Translation Model Adaptation Using Genre-Revealing Text Features

Marlies van der Wees **Arianna Bisazza** **Christof Monz**
Informatics Institute, University of Amsterdam
{m.e.vanderwees,a.bisazza,c.monz}@uva.nl

Abstract

Research in domain adaptation for statistical machine translation (SMT) has resulted in various approaches that adapt system components to specific translation tasks. The concept of a *domain*, however, is not precisely defined, and most approaches rely on provenance information or manual subcorpus labels, while genre differences have not been addressed explicitly. Motivated by the large translation quality gap that is commonly observed between different genres in a test corpus, we explore the use of document-level genre-revealing text features for the task of translation model adaptation. Results show that automatic indicators of genre can replace manual subcorpus labels, yielding significant improvements across two test sets of up to 0.9 BLEU. In addition, we find that our genre-adapted translation models encourage document-level translation consistency.

1 Introduction

Statistical machine translation (SMT) systems use large bilingual corpora to train translation models, which can be used to translate unseen test sentences. Training corpora are typically collected from a wide variety of sources and therefore have varying textual characteristics such as writing style and vocabulary. The test set, on the other hand, is much smaller and usually more homogeneous. As a result, there is often a mismatch between the test data and the majority of the training data. In such situations, it is beneficial to adapt the translation system to the translation task at hand, which is exactly the challenge of domain adaptation in SMT.

The concept of a *domain*, however, is not precisely defined across existing domain adaptation methods. Different domains typically correspond to different subcorpora, in which documents exhibit a particular combination of genre and topic, and optionally other textual characteristics such as dialect and register. This definition, however, has two major shortcomings. First, subcorpus-based domains depend on provenance information, which might not be available, or on manual grouping of documents into subcorpora, which is labor intensive and often carried out according to arbitrary criteria. Second, the commonly used notion of a domain neglects the fact that topic and genre are two distinct properties of text (Stein and Meyer Zu Eissen, 2006). While this distinction has long been acknowledged in text classification literature (Lee, 2001; Dewdney et al., 2001; Lee and Myaeng, 2002), most work on domain adaptation in SMT uses in-domain and out-of-domain data that differs on both the topic and the genre level (e.g., Europarl political proceedings (Koehn, 2005) versus EMEA medical text (Tiedemann, 2009)), making it unclear whether the proposed solutions address topic or genre differences.

In this work, we follow text classification literature for definitions of the concepts topic and genre. While *topic* refers to the general subject (e.g., sports, politics or science) of a document, *genre* is harder to define since existing definitions vary. Swales (1990), for example, refers to genre as a class of communicative events with a shared set of communicative purposes, and Karlgren (2004) calls it a grouping of documents that are stylistically consistent. Based on previous definitions, Santini (2004) concludes that the term genre is pri-

Proceedings of the Second Workshop on Discourse in Machine Translation (DiscoMT), pages 132–141,
Lisbon, Portugal, 17 September 2015.

marily used as a concept complementary to topic, covering the non-topical text properties function, style, and text type. Examples of genres include editorials, newswire, or user-generated (UG) text, i.e., content written by lay-persons that has not undergone any editorial control. Within the latter we can distinguish more fine-grained subclasses, such as dialog-oriented content (e.g., SMS or chat messages), weblogs, or commentaries to news articles, all of which pose different challenges to SMT (van der Wees et al., 2015a).

Recently, we studied the impact of topic and genre differences on SMT quality using the Gen&Topic benchmark set, an Arabic-English evaluation set with controlled topic distributions over two genres; newswire and UG comments (van der Wees et al., 2015b). Motivated by the observation that translation quality varies more between the two genres than across topics, we explore in this paper the task of genre adaptation. Concretely, we incorporate genre-revealing features, inspired by previous findings in genre classification literature, into a competitive translation model adaptation approach with the aim of improving translation quality across two test sets; the first containing newswire and UG comments, and the second containing newswire and UG weblogs.

In a series of translation experiments we show that automatic indicators of genre can replace manual subcorpus labels, yielding improvements of up to 0.9 BLEU over a strong unadapted baseline. In addition, we observe small but mostly significant improvements when using the automatic genre indicators on top of manual subcorpus labels. We also find that our genre-revealing feature values can be computed on either side of the training bitext, indicating that the proposed features are to a large extent language independent. Finally, we notice that our genre-adapted translation models encourage document-level translation consistency with respect to the unadapted baseline.

2 Related work

In recent years, domain adaptation for SMT has been studied actively. Outside of SMT research, text genre classification has received considerable attention, resulting in various sets of genre-revealing features. To our knowledge, the fields have not been combined in any previous work.

2.1 Domain adaptation for SMT

Most existing domain adaptation approaches can be grouped into two categories, depending on where in the SMT pipeline they adapt the system. First, *mixture modeling* approaches learn models from different subcorpora and interpolate these linearly (Foster and Kuhn, 2007) or log-linearly (Koehn and Schroeder, 2007). Sennrich (2012) enhances the approach by interpolating up to ten models, and Bertoldi and Federico (2009) use in-domain monolingual data to automatically generate in-domain bilingual data.

Second, *instance weighting* methods prioritize training instances that are most relevant to the test data, by assigning weights to sentence pairs (Matsoukas et al., 2009) or phrase pairs (Foster et al., 2010; Chen et al., 2013). In the most extreme case, weights are binary and training instances are either selected or discarded (Moore and Lewis, 2010; Axelrod et al., 2011).

In most previous work, domains are typically hard-labeled concepts that correspond to provenance or particular topic-genre combinations. In recent years, some work has explicitly addressed *topic* adaptation for SMT (Eidelman et al., 2012; Hewavitharana et al., 2013; Hasler et al., 2014a; Hasler et al., 2014b) using latent Dirichlet allocation (Blei et al., 2003). Surprisingly, *genre* (or style) adaptation has only been addressed to a limited extent (Bisazza and Federico, 2012; Wang et al., 2012), with methods requiring the availability of clearly separable in-domain and out-of-domain training corpora.

2.2 Text genre classification

Work on text genre classification has resulted in various methods that use different sets of genre-specific text features. Karlgren and Cutting (1994) were among the first to use simple document statistics, such as common word frequencies, first-person pronoun count, and average sentence length. Kessler et al. (1997) categorize four types of genre-revealing cues: *structural cues* (e.g., part-of-speech (POS) tag counts), *lexical cues* (specific words), *character-level cues* (e.g., punctuation marks), and *derivative cues* (ratios and variation measures based on other types of cues). Dewdney et al. (2001) compare a large number of document features and show that these outperform bag-of-words approaches, which are traditionally used in topic-based text classifica-

tion. Finn and Kushmerick (2006) also compare the bag-of-words approach with simple text statistics and conclude that both methods achieve high classification accuracy on fixed topic-genre combinations but perform worse when predicting topic-independent genre labels.

While mostly focused on the English language, some work has addressed language-independent (Sharoff, 2007; Sharoff et al., 2010) or cross-lingual genre classification (Gliozzo and Strapparava, 2006; Petrenz, 2012; Petrenz and Webber, 2012), indicating that a single set of genre-revealing features can generalize across multiple languages. In this paper, we examine whether genre-revealing features are also language independent when applied to translation model genre adaptation for SMT.

3 Translation model genre adaptation

For the task of genre adaptation to the genres newswire (NW) and UG comments or weblogs, we use a flexible translation model adaptation approach based on phrase pair weighting using a vector space model (VSM) inspired by Chen et al. (2013). The reason we choose an instance-weighting method rather than a mixture modeling approach is twofold: First, mixture modeling approaches intrinsically depend on subcorpus boundaries, which resemble provenance or require manual labeling. Second, Irvine et al. (2013) have shown that including relevant training data in a mixture modeling approach solves many coverage errors, but also introduces substantial amounts of new scoring errors. With phrase-pair weighting we aim to optimize phrase translation selection while keeping our training data fixed, and we can thus compare the impact of several methodological variants on genre adaptation for SMT.

3.1 VSM adaptation framework

In the selected adaptation method, each phrase pair in the training data is represented by a vector capturing information about the phrase:

$$V(\bar{f},\bar{e}) = < w_1(\bar{f},\bar{e}), \ldots, w_N(\bar{f},\bar{e}) > . \quad (1)$$

Here, $w_i(\bar{f},\bar{e})$ is the weight for phrase pair $(\bar{f},\bar{e})$ of dimension $i \in N$ in the vector space. The exact definition of dimensions $i \in N$, and hence the information captured by the vector, depends on the definition of the vector space, for which we describe different variants in Sections 3.2–3.4.

In addition to the phrase pair vectors, a single vector is created for the development set which is assumed to be similar to the test data:

$$V(dev) = < w_1(dev), \ldots, w_N(dev) >, \quad (2)$$

where weights $w_i(dev)$ are computed for the entire development set, summing over the vectors of all phrase pairs that occur in the development set:

$$w_i(dev) = \sum_{(\bar{f},\bar{e}) \in P_{dev}} c_{dev}(\bar{f},\bar{e})\ w_i(\bar{f},\bar{e}). \quad (3)$$

Here P_{dev} refers to the set of phrase pairs that can be extracted from the development set, $c_{dev}(\bar{f},\bar{e})$ is the count of phrase pair $(\bar{f},\bar{e})$ in the development set, and $w_i(\bar{f},\bar{e})$ is the phrase pair's weight for dimension i in the vector space.

Next, for each phrase pair in the training corpus, we compute the Bhattacharyya Coefficient (BC) (Bhattacharyya, 1946) as a similarity score[1] between its vector and the development vector:

$$BC(dev;\bar{f},\bar{e}) = \sum_{i=0}^{i=N} \sqrt{p_i(dev) \cdot p_i(\bar{f},\bar{e})}, \quad (4)$$

where $p_i(dev)$ and $p_i(\bar{f},\bar{e})$ are probabilities representing smoothed normalized vector weights $w_i(dev)$ and $w_i(\bar{f},\bar{e})$, respectively.

The computed similarity is assumed to indicate the relevance of the phrase pair with respect to the development and test set and is added to the decoder as a new feature. In a similar fashion, two similarity-based decoder features $BC(dev;\bar{f},\bullet)$ and $BC(dev;\bullet,\bar{e})$ are added for the marginal counts of the source and target phrases, respectively. Further technical details can be found in (Chen et al., 2013).

The presented framework for translation model adaptation allows us to empirically compare various sets of VSM features, of which we present three in the following sections.

3.2 Genre adaptation with subcorpus labels

First, we adhere to the commonly used scenario in which adaptation is guided by manual subcorpus labels that resemble provenance of training documents. In this formulation, each weight $w_i(\bar{f},\bar{e})$ in Equation (1) is a standard *tf-idf* weight capturing the relative occurrence of phrase pair $(\bar{f},\bar{e})$ in

[1] Chen et al. (2013) compared three similarity measures and observed that the BC similarity performed best.

different subcorpora. Since our aim is to adapt to multiple genres in a test corpus, we follow Chen et al. (2013) and manually group our training data into subcorpora that reflect various genres (see Table 3). While this definition of the vector space can approximate genres at different levels of granularity, manual subcorpus labels are labor intensive to generate, particularly in the scenario where provenance information is not available, and may not generalize well to new translation tasks.

3.3 Genre adaptation with genre features

To move away from manually assigned subcorpus labels, we explore the use of genre-revealing features that have proven successful for distinguishing genres in classification tasks (Section 2.2). To this end, we construct a list of features that are directly observable in raw text, see Table 1. For each genre feature i, we first compute its raw count at the document level $c_i(d)$, which we then normalize for document length and scale to a value in range $[0, 1]$ to obtain the final document-level feature value $w_i(d)$. Next, each vector weight $w_i(\bar{f}, \bar{e})$ in Equation (1) equals the weighted average of the document-level values of genre feature i for all training instances of phrase pair $(\bar{f}, \bar{e})$:

$$w_i(\bar{f}, \bar{e}) = \frac{1}{c_{train}(\bar{f}, \bar{e})} \sum_{d \in D} c_d(\bar{f}, \bar{e})\ w_i(d). \quad (5)$$

Here, $c_{train}(\bar{f}, \bar{e})$ is the total count of phrase pair $(\bar{f}, \bar{e})$ in the training corpus, D is the number of documents in the training corpus, $c_d(\bar{f}, \bar{e})$ is the count of $(\bar{f}, \bar{e})$ in document d, and $w_i(d)$ is the document-level value of genre feature i for document d. Note that this definition differs from the standard *tf-idf* weight that is used in Section 3.2 since each genre feature has exactly one score per document, and we do not have to normalize for dissimilar subcorpus sizes.

We determine the most genre-discriminating features with a Mann-Whitney U test (Mann and Whitney, 1947) on the observed feature values for each genre in the development set. The seven most discriminative features between the genres NW and UG which we use in the remainder of this paper are shown in the top part of Table 1. The main goal of this paper is to investigate whether this type of genre-revealing features can be useful for the task of translation model genre adaptation, hence we do not attempt to fully exploit the set of possible features. Since genre-discriminating features potentially generalize across languages (Petrenz and Webber, 2012), we compute the document-level feature values $w_i(d)$ on the source as well as the target sides of our bitext, and we examine whether both are equally suitable for translation model genre adaptation.

Feature
First person pronoun count
Second person pronoun count
Repeating punctuation count ("...", "?!", etc.)
Exclamation mark count
Question mark count
Emoticons count
Numbers count
Third person pronoun count
Plural pronoun count
Average word length
Average sentence length
Total punctuation count
Quote count
Dates count
Percentages count
Long words (> 7 characters) count
Stopwords count
Unique words count

Table 1: Selection of document-level features inspired by genre-classification literature. The top seven features are most discriminative between the genres NW and UG, and are used in the genre-specific VSM approaches.

3.4 Genre adaptation with LDA

Another type of feature that does not depend on provenance information is Latent Dirichlet allocation (LDA) (Blei et al., 2003), an unsupervised word-based approach that infers a preset number of latent dimensions in a corpus and represents documents as distributions over those dimensions. Despite its recent successes in topic adaptation for SMT, we expect such a bag-of-words approach to be insufficient to model genre accurately. Nevertheless, since many of the proposed genre-revealing features are in fact lexical features, it is worth verifying whether LDA can infer genre differences directly from raw text.

To this end, we use LDA-inferred document distributions as a third vector representation in the adaptation framework. Weights $w_i(\bar{f}, \bar{e})$ in Equation (1) are now average probabilities of latent dimension i for all training instances of phrase pair $(\bar{f}, \bar{e})$, computed as in Equation (5). We implement LDA using Gensim (Řehůřek and Sojka,

Benchmark			NW	UG	Total
Gen&Topic (1 reference)	Dev	#Sent	997	1,127	2,124
		#Tok	26.9K	25.8K	52.7K
	Test	#Sent	1,567	1,749	3,316
		#Tok	46.3K	45.5K	91.8K
NIST (4 references)	Dev	#Sent	1,033	764	1,797
		#Tok	34.4K	14.6K	49.0K
	Test	#Sent	1,399	1,274	2,673
		#Tok	46.6K	39.9K	86.6K

Table 2: Corpus statistics of the evaluation sets. Numbers of tokens are counted on the Arabic side. Note that Gen&Topic contains one reference translation per sentence, while NIST has four sets of reference translations.

2010), with varying numbers of latent dimensions (5, 10, 20, and 50). Of these, LDA with 10 dimensions yields the best translation performance, which is consistent with findings in a related topic adaptation approach by Eidelman et al. (2012). The LDA features in this VSM variant are inferred from the source side of the training data.

4 Experimental setup

We evaluate the methods described in Section 3 on two Arabic-to-English translation tasks, both comprising the NW and UG. The first evaluation set is the Gen&Topic benchmark (van der Wees et al., 2015b), which consists of manually translated web-crawled news articles and their respective manually translated user comments, both covering five different topics. Since this evaluation set has controlled topic distributions per genre, differences in translation quality between genres can be entirely attributed to actual genre differences. The second evaluation set contains NIST OpenMT Arabic-English test sets, using NIST 2006 for tuning, and NIST 2008 and NIST 2009 combined for testing. These data sets cover the genres NW and UG weblogs but are not controlled for topic distributions. Specifications for both evaluation sets are shown in Table 2. Note that Gen&Topic contains one reference translation per sentence, while NIST has four sets of reference translations.

We perform our experiments using an in-house phrase-based SMT system similar to Moses (Koehn et al., 2007). All runs use lexicalized reordering, distinguishing between monotone, swap, and discontinuous reordering, with respect to the previous and next phrase (Koehn et al., 2005).

Subcorpus	Genre	#Sentences	#Tokens
NIST broadcast conv.	BC	48K	1,071K
NIST broadcast news	BN	41K	923K
NIST newsgroup	NG	15K	392K
NIST newswire	NW	133K	4,545K
NIST weblog	WL	7.7K	126K
ISI newswire	NW	699K	22,231K
Web newswire	NW	376K	11,107K
Web UG comments	CM	203K	5,985K
Web editorials	ED	127K	4,341K
Web Ted talks	SP	98K	2,168K
Total	All	1.75M	52.9M

Table 3: Corpus statistics of the Arabic-English parallel training data. Tokens are counted on the Arabic side. Genre mapping: BC=broadcast conversation, BN=broadcast news, NG=newsgroup, NW=newswire, WL=UG weblogs, CM=UG comments, ED=editorials, SP=speech transcripts.

Other features include linear distortion with limit 5, lexical weighting (Koehn et al., 2003), and a 5-gram target language model trained with Kneser-Ney smoothing (Chen and Goodman, 1999). The feature weights are tuned using pairwise ranking optimization (PRO) (Hopkins and May, 2011). For all experiments, tuning is done separately for the two genre-specific development sets.

All runs use parallel corpora made available for NIST OpenMT 2012, excluding the UN data. While LDC-distributed data sets contain substantial portions of documents within the NW genre, they only contain small portions of UG documents. To alleviate this imbalance we augment our LDC-distributed training data with a variety of web-crawled manually translated documents, containing user comments that are of a similar nature as the UG documents in the Gen&Topic, set as well as a number of other genres. Table 3 lists the corpus statistics of the training data, split by manual subcorpus labels as used for the subcorpus VSM variant (see Section 3.2). While our manually grouped subcorpora approximate those used by Chen et al. (2013), exact agreement was impossible to obtain, illustrating that it is not trivial to manually generate optimal subcorpus labels.

We tokenize all Arabic data using MADA (Habash and Rambow, 2005), ATB scheme. Word alignment was performed by running GIZA++ in both directions and generating the symmetric alignments using the 'grow-diag-final-and' heuristics. We use an adapted language model which

Method		Gen&Topic (1 reference) NW	UG	All	NIST (4 references) NW	UG	All
Baseline		21.5	17.2	19.3	55.3	40.4	48.5
VSM variants using automatic indicators of genre:							
LDA 10 topics		21.7 (+0.2)	17.3 (+0.1)	19.4$^{\triangle}$(+0.1)	55.9$^{\blacktriangle}$(+0.6)	40.7$^{\triangle}$(+0.3)	49.0$^{\blacktriangle}$(+0.5)
Genre features	Source	21.9$^{\blacktriangle}$(+0.4)	17.4$^{\triangle}$(+0.2)	19.6$^{\blacktriangle}$(+0.3)	55.7$^{\blacktriangle}$(+0.4)	41.0$^{\blacktriangle}$(+0.6)	49.0$^{\blacktriangle}$(+0.5)
	Target	21.7 (+0.2)	17.5$^{\blacktriangle}$(+0.3)	19.6$^{\blacktriangle}$(+0.3)	55.9$^{\blacktriangle}$(+0.6)	41.2$^{\blacktriangle}$(+0.8)	49.1$^{\blacktriangle}$(+0.6)
Genre+LDA	Source	**21.9$^{\blacktriangle}$(+0.4)**	**17.5$^{\blacktriangle}$(+0.3)**	**19.7$^{\blacktriangle}$(+0.4)**	56.1$^{\blacktriangle}$(+0.8)	41.2$^{\blacktriangle}$(+0.8)	49.2$^{\blacktriangle}$(+0.7)
	Target	21.8$^{\blacktriangle}$(+0.3)	17.5$^{\blacktriangle}$(+0.3)	19.6$^{\blacktriangle}$(+0.3)	**56.2$^{\blacktriangle}$(+0.9)**	**41.2$^{\blacktriangle}$(+0.8)**	**49.2$^{\blacktriangle}$(+0.7)**

Table 4: BLEU scores of the baseline system and all VSM variants using automatic indicators of genre. Significance is tested against the baseline, and the best performing VSM variant per test set is bold-faced.

is trained on 1.6B tokens and linearly interpolates different English Gigaword subcorpora with the English side of our bitext. The resulting model covers both genres in the benchmark sets, but is not varied between experiments since we want to investigate the effects of different features on translation model adaptation.

5 Results

In this section we compare a number of variants of the general VSM framework, differing in the way vectors are defined and constructed (see Sections 3.2–3.4). Translation quality of all experiments is measured with case-insensitive BLEU (Papineni et al., 2002) using the closest-reference brevity penalty. We use approximate randomization (Noreen, 1989) for significance testing (Riezler and Maxwell, 2005). Statistically significant differences are marked by $\triangle$ and $\blacktriangle$ for the $p \leq 0.05$ and the $p \leq 0.01$ level, respectively.

VSM using intrinsic text features. We first test various VSM variants that use automatic indicators of genre and do not depend on the availability of provenance information or manual subcorpus labels (Table 4). Of these, genre adaptation with LDA-based features (Section 3.4) achieves strongly significant improvements over the unadapted baseline for the NIST-NW and the complete NIST test sets, however improvements on the other test portions are very small. When manually inspecting the LDA-inferred latent dimensions, we observe that LDA is overly aggressive in considering all of the UG genre as a single thread, while latent dimensions inferred for NW are more fine-grained. While this finding can be explained by the unbalanced amount of training data per genre, it also illustrates that LDA-based features seem less suitable to capture low-resource genres.

Next, we evaluate the VSM variant that uses genre-revealing text features inspired by genre classification research (Section 3.3). This approach achieves statistically significant improvements over the baseline in all runs except one (i.e., target-side features on Gen&Topic NW). We also see that translation quality is fairly similar for features computed on either side of the bitext, indicating that the proposed genre features can generalize across languages.

Our last VSM variant in Table 4 combines genre-revealing and LDA features by using VSM similarities from both approaches as additional decoder features. This combined setting yields the largest improvements, which are all strongly significant and always equal to or better than the performance achieved by either individual feature type, suggesting that the two vector representations are to some extent complementary. Again, source and target genre feature values perform alike, with source-side genre features performing best for Gen&Topic, and target-side genre features obtaining slightly better overall results for NIST.

VSM using manual subcorpus labels. Next we compare our best performing VSM variant per test set (bold-faced in Table 4) to the originally proposed VSM variant using manual subcorpus labels (Section 3.2). The latter can be considered as an adapted baseline, however with the disadvantage that it relies on the availability of provenance information or manual grouping of documents into informative subcorpora.

Table 5 first shows the performance of VSM with manual subcorpus labels, which works well

Method	Gen&Topic (1 reference)			NIST (4 references)		
	NW	UG	All	NW	UG	All
VSM manual subcorpora	21.6	17.3	19.3	56.3	41.1	49.2
△ *wrt unadapted baseline*	*(+0.1)*	*(+0.1)*	*(±0.0)*	*(+1.0)*$^{\blacktriangle}$	*(+0.7)*$^{\blacktriangle}$	*(+0.7)*$^{\blacktriangle}$
VSM automatic genre	21.9$^{\blacktriangle}$(+0.3)	17.5$^{\triangle}$(+0.2)	19.7$^{\blacktriangle}$(+0.4)	56.2 (−0.1)	41.2 (+0.1)	49.2 (±0.0)
VSM manual+automatic	21.9$^{\blacktriangle}$(+0.2)	17.4 (+0.1)	19.6$^{\blacktriangle}$(+0.3)	56.4 (+0.1)	41.4$^{\triangle}$(+0.3)	49.5$^{\blacktriangle}$(+0.3)

Table 5: BLEU scores of VSM with manual subcorpus labels in comparison to the best performing VSM with automatic indicators of genre per test corpus (see bold-faced results in Table 4), and the combination of manual subcorpus labels and automatic features. BLEU differences and significance for the bottom two variants are measured with respect to VSM manual subcorpora.

on NIST, confirming previously published results (Chen et al., 2013), but does not lead to significant improvements on Gen&Topic with respect to the unadapted baseline. This suggests that the success of this approach depends on a good fit between the test data distribution and the partitioning of training data into subcorpora, and that a single set of manual subcorpus labels is not guaranteed to generalize to new translation tasks.

The bottom half of the table shows that similar (for NIST) or larger (for Gen&Topic) improvements can be achieved when using the most competitive VSM variant that uses intrinsic text properties instead of manual subcorpus labels. Finally, we use intrinsic text features on top of manual subcorpus labels, i.e., we add all three proposed VSM feature types as additional decoder features. For NIST, this approach yields weakly significant improvements over the runs with only manual subcorpus labels, indicating that the automatic genre features capture additional genre information that is not contained in the manually grouped subcorpora. For Gen&Topic, including manual subcorpus labels does not increase translation performance with respect to VSM with genre and LDA features only, confirming the poor generalization of manual subcorpus labels to new translation tasks.

6 Translation consistency analysis

In the proposed translation model adaptation approach lexical choice is more tailored towards the different genres than in the baseline. We therefore hypothesize that the adapted system increases consistency of output translations within genres. To test this hypothesis, we measure translation consistency following Carpuat and Simard (2012). Their approach studies *repeated phrases*, defined as source phrases p in the phrase table that occur more than once in a single test document d and contain at least one content word. For each repeated phrase, all of its 1-best output translations are compared. If these are identical except for punctuation or stopword differences, the repeated phrase is deemed *consistent*.

Test set	Genre	# Repeated phrases	% Consistent phrases	
			Base	VSM auto. genre
G&T	NW	7,318	43.2	47.4 (+4.2)
	UG	6,024	55.5	58.2 (+2.7)
	All	13,342	48.7	52.3 (+3.6)
NIST	NW	7,412	40.5	40.6 (+0.1)
	UG	5,431	54.5	57.1 (+2.6)
	All	12,843	46.5	47.6 (+1.1)

Table 6: Document-level translation consistency values for the baseline and best performing VSM variant using automatic genre indicators.

The results of the consistency analysis for the unadapted baseline and the best performing VSM genre+LDA variants are shown in Table 6. We observe that for both benchmark sets translation consistency is clearly lower in NW than in UG documents. This is likely due to the lower coverage of UG in the training data, which is in agreement with the finding by Carpuat and Simard that translation consistency increases for weaker systems trained on smaller amounts of training data. In line with our expectation, the results also show that document-level translation consistency increases when using the adapted system. Although Carpuat and Simard show that translation consistency does not imply higher quality, they also conclude that consistently translated phrases are more often translated correctly than inconsistently translated phrases.

Table 7 shows some examples of phrases that

Genre	Source phrase	Baseline translation(s)	VSM automatic genre translation(s)
Inconsistent in baseline, consistent in adapted system:			
UG	و هذا يدل	and this indicates / and this shows that	and this shows
UG	و الاجهاد	fatigue and stress / and the stress	and the stress
NW	القطاع الصحي	the health sector / workers in the health sector	the health sector
NW	المائة من	percent of egyptians / percent of them	percent of
Consistent in baseline, inconsistent in adapted system:			
UG	مليار دولار سنويا	billion dollars annually	billion dollars annually / billion dollars a year
UG	التطعيم	immunization	immunization / vaccination
NW	شرق افريقيا	east african countries	east african countries / east africa
NW	عالميا	worldwide	worldwide / global

Table 7: Examples of source phrases that generate inconsistent translations in the baseline and consistent translations in the adapted system (top), and vice versa (bottom).

were translated consistently in one system, but inconsistently in the other. While more phrases moved from being translated inconsistently in the baseline to consistently in the adapted system, the opposite was also observed for all benchmark sets. Looking at the examples for UG, we see that the adapted system often favors translations that are more colloquial or simplified than (some of) their counterparts in the baseline system, e.g., "shows" instead of "indicates", "a year" instead of "annually", and "vaccination" instead of "immunization". For NW, on the other hand, translations in the adapted system are often more formal (e.g., "global" instead of "worldwide") or more concise (e.g., "the health sector" instead of "workers in the health sector", and "east africa" instead of "east african countries") than in the baseline.

7 Conclusions

Domain adaptation is an active field for statistical machine translation (SMT), and has resulted in various approaches that adapt system components to specific translation tasks. However, the concept of a *domain* is not precisely defined and often confuses the notions of *topic*, *genre*, and *provenance*. Motivated by the large translation quality gap that is commonly observed between different genres, we have explored the task of translation model genre adaptation. To this end, we incorporated document-level genre-revealing features, inspired by genre classification research, into a competitive adaptation framework.

In a series of experiments across two test sets with two genres we show that automatic indicators of genre can replace manual subcorpus labels, yielding significant improvements of up to 0.9 BLEU over an unadapted baseline. In addition, we observe small improvements when using automatic genre features on top of manual subcorpus labels. We also find that the genre-revealing feature values can be computed on either side of the training bitext, indicating that our proposed features are language independent. Therefore, the advantages of using the proposed method are twofold: (i) manual subcorpus labels are not required, and (ii) the same set of features can be used successfully across different test sets and languages. Finally, we find that our genre-adapted translation models encourage document-level translation consistency with respect to the unadapted baseline.

Future work includes developing other methods for genre adaptation, on both the translation and language model level; possibly eliminating the need of a development set that is representative of the test set's genre distribution; scaling to more than two genres; and finally improving model coverage in addition to scoring.

Acknowledgments

This research was funded in part by the Netherlands Organization for Scientific Research (NWO) under project number 639.022.213.

References

Amittai Axelrod, Xiaodong He, and Jianfeng Gao. 2011. Domain adaptation via pseudo in-domain data selection. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 355–362.

Nicola Bertoldi and Marcello Federico. 2009. Domain adaptation for statistical machine translation with monolingual resources. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 182–189.

Anil Bhattacharyya. 1946. On a measure of divergence between two multinomial populations. *Sankhyā: The Indian Journal of Statistics*, pages 401–406.

Arianna Bisazza and Marcello Federico. 2012. Cutting the long tail: Hybrid language models for translation style adaptation. In *Proceedings of the 13th Conference of the European Chapter of the ACL*, pages 439–448.

David Blei, Andrew Ng, and Michael Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

Marine Carpuat and Michel Simard. 2012. The trouble with SMT consistency. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 442–449.

Stanley F. Chen and Joshua Goodman. 1999. An empirical study of smoothing techniques for language modeling. *Computer Speech and Language*, 13(4):359–393.

Boxing Chen, Roland Kuhn, and George Foster. 2013. Vector space model for adaptation in statistical machine translation. In *Proceedings of the 51st Annual Meeting of the ACL*, pages 1285–1293.

Nigel Dewdney, Carol VanEss-Dykema, and Richard MacMillan. 2001. The form is the substance: classification of genres in text. In *Proceedings of the Workshop on Human Language Technology and Knowledge Management*.

Vladimir Eidelman, Jordan Boyd-Graber, and Philip Resnik. 2012. Topic models for dynamic translation model adaptation. In *Proceedings of the 50th Annual Meeting of the ACL*, pages 115–119.

Aidan Finn and Nicholas Kushmerick. 2006. Learning to classify documents according to genre. *Journal of the American Society for Information Science and Technology*, 57:1506–1518.

George Foster and Roland Kuhn. 2007. Mixture-model adaptation for SMT. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 128–135.

George Foster, Cyril Goutte, and Roland Kuhn. 2010. Discriminative instance weighting for domain adaptation in statistical machine translation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 451–459.

Alfio Gliozzo and Carlo Strapparava. 2006. Exploiting comparable corpora and bilingual dictionaries for cross-language text categorization. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL*, pages 553–560.

Nizar Habash and Owen Rambow. 2005. Arabic tokenization, part-of-speech tagging and morphological disambiguation in one fell swoop. In *Proceedings of the 43rd Annual Meeting of the ACL*, pages 573–580.

Eva Hasler, Phil Blunsom, Philipp Koehn, and Barry Haddow. 2014a. Dynamic topic adaptation for phrase-based MT. In *Proceedings of the 14th Conference of the European Chapter of the ACL*, pages 328–337.

Eva Hasler, Barry Haddow, and Philipp Koehn. 2014b. Dynamic topic adaptation for SMT using distributional profiles. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 445–456.

Sanjika Hewavitharana, Dennis Mehay, Sankaranarayanan Ananthakrishnan, and Prem Natarajan. 2013. Incremental topic-based translation model adaptation for conversational spoken language translation. In *Proceedings of the 51st Annual Meeting of the ACL*, pages 697–701.

Mark Hopkins and Jonathan May. 2011. Tuning as ranking. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1352–1362. ACL.

Ann Irvine, John Morgan, Marine Carpuat, Hal Daumé III, and Dragos Stefan Munteanu. 2013. Measuring machine translation errors in new domains. *Transactions of the ACL*, 1:429–440.

Jussi Karlgren and Douglas Cutting. 1994. Recognizing text genres with simple metrics using discriminant analysis. In *Proceedings of the 15th International conference on Computational Linguistics (COLING 94)*, pages 1071–1075.

Jussi Karlgren. 2004. The wheres and whyfores for studying text genre computationally. In *Workshop on Style and Meaning in Language, Art, Music, and Design*.

Brett Kessler, Geoffrey Numberg, and Hinrich Schütze. 1997. Automatic detection of text genre. In *Proceedings of the eighth conference of the European chapter of the ACL*, pages 32–38.

Philipp Koehn and Josh Schroeder. 2007. Experiments in domain adaptation for statistical machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 224–227.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the 2003 Conference of the North American Chapter of the ACL on Human Language Technology*, pages 48–54.

Philipp Koehn, Amittai Axelrod, Alexandra Birch Mayne, Chris Callison-Burch, Miles Osborne, and David Talbot. 2005. Edinburgh system description for the 2005 IWSLT speech translation evaluation. In *Proceedings of the International Workshop on Spoken Language Translation*.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 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*, pages 177–180.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Machine Translation Summit X*, pages 79–86.

Yong-Bae Lee and Sung Hyon Myaeng. 2002. Text genre classification with genre-revealing and subject-revealing features. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 145–150.

David Y.W. Lee. 2001. Genres, registers, text types, domains and styles: Clarifying the concepts and navigating a path through the BNC jungle. *Language Learning & Technology*, 5(3):37–72.

Henry B. Mann and Donald R. Whitney. 1947. On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics*, pages 50–60.

Spyros Matsoukas, Antti-Veikko I. Rosti, and Bing Zhang. 2009. Discriminative corpus weight estimation for machine translation. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 708–717.

Robert C. Moore and William Lewis. 2010. Intelligent selection of language model training data. In *Proceedings of the ACL 2010 Conference (Short Papers)*, pages 220–224.

Eric W. Noreen. 1989. *Computer-Intensive Methods for Testing Hypotheses: An Introduction*. Wiley-Interscience.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the ACL*, pages 311–318.

Philipp Petrenz and Bonnie Webber. 2012. Robust cross-lingual genre classification through comparable corpora. In *The 5th Workshop on Building and Using Comparable Corpora*, pages 1–9.

Philipp Petrenz. 2012. Cross-lingual genre classification. In *Proceedings of the Student Research Workshop at the 13th Conference of the European Chapter of the ACL*, pages 11–21.

Radim Řehůřek and Petr Sojka. 2010. Software framework for topic modelling with large corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50.

Stefan Riezler and John T. Maxwell. 2005. On some pitfalls in automatic evaluation and significance testing for MT. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 57–64.

Marina Santini. 2004. State-of-the-art on automatic genre identification. Technical Report ITRI-04-03, Information Technology Research Institute, University of Brighton.

Rico Sennrich. 2012. Perplexity minimization for translation model domain adaptation in statistical machine translation. In *Proceedings of the 13th Conference of the European Chapter of the ACL*, pages 539–549.

Serge Sharoff, Zhili Wu, and Katja Markert. 2010. The web library of babel: evaluating genre collections. In *Proceedings of the Seventh conference on International Language Resources and Evaluation*, pages 3063–3070.

Serge Sharoff. 2007. Classifying web corpora into domain and genre using automatic feature identification. In *Proceedings of the 3rd Web as Corpus Workshop*.

Benno Stein and Sven Meyer Zu Eissen. 2006. Distinguishing topic from genre. In *Proceedings of the 6th International Conference on Knowledge Management (I-KNOW 06)*, pages 449–456.

John M. Swales. 1990. *Genre Analysis*. Cambridge University Press., Cambridge, UK.

Jörg Tiedemann. 2009. News from OPUS - A collection of multilingual parallel corpora with tools and interfaces. In *Recent Advances in Natural Language Processing*, pages 237–248.

Marlies van der Wees, Arianna Bisazza, and Christof Monz. 2015a. Five shades of noise: Analyzing machine translation errors in user-generated text. In *Proceedings of the ACL 2015 Workshop on Noisy User-generated Text*, pages 28–37.

Marlies van der Wees, Arianna Bisazza, Wouter Weerkamp, and Christof Monz. 2015b. What's in a domain? Analyzing genre and topic differences in statistical machine translation. In *Proceedings of the Joint Conference of the 53rd Annual Meeting of the ACL and the 7th International Joint Conference on Natural Language Processing of the AFNLP (Short Papers)*, pages 560–566.

Wei Wang, Klaus Macherey, Wolfgang Macherey, Franz Och, and Peng Xu. 2012. Improved domain adaptation for statistical machine translation. In *Proceedings of the Tenth Conference of the Association for Machine Translation in the Americas*.