation hypotheses for each translation input. In practice, the normalization factor is computed using an N-best list of top-N translation hypotheses or a lattice (Macherey et al. 2008).[8]

Optimizing an SMT system using maximum likelihood or related criteria has a loose relation to the translation quality on unseen text (Och 2003). There is a need to directly incorporate translation accuracy on a held-out development set into the optimization, now a fundamental part of modern SMT systems. Numerous optimization methods have been proposed in the literature, such as MERT (Och 2003), MIRA (Watanabe et al. 2007; Chiang et al. 2008; Cherry and Foster 2012), and Pairwise Ranked Optimization (PRO: Hopkins and May (2011)). Readers may refer to Neubig and Watanabe (2016) for a comprehensive survey of system optimization methods in general.

The latter SMT framework has two notable shortcomings that make the problem of domain adaptation for SMT even more challenging:

- First, having more translation features significantly increases the difficulty of the optimization. Specifically, having more feature dimensions requires a much larger held-out development set for system optimization, as shown in Waite and Byrne (2015). This is an issue in domain adaptation for SMT because creating such an in-domain held-out development dataset is expensive.
- Second, log-linear models try to separate good and bad translation hypotheses using a linear hyper-plane. This is potentially problematic, as interactions between domain-specific features can be complex. It may be necessary to perform preprocessing steps over the feature space to produce a feature set that is less prone to non-linearities (Liu et al. 2013; Clark et al. 2014). However, methods tailored to such a special treatment are quite sophisticated and not widely deployed in practice.

3 Phrase-based SMT system

There are many types of translation systems that have been built in the past, for example:

- Syntax-based translation systems (Yamada and Knight 2001),
- Phrase-based SMT systems (Och and Ney 2002; Koehn et al. 2003),
- Hierarchical phrase-based SMT systems (Chiang 2005, 2007),

[8] As a side note, the size of the N-best list does not seem to have a significant impact on adaptation [cf. Bertoldi and Federico (2009)].

– Syntactic phrase-based SMT systems (Quirk et al. 2005; Quirk and Menezes 2006).

This paper focuses on phrase-based SMT systems (Och and Ney 2002; Koehn et al. 2003).

3.1 Model

A standard phrase-based SMT system has various dense feature functions (i.e. highly informative feature functions) estimated at phrase level. Three of the most important translation models are a phrase-based model $\phi_{TM}(\mathbf{e}, \mathbf{f})$, lexical weighting $\phi_{LW}(\mathbf{e}, \mathbf{f})$, and reordering model $\phi_{RM}(\mathbf{e}, \mathbf{f})$. A common domain-adaptation strategy for SMT is to directly adapt these models. We thus describe them in detail below.

– *Phrase-based model* At the core of a phrase-based SMT system is the phrase-based model, which aims at modeling translation of sentence pairs at phrase level. Given an input sentence $\mathbf{f}$, let us assume that a sequence of target-language phrases $\mathbf{e} = (\tilde{e}_1, \tilde{e}_2, \cdots, \tilde{e}_n)$ is currently hypothesized by the decoder. Let us also assume we are provided with a phrase alignment $\mathbf{a} = (a_1, a_2, \cdots, a_n)$ that defines a source $\tilde{f}_{a_i}$ for each translated phrase $\tilde{e}_i$. The model is estimated as in (4):

$$\begin{aligned}\phi_{TM}(\mathbf{e}, \mathbf{f}) &= \log P_{TM}(\mathbf{e}|\, \mathbf{f}) = \log \prod_{i=1}^{n} P(\tilde{e}_i|\, \tilde{f}_{a_i}) \\ &= \sum_{i=1}^{n} \log P(\tilde{e}_i|\, \tilde{f}_{a_i}) \end{aligned} \tag{4}$$

– *lexical weighting* The lexical weighting provides smoother estimates for probabilities of phrase pairs. The model is estimated as in (5):

$$\begin{aligned}\phi_{LW}(\mathbf{e}, \mathbf{f}) &= \log P_{LW}(\mathbf{e}|\, \mathbf{f}) = \log \prod_{i=1}^{n} P(\tilde{e}_i|\, \tilde{f}_{a_i},\ a_i) \\ &= \sum_{i=1}^{n} \log P(\tilde{e}_i|\, \tilde{f}_{a_i},\ a_i) \end{aligned} \tag{5}$$

Here, the distribution $P(\tilde{e}_i|\, \tilde{f}_{a_i},\ a_i)$ is computed based on lexical probabilities $P(e \mid f)$ between words $\langle e,\ f \rangle$ in a phrase pair $\langle \tilde{e}_i,\ \tilde{f}_{a_i} \rangle$. Different models have a slightly different way of computing $P(\tilde{e}_i|\, \tilde{f}_{a_i},\ a_i)$. A typical estimate of $P(\tilde{e}_i|\, \tilde{f}_{a_i},\ a_i)$ (Koehn et al. 2003) is as in (6):

$$P(\tilde{e}_i|\, \tilde{f}_{a_i},\ a_i) = \prod_{i=1}^{|\tilde{e}_i|} \frac{1}{|\{j|(j,\ k) \in a_i\}|} \sum_{(j,\ k) \in a_i} P(\tilde{e}_i^k|\, \tilde{f}_{a_i}^j). \tag{6}$$

Here,
- $\tilde{e}_i^k$: word at position k in target phrase $\tilde{e}_i$,
- $\tilde{f}_{a_i}^j$: word at position j in source phrase $\tilde{f}_{a_i}$.
- $|\tilde{e}_i|$: length of phrase $\tilde{e}_i$
- $|\{j|(j,\ k) \in a_i\}|$: the number of source words that each target word at position k in phrase $\tilde{e}_i$ aligns to.

– *Reordering model* Such phrase-based models and lexical weighting are not meant for handling word/phrase order phenomena between languages. For state-of-the-art phrase-based SMT systems, integrating *lexicalized reordering models* (Tillmann 2004; Koehn et al. 2007; Galley and Manning 2008) must be considered. These models estimate the probability of a sequence of orientations $\mathbf{O} = (o_1,\ o_2,\ \ldots,\ o_n)$ as in (7):

$$\begin{aligned}\phi_{RM}(\mathbf{e}, \mathbf{f}, \mathbf{O}) &= \log P_{RM}(\mathbf{O}|\ \mathbf{e},\ \mathbf{f}) = \log \prod_{i=1}^{n} P(o_i|\tilde{e}_i,\ \tilde{f}_{a_i},\ a_{i-1},\ a_{i-2}) \\ &= \sum_{i=1}^{n} \log P(o_i|\ \tilde{e}_i,\ \tilde{f}_{a_i},\ a_{i-1},\ a_{i-2}) \end{aligned} \tag{7}$$

Here, each orientation o_i takes possible values $\{M, S, D\}$, representing how likely a phrase is to directly follow a previous phrase (***M****onotone*), to swap positions with it (***S****wap*), or to be not adjacent to it (***D****iscontinous*).

Beside these three types of dense translation features, there are also penalties for word, phrase and distance-based reordering. Those are the basic translation features that form a phrase-based SMT system (beside the language model).

A phrase-based SMT system can be also augmented with millions of sparse feature functions (e.g. phrase features (Chiang et al. 2009; Simianer et al. 2012), lexical features (Watanabe et al. 2007; Chiang et al. 2009), or syntax-based features (Blunsom and Osborne 2008; Marton and Resnik 2008)). It is possible to induce sparse features using a large portion of the parallel training data. However, scaling training to large data requires extensive additional efforts [cf. Yu et al. (2013)]. Models employing sparse features are often trained using a small held-out development set in practice.

3.2 Training

The most common approach to training a phrase-based SMT system is using relative frequency estimation. We take phrase translation scores as an example. To compute $P(\tilde{e}|\ \tilde{f})$, we first count the number of times phrase $\tilde{e}$ aligns to phrase $\tilde{f}$ in the parallel training data, before normalizing into probability by dividing by the total number of possible alignments to $\tilde{f}$, as in (8):

$$P(\tilde{e}|\ \tilde{f}) = \frac{c(\tilde{e},\ \tilde{f})}{\sum_{\tilde{e}'} c(\tilde{e}',\ \tilde{f})} \tag{8}$$

This distribution, however, does not necessarily maximize the likelihood of the parallel training data. This is similar to the Data Oriented Parsing (DOP) method (Bod et al. 2003) in parsing, which hypothesizes a distribution over many possible derivations of each training example from subtrees of varying sizes.

The key to the training is extracting bilingual phrases from bilingual data. The standard way is to rely on the word-aligned training data, using a heuristic method such as *grow-diag-final-and*, *grow-diag-final* or *final* (Koehn et al. 2003).

3.2.1 Word alignment

We now discuss how to create word-aligned training data. Given a parallel sentence, we look for the most probable alignment between words, $\hat{\mathbf{a}}$, as in (9):

$$\hat{\mathbf{a}} = \underset{\mathbf{a}}{\operatorname{argmax}}\, P(\mathbf{f},\ \mathbf{a}|\ \mathbf{e}). \tag{9}$$

The idea of word alignment can be traced back to Brown et al. (1990). The degree of difficulty of the search in Eq. (9) depends on the underlying independence assumptions. Even now, over twenty years since the IBM Models (Brown et al. 1993) and the HMM-based alignment model (Vogel et al. 1996), word alignment is still an active research topic (Simion et al. 2013; Chang et al. 2014; Tamura et al. 2014; Liu et al. 2015; Shen et al. 2015; Wang et al. 2015).

We now briefly review the HMM alignment model (Vogel et al. 1996), which is one of the most popular and widely used alignment models. The generative story of the model is shown in Fig. 1. The latent states rely on the target-language words and generate source-language words.

Formally, let us assume the target sentence $\mathbf{e}$ contains I words $\mathbf{e} = (e_1, \ldots, e_I)$ and the source sentence $\mathbf{f}$ contains J words $\mathbf{f} = (f_1, \ldots, f_J)$. For an alignment $\mathbf{a} = (a_1, \ldots, a_J)$ of the sentence pair $\langle \mathbf{e}, \mathbf{f} \rangle$, the model factors $P(\mathbf{f},\ \mathbf{a}|\ \mathbf{e})$ into the word-translation and transition probabilities as in (10):

$$P(\mathbf{f},\ \mathbf{a}|\ \mathbf{e}) = \prod_{j=1}^{J} P(f_j|\ e_{a_j}) P(a_j|\ a_{j-1}). \tag{10}$$

Here, $P(f_j|\ e_{a_j})$ represents word-translation probabilities and $P(a_j|\ a_{j-1})$ represents word-transition probabilities. In practice $P(a_j|\ a_{j-1})$ depends only on the distance $(a_j - a_{j-1})$. Note also that the first-order dependency model is an extension of the uniform dependency model of IBM Model 1 and zero-order dependency model of IBM model 2. With the HMM alignment model, the most probable alignment $\hat{\mathbf{a}}$ for each sentence pair can be computed efficiently using the Viterbi algorithm.

The HMM alignment model has two kinds of parameters: word-translation probabilities and transition probabilities. Adapting the expectation maximization (EM) algorithm (Dempster et al. 1977) for training the model is straightforward (Vogel et al. 1996). For the sake of completeness we present the algorithm in detail. We use $c(f|\ e;\ \mathbf{f},\ \mathbf{e})$ to denote the expected count of word e aligning to word f. We also use $c(i|\ i';\ \mathbf{f},\ \mathbf{e})$ to denote the expected counts of two certain consecutive source words

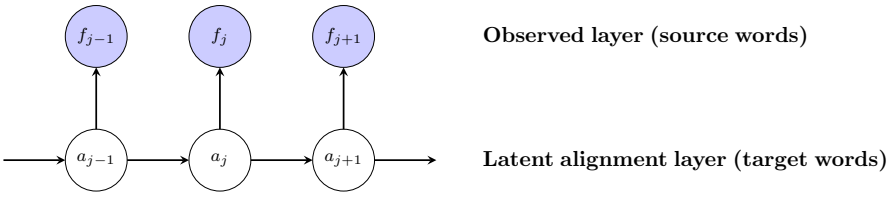

Fig. 1 HMM alignment model with observed and latent alignment layers

E-step

$$c(f|\ e;\ \mathbf{f},\ \mathbf{e}) \ = \ \sum\nolimits_{\mathbf{a}} \frac{P^{(c)}(\mathbf{f},\ \mathbf{a}|\ \mathbf{e})}{P^{(c)}(\mathbf{f}|\ \mathbf{e})} \sum\nolimits_{j=1}^{J} \delta(f,\ f_j) \sum\nolimits_{i=0}^{I} \delta(e,\ e_i) \tag{11}$$

$$c(i|\ i';\ \mathbf{f},\ \mathbf{e}) \ = \ \sum\nolimits_{\mathbf{a}} \frac{P^{(c)}(\mathbf{f},\ \mathbf{a}|\ \mathbf{e})}{P^{(c)}(\mathbf{f}|\ \mathbf{e})} \sum\nolimits_{j=1}^{J} \delta(a_j,\ i)\delta(a_{j-1},\ i') \tag{12}$$

M-step

$$P^{(+)}(f|e) = \frac{\sum_{\langle \mathbf{f},\ \mathbf{e}\rangle} c(f|e;\mathbf{f},\ \mathbf{e})}{\sum_{f} \sum_{\langle \mathbf{f},\ \mathbf{e}\rangle} c(f|\ e;\ \mathbf{f},\ \mathbf{e})},\ P^{(+)}(i|i') = \frac{\sum_{\langle \mathbf{f},\ \mathbf{e}\rangle} c(i|i';\mathbf{f},\ \mathbf{e})}{\sum_{i} \sum_{\langle \mathbf{f},\ \mathbf{e}\rangle} c(i|\ i';\ \mathbf{f},\ \mathbf{e})} \tag{13}$$

Fig. 2 Pseudocode for the training algorithm for the HMM alignment model. Note that $P^{(c)}$ denotes current iteration estimates, $P^{(+)}$ denotes the re-estimates and δ denotes the Kronecker delta function. Note that $P(\cdot|\ \cdot) = \sum_{\mathbf{a}} P(\cdot,\ \mathbf{a}|\ \cdot)$ which can be computed efficiently using dynamic programming

j and $j-1$ aligning to two target words i and i', respectively. Figure 2 presents the algorithm.

Does word alignment suffer from domain mismatch? A domain mismatch could have a negative impact on word-alignment accuracy, for example:

- Word-alignment models, like any statistical models, suffer from lack of in-domain data for training (Duh et al. 2010; Shah et al. 2010; Gao et al. 2011).
- The insensitivity of existing word-alignment models to domains often yields sub-optimal results on large heterogeneous data (Gao et al. 2011; Cuong and Sima'an 2015).

In Sect. 8.2 we discuss this aspect in detail.

3.3 Decoding

Decoding for phrase-based SMT system is a difficult problem. The search can be done by various approaches, such as beam search (Koehn 2004) or exact decoding (Chang and Collins 2011; Aziz et al. 2014). Among these competing approaches, beam search is probably the most popular decoding framework for phrase-based SMT systems. Starting from an initial hypothesis, given an input string of words, a number of phrase translations are applied to expand the current hypothesis until all words are marked as translated.

Beam search heuristically prunes the search space, and as a result, the search is inexact and search errors can occur as the best-scoring hypothesis is not necessarily optimal in terms of the given model parameters. Extensive prior work on minimum Bayes risk (MBR) objectives [cf. Kumar and Byrne (2004)] can potentially mitigate this issue. MBR methods select translations that are less 'risky' by taking the uncertainty in model predictions into account. Sect. 8.6 discusses a link between MBR and domain adaptation for SMT.

Table 1 Translation errors on an unseen domain

English–Spanish (task: consumer and industrial electronics)	
Input	*El reproductor puede reproducir señales de audio grabadas en mix-mode cd, cd-g, cd-extra y cd text*
Human translation	*The player **can play back audio signals** recorded in mix-mode cd, cd-g, cd-extra and cd text*
SMT output	*The player **can reproduce signs of audio** recorded in mix-mode cd, cd-g, cd-extra and cd text*
Input	*Se puede crear un archivo autodescodificable cuando el archivo codificado se abre con la contraseña maestra*
Human translation	*A self-decrypting file can be created when the encrypted file is **opened with** the **master password***
SMT output	*To create an file autodescodificable when the file codified **commenced with** the **password teacher***
Input	*Repite todas las pistas (únicamente cds de vídeo sin pbc)*
Human translation	*Repeat **all tracks** (non-pbc video cds only)*
SMT output	*Repeated **all avenues** (only cds video without pbc)*

4 Translation errors when applied to new domains

Applying a phrase-based SMT system to new domains produces suboptimal translation in practice, e.g. Newswire (Foster et al. 2013), Medical (Irvine et al. 2013b), Patents (Wäschle and Riezler 2012), Transcribed Lectures (Federico et al. 2012), Web Blogs (Su et al. 2012; Foster et al. 2013), TED Talks (Duh et al. 2010; Mansour et al. 2011; Hasler et al. 2014), Subtitles (Irvine et al. 2013b), or Web Queries (Nikoulina et al. 2012). This section reviews different sources of translation errors when applied to new domains.

4.1 Lexical selection

Lexical selection appears to be the most common source of errors (Irvine et al. 2013a; Van Der Wees et al. 2015). We present some examples in Table 1. Here, we train a standard phrase-based SMT system for English–Spanish on a large dataset combined from multiple resources including EuroParl, Common Crawl Corpus, UN Corpus, and News Commentary. We then apply the system to a new domain of "*Consumer and Industrial Electronics*". As shown in Table 1, incorrect translations are "*can reproduce signs of audio*" instead of "*can play back audio signals*", "*password teacher*" instead of "*master password*", "*commenced with*" instead of "*opened with*" file, and "Repeated *all avenues*" instead of "Repeat *all tracks*".

An important question is what went wrong with lexical selection, i.e. what made the phrase-based SMT system suffer from degradation in lexical translation quality on

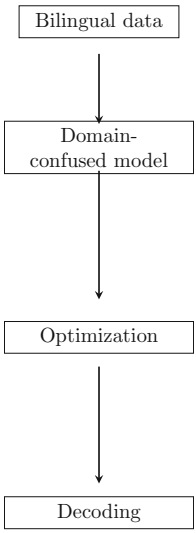

Fig. 3 Statistical translation framework

new domains? Two main different error types that cause the degradation are as follows (Irvine et al. 2013a)[9]:

- SEEN/SENSE: an incorrect translation for unobserved source-language words and an incorrect translation because of known source-language words but with unobserved target words in the parallel training data.
- SCORE: an incorrect translation for which the system goes for an incorrect translation path (i.e. incorrect ranking).

The majority of cases where degradation in lexical translation quality is seen are due to SEEN and SENSE errors. However, it is important to understand that improving coverage does not necessarily result in better translation quality. This leads to the error type SCORE, which is perhaps a much harder problem to address.

To provide a better understanding of the SCORE error, let us step back and reconsider how SMT models are estimated (Fig. 3). Statistical translation models are trained without integrating (likely hidden) domain information of the bilingual data. This results in coarse and domain-confused translation statistics that reflect translation preferences *aggregated* over different translation options with respect to different domains. Some translation options are more popular than others for a specific word or phrase in general. When it comes to a specific domain, however, it is likely that one of the rare translation options would be the most relevant one. A standard phrase-based SMT system is unlikely to be able to provide such a translation in this case, given that resulting domain-confused statistics are not expressive enough as they do not take domain information into account.

[9] In principle, search errors caused by a decoding algorithm can be a factor. The contribution of this factor to degradation of lexical translation quality, however, is minor, as shown in Irvine et al. (2013a).

4.2 Reordering

Different from the lexical selection, it is not clear that reordering model adaptation improves translation. There is some evidence supporting this hypothesis, notably from Chen et al. (2013a) and Zhang et al. (2015). Chen et al. (2013a) show that there are two potential reasons for an improvement in translation quality caused by reordering model adaptation:

- some corpora may be better for training reordering models than others, and
- there exists domain-dependent differences in reordering.

The first statement is intuitively plausible. Some data may contain noisy parallel sentences (e.g. comparable data), or simply too short parallel examples (e.g. Subtitles, Search Queries), which have a negative impact on parameter estimates (i.e. less accurate estimates).

Meanwhile, it is not at all obvious that reordering of phrase pairs is particularly domain-specific. Chen et al. (2013a) suggest that this is the case for Chinese–English and Arabic–English. They train lexicalized reordering models (Tillmann 2004; Koehn et al. 2007; Galley and Manning 2008) on different but high-quality parallel training data with specific genres. Their results show that the estimates of reordering parameters are significantly different between the corpora (e.g. the reordering probabilities estimated from News bilingual training data are different from those estimated from Legal bilingual data). It is, therefore, unsurprising that domain adaptation can help phrase-based SMT systems to improve reordering for English–Chinese as in Chen et al. (2013a).

However, it is unlikely that this would happen for all language pairs. Taking English–Spanish as an example, Cuong and Sima'an (2014a) train different lexicalized reordering models on a somewhat similar scenario with News parallel training data, including four sub-corpora: EuroParl, Common Crawl Corpus, UN Corpus, and News Commentary. They show that adapting reordering models for a new domain of *Consumer and Industrial Electronics* contributes only a minor translation improvement for this domain. Cuong et al. (2016a) show similar examples with English–Dutch.

As a side note, it is likely the case that dialect contributes to reordering behaviour, cf. Chen et al. (2013a) for Chinese, and Jeblee et al. (2014) for Egyptian Arabic. Domain adaptation with respect to this aspect (e.g. training lexicalized reordering models on different dialect bilingual training data) might, therefore, contribute reordering improvements.

4.3 Optimization

Domain mismatch between held-out development and test data is also an important source of errors. This is widely observed in many studies, e.g. Nikoulina et al. (2012), Pecina et al. (2012). In Table 2, we show a qualitative example. Specifically, we first train a phrase-based SMT system for English–German on a large dataset combined from multiple resources including EuroParl, Common Crawl Corpus and News Commentary. We then apply the system to a new domain of "*Legal Service*", but with three different scenarios for system optimization:

Table 2 Degradation in translation quality on a domain-specific translation task with different tuning scenarios

English–German (task: legal) Tuning scenario	BLEU↑
In-domain (Legal)	28.8
Mixed-domains (including legal)	28.5
Mixed-domains (exclude legal)	28.3

1. we optimize the system on an in-domain (*Legal*) held-out development set with $2K$ sentence pairs;
2. we optimize the system on a mixed-domain held-out development set with $8K$ sentence pairs from a combination of different domains: The in-domain *Legal* held-out development set itself, plus three different held-out development sets of *Software*, *Hardware* and *Professional & Business Services*;
3. we optimize the system on another mixed-domain held-out development set with $6K$ sentence pairs of *Software*, *Hardware* and *Professional & Business Services* in the third setting. This is the mixed-domain held-out development set in the second setting, but excludes the in-domain development set part.

Note that there is no prior knowledge about the domain's provenance of the mixed-domain held-out development set in the second and third settings.

Table 2 presents the translation performance of the phrase-based SMT system with respect to the different tuning scenarios. It can be seen quite clearly from the lower BLEU scores (Papineni et al. 2002) that moving to a new domain without having an in-domain held-out development set for system optimization can degrade the translation quality of a phrase-based SMT system. Note that our comparison may favour mixed-domain tuning scenarios: the mixed-domain held-out development sets are at least three times larger than the in-domain set, which presumably improves system optimization. In practice, the degradation in translation quality may be much more substantial, especially in a setting where the desired task is different from the held-out development set (e.g. Subtitles, Search Queries).

5 Domain adaptation: a general picture

A typical phrase-based SMT system contains various components, such as word alignment, language, translation and reordering models. This distinguishes SMT from most other Natural Language Processing tasks, and makes application of standard domain-adaptation methods less straightforward.

In general, the most popular approach to domain adaptation for SMT is to induce domain-focused translation statistics from seed in-domain data. Domain-focused translation statistics are typically domain-specific phrase translation probability distributions, lexical weighting and reordering probabilities. In the end, we can combine them together with the baseline 'domain-confused' translation features, or even replace the baseline features. This results in a statistical translation framework with a combination of multiple (sub-)models for translation. Figure 4 provides an illustration of the standard approach to domain adaptation for SMT.

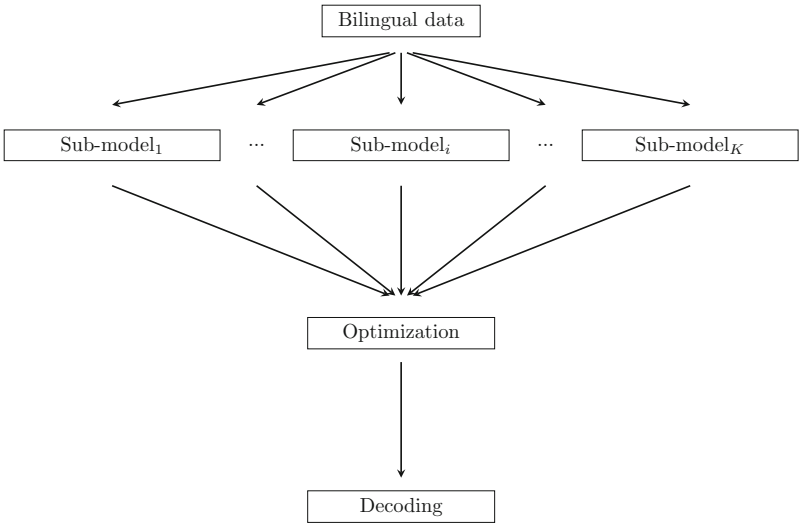

Fig. 4 Statistical translation framework with a combination of multiple K sub-models for translation

Implementing such a framework, however, is non-trivial. Two main technical challenges are as follows:

- The induction of domain-focused translation statistics: specific prior knowledge (e.g. in-domain bilingual corpora, comparable corpora, monolingual corpora) requires a different model for inducing domain-focused translation statistics. Section 6 provides a systematic overview of previous approaches to the problem.
- The combination of multiple (sub-)models for translation: the main object is a combination model tailored to high-dimensional feature spaces, which is surprisingly hard to achieve. Sect. 7 reviews different combination models for adaptation.

Beside the two main research lines, previous work also considers other adaptation scenarios. This survey covers several adaptation trends (Sect. 8). We first review the induction of domain-focused sparse features and word-alignment probabilities (Sects. 8.1, 8.2). We also show how an existing system can be adapted to multiple specific domains at the same time (Sect. 8.3). Another scenario is applying an SMT system to web search queries (Sect. 8.4). We also discuss how web-based translation services can be improved when domain of a new request is not known in advance (Sect. 8.5, 8.6).

6 Domain-specific translation induction for SMT

We start with induction using in-domain parallel data, and continue with comparable and monolingual corpora. We also discuss the induction with the domain's provenance, which is special in that we are provided with a large corpus consisting of different domain-specific subcorpora that are not necessarily strictly related to the desired task.

6.1 Induction with in-domain parallel data

In many studies, a seed in-domain parallel corpus ($\mathcal{C}_{IN}$) exemplifying the target translation task is used as a form of prior knowledge for domain adaptation for SMT. The data, however, is very small compared with a mixed of domains corpus $\mathcal{C}_{OUT}$. The main goal of translation induction with in-domain parallel corpora is inducing a phrase-based model from $\mathcal{C}_{OUT}$ for adaptation. We now review the two most popular approaches to domain adaptation in this scenario: instance weighting and data selection.

Instance weighting Instance weighting is perhaps the most effective approach to learning domain-focused translation statistics. To give some intuition about how instance weighting addresses the problem, in this general exposition we introduce a *latent domain variable* z to mark whether a phrase is in-domain (z_1) or out-of-domain (z_0). With the introduction of the latent variable, we expect to extend the translation tables in phrase-based models from domain-confused $P(\tilde{e}|\ \tilde{f})$ to domain-focused by conditioning them on z, i.e. $P(\tilde{e}|\ \tilde{f},\ z)$. Note how $P(\tilde{e}|\ \tilde{f},\ z)$ contains $P(\tilde{e}|\ \tilde{f})$ as a special case, as in (14):

$$P(\tilde{e}|\ \tilde{f},\ z) = \frac{P(\tilde{e}|\ \tilde{f})P(z|\ \tilde{e},\ \tilde{f})}{\sum_{\tilde{e}'} P(\tilde{e}'|\ \tilde{f})P(z|\ \tilde{e}',\ \tilde{f})}. \tag{14}$$

Here $P(z|\ \tilde{e},\ \tilde{f})$ is viewed as a latent phrase-relevance model, i.e. the probability that a phrase pair is in- (z_1) or out-of-domain (z_0). In the end, the adaptation can be performed by replacing the domain-confused tables $P(\tilde{e}|\ \tilde{f})$ with the in-domain-focused ones $P(\tilde{e}|\ \tilde{f},\ z_1)$, or simply by using these domain-focused models as additional features for the baseline phrase-based SMT system.

From Eq (14), the main challenge of inducing $P(\tilde{e}|\ \tilde{f},\ z)$ is inducing the latent phrase-relevance model $P(z|\ \tilde{e},\ \tilde{f})$. Following Matsoukas et al. (2009), a fairly large body of work on domain adaptation for SMT embeds $P(z|\ \tilde{e},\ \tilde{f})$ in an asymmetric sentence-level model $P(z|\ \mathbf{e},\ \mathbf{f})$ for sentence pairs $\langle \mathbf{e}, \mathbf{f} \rangle$. Specifically, the estimation of $P(z|\ \tilde{e},\ \tilde{f})$ for phrases $\tilde{e}$ and $\tilde{f}$ can be simplified by computing $P(z|\ \mathbf{e},\ \mathbf{f})$ for sentence pairs $\langle \mathbf{e}, \mathbf{f} \rangle$ as in (15):

$$P(z|\ \tilde{e},\ \tilde{f}) = \frac{\sum_{\mathbf{e},\ \mathbf{f}} P(z|\ \mathbf{e},\ \mathbf{f})\ c(\tilde{e};\ \mathbf{e})\ c(\tilde{f};\ \mathbf{f})}{\sum_{z' \in \{z_1,\ z_0\}} \sum_{\mathbf{e},\ \mathbf{f}} P(z'|\ \mathbf{e},\ \mathbf{f})\ c(\tilde{e};\ \mathbf{e})\ c(\tilde{f};\ \mathbf{f})}. \tag{15}$$

Here, $c(\tilde{e},\ \mathbf{e})$ and $c(\tilde{f},\ \mathbf{f})$ are the count of phrases $\tilde{e}$ and $\tilde{f}$ in sentence pairs $\langle \mathbf{e},\ \mathbf{f} \rangle$ in the training corpus.

But how can the asymmetric sentence level model be learned? A simple and straightforward way proposed by Cuong and Sima'an (2014a) is to devise an EM algorithm for learning (Fig. 5). At every iteration, in- or out-of-domain estimates provide full sentence pairs $\langle \mathbf{e},\ \mathbf{f} \rangle$ with probabilities $P(z|\ \mathbf{e},\ \mathbf{f})$. The latent phrase-relevance model parameters are then re-estimated using these expectations. Metaphorically, during each EM iteration the current in- or out-of-domain phrase pairs compete in *inviting* $\mathcal{C}_{OUT}$

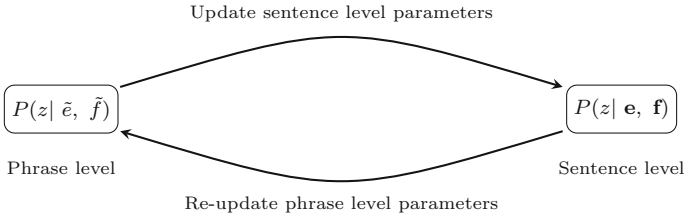

Fig. 5 The EM-based training algorithm for learning $P(z|\ \tilde{e},\ \tilde{f})$ and $P(z|\ \mathbf{e},\ \mathbf{f})$ simultaneously

sentence pairs to be in- or out-of-domain, which bring in new (weights for) in- and out-of-domain phrases.

Another approach is directly building a logistic weighting model for the asymmetric sentence-level model. Specifically, a logistic weighting model maps a set of features $\phi(\mathbf{e},\ \mathbf{f})$ with the parameter vector $\mathbf{w}$ to a scalar weight in $(0, 1)$. There are numerous types of sentence-level features that can be used, such as manual sub-corpus and genre membership, number of source and target token, and ratio of number of the tokens on both sides. Interestingly, the parameter vector $\mathbf{w}$ can be learned directly simultaneously with the log-linear model weight parameters so as to optimize the translation accuracy on a held-out development set. This approach was first proposed by Matsoukas et al. (2009).

An alternative approach to learning domain-focused translation statistics is directly building a discriminative model at phrase level. This approach is intuitively plausible, as a sentence itself may often contain a mixture of domains. In the work of Foster et al. (2010), the estimation of domain-focused phrase translation probabilities can be directly computed as in (16):

$$P(\tilde{e}|\ \tilde{f},\ IN) = \frac{c_{\mathbf{w}}(\tilde{e},\ \tilde{f})}{\sum_{\tilde{e}'} c_{\mathbf{w}}(\tilde{e}',\ \tilde{f})}, \tag{16}$$

where the modified count $c_{\mathbf{w}}(\tilde{e},\ \tilde{f})$ is computed as in (17):

$$c_{\mathbf{w}}(\tilde{e},\ \tilde{f}) = \frac{1}{1 + \exp(-\mathbf{w} \cdot \boldsymbol{\phi}(\tilde{e}, \tilde{f}))} c(\tilde{e},\ \tilde{f}). \tag{17}$$

Learning the weight parameters $\mathbf{w} = \{w_1, \ldots, w_K\}$ of K features for the logistic weighting model can be done using maximum likelihood or related criteria. More specifically, let us assume a held-out development set in which each sentence $\langle \mathbf{e},\ \mathbf{f} \rangle$ contains a (multi-)set $\mathcal{A}(\mathbf{e},\ \mathbf{f})$ of extracted phrases $\langle \tilde{e},\ \tilde{f} \rangle$. The objective function is the maximization of the likelihood over $\mathcal{A}(\mathbf{e},\ \mathbf{f})$ for all parallel sentences $\langle \mathbf{e},\ \mathbf{f} \rangle$ in the development set with respect to $\mathbf{w}$, as in (18):

$$\hat{\mathbf{w}} = \underset{\mathbf{w}}{\operatorname{argmax}} \sum_{\langle \mathbf{e},\ \mathbf{f} \rangle} \sum_{\langle \tilde{e},\ \tilde{f} \rangle \in \mathcal{A}(\mathbf{e},\mathbf{f})} \tilde{P}(\tilde{e}, \tilde{f}) \log P(\tilde{e}|\ \tilde{f},\ IN). \tag{18}$$

Here, note that $\tilde{P}(\tilde{e},\ \tilde{f})$ is computed from all phrase pairs extracted from the held-out development set. The optimization problem can be solved using the popular L-BFGS algorithm, as shown in Foster et al. (2010). The algorithm requires computing the gradient $\frac{\partial P(\tilde{e}|\ \tilde{f},\ IN)}{\partial w_i}$, which is done as in (19):

$$\frac{\partial P(\tilde{e}|\ \tilde{f},\ IN)}{\partial \lambda_i} = \frac{1}{P(\tilde{e}|\ \tilde{f},\ IN)} \left[\frac{c_{\mathbf{w}_i}(\tilde{e},\ \tilde{f})}{\sum_{\tilde{f}'} c_{\mathbf{w}}(\tilde{e},\ \tilde{f}')} - \frac{c_{\mathbf{w}}(\tilde{e},\ \tilde{f}) \sum_{\tilde{f}'} c_{\mathbf{w}_i}(\tilde{e},\ \tilde{f}')}{\left(\sum_{\tilde{f}'} c_{\mathbf{w}}(\tilde{e},\ \tilde{f}')\right)^2} \right]. \tag{19}$$

where:

$$c_{\mathbf{w}_i}(\tilde{e},\ \tilde{f}) = c_{\mathbf{w}}(\tilde{e},\ \tilde{f}) f_i(\tilde{e},\ \tilde{f}) \left(\frac{\exp(-\mathbf{w} \cdot \boldsymbol{\phi}(\tilde{e},\ \tilde{f}))}{1 + \exp(-\mathbf{w} \cdot \boldsymbol{\phi}(\tilde{e},\ \tilde{f}))} \right). \tag{20}$$

Both approaches have their own advantages and disadvantages. The EM-based approach strives for simplicity and is accordingly much easier to implement. However, using a discriminative model to learn relevance of sentence pairs and phrases in the parallel training data would perhaps be much more effective, but requires feature engineering, so is more difficult to implement. An empirical comparison of the approaches, however, has yet to be thoroughly conducted, to the best of our knowledge.

Note that using the same algorithm we can also adapt all other core translation components in tandem, including lexical weighting and lexicalized reordering models.

Data selection Another approach to learning domain-focused translation statistics is selecting training data from a large corpus. Then, we can simply train a phrase-based SMT system on the selected data. The resulting translation statistics are presumably domain-focused. Data selection would naturally be less effective than instance weighting, as we strictly remove a lot of bilingual data that are (presumably) not relevant to a desired task. However, data selection has received considerable attention in the past years for two main reasons:

1. Large bilingual training data comes with a cost: training phrase-based SMT systems on large data is extremely expensive and time-consuming;
2. A small, well-selected subset of the data often outperforms the full dataset for training a phrase-based SMT system (Axelrod et al. 2011; Biçici and Yuret 2011; Mansour et al. 2011; Duh et al. 2013; Cuong and Sima'an 2014b; Kirchhoff and Bilmes 2014; Mansour and Ney 2014; Zhang and Chiang 2014; Poncelas et al. 2017).

Existing work can be roughly classified depending on what kind of information is used for selection. The most popular approach (Axelrod et al. 2011) selects sentence pairs using the cross-entropy difference between in- and out-of-domain language models (both source and target sides), as in (21):

$$rank(\mathbf{f},\ \mathbf{e}) = \underbrace{\left(H_{LM_{IN}}(\mathbf{f}) - H_{LM_{OUT}}(\mathbf{f})\right)}_{source\ side} + \underbrace{\left(H_{LM_{IN}}(\mathbf{e}) - H_{LM_{OUT}}(\mathbf{e})\right)}_{target\ side}. \quad (21)$$

The cross-entropy is defined as in (22)–(23):

$$H_{LM}(\mathbf{f}) = -\frac{1}{m}\sum_{i=1}^{m} \log P(f_i|\ f_1^{i-1}) \quad (22)$$

$$H_{LM}(\mathbf{e}) = -\frac{1}{l}\sum_{i=1}^{l} \log P(e_i|\ e_1^{i-1}) \quad (23)$$

The method itself is a modification of the method proposed in Moore and Lewis (2010), which was introduced to address exactly the same problem we are discussing, but for only one side (i.e. monolingual data).

More recent approaches (Mansour et al. 2011; Cuong and Sima'an 2014b; Mansour and Ney 2014) use translation model information. The idea is intuitively plausible: in the translation context, a source phrase often has different translations in different domains, which cannot be distinguished with monolingual language models. But how much should data selection depend on bilingual vs. monolingual factors? Cuong and Sima'an (2014b) present a comprehensive study of the contribution of these factors, showing that they actually complement each other for data selection.

One of the most difficult problems in data selection is to jointly learn translation and language models. An EM-based learning algorithm was first proposed by Cuong and Sima'an (2014b) to address the problem. However, a joint bilingual neural network model proposed by Devlin et al. (2014) might be a more powerful solution to the problem. Chen et al. (2016) were the first to deploy the joint bilingual neural network model to address the problem. In their work, promising data-selection performance is observed.

As a side note, data selection often goes hand in hand with data reduction for SMT (Eck et al. 2005; Lewis and Eetemadi 2013). Data reduction aims at reducing the size of data that is used for training, while at the same time impacting very little on quality.

6.2 Induction with comparable corpora

Creating an in-domain dataset is extremely expensive in practice. A cheaper approach to domain adaptation for SMT is mining comparable corpora (Snover et al. 2008; Daumé and Jagarlamudi 2011; Irvine et al. 2013b).

We now present two notable approaches as examples. The first approach is mining unseen words for an adaptation task (Daumé and Jagarlamudi 2011), which extends the approach described in Haghighi et al. (2008) to mining translations from comparable corpora. Learning bilingual lexicons from comparable corpora is obviously not an easy task [cf. Koehn and Knight (2002), Haghighi et al. (2008), Tamura et al. (2012)], and their mining is "bootstrapped" based on a bilingual dictionary that is created automatically from out-of-domain corpora. The output of the dictionary-mining approach is normally a list of (source and target) word pairs, with corresponding scores represent-

ing the word-translation probability. Perhaps surprisingly, a straightforward approach to incorporating the induced word pairs by having an additional feature representing dictionary-mining translation probability may not be helpful. A more effective way, as described in Daumé and Jagarlamudi (2011), is to have not only the dictionary-mining translation-probability feature, but also an additional feature to mark whether a phrase pair is seen in the source and target data or not.

The second approach, proposed by Irvine et al. (2013b), directly recovers the joint probability distribution of source and target word pairs on a new domain. Specifically, assume we have access to a joint distribution $P_{OUT}(f,\ e)$ over source and target word pairs $\langle f,\ e\rangle$, estimated from an out-of-domain corpus. Let $\tilde{P}(f)$ and $\tilde{P}(e)$ be the empirical marginal distributions estimated from comparable corpora (i.e. we extract raw word frequencies from the corpora). Irvine et al. (2013b) cast the learning of the joint probability distribution of source and target word pairs on a new domain as a linear programming problem, as in (24):

$$\hat{P}_{IN} = \underset{P_{IN}}{\operatorname{argmin}} \| \sum_{\langle f,\ e\rangle} P_{IN}(f,e)\ -\ P_{OUT}(f,e)\|_1, \tag{24}$$

subject to:

$$\sum_{\langle f,e\rangle} P_{IN}(f,e) = 1, \sum_{e} P_{IN}(f,e) = \tilde{P}(f),$$
$$\sum_{f} P_{IN}(f) = \tilde{P}(e),\ \text{and}\ P_{IN}(f,e) \geq 0.$$

Here, l_1-norm ($\|\cdot\|_1$) is used to measure the distance between two distributions. Regularization terms are usually added into Eq. (24) so that the solution would be as sparse as possible. A linear programming solver can be used to learn $P_{IN}(f,\ e)$ from Eq. (24).

The method is perhaps one of the most elegant approaches to domain adaptation for SMT. It exploits cheap resources and shows significant improvement in translation quality on new domains.

6.3 Induction with monolingual data

Exploiting in-domain monolingual data is also an effective approach to domain adaptation for SMT. In general, synthetic bilingual data is first generated by using a phrase-based SMT system. Then, we can use the created data to induce domain-focused translation statistics (Schwenk 2008; Wu et al. 2008; Bertoldi and Federico 2009; Schwenk and Senellart 2009). Empirical results show that having in-domain monolingual data could substantially improve translation quality for a new domain, especially with in-domain monolingual data on the target side (Lambert et al. 2011).

Surprisingly, we can still derive improvements from incorporating induced domain-focused translation features to the baseline, given that the baseline is already

augmented with induced domain-focused language-model features. As a side note, the adaptation of reordering model gives consistent but modest improvements in this scenario (Schwenk 2008; Bertoldi and Federico 2009).

6.4 Induction with monolingual data and meta-information

Beside generating synthetic bilingual data, are there any other ways of adapting translation models with monolingual corpora? There has been an intensive line of research that focuses on translation-model adaptation using topic models (Gong et al. 2011; Eidelman et al. 2012; Su et al. 2012; Hewavitharana et al. 2013; Hasler et al. 2014; Hu et al. 2014). Such studies interchangeably use the term "topic" and "domain".

Assume we are provided with an out-of-domain parallel corpus $\mathcal{C}_{OUT} = \{\mathcal{S}_{OUT}, \mathcal{T}_{OUT}\}$, together with an in-domain monolingual corpus on the source side $\mathcal{S}_{IN}$ only. Given the data, a general approach is building an adapted translation model in the following steps:

- Step 1: Estimating topic models (e.g. Probabilistic Latent Semantic Analysis (Hofmann 1999), Latent Dirichlet Allocation (Blei et al. 2003), or Hidden Topic Markov Models (Gruber et al. 2007)) at document level in monolingual corpora;
- Step 2: Estimating topic-specific translation models (i.e. conditioning the translation of phrase pairs on the topic information of source phrases);
- Step 3: Estimating topic posterior distributions of phrases;
- Step 4: Estimating phrase-translation probabilities using predefined topic-specific translation models and topic posterior distributions of phrases.

More formally, let us use $P(z_{\mathbf{f}_{IN}}|\,\mathbf{f})$ and $P(z_{\mathbf{f}_{OUT}}|\,\mathbf{f})$ to indicate how a sentence $\mathbf{f}$ expresses a specific source-side topic in in- and out-of-domain monolingual corpora. The sentence-topic distributions are provided by topic models (Step 1).

Let us use $P(\tilde{e}|\,\tilde{f}, z_{\tilde{f}_{OUT}})$ to indicate the probability of translating a phrase $\tilde{f}$ as a phrase $\tilde{e}$ given the source-side topic $z_{\tilde{f}_{OUT}}$. The topic-specific translation models are estimated as in (25) (Step 2):

$$P(\tilde{e}|\,\tilde{f},\, z_{\tilde{f}_{OUT}}) = \frac{\sum_{\mathbf{e},\,\mathbf{f}\,\in\,\mathcal{C}_{OUT}} P(z_{\tilde{f}_{OUT}}|\,\mathbf{f})\; c(\tilde{f};\;\mathbf{f})\; c(\tilde{e};\;\mathbf{e})}{\sum_{\tilde{e}'}\sum_{\mathbf{e},\,\mathbf{f}\,\in\,\mathcal{C}_{OUT}} P(z_{\tilde{f}_{OUT}}|\,\mathbf{f})\; c(\tilde{f};\;\mathbf{f})\; c(\tilde{e}';\;\mathbf{e})}. \tag{25}$$

Let us use $P(z_{\tilde{f}_{IN}}|\,\tilde{f})$ and $P(z_{\tilde{f}_{OUT}}|\,\tilde{f})$ to denote the phrase-topic distributions. The distributions can be computed as in (26)–(27) (Step 3):[10]

$$P(z_{\tilde{f}_{IN}}|\,\tilde{f}) = \frac{\sum_{\mathbf{f}\,\in\,\mathcal{S}_{IN}} P(z_{\tilde{f}_{IN}}|\,\mathbf{f})\; c(\tilde{f};\;\mathbf{f})}{\sum_{z'_{\tilde{f}_{IN}}}\sum_{\mathbf{f}\,\in\,\mathcal{S}_{IN}} P(z'_{\tilde{f}_{OUT}}|\,\mathbf{f})\; c(\tilde{f};\;\mathbf{f})}. \tag{26}$$

[10] In Su et al. (2012), an interpolation model is computed for $P_{IN}(z_{\tilde{f}_{IN}}|\,\tilde{f})$, which is decomposed into the topic posterior distribution at word level for smoothing.

$$P(z_{\tilde{f}_{OUT}} | \tilde{f}) = \frac{\sum_{\mathbf{f} \in \mathcal{S}_{OUT}} P(z_{\tilde{f}_{OUT}} | \mathbf{f})\, c(\tilde{f};\, \mathbf{f})}{\sum_{z'_{\tilde{f}_{OUT}}} \sum_{\mathbf{f} \in \mathcal{C}_{OUT}} P(z'_{\tilde{f}_{OUT}} | \mathbf{f})\, c(\tilde{f};\, \mathbf{f})}. \quad (27)$$

Finally, phrase-translation probabilities can be computed as in (28) (Step 4):

$$P(\tilde{e} | \tilde{f}) = \sum_{z_{\tilde{f}_{IN}}} \sum_{z_{\tilde{f}_{OUT}}} P(\tilde{e} | \tilde{f},\, z_{\tilde{f}_{OUT}}) P(z_{\tilde{f}_{OUT}} | z_{\tilde{f}_{IN}}) P(z_{\tilde{f}_{IN}} | \tilde{f}), \quad (28)$$

where the topic-mapping probability distribution $P(z_{\tilde{f}_{OUT}} | z_{\tilde{f}_{IN}})$ can be computed as in (29):[11]

$$P(z_{\tilde{f}_{OUT}} | z_{\tilde{f}_{IN}}) = \sum_{\tilde{f} \in \mathcal{S}_{IN} \cap \mathcal{S}_{OUT}} P_{IN}(z_{\tilde{f}_{IN}} | \tilde{f}) P_{OUT}(z_{\tilde{f}_{OUT}} | \tilde{f}). \quad (29)$$

The estimate of $P(\tilde{e} | \tilde{f})$ as in Eq. (28) can be used to replace the domain-confused translation probability. It can also simply serve as an additional feature to the baseline.

In practice, it is also possible that instead of having only the source side $\mathcal{S}_{IN}$ monolingual data, we are provided with an in-domain parallel corpus $\mathcal{C}_{IN} = \{\mathcal{S}_{IN}, \mathcal{T}_{IN}\}$. In that case, bilingual topic inference should be preferred to monolingual topic inference (Mimno et al. 2009; Hasler et al. 2014; Hu et al. 2014).

Using topic models for domain adaptation for SMT provides an effective way of quantifying the effect of the topical context information on translation selection. Using the same approach, we can adapt all other core translation components in tandem, including lexical weighting and lexicalized reordering models.

Meanwhile, the model has a potential drawback: most parallel corpora lack the annotation of document boundaries. Of course, a single sentence can be considered as a short pseudo-document, but it is questionable whether such a corpus with short pseudo-documents is topic-model 'friendly' (Tang et al. 2014).

6.5 Induction using a domain's provenance

In practice, there are adaptation scenarios where we are provided with a large corpus consisting of different domain-specific subcorpora, where the subcorpora are manually grouped/annotated, but not necessarily strictly related to the desired task. In that scenario, it is still very useful to condition the lexical weighting features on provenance (Chiang et al. 2011). In the end, we can simply optimize the system with different types of domain-focused translation statistics on an in-domain held-out development set.

Another simple and elegant approach is to use a vector space model. Specifically, let us assume we are provided with a corpus consisting of N different domain-specific subcorpora. First, we create a vector profile for every phrase pair extracted from the training data, as in (30):

[11] Joint inference of topic models on a concatenation of $\mathcal{S}_{IN}$ and $\mathcal{S}_{OUT}$ would drop the requirement of computing the topic-mapping probability distribution [cf. Gong et al. (2011) and Hewavitharana et al. (2013)]. An empirical comparison of the approaches, however, has yet to be thoroughly conducted, to the best of our knowledge.

$$V_{training}(\tilde{f},\ \tilde{e}) = \left[w_1(\tilde{f},\ \tilde{e}), \ldots, w_N(\tilde{f},\ \tilde{e}) \right] \tag{30}$$

Another vector profile is created for every phrase pair extracted from the in-domain held-out development set, as in (31):

$$V_{dev}(\tilde{f},\ \tilde{e}) = \left[w_1(\tilde{f},\ \tilde{e}), \ldots, w_N(\tilde{f},\ \tilde{e}) \right] \tag{31}$$

In principle, each element of the vector $w(\tilde{f},\ \tilde{e})$ can simply be the count of a phrase pair. A better approach proposed by Chen et al. (2013b) is adapting standard *tf-idf* statistics, a standard technique in IR.

Then, we simply use the similarity score between these two types of vectors as additional feature functions (e.g. the Bhattacharyya distance (Bhattacharyya 1946), the Kullback-Leibler distance (Kullback and Leibler 1951), and the cosine distance), which reward phrase pairs that are relevant to the desired task.

The vector space model approach was first proposed by Chen et al. (2013b), and is a very effective adaptation technique for SMT. However, a domain's provenance is not always available in practice. Despite the fact that topic models can automatically provide meta-information, experiments in this setting show only a modest improvement [cf. Hewavitharana et al. (2013)].

7 Model combination for adaptation

Domain-focused translation statistics, once induced, need to be combined together in an appropriate way. The main desire is to have a combination model tailored to high-dimensional feature spaces.

7.1 Log-linear mixture

Log-linear translation model mixtures (Birch et al. 2007; Koehn and Schroeder 2007) are of the form in (32):

$$\begin{aligned} \phi_{TM}(\mathbf{e},\ \mathbf{f}) = \lambda \sum\nolimits_{i=1}^{n} \log P(\tilde{e}_i |\ \tilde{f}_{a_i}, IN) \\ + (1-\lambda) \sum\nolimits_{i=1}^{n} \log P(\tilde{e}_i |\ \tilde{f}_{a_i},\ OUT). \end{aligned} \tag{32}$$

Here, $P(\tilde{e}_i |\ \tilde{f}_{a_i}, IN)$ and $P(\tilde{e}_i |\ \tilde{f}_{a_i}, IN)$ represent different types of domain-focused translation statistics with respect to IN and OUT. As in Eq. (32), they can be added to the baseline as additional features. There is also no further effort needed for training: the respective weights are set with any weight optimization method (e.g. MERT, MIRA, PRO).

The implementation of a log-linear translation mixture model for adaptation can be slightly different in practice. It is common to leave the decoder as is (Razmara

et al. 2012), but it is also possible to put constraints on hypotheses generated by the decoder (Birch et al. 2007; Koehn and Schroeder 2007). For instance, the decoder may only generate hypotheses that are contained in both in-domain and out-of-domain translation tables. The decoder may also generate hypotheses that are contained in each of the tables. An empirical comparison of the implementations, however, has yet to be thoroughly conducted, to the best of our knowledge.

This model has two potential drawbacks:

1. In practice, it is common to have many sub-models, which leads to significantly longer search and potentially more search errors. This also makes system optimization even more challenging. It is not uncommon for such a log-linear mixture model to perform significantly worse than a system trained on a concatenation of all the data (Sennrich 2012; Wäschle and Riezler 2012);
2. Having high-dimensional feature spaces requires a much larger held-out development set for system optimization (Waite and Byrne 2015). This is unrealistic in practice, as in-domain data is very expensive to annotate.

7.2 Linear mixture

Linear translation model mixtures are of the form in (33):

$$\phi_{TM}(\mathbf{e},\mathbf{f}) = \sum_{i=1}^{n} \log\left(\lambda P(\tilde{e}_i|\ \tilde{f}_{a_i},\ IN) + (1-\lambda)P(\tilde{e}_i|\ \tilde{f}_{a_i},\ OUT)\right) \tag{33}$$

An alternative form of linear combination is a maximum a posteriori (MAP) combination, as in (34):

$$\phi_{TM}(\mathbf{e},\mathbf{f}) = \sum_{i=1}^{n} \log\left(\frac{c_{IN}(\tilde{e}_i,\ \tilde{f}_{a_i}) + \lambda P(\tilde{e}_i|\ \tilde{f}_{a_i}, OUT)}{\sum_{\tilde{e}'} c_{IN}(\tilde{e}',\ \tilde{f}_{a_i}) + \lambda}\right) \tag{34}$$

This model was first proposed by Foster and Kuhn (2007), but training the model is not straightforward. It is desirable to directly optimize the weights of the baseline system $\mathbf{w} = \{w_1, \ldots, w_M\}$ and interpolation weight λ directly for BLEU. This is possible (Foster et al. 2013; Haddow 2013), but very challenging to implement.[12] In practice, the most common approach is performing system optimization with a two-step procedure as follows:

– First, we learn the interpolation weight by maximum likelihood or related criteria;
– We hold the interpolation weight as constant, and optimize the log-linear weights as normal with any optimization method.

By isolating the task of learning log-linear weights, the problem of learning the interpolation weight is not hard (Foster et al. 2010; Sennrich 2012). Specifically, let us

[12] There has not been any attempt at such an implementation for combining multiple sub-models, as far as we are aware.

assume a held-out development set, in which each sentence pair $\langle \mathbf{e}, \mathbf{f} \rangle$ contains a (multi-)set $\mathcal{A}(\mathbf{e}, \mathbf{f})$ of extracted phrases $\langle \tilde{e}, \tilde{f} \rangle$. The objective function is the maximization of the likelihood over $\mathcal{A}(\mathbf{e}, \mathbf{f})$ for all pairs $\langle \mathbf{e}, \mathbf{f} \rangle$ with respect to λ, as in (35):

$$\hat{\lambda} = \underset{\lambda}{\operatorname{argmax}} \sum_{\langle \mathbf{e}, \mathbf{f} \rangle} \sum_{\langle \tilde{e}, \tilde{f} \rangle \in \mathcal{A}(\mathbf{e}, \mathbf{f})} \tilde{P}(\tilde{e}, \tilde{f}) \log \Big(\lambda P(\tilde{e} | \tilde{f}, IN) + (1 - \lambda) P(\tilde{e} | \tilde{f}, OUT) \Big) \tag{35}$$

If we are using MAP, the objective function of training is as in (35):

$$\hat{\lambda} = \underset{\lambda}{\operatorname{argmax}} \sum_{\langle \mathbf{e}, \mathbf{f} \rangle} \sum_{\langle \tilde{e}, \tilde{f} \rangle \in \mathcal{A}(\mathbf{e}, \mathbf{f})} \tilde{P}(\tilde{e}, \tilde{f}) \log \frac{c_{IN}(\tilde{e}, \tilde{f}) + \lambda P(\tilde{e} | \tilde{f}, OUT)}{\sum_{\tilde{e}'} c_{IN}(\tilde{e}, \tilde{f}) + \lambda} \tag{36}$$

Note that $\tilde{P}(\tilde{e}, \tilde{f})$ in both cases is computed from all phrase pairs extracted from the held-out development set.

Since the objective function is convex, the optimization can be done efficiently with EM (Carpuat et al. 2014) or Limited-memory BFGS algorithm (Sennrich 2012).[13] Both algorithms require computing the gradient $\frac{\partial}{\partial \lambda}$. The gradient is easy to compute in the first case, as in (37):

$$\frac{\partial}{\partial \lambda} = \left[\frac{P(\tilde{e} | \tilde{f}, IN) - P(\tilde{e} | \tilde{f}, OUT)}{\lambda P(\tilde{e} | \tilde{f}, IN) + (1 - \lambda) P(\tilde{e} | \tilde{f}, OUT)} \right] \tag{37}$$

If we are using MAP, the gradient is slightly different, as in (38):

$$\frac{\partial}{\partial \lambda} = \frac{-\sum_{\tilde{e}'} c_{IN}(\tilde{e}', \tilde{f})}{\left(\sum_{\tilde{e}'} c_{IN}(\tilde{e}', \tilde{f}) + \lambda \right)^2} \left[\frac{P(\tilde{e} | \tilde{f}, IN) - P(\tilde{e} | \tilde{f}, OUT)}{\frac{c_{IN}(\tilde{e}, \tilde{f}) + \lambda \tilde{P}(\tilde{e} | \tilde{f}, OUT)}{\sum_{\tilde{e}'} c_{IN}(\tilde{e}', \tilde{f}) + \lambda}} \right] \tag{38}$$

A linear translation model is perhaps the most common combination model for adaptation. Compared with the log-linear translation model, it often works better with high-dimensional feature spaces. However, the model has two potential drawbacks:

1. The maximum likelihood or related criteria may not correlate well with translation accuracy. It is not uncommon that assigning optimized weights underperforms compared to uniform weights;
2. The performance would likely suffer from combining too many (e.g. more than 10) sub-models, leaving an open question of how best to design a combination model tailored to very high-dimensional feature spaces.

[13] The EM algorithm often gives a more efficient and stable performance in practice [cf. Razmara et al. (2012)].

7.3 Fill-up

A very simple approach that provides a competitive performance to log-linear and linear translation model mixtures is Fill-up. The idea of Fill-up was first proposed by Besling and Meier (1995) for addressing the problem of language model adaptation for speech recognition. It was first introduced in SMT by Nakov (2008), and first used in domain adaptation for SMT in the work of Bisazza et al. (2011).

Let us assume we have two translation tables T_{IN} and T_{OUT}, with their corresponding phrase translation probabilities $P(\tilde{e}|\ \tilde{f},\ IN)$ and $P(\tilde{e}|\ \tilde{f},\ OUT)$, respectively. A Fill-up table T_{FILLUP} is defined as in (39):

$$\forall(\tilde{f},\tilde{e}) \in T_{IN}\ \cup\ T_{OUT}:$$
$$T_{FILLUP}(\tilde{f},\ \tilde{e}) = \begin{cases} \{P(\tilde{e}|\ \tilde{f},\ IN), \exp(0)\} & \text{if } (\tilde{f},\ \tilde{e}) \in T_{IN} \\ \{P(\tilde{e}|\ \tilde{f},\ OUT), \exp(1)\} & \text{otherwise.} \end{cases} \tag{39}$$

Here, the entries of T_{FILLUP} correspond to the union of the two phrase tables, in which we consider T_{IN} as the more reliable source and use it whenever possible. The exponential function (i.e. exp(0) and exp(1)) is to mark whether a phrase pair is in-domain (T_{IN}) or out-of-domain (T_{OUT}).[14]

Simplicity is perhaps the main advantage of Fill-up. The model, however, has two potential drawbacks:

- It remains unclear whether the approach is able to scale to many sub-models. Such an empirical evaluation has yet to be thoroughly conducted, to the best of our knowledge.
- Translation probabilities in T_{FILLUP} do not form a full probability distribution. This is potentially problematic: interactions between features can be complex and log-linear models may not be able to handle the interactions.

8 Other trends in domain adaptation

This survey covers several other adaptation trends. We first review the induction of domain-focused sparse features and word-alignment probabilities (Sects. 8.1, 8.2). We also show how an existing system can be adapted to multiple specific domains at the same time (Sect. 8.3). Another scenario is applying an SMT system to web search queries (Sect. 8.4). We also discuss how web-based translation services can be improved when domain of a new request is not known in advance (Sects. 8.5, 8.6).

[14] Zhang et al. (2014a) improve upon the binary fill-up model of Bisazza et al. (2011) with a probability distribution over phrase pairs to signify the extent to which a phrase pair is considered in-domain or out-of-domain.

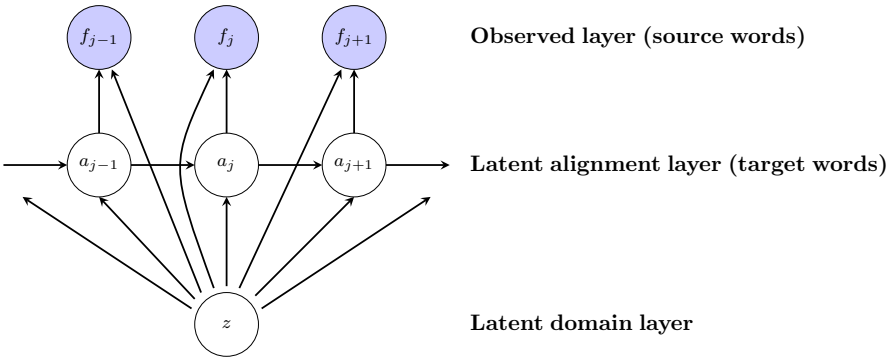

Fig. 6 Latent domain HMM alignment model. An additional latent layer representing domains has been conditioned on by both the remaining layers

8.1 Adaptation with sparse features

Having in-domain sparse feature functions is particularly useful when applying a phrase-based SMT system to new domains (Bertoldi and Federico 2009; Hasler et al. 2012; Green et al. 2013, 2014). This is because sparse features allow for more flexibility than dense features, but at the risk of increasing the difficulty of the optimization. Applying cross-validation techniques (e.g. jackknife training (Hasler et al. 2012)) is often very useful to avoid overfitting.

8.2 Domain adaptation for word alignment

There is some evidence to support the claim that like any statistical models, word-alignment models suffer significantly from a lack of in-domain data for training. Wu et al. (2005) train different alignment models independently on different domain-specific subcorpora. In the end, they show that an interpolation of the alignment models improves word-alignment accuracy.

Similar findings are reported in Duh et al. (2010) and Gao et al. (2011). Duh et al suggest that training a phrase-based SMT system might benefit from using the following simple trick: they first train statistical alignment models on a concatenation of both in-domain and a much larger out-of-domain dataset, and then exclude out-of-domain data during phrase extraction. Gao at al show that an interpolation of domain-specific and general-domain alignment models improves translation accuracy.

As a side note, Shah et al. (2010) show that weighting sentence pairs according to their relevance to a new domain benefits word-alignment training.

Recently, Cuong and Sima'an (2015) provide an in-depth study of domain adaptation for word alignment. They focus on the insensitivity of existing word-alignment models to domain differences, which often yields suboptimal results on heterogeneous corpora (e.g. EuroParl, Common Crawl Corpus, UN Corpus, and News Commentary). A latent domain word-alignment model is proposed, which explicitly incorporates latent domain information in learning domain-focused lexical and alignment statis-

tics. Figure 6 presents such a case with a latent domain HMM alignment model. Cuong and Sima'an (2015) train the model on a heterogeneous corpus, using a small number of seed samples from different domains. Their experiments show that the derived domain-focused statistics, once combined together, produce significant improvements both in word alignment and translation.

8.3 Multi-domain adaptation for SMT

A common scenario in practice is adapting an existing system to multiple domain-specific tasks at the same time, which is clearly a challenging problem.

8.3.1 Adaptation with multi-task learning

The main approach is to optimize an SMT system in the way that exploits commonalities shared among different tasks (Wäschle and Riezler 2012). More formally, let us use $\{\hat{\mathbf{w}}_1, \ldots, \hat{\mathbf{w}}_K\}$ to denote a set of model parameters with respect to K different domains. The commonalities shared among different tasks are modeled as in (40):

$$\mathbf{w}_{AVG} = \frac{1}{K} \sum_{d=1}^{K} \mathbf{w}_d. \tag{40}$$

In the end, the goal is to learn model parameters that maximize the objective function, as in (41):

$$\begin{aligned} \{\hat{\mathbf{w}}_1, \ldots, \hat{\mathbf{w}}_K\} = \underset{\mathbf{w}_1,\ldots,\mathbf{w}_K}{\operatorname{argmin}} & \sum_{d=1}^{K} loss_d(\mathbf{w}_d) \\ & + \lambda \sum_{d=1}^{K} \|\mathbf{w}_d - \mathbf{w}_{AVG}\|_1. \end{aligned} \tag{41}$$

Here, the parameter λ controls the influence of the regularization, which trades off between task-specific parameter vectors and their distance to the average. Meanwhile, we use $loss_d(\mathbf{w}_d)$ to represent a translation loss function on the held-out development set from task d. The optimization problem can be solved using gradient-descent optimization with l_1-regularization (Tsuruoka et al. 2009; Wäschle and Riezler 2012).

While this method is intuitively plausible, it gives only a modest translation improvement (Wäschle and Riezler 2012). Different variants are proposed in the literature (Simianer et al. 2012; Cui et al. 2013) which show a potentially more promising performance.

8.3.2 Adaptation with genre-aware decoding

Another interesting approach to multi-domain adaptation for SMT is using a genre-aware classifier (Wang et al. 2012). The core to this approach is a source-sentence genre classifier that signals the most relevant domain to source sentences. In this way, the MT system is configured to use the proper domain feature weights and appropriate domain language model. Note that in the work of Wang et al. (2012), their system

uses a single translation model to serve different domains. This allows the system to scale more easily to many domains, but makes tuning and decoding more difficult. Wang et al. (2012) introduce simple genre-aware decoding and tuning techniques to address the problem. Their experiments show that the proposed system is capable of producing better domain-specific translations while simultaneously preserving the quality of general-domain translations.

8.4 Cross-lingual information retrieval

A practical real-world problem is translating web search queries into several target languages, so that a search engine can return search results in the corresponding languages. The quality of a translation component thus plays a crucial role. The problem, however, is particularly difficult for three specific reasons:

1. Translation quality degrades substantially when applying a generic phrase-based SMT system to a domain-specific task. This is particularly true for search queries, due to their unique characteristics: search queries are very short (just a couple of words per query) and the word order is typically different to a typical sentence in natural language.
2. Second, a phrase-based SMT system is usually trained to optimize the quality of the translation, which is not necessarily correlated with the retrieval quality (especially for the short queries) (Kettunen 2009; Nikoulina et al. 2012). For example, the word order which is crucial for translation quality is often ignored by IR models. In contrast, retrieval systems often use bag-of-word representations in document-scoring models, and queries are rarely grammatical natural language sentences.
3. Finally, there are only a few tiny corpora of parallel queries (e.g. CLEF tracks) that can be obtained.

A very simple, yet effective approach to improving adaptation for CLIR is reranking the N-best translation candidates generated by a baseline system (Nikoulina et al. 2012). Note that a re-ranker should be optimized to maximize a retrieval metric rather than translation accuracy. Putting constraints on hypotheses generated by the decoder is another approach to improving adaptation for CLIR (Dong et al. 2014; Hieber and Riezler 2015). While the latter approach may be more efficient, such an implementation is obviously far more complicated.

8.5 Cache-based adaptive models for translation adaptation

A common scenario in practice, particularly for web-based translation services such as Bing Translator and Google Translate, is that translation requests are unknown as to their domain. A common approach is to exploit two general phenomena in natural language and translation:

1. Repetition and recency effects of words: many words, especially content words, are repeated in close context;
2. Consistency of translations: the translation of content words is consistent given a specific context.

The two phenomena provide us with a natural way to perform fully unsupervised domain adaptation on a new domain: a phrase-based SMT system performs the translation of a sentence by not only considering the sentence itself, but also taking the translation history of recent input sentences into account.

Accounting for these phenomena in translation is fairly simple, using a cache-based adaptive model (Kuhn and De Mori 1992). More specifically, Tiedemann (2010) develops two cache-based adaptive models, that we now describe.

Cache-based adaptive language model Tiedemann (2010) uses a dynamic cache-based adaptive language model in the form of a linear mixture as in (42):

$$\begin{aligned} P(e_n|\, e_{n-k},\, \ldots,\, e_{n-1}) = (1\, -\, \lambda) P(e_n|\, e_{n-k},\, \ldots,\, e_{n-1},\, OUT) \\ +\, \lambda P(e_n|\, e_{n-k},\, \ldots,\, e_{n-1},\, CACHE) \end{aligned} \tag{42}$$

Here, the cache stores the best translation hypotheses of previous sentences. Of course the size of the cache is very small (e.g. 100-5000 words). The value of the interpolation weight λ can be set manually. The EM algorithm can also be used to learn the weight automatically.

Implementing the model as a simple unigram model is a good option, but a better solution in practice would be introducing a decay factor in the estimation of cache probabilities, as in (43):

$$P(e_n|\, e_{n-k},\, \ldots,\, e_{n-1},\, CACHE) \propto \sum_{i=n-k}^{n-1} \delta(e_n\, =\, e_i) \exp\Big(-\alpha(n-i)\Big) \tag{43}$$

This approach was first introduced by Clarkson and Robinson (1997). Here, δ is the Kronecker delta function. The decay rate α is normally set to a very small value [e.g. 0.005 as in Clarkson and Robinson (1997)].

Cache-based adaptive translation model Tiedemann (2010) develops a cache-based adaptive translation model in a similar manner, using a decay factor to compute translation model scores from the cache, as in (44):

$$P(\tilde{e}_n|\, \tilde{f}_n,\, CACHE) \propto \sum_{i=1}^{K} \delta\Big(\langle \tilde{e}_n,\, \tilde{f}_n\rangle\, =\, \langle \tilde{e}_i,\, \tilde{f}_i\rangle\Big) \exp\Big(-\alpha i\Big) \tag{44}$$

The cache-based adaptive models can be integrated into a phrase-based SMT system in a straightforward manner: both can be used to replace the language and translation models, or to serve as additional feature functions within a log-linear model. In the end, the decoder is forced to prefer identical translations for repeated terms.

While using cache-based adaptive models is an elegant approach, Tiedemann observes that the adaptation effect is rather modest. Nor is it terribly robust; it is not uncommon that an augmented SMT system produces a rather suboptimal translation. There are two potential reasons for this:

– First, it would be risky to assume that previous translation hypotheses are good enough to be cached [cf. the risk of error propagation (Tiedemann 2010)].

– Second, using the translation of initial sentences in the input stream may not be so beneficial.

Potential solutions to these problems are quite straightforward (Gong et al. 2011; Louis and Webber 2014). For instance, in the work of Gong et al. (2011), the cache stores similar target sentence pairs in the bilingual training data to the translation hypotheses, instead of the translation hypotheses by themselves. As a side note, other types of cache can be developed to improve adaptation, e.g. caching not only phrase pairs but also topic caches, as in Gong et al. (2011).

8.6 Rewarding domain invariance for adaptation

When the target domain is unknown at training time, the system could also be trained to make safer choices, preferring translations which are likely to work across different domains. For exampleAs we pointed out early on, when translating from English to Russian, the most natural translation for the word 'code' would be highly dependent on the domain (and the corresponding word sense). Russian words 'шифр' ('cipher'), 'закон' ('law') or 'программа' ('program') would perhaps be optimal choices if we consider cryptography, legal and software development domains, respectively. However, the translation 'код' ('code') is also acceptable across all these domains and, as such, would be a safer choice when the target domain is unknown. Note that such a translation may not be the most frequent overall and, consequently, might not be proposed by a standard (i.e. domain-agnostic) phrase-based translation system.

In order to encode a preference for domain-invariant translations, we can first *project* phrases onto a compact $(K-1)$ dimensional simplex of subdomains with vectors, as in (45)–(46):

$$\tilde{\mathbf{e}} = \left[P(z=1|\,\tilde{e}),\ \ldots,\ P(z=K|\,\tilde{e}) \right] \tag{45}$$

$$\tilde{\mathbf{f}} = \left[P(z=1|\,\tilde{f}),\ \ldots,\ P(z=K|\,\tilde{f}) \right]. \tag{46}$$

See Fig. 7 for an illustration of the projection framework.

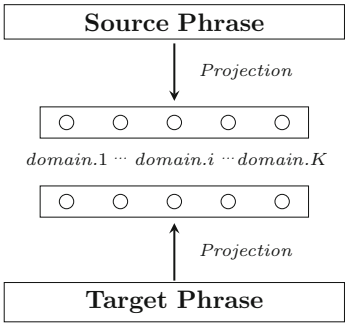

Fig. 7 The projection framework of phrases into a K-dimensional vector space of probabilistic latent subdomains

Of course, the subdomains are usually not specified in the heterogeneous training data. We can treat the subdomains as latent, and induce them automatically (Cuong et al. 2016b). In the end, we can use a relevant measure to quantify how likely a phrase (or a phrase-pair) is to be "domain-invariant", for instance:

– *Domain-specificity of phrases* A rule with source and target phrases having a *peaked* distribution over latent subdomains is likely domain-specific. Technically speaking, entropy is a natural choice for quantifying domain specificity. Here, we opt for the Renyi entropy and define the domain specificity as in (47)–(48):

$$D_\alpha(\tilde{\mathbf{e}}) = \frac{1}{1-\alpha} \log \left(\sum\nolimits_{i=1}^{K} P(z=i|\,\tilde{e})^\alpha \right) \tag{47}$$

$$D_\alpha(\tilde{\mathbf{f}}) = \frac{1}{1-\alpha} \log \left(\sum\nolimits_{i=1}^{K} P(z=i|\,\tilde{f})^\alpha \right) \tag{48}$$

Normally, the value of α is set to 2 by default (also known as the Collision entropy).

– *Source-target coherence across subdomains* A translation rule with source and target phrases having two similar distributions over the latent subdomains is likely to be safer to use. We can use the Chebyshev distance to measure the similarity between two distributions. The divergence of two vectors $\tilde{\mathbf{e}}$ and $\tilde{\mathbf{f}}$ is defined as in (49):

$$D(\tilde{\mathbf{e}},\ \tilde{\mathbf{f}}) = \max_{i=\{1,\ \ldots,\ K\}} \left| P(z=i|\,\tilde{e}) - P(z=i|\,\tilde{f}) \right| \tag{49}$$

Once integrated into a phrase-based SMT system as feature functions, the measures force the decoder to give higher preference to domain-invariant translations, which work well across domains, over risky domain-specific alternatives. The translation improvement is quite robust; it is obtained without tuning specifically for the target domain or using other domain-related meta-information in the training corpus (Cuong et al. 2016b).

A similar idea has been deployed in Zhang et al. (2014b), which exploits topic-insensitivity that is learned over documents for translation. There is a link between this line of work and extensive prior work on minimum Bayes risk (MBR) objectives (used either at test time (Kumar and Byrne 2004) or during training (Goodman 1998; Sima'an 2003; Smith and Eisner 2006; Pauls et al. 2009)). The goal of MBR minimization is to select translations that are less 'risky', but there is a degree of uncertainty in modelling such predictions, and some of this uncertainty may indeed be associated with domain-variability of translations. Still, a system trained with an MBR objective will tend to output the most frequent translation rather than the most domain-invariant one, and this, as we argued in the introduction, might not be the right decision when applying it across domains. We believe that the two classes of methods are largely complementary.

9 Conclusion

This paper contributes a comprehensive survey of domain adaptation for SMT. We first introduce preliminaries regarding SMT in general, with a focus on aspects of SMT relevant to domain adaptation. We present an in-depth discussion where we explain what may go wrong with translation when applying a phrase-based SMT system to new domains.

The question of "what constitutes a domain?" is an open one which has not been well defined in the literature. Each different view of factors contributing to defining the domain leads to a different approach to domain adaptation. We provide a general picture of domain adaptation, and show how different research lines fall into a specific part of the general picture, as well as how they relate to each other. Providing such a comprehensive survey is, to the best of our knowledge, a novel contribution.

As discussed, SMT is just one among data-driven approaches to modeling translation. Other approaches can be deployed, e.g. example-based machine translation (Nagao 1984; Carl and Way 2003) and neural MT (Bahdanau et al. 2015). While it is pretty clear that example-based machine translation can benefit from what the domain-adaptation literature for SMT offers, it would be less clear whether neural MT can learn from that or not. Recent studies suggest this is the case, where classic techniques in domain adaptation for SMT can be used to perform adaptation for neural translation models [cf. Durrani et al. (2015), Joty et al. (2015)]. More specifically, Durrani et al. (2015) shows that EM-based mixture modeling and data-selection techniques also give a significant improvement in adaptation. Joty et al. (2015) reveal that regularizing the loss function towards the in-domain neural network joint model also improves translation.

Acknowledgements We thank the editor, anonymous reviewers and Ivan Titov for their inputs. The work is performed at ILLC, University of Amsterdam. The authors are supported by EU FP7 Marie Curie ITN Project (nr. 317471) and QT21 Project (H2020 nr. 645452).

Funding Funding was provided by VICI (Grant No. 277-89-002).

References

Axelrod A, He X, Gao J (2011) Domain adaptation via pseudo in-domain data selection. In: EMNLP '11: proceedings of the conference on empirical methods in natural language processing, Edinburgh, UK, pp 355–362

Aziz W, Dymetman M, Specia L (2014) Exact decoding for phrase-based statistical machine translation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), Doha, Qatar, pp 1237–1249

Bahdanau D, Cho K, Bengio Y (2015) Neural machine translation by jointly learning to align and translate. In: Proceedings of the international conference on learning representations, San Diego, CA

Bertoldi N, Federico M (2009) Domain adaptation for statistical machine translation with monolingual resources. In: Proceedings of the fourth workshop on statistical machine translation, Athens, Greece, pp 182–189

Besling S, Meier HG (1995) Language model speaker adaptation. In: Fourth European conference on speech communication and technology (EUROSPEECH '95), Madrid, Spain, pp 1755–1758

Bhattacharyya A (1946) On a measure of divergence between two multinomial populations. Sankhya Indian J Stat 7(4):401–406

Birch A, Osborne M, Koehn P (2007) CCG supertags in factored statistical machine translation. In: Proceedings of the second workshop on statistical machine translation, Prague, Czech Republic, pp 9–16
Bisazza A, Ruiz N, Federico M (2011) Fill-up versus interpolation methods for phrase-based SMT adaptation. In: 2011 international workshop on spoken language translation, IWSLT, San Francisco, CA, USA, pp 136–143
Biçici E, Yuret D (2011) Instance selection for machine translation using feature decay algorithms. In: WMT 2011: Proceedings of the 6th workshop on statistical machine translation, Edinburgh, Scotland, UK, pp 272–283
Blei DM, Ng AY, Jordan MI (2003) Latent dirichlet allocation. J Mach Learn Res 3:993–1022
Blunsom P, Osborne M (2008) Probabilistic inference for machine translation. In: Proceedings of the 2008 conference on empirical methods in natural language processing, Honolulu, Hawaii, pp 215–223
Bod R, Scha R, Sima'an K (2003) Data-oriented parsing. Center for the Study of Language and Information—Lecture Notes, Amsterdam, The Netherlands
Brown PF, Cocke J, Della Pietra SA, Della Pietra VJ, Jelinek F, Lafferty JD, Mercer RL, Roossin PS (1990) A statistical approach to machine translation. Comput Linguist 16(2):79–85
Brown PF, Della Pietra VJ, Della Pietra SA, Mercer RL (1993) The mathematics of statistical machine translation: Parameter estimation. Comput Linguist 19(2):263–311
Carl M, Way A (eds) (2003) Recent advances in example-based machine translation. Kluwer Academic Publishers, Dordrecht
Carpuat M, Goutte C, Foster G (2014) Linear mixture models for robust machine translation. In: Proceedings of the ninth workshop on statistical machine translation, Baltimore, Maryland, USA, pp 499–509
Chang YW, Collins M (2011) Exact decoding of phrase-based translation models through lagrangian relaxation. In: Proceedings of the conference on empirical methods in natural language processing, Edinburgh, United Kingdom, pp 26–37
Chang YW, Rush AM, DeNero J, Collins M (2014) A constrained viterbi relaxation for bidirectional word alignment. In: Proceedings of the 52nd annual meeting of the association for computational linguistics (volume 1: long papers), Baltimore, Maryland, pp 1481–1490
Chen B, Foster G, Kuhn R (2013) Adaptation of reordering models for statistical machine translation. In: 2013 conference of the North American Chapter of the Association for computational linguistics: human language technologies. Atlanta, Georgia, pp 938–946
Chen B, Kuhn R, Foster GF (2013b) Vector space model for adaptation in statistical machine translation. In: Proceedings of the 51st annual meeting of the association for computational linguistics, volume 1: long papers, Sofia, Bulgaria, pp 1285–1293
Chen B, Kuhn R, Foster G, Cherry C, Huang F (2016) Bilingual methods for adaptive training data selection for machine translation. In: Conference of the association for machine translation in the Americas, the twelfth conference of the association for machine translation in the Americas, Austin, Texas, pp 93–106
Cherry C, Foster G (2012) Batch tuning strategies for statistical machine translation. In: Proceedings of the 2012 conference of the North American chapter of the association for computational linguistics: human language technologies, Montreal, Canada, pp 427–436
Chiang D (2005) A hierarchical phrase-based model for statistical machine translation. In: Proceedings of the 43rd annual meeting on association for computational linguistics, Ann Arbor, Michigan, pp 263–270
Chiang D (2007) Hierarchical phrase-based translation. Comput Linguist 33(2):202–228
Chiang D, Marton Y, Resnik P (2008) Online large-margin training of syntactic and structural translation features. In: Proceedings of the 2008 conference on empirical methods in natural language processing, Honolulu, Hawaii, pp 224–233
Chiang D, Knight K, Wang W (2009) 11,001 new features for statistical machine translation. In: Proceedings of Human Language Technologies: The 2009 annual conference of the North American chapter of the association for computational linguistics, Boulder, Colorado, pp 218–226
Chiang D, DeNeefe S, Pust M (2011) Two easy improvements to lexical weighting. In: Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies: short papers, Portland, Oregon, vol 2, pp 455–460
Clark J, Dyer C, Lavie A (2014) Locally non-linear learning for statistical machine translation via discretization and structured regularization. Trans Assoc Comput Linguist 2:393–404

Clarkson P, Robinson A (1997) Language model adaptation using mixtures and an exponentially decaying cache. In: IEEE international conference on acoustics, speech, and signal processing, ICASSP-97. Munich, Germany, pp 799–802

Cui L, Chen X, Zhang D, Liu S, Li M, Zhou M (2013) Multi-domain adaptation for SMT using multi-task learning. In: Proceedings of the 2013 conference on empirical methods in natural language processing, Seattle, Washington, USA, pp 1055–1065

Cuong H, Sima'an K (2014a) Latent domain phrase-based models for adaptation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), Doha, Qatar, pp 566–576

Cuong H, Sima'an K (2014b) Latent domain translation models in mix-of-domains haystack. In: Proceedings of COLING 2014, the 25th international conference on computational linguistics: technical papers, Dublin, Ireland, pp 1928–1939

Cuong H, Sima'an K (2015) Latent domain word alignment for heterogeneous corpora. In: Proceedings of the 2015 conference of the north american chapter of the association for computational linguistics: human language technologies, Denver, Colorado, USA, pp 398–408

Cuong H, Frank S, Sima'an K (2016a) ILLC-UvA adaptation system (scorpio) at WMT'16 IT-DOMAIN Task. In: Proceedings of the first conference on machine translation, shared task papers, Berlin, Germany, vol 2, pp 423–427

Cuong H, Sima'an K, Titov I (2016b) Adapting to all domains at once: rewarding domain invariance in SMT. Trans Assoc Comput Linguist 4:99–112

Daumé H III, Jagarlamudi J (2011) Domain adaptation for machine translation by mining unseen words. In: Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies: short papers, Portland, Oregon, vol 2, pp 407–412

Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm. J R Stat Soc Ser B 39(1):1–38

Devlin J, Zbib R, Huang Z, Lamar T, Schwartz R, Makhoul J (2014) Fast and robust neural network joint models for statistical machine translation. In: Proceedings of the 52nd annual meeting of the association for computational linguistics (volume 1: long papers), Baltimore, Maryland, pp 1370–1380

Dong M, Cheng Y, Liu Y, Xu J, Sun M, Izuha T, Hao J (2014) Query lattice for translation retrieval. In: Proceedings of COLING 2014, the 25th international conference on computational linguistics: technical papers, Dublin, Ireland, pp 2031–2041

Duh K, Sudoh K, Tsukada H (2010) Analysis of translation model adaptation in statistical machine translation. In: 2010 international workshop on spoken language translation, IWSLT 2010. France, Paris, pp 243–250

Duh K, Neubig G, Sudoh K, Tsukada H (2013) Adaptation data selection using neural language models: experiments in machine translation. In: 51st annual meeting of the association for computational linguistics (short papers). Sofia, Bulgaria, vol 2, pp 678–683

Durrani N, Sajjad H, Joty S, Abdelali A, Vogel S (2015) Using joint models for domain adaptation in statistical machine translation. In: Proceedings of the MT summit XV, MT researchers' track, Miami, Florida, USA, vol. 1, pp 117–130

Eck M, Vogel S, Waibel A (2005) Low cost portability for statistical machine translation based on n-gram coverage. In: MT Summit X, conference proceedings: the tenth machine translation summit, Phuket, Thailand, pp 227–234

Eidelman V, Boyd-Graber J, Resnik P (2012) Topic models for dynamic translation model adaptation. In: Proceedings of the 50th annual meeting of the association for computational linguistics: short papers, Jeju Island, Korea, vol 2, pp 115–119

Federico M, Cettolo M, Bentivogli L, Paul M, Stüker S (2012) Overview of the IWSLT 2012 evaluation campaign. In: 2012 international workshop on spoken language translation, Hong Kong, pp 12–33

Foster G, Kuhn R (2007) Mixture-model adaptation for smt. In: Proceedings of the second workshop on statistical machine translation, Prague, Czech Republic, pp 128–135

Foster G, Goutte C, Kuhn R (2010) Discriminative instance weighting for domain adaptation in statistical machine translation. In: 2010 conference on empirical methods in natural language processing, Massachusetts, Cambridge, pp 451–459

Foster G, Chen B, Kuhn R (2013) Simulating discriminative training for linear mixture adaptation in statistical machine translation. In: Proceedings of the XIV machine translation summit, Nice, France, pp 183–190

Galley M, Manning CD (2008) A simple and effective hierarchical phrase reordering model. In: Proceedings of the 2008 conference on empirical methods in natural language processing, Honolulu, Hawaii, pp 848–856

Gao Q, Lewis W, Quirk C, Hwang MY (2011) Incremental training and intentional over-fitting of word alignment. In: Proceedings of the 13th machine translation summit (MT summit XIII), Xiamen, China, pp 106–113

Gong Z, Zhang M, Zhou G (2011) Cache-based document-level statistical machine translation. In: Proceedings of the conference on empirical methods in natural language processing, Edinburgh, United Kingdom, pp 909–919

Goodman JT (1998) Parsing inside-out. PhD thesis, Harvard University, Cambridge, MA

Green S, Wang S, Cer D, Manning CD (2013) Fast and adaptive online training of feature-rich translation models. In: Proceedings of the 51st annual meeting of the association for computational linguistics (volume 1: long papers), Sofia, Bulgaria, pp 311–321

Green S, Cer DM, Manning CD (2014) An empirical comparison of features and tuning for phrase-based machine translation. In: Proceedings of the ninth workshop on statistical machine translation, WMT@ACL 2014, Baltimore, Maryland, USA, pp 466–476

Gruber A, Weiss Y, Rosen-Zvi M (2007) Hidden topic markov models. In: Proceedings of the eleventh international conference on artificial intelligence and statistics, San Juan, Puerto Rico, pp 163–170

Haddow B (2013) Applying pairwise ranked optimisation to improve the interpolation of translation models. In: Proceedings of the human language technologies: conference of the North American chapter of the association of computational linguistics, Atlanta, Georgia, USA, pp 342–347

Haghighi A, Liang P, Berg-Kirkpatrick T, Klein D (2008) Learning bilingual lexicons from monolingual corpora. In: Proceedings of ACL-08: HLT, Columbus, Ohio, pp 771–779

Hasler E, Haddow B, Koehn P (2012) Sparse lexicalised features and topic adaptation for SMT. In: 2012 international workshop on spoken language translation. IWSLT, Hong Kong, pp 268–275

Hasler E, Blunsom P, Koehn P, Haddow B (2014) Dynamic topic adaptation for phrase-based MT. In: Proceedings of the 14th conference of the European chapter of the association for computational linguistics, Gothenburg, Sweden, pp 328–337

Hewavitharana S, Mehay D, Ananthakrishnan S, Natarajan P (2013) Incremental topic-based translation model adaptation for conversational spoken language translation. In: Proceedings of the 51st annual meeting of the association for computational linguistics (volume 2: short papers), Sofia, Bulgaria, pp 697–701

Hieber F, Riezler S (2015) Bag-of-words forced decoding for cross-lingual information retrieval. In: NAACL HLT 2015, the 2015 conference of the North American chapter of the association for computational linguistics: human language technologies. Denver, Colorado, USA, pp 1172–1182

Hofmann T (1999) Probabilistic latent semantic analysis. In: Proceedings of the fifteenth conference on uncertainty in artificial intelligence, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, pp 289–296

Hopkins M, May J (2011) Tuning as ranking. In: Proceedings of the conference on empirical methods in natural language processing, Edinburgh, United Kingdom, pp 1352–1362

Hu Y, Zhai K, Eidelman V, Boyd-Graber J (2014) Polylingual tree-based topic models for translation domain adaptation. In: Proceedings of the 52nd annual meeting of the association for computational linguistics (volume 1: long papers), Baltimore, Maryland, pp 1166–1176

Irvine A, Morgan J, Carpuat M, Daumé H III, Munteanu D (2013a) Measuring machine translation errors in new domains. Trans Assoc Comput Linguist 1:429–440

Irvine A, Quirk C, Daumé III H (2013b) Monolingual marginal matching for translation model adaptation. In: Proceedings of the 2013 conference on empirical methods in natural language processing, Seattle, Washington, USA, pp 1077–1088

Jeblee S, Feely W, Bouamor H, Lavie A, Habash N, Oflazer K (2014) Domain and dialect adaptation for machine translation into egyptian arabic. In: Proceedings of the EMNLP 2014 workshop on arabic natural language processing (ANLP), Doha, Qatar, pp 196–206

Joty S, Sajjad H, Durrani N, Al-Mannai K, Abdelali A, Vogel S (2015) How to avoid unwanted pregnancies: Domain adaptation using neural network models. In: Proceedings of the 2015 conference on empirical methods in natural language processing, Lisbon, Portugal, pp 1259–1270

Kettunen K (2009) Choosing the Best MT Programs for CLIR purposes—can MT metrics be helpful? In: Proceedings of the 31st European conference on information retrieval research: advances in infor-

mation retrieval, Springer International Publishing, Heidelberg/Berlin, Germany. Lecture Notes in Computer Science, vol 5478, pp 706–712
Kirchhoff K, Bilmes J (2014) Submodularity for data selection in machine translation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), Doha, Qatar, pp 131–141
Koehn P (2004) Pharaoh: A beam search decoder for phrase-based statistical machine translation models. In: Proceedings of the machine translation: from real users to research: 6th conference of the association for machine translation in the Americas, Springer, Berlin/Heidelberg, Germany, pp 115–124
Koehn P (2005) Europarl: A parallel corpus for statistical machine translation. In: MT Summit X, conference proceedings: the tenth machine translation summit, Phuket, Thailand, pp 79–86
Koehn P (2010) Statistical machine translation. Cambridge University Press, New York, NY, USA
Koehn P, Knight K (2002) Learning a translation lexicon from monolingual corpora. In: Proceedings of the ACL-02 workshop on unsupervised lexical acquisition, Philadelphia, Pennsylvania, pp 9–16
Koehn P, Schroeder J (2007) Experiments in domain adaptation for statistical machine translation. In: Proceedings of the second workshop on statistical machine translation, Prague, Czech Republic, pp 224–227
Koehn P, Och FJ, Marcu D (2003) Statistical phrase-based translation. In: Proceedings of the 2003 conference of the North American chapter of the association for computational linguistics on human language technology, vol 1, Edmonton, Canada, pp 48–54
Koehn P, Hoang H, Birch A, Callison-Burch C, Federico M, Bertoldi N, Cowan B, Shen W, Moran C, Zens R, Dyer C, Bojar O, Constantin A, Herbst E (2007) Moses: open source toolkit for statistical machine translation. In: Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions, Prague, Czech Republic, pp 177–180
Kuhn R, De Mori R (1992) Corrections to "a cache-based language model for speech recognition". IEEE Trans Pattern Anal Mach Intell 14(6):691–692
Kullback S, Leibler RA (1951) On information and sufficiency. Ann Math Stat 22(1):79–86
Kumar S, Byrne W (2004) Minimum Bayes-risk decoding for statistical machine translation. In: Proceedings of the human language technology conference of the North American chapter of the association for computational linguistics: HLT-NAACL 2004, Boston, Massachusetts, USA, pp 169–176
Lambert P, Schwenk H, Servan C, Abdul-Rauf S (2011) Investigations on translation model adaptation using monolingual data. In: Proceedings of the sixth workshop on statistical machine translation, Edinburgh, Scotland, pp 284–293
Lewis W, Eetemadi S (2013) Dramatically reducing training data size through vocabulary saturation. In: Proceedings of the eighth workshop on statistical machine translation, Sofia, Bulgaria, pp 281–291
Liu C, Liu Y, Sun M, Luan H, Yu H (2015) Generalized agreement for bidirectional word alignment. In: Proceedings of the 2015 conference on empirical methods in natural language processing, Lisbon, Portugal, pp 1828–1836
Liu L, Watanabe T, Sumita E, Zhao T (2013) Additive neural networks for statistical machine translation. In: 51st annual meeting of the association for computational linguistics (long papers), Sofia, Bulgaria, vol 1, pp 791–801
Lopez A (2008) Statistical machine translation. ACM Comput Surv 40(3):1–49
Louis A, Webber B (2014) Structured and unstructured cache models for smt domain adaptation. In: Proceedings of the 14th conference of the European chapter of the association for computational linguistics, Gothenburg, Sweden, pp 155–163
Macherey W, Och FJ, Thayer I, Uszkoreit J (2008) Lattice-based minimum error rate training for statistical machine translation. In: Proceedings of the 2008 conference on empirical methods in natural language processing, Honolulu, Hawaii, pp 725–734
Mansour S, Ney H (2014) Unsupervised adaptation for statistical machine translation. In: Proceedings of the ninth workshop on statistical machine translation, Baltimore, Maryland, USA, pp 457–465
Mansour S, Wuebker J, Ney H (2011) Combining translation and language model scoring for domain-specific data filtering. In: International workshop on spoken language translation, CA, USA, San Francisco, pp 222–229
Marton Y, Resnik P (2008) Soft syntactic constraints for hierarchical phrased-based translation. In: Proceedings of ACL-08: HLT, Columbus, Ohio, pp 1003–1011
Matsoukas S, Rosti AVI, Zhang B (2009) Discriminative corpus weight estimation for machine translation. In: Proceedings of the 2009 conference on empirical methods in natural language processing, Singapore, vol 2, pp 708–717

Mimno D, Wallach HM, Naradowsky J, Smith DA, McCallum A (2009) Polylingual topic models. In: Proceedings of the 2009 conference on empirical methods in natural language processing, Singapore, vol 2, pp 880–889

Moore RC, Lewis W (2010) Intelligent selection of language model training data. In: Proceedings of the ACL 2010 conference short papers, Uppsala, Sweden, pp 220–224

Nagao M (1984) A framework of a mechanical translation between japanese and english by analogy principle. In: Elithorn A, Banerji R (eds) Artif Hum Intell. North-Holland, Amsterdam, pp 173–180

Nakov P (2008) Improving english-spanish statistical machine translation: Experiments in domain adaptation, sentence paraphrasing, tokenization, and recasing. In: Proceedings of the third workshop on statistical machine translation, Columbus, Ohio, pp 147–150

Neubig G, Watanabe T (2016) Optimization for statistical machine translation: a survey. Comput Linguist 42(1):1–54

Nikoulina V, Kovachev B, Lagos N, Monz C (2012) Adaptation of statistical machine translation model for cross-lingual information retrieval in a service context. In: Proceedings of the 13th conference of the European chapter of the association for computational linguistics, Avignon, France, pp 109–119

Och FJ (2003) Minimum error rate training in statistical machine translation. In: Proceedings of the 41st annual meeting on association for computational linguistics, Sapporo, Japan, vol 1, pp 160–167

Och FJ, Ney H (2002) Discriminative training and maximum entropy models for statistical machine translation. In: Proceedings of the 40th annual meeting on association for computational linguistics, Philadelphia, Pennsylvania, pp 295–302

Ozdowska S, Way A (2009) Optimal bilingual data for French-English PB-SMT. In: Proceedings of the 13th annual meeting of the European association for machine translation, Barcelona, Spain, pp 96–103

Papineni K, Roukos S, Ward T, Zhu WJ (2002) Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting on association for computational linguistics, Philadelphia, Pennsylvania, pp 311–318

Pauls A, DeNero J, Klein D (2009) Consensus training for consensus decoding in machine translation. In: Proceedings of the 2009 conference on empirical methods in natural language processing, vol 3, Singapore, pp 1418–1427

Pecina P, Toral A, van Genabith J (2012) Simple and effective parameter tuning for domain adaptation of statistical machine translation. In: Proceedings of the 24th international conference on computational linguistics, Mumbai, India, pp 2209–2224

Poncelas A, de Buy Maillette, Wenniger G, Way A (2017) Applying n-gram alignment entropy to improve feature decay algorithms. Prague Bull Math Linguist 108:245–256

Quirk C, Menezes A (2006) Dependency treelet translation: the convergence of statistical and example-based machine-translation? Mach Transl 20(1):43–65

Quirk C, Menezes A, Cherry C (2005) Dependency treelet translation: syntactically informed phrasal SMT. In: Proceedings of the 43rd annual meeting on association for computational linguistics, Ann Arbor, Michigan, pp 271–279

Razmara M, Foster G, Sankaran B, Sarkar A (2012) Mixing multiple translation models in statistical machine translation. In: Proceedings of the 50th annual meeting of the association for computational linguistics, long papers, Jeju Island, Korea, vol 1, pp 940–949

Schwenk H (2008) Investigations on large-scale lightly-supervised training for statistical machine translation. In: 2008 international workshop on spoken language translation, Honolulu, Hawaii, USA, pp 182–189

Schwenk H, Senellart J (2009) Translation model adaptation for an Arabic/French news translation system by lightly-supervised training. In: MT Summit XII: proceedings of the twelfth machine translation summit, Ottawa, Ontario, Canada, pp 308–315

Sennrich R (2012) Perplexity minimization for translation model domain adaptation in statistical machine translation. In: Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics, Avignon, France, pp 539–549

Shah K, Barrault L, Schwenk H (2010) Translation model adaptation by resampling. In: Proceedings of the joint fifth workshop on statistical machine translation and MetricsMATR, Uppsala, Sweden, pp 392–399

Shah K, Barrault L, Schwenk H (2012) A general framework to weight heterogeneous parallel data for model adaptation in statistical machine translation. In: Proceedings of the AMTA-2012: the tenth biennial conference of the association for machine translation in the Americas, San Diego, CA, 10pp

Shen S, Liu Y, Sun M, Luan H (2015) Consistency-aware search for word alignment. In: Proceedings of the 2015 conference on empirical methods in natural language processing, Lisbon, Portugal, pp 1228–1237

Sima'an K (2003) On maximizing metrics for syntactic disambiguation. In: Proceedings of the 8th international workshop on parsing technologies (IWPT), Nancy, France, pp 183–194

Simianer P, Riezler S, Dyer C (2012) Joint feature selection in distributed stochastic learning for large-scale discriminative training in smt. In: Proceedings of the 50th annual meeting of the association for computational linguistics: long papers, vol 1, Jeju Island, Korea, pp 11–21

Simion A, Collins M, Stein C (2013) A convex alternative to IBM model 2. In: Proceedings of the 2013 conference on empirical methods in natural language processing, EMNLP 2013, Seattle, Washington, USA, pp 1574–1583

Smith DA, Eisner J (2006) Minimum risk annealing for training log-linear models. In: Proceedings of the COLING/ACL 2006 main conference poster sessions, Sydney, Australia, pp 787–794

Snover M, Dorr B, Schwartz R (2008) Language and translation model adaptation using comparable corpora. In: Proceedings of the conference on empirical methods in natural language processing, Honolulu, Hawaii, pp 857–866

Su J, Wu H, Wang H, Chen Y, Shi X, Dong H, Liu Q (2012) Translation model adaptation for statistical machine translation with monolingual topic information. In: Proceedings of the 50th annual meeting of the association for computational linguistics (long papers), Jeju Island, Korea, vol 1, pp 459–468

Tamura A, Watanabe T, Sumita E (2012) Bilingual lexicon extraction from comparable corpora using label propagation. In: Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning, Jeju Island, Korea, pp 24–36

Tamura A, Watanabe T, Sumita E (2014) Recurrent neural networks for word alignment model. In: Proceedings of the 52nd annual meeting of the association for computational linguistics (long papers), Baltimore, Maryland, vol 1, pp 1470–1480

Tang J, Meng Z, Nguyen X, Mei Q, Zhang M (2014) Understanding the limiting factors of topic modeling via posterior contraction analysis. In: Proceedings of the 31st international conference on machine learning (ICML-14), Beijing, China, pp 190–198

Tiedemann J (2010) Context adaptation in statistical machine translation using models with exponentially decaying cache. In: Proceedings of the 2010 workshop on domain adaptation for natural language processing, Uppsala, Sweden, pp 8–15

Tillmann C (2004) A unigram orientation model for statistical machine translation. In: Proceedings of HLT-NAACL 2004: short papers, Boston, Massachusetts, pp 101–104

Tsuruoka Y, Tsujii J, Ananiadou S (2009) Stochastic gradient descent training for l1-regularized log-linear models with cumulative penalty. In: Proceedings of the joint conference of the 47th annual meeting of the ACL and the 4th international joint conference on natural language processing of the AFNLP, Singapore, vol 1, pp 477–485

Vogel S, Ney H, Tillmann C (1996) HMM-based word alignment in statistical translation. In: Proceedings of the coling 1996: the 16th international conference on computational linguistics, Denmark, Copenhagen, pp 836–841

Waite A, Byrne B (2015) The geometry of statistical machine translation. In: Proceedings of the 2015 conference of the North American chapter of the association for computational linguistics: human language technologies, Denver, Colorado, pp 376–386

Wang W, Macherey K, Macherey W, Och F, Xu P (2012) Improved domain adaptation for statistical machine translation. In: Proceedings of the AMTA-2012: the tenth biennial conference of the association for machine translation in the Americas, San Diego, CA

Wang X, Utiyama M, Finch A, Watanabe T, Sumita E (2015) Leave-one-out word alignment without garbage collector effects. In: Proceedings of the 2015 conference on empirical methods in natural language processing, Lisbon, Portugal, pp 1817–1827

Wäschle K, Riezler S (2012) Structural and topical dimensions in multi-task patent translation. In: Proceedings of the 13th conference of the European chapter of the association for computational linguistics, Avignon, France, pp 818–828

Watanabe T, Suzuki J, Tsukada H, Isozaki H (2007) Online large-margin training for statistical machine translation. In: EMNLP-CoNLL 2007, Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning, Prague, Czech Republic, pp 764–773

Van Der Wees M, Bisazza A, Weerkamp W, Monz C (2015) What's in a Domain? Analyzing Genre and Topic Differences in Statistical Machine Translation. In: Proceedings of the 53rd annual meeting of the association for computational linguistics and the 7th international joint conference on natural language processing of the asian federation of natural language processing, Short Papers, Beijing, China, vol 2, pp 560–566

Wu H, Wang H, Liu Z (2005) Alignment model adaptation for domain-specific word alignment. In: Proceedings of the 43rd annual meeting on association for computational linguistics, Ann Arbor, Michigan, pp 467–474

Wu H, Wang H, Zong C (2008) Domain adaptation for statistical machine translation with domain dictionary and monolingual corpora. In: Proceedings of the 22nd international conference on computational linguistics, Manchester, United Kingdom, vol 1, pp 993–1000

Yamada K, Knight K (2001) A syntax-based statistical translation model. In: Proceedings of the 39th annual meeting on association for computational linguistics, Toulouse, France, pp 523–530

Yu H, Huang L, Mi H, Zhao K (2013) Max-violation perceptron and forced decoding for scalable MT training. In: Proceedings of the 2013 conference on empirical methods in natural language processing, Seattle, Washington, USA, pp 1112–1123

Zhang B, Su J, Xiong D, Duan H, Yao J (2015) Discriminative reordering model adaptation via structural learning. In: Proceedings of the 24th international conference on artificial intelligence, Buenos Aires, Argentina, pp 1040–1046

Zhang H, Chiang D (2014) Kneser-Ney smoothing on expected counts. In: Proceedings of the 52nd annual meeting of the association for computational linguistics (long papers), Baltimore, Maryland, vol 1, pp 765–774

Zhang J, Li L, Way A, Liu Q (2014a) A probabilistic feature-based fill-up for smt. In: Proceedings of the 11th conference of the association for machine translation in the Americas, MT Researchers Track, Vancouver, Canada, vol 1, pp 96–109

Zhang M, Xiao X, Xiong D, Liu Q (2014b) Topic-based dissimilarity and sensitivity models for translation rule selection. J Artif Intell Res 50(1):1–30