Measuring Machine Translation Errors in New Domains

Ann Irvine
Johns Hopkins University
anni@jhu.edu

John Morgan
University of Maryland
jjm@cs.umd.edu

Marine Carpuat
National Research Council Canada
marine.carpuat@nrc.gc.ca

Hal Daumé III
University of Maryland
me@hal3.name

Dragos Munteanu
SDL Research
dmunteanu@sdl.com

Abstract

We develop two techniques for analyzing the effect of porting a machine translation system to a new domain. One is a macro-level analysis that measures how domain shift affects corpus-level evaluation; the second is a micro-level analysis for word-level errors. We apply these methods to understand what happens when a Parliament-trained phrase-based machine translation system is applied in four very different domains: news, medical texts, scientific articles and movie subtitles. We present quantitative and qualitative experiments that highlight opportunities for future research in domain adaptation for machine translation.

1 Introduction

When building a statistical machine translation (SMT) system, the expected use case is often limited to a specific domain, genre and register (henceforth "domain" refers to this set, in keeping with standard, imprecise, terminology), such as a particular type of legal or medical document. Unfortunately, it is expensive to obtain enough parallel data to reliably estimate translation models in a new domain. Instead, one can hope that large amounts of data from another, "old domain," might be close enough to stand as a proxy. This is the defacto standard: we train SMT systems on Parliament proceedings, but then use them to translate all sorts of new text. Unfortunately, this results in significantly degraded translation quality. In this paper, we present two complementary methods for *quantifiably measuring* the source of translation errors (§5.1 and §5.2) in a novel taxonomy (§4). We show quantitative (§7.1) and qualitative (§7.2) results obtained from our methods on four very different new domains: newswire, medical texts, scientific abstracts, and movie subtitles.

Old Domain (Hansard)	
Inp	monsieur le président, les pêcheurs de homard de la région de l'atlantique sont dans une situation catastrophique.
Ref	mr. speaker, lobster fishers in atlantic canada are facing a disaster.
Out	mr. speaker, the lobster fishers in atlantic canada are in a mess.
New Domain (Medical)	
Inp	mode et voie(s) d'administration
Ref	method and route(s) of administration
Out	fashion and voie(s) of directors

TABLE 1: Example inputs, references and system outputs. There are three types of errors: unseen words (blue), incorrect sense selection (red) and unknown sense (green).

Our basic approach is to think of translation errors in the context of a novel taxonomy of error categories, "S^4." Our taxonomy contains categories for the errors shown in Table 1, in which an SMT system trained on the Hansard parliamentary proceedings is applied to a new domain (in this case, medical texts). Our categorization focuses on the following: new French words, new French senses, and incorrectly chosen translations. The first methodology we develop for studying such errors is a micro-level study of the frequency and distribution of these error types in real translation output at the level of individual words (§5.1), without respect to how these errors affect overall translation quality. The second is a macro-level study of how these errors affect translation performance (measured by BLEU; §5.2). One important feature of our methodologies is that we focus on errors that *could possibly be fixed* given access to data from a new domain, rather than all errors that might arise because the particular translation model used is inadequate to capture the required

Transactions of the Association for Computational Linguistics, 1 (2013) 429–440. Action Editor: Philipp Koehn.
Submitted 3/2013; Revised 8/2013; Published 10/2013.

translation task (formally: we measure estimation error, not approximation error).

Our goal is *neither* to build better SMT systems nor to develop novel domain adaptation methods. We take an *ab initio* approach and ask: given a large unadapted, out of the box SMT system, what happens when it is applied in a new domain ? In order to answer this question, we will use parallel data in new domains, but only for testing purposes. The baseline SMT system is not adapted, except for the use of (1) a language model trained on *monolingual* new-domain language data, [1] and (2) a few thousand parallel sentences of tuning data in the new domain.

2 Summary of Results

We conduct experiments across a variety of domains (described in §6). [2] As in any study, our results are limited by assumptions about language, domains, and MT systems: these assumptions and their consequences are discussed in §8. Our high-level conclusions on the domains we study are summarized below (details may be found in §7).

1. Adapting an SMT system from the Parliament domain to the news domain is not a representative adaptation task; there are a very small number of errors due to unseen words, which are minor in comparison to all other domains. (Despite the fact that most previous work focuses exclusively on using news as a "new" domain, §3).

2. For the remaining domains, unseen words have a significant effect, both in terms of BLEU scores as well as fine-grained translation distinctions. However, many of these words have multiple translations, and a system must be able to correctly select which one to use in a particular context.

3. Likewise, words that gain new senses account for approximately as much error as unseen words, suggesting a novel avenue for research in sense induction. Unfortunately, it appears that choosing the right sense for these at translation time is even more difficult than in the unseen word case.

4. The story is more complicated for seen words with known translations: if we limit ourselves to "high confidence" translations, there is a lot to be gained by improving the scores in translation models. However, for an entire phrase table, manipulating scores can hurt as often as it helps.

1. We use old/new to refer to domains and source/target to refer to languages, to avoid ambiguity (we stay away from in-domain and out-of-domain, which is itself ambiguous).

2. All source data, methodological code and outputs are available at `http://hal3.name/damt`.

3 Related Work

Most related work has focused on either (a) analyzing errors made by machine translation systems in a non-adaptation setting (Popović and Ney, 2011), or (b) trying to directly improve machine translation performance. A small amount of work (discussed next) addresses issues of analyzing MT systems in a domain adaptation setting.

3.1 Analysis of Domain Effects

To date, work on domain adaptation in SMT mostly proposed methods to efficiently combine data from multiple domains. To the best of our knowledge, there have been only a few studies to understand how domain shifts affect translation quality (Duh et al., 2010; Bisazza et al., 2011; Haddow and Koehn, 2012). However, these start from different premises than this paper, and as a result, ask related but complementary questions. These previous analyses focus on how to improve a particular MT architecture (trained on new domain data) by injecting old domain data into a specific part of the pipeline in order to improve BLEU score. In comparison to this work, we focus on finer-grained phenomena. We distinguish between effects previously lumped together as "missing phrase-table entries."

Despite different starting assumptions, language pairs and data, some of our conclusions are consistent with previous work: in particular, we highlight the importance of differences in *coverage* in an adaptation setting. However, our fine-grained analysis shows that correctly scoring translations for previously unseen words and senses is a complex issue. Finally, these other studies suggest potential directions for refining our error categories: for instance, Haddow and Koehn (2012) show that the impact of additional new or old domain data is different for rare vs. frequent phrases.

3.2 Domain Adaptation for MT

Prior work focuses on methods combining data from old and new domains to learn translation and language models.

Many filtering techniques have been proposed to select OLD data that is similar to NEW . Information retrieval techniques have been used to improve the language model (Zhao et al., 2004), the translation model (Hildebrand et al., 2005; Lu et al., 2007; Gong et al., 2011; Duh et al., 2010; Banerjee et al., 2012), or both (Lu et al., 2007); language model cross-entropy has also been used for data selection (Axelrod et al., 2011; Mansour et al., 2011; Sennrich, 2012).

Another research thread addresses corpora weighting, rather than hard filtering. Weighting has been applied at different levels of granularity: sentence pairs (Matsoukas et al., 2009), phrase pairs (Foster et al., 2010), n-grams (Ananthakrishnan et al., 2011), or sub-corpora through factored models (Niehues and Waibel, 2010). In particular, Foster et al. (2010) show that adapting at the phrase pair levels outperform earlier coarser corpus level combination approaches (Foster and Kuhn, 2007). This is consistent with our analysis: domain shifts have a fine-grained impact on translation quality.

Finally, strategies have been proposed to combine sub-models trained independently on different sub-corpora. Linear interpolation is widely used for mixing language models in speech recognition, and it has also been used for adapting translation and language models in MT(Foster and Kuhn, 2007; Tiedemann, 2010; Lavergne et al., 2011). Log-linear combination fits well in existing SMT architectures (Foster and Kuhn, 2007; Koehn and Schroeder, 2007). Koehn and Schroeder (2007) consider both an intersection setting (where only entries occurring in all phrase-tables combined are considered), and a union setting (where entries which are not in the intersection are given an arbitrary null score). Razmara et al. (2012) take this approach further and frame combination as ensemble decoding.

3.3 Targeting Specific Error Types

The experiments conducted in this article motivated follow-up work on identifying when a word has gained a new sense in a new domain (Carpuat et al., 2013), as well as learning joint word translation probability distributions from comparable new domain corpora (Irvine et al., 2013). Earlier, Daumé III and Jagarlamudi (2011) showed how mining translations for unseen words from comparable corpora can improve SMT in a new domain.

4 The S^4 Taxonomy

We begin with a simple question: when we move an SMT system from an old domain to a new domain, what goes wrong ? We employ a set of *four* error types as our taxonomy. We refer to these error types as SEEN, SENSE, SCORE and SEARCH, and together as the S^4 taxonomy:

SEEN: an attempt to translate a source word or phrase that has never been seen before. For example, "voie(s)" in Table 1.

SENSE: an attempt to translate a previously seen source word or phrase, but for which the correct target language *sense* has never been observed. [3] In Table 1, the Hansard-trained system had never seen "mode" translated as "method."

SCORE: an incorrect translation for which the system *could* have succeeded but did not because an incorrect alternative outweighed the correct translation. In a conventional translation system, this could be due to errors in the language model, translation model, or both. In Table 1, the Hansard-trained system had seen "administration" translated as "administration," but "directors" had a higher probability.

SEARCH: an error due to pruning in beam search.

When limiting oneself to issues of *lexical selection*, this set is exhaustive and disjoint: any lexical selection error made by an MT system can be attributed to *exactly one* of these error categories. This observation is important for developing methodologies for measuring the impact of each of these sources of error. Partitions of the set of errors that focus on categories other than lexical choice have been investigated by Vilar et al. (2006).

5 Methodology for Analyzing MT Systems

Given the S^4 taxonomy for categorizing SMT errors, it would be possible (if painstaking) to manually annotate SMT output with error types. We prefer automated methods. In this section we describe two such methods: a micro-level analysis to

3. We define "sense" as a particular translation into a target language, in line with Carpuat & Wu (2007) or Mihalcea et al. (2010). This means both traditional word sense errors and other translation errors (like morphological variants) are included.

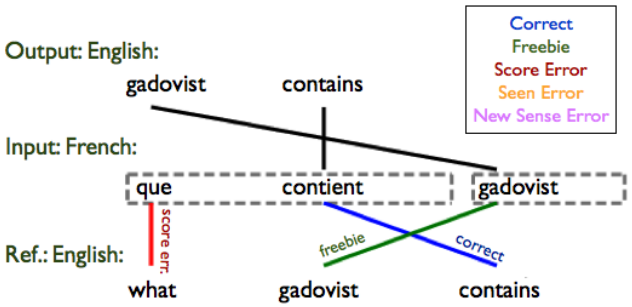

FIGURE 1: Example of WADE visualization. Dashed boxes around the French input mark the phrase spans used by the decoder.

see what happens at the word level (regardless of how it affects translation performance) and a macro-level analysis to discover impact on corpus translation performance. We focus on the first three S^4 categories and separately discuss search errors (§7). In both cases, we use exact string match to detect translation equivalences, as has been done previously in other settings that also use word alignments to inspect errors or automatically generate data for other tasks (Blatz et al., 2004; Carpuat and Wu, 2007; Popović and Ney, 2011; Bach et al., 2011, among others).

5.1 Micro-analysis: WADE

We define *Word Alignment Driven Evaluation*, or WADE, which is a technique for analyzing MT system output at the word level, allowing us to (1) manually browse visualizations of MT output annotated with S^4 error types, and (2) aggregate counts of errors. WADE is based on the fact that we can automatically word-align a French test sentence and its English reference translation, and the MT decoder naturally produces a word alignment between a French sentence and its machine translation. We can then check whether the MT output has the same set of English words aligned to each French word that we would hope for, given the reference.

In some ways, WADE is similar to the word-based analysis technique of Popović and Ney (2011). However, in contrast to that work, we do not directly align the hypothesis and reference translations but, rather, pivot through the source text. Additionally, we use WADE to annotate S^4 errors, which are driven more by how lexical choice is made within the SMT framework than by linguistic properties of words in the reference and hypothesis translations. For example, in the case of domain adaptation, we do not expect the rate of inflectional errors to be affected by domain shift.

In WADE, the unit of analysis is each word alignment between a French word, f_i, and a reference English word, e_j. To annotate the aligned pair, $a_{i,j}$, we consider the word(s), H_i, in the output English sentence which are aligned (by the decoder) to f_i. If e_j appears in the set H_i, then the alignment $a_{i,j}$ is marked correct. If not, the alignment is categorized with one of the S^4 error types. If the French word f_i does not appear in the phrase table used for translation, then the alignment is marked as a SEEN error. If f_i does appear in the phrase table, but it is never observed translating as e_j, then the alignment is marked as a SENSE error. If f_i had been observed translating as e_j, but the decoder chose an alternate translation, then the alignment is marked as a SCORE error. Our results in §7 show that SEARCH errors are very infrequent, so we mark all errors other than SEEN and SENSE as SCORE errors. We make use of one additional category: Freebie. Our MT system copies unseen (aka "OOV") French words into the English output, and "freebies" are French words for which this is correct.

For WADE analysis only, we use the alignments yielded by a model trained over our train and test datasets and the grow-diag-final heuristic. Because WADE's unit of analysis is each *alignment* link between the source text and its reference, it ignores unaligned words in the input source text.

Figure 1 shows an example of a WADE-annotated sentence. In addition to providing an easy way to visualize and browse the errors in MT output, WADE allows us to aggregate counts over the S^4 error types. In our analysis (§7), we present results that show not only total numbers of each error type but also how WADE-annotations change when we introduce some NEW-domain parallel training data. For example, SEEN errors could remain SEEN errors, become correct, or become SENSE or SCORE errors when we introduce additional training data.

5.2 Macro-analysis: TETRA

In this section, we discuss an approach to measuring the effect of each potential source of error when a translation system is considered in full. The key

idea is to *enhance* the translation model of OLD, an MT system trained on old domain parallel text, to compare the impact of potential sources of *improvement*. We use parallel new domain data to propose enhancements to the OLD system. This provides a realistic measure of what could be achieved *if* one had access to parallel data in the new domain. The specific system we build, called MIXED, is a linear interpolation of a translation model trained only on old domain data and a model trained only on new domain data (Foster and Kuhn, 2007). The mixing weights are selected via grid search on a tuning set, selecting for BLEU. We call our approach TETRA: Table Enhancement for Translation Analysis.

Below, we design experiments to tease apart the differences in domains by adjusting the models and enhancing OLD to be more like MIXED. We perform different enhancements depending on the error category we are targeting. As discussed in §6, our experiments are conducted using phrase-based SMT systems, so the translation models (TM) that are enhanced are the phrase table and reordering table.

Seen In order to estimate the effect of SEEN errors, we enhance the TM of OLD by adding phrase pairs that translate words found only in the new-domain data, and we measure the BLEU improvement. More precisely, we identify the set of phrase pairs in the TM of MIXED, for which the French side contains at least one word that does not appear in the old-domain training data. These are the phrases responsible for the SEEN errors. We build system TETRA+SEEN by adding these phrases to the TM of OLD. When adding these phrases, we add them together with their feature value scores.

Sense Analogously, the phrases responsible for SENSE errors are those from MIXED where the French side exists in the phrase table of OLD, but their English translations do not. We build TETRA+SENSE by adding these phrases to OLD.

Score To isolate and measure the effect of phrase scores, we consider the phrases that our OLD and MIXED systems have in common: the intersection of their translation tables. We build two systems, OLD SCORE and NEW SCORE, with identical phrase pairs; in OLD SCORE, the feature values are taken from the OLD system's tables; in NEW SCORE the feature values are taken from the MIXED system's tables.

Domain	Sentences	L	Tokens	Types	# Phrases
Hansard	8,107,356	fr	161.7m	192k	479.0m
		en	144.5m	187k	
News	135,838	fr	3.9m	63k	12.4m
		en	3.3m	52k	
EMEA	472,231	fr	6.5m	35k	4.4m
		en	5.9m	30k	
Science	139,215	fr	4.3m	118k	8.4m
		en	3.6m	114k	
Subs	19,239,980	fr	155.0m	362k	364.7m
		en	174.4m	293k	

TABLE 2: Basic characteristics of the training data: Number of sentences, tokens, word types and number of phrase pairs in the phrase tables.

6 Experimental conditions

6.1 Domains and Data

We conduct our study on French-English datasets. We consider five very different domains for which large corpora are publicly available. The largest corpus is the Hansard parliamentary proceedings. Corpora in the four other domains are smaller and more specialized, and, thus, more naturally serve as new domains. For each new domain, we use all available data. We do not attempt to hold the amount of new domain data constant, as we suspect that such artificial constraints would not be sufficient to control for the very different natures of the domains. Detailed statistics for the parallel corpora are given in Table 2.

Hansard: Canadian parliamentary proceedings, consists of manual transcriptions and translations of meetings of Canada's House of Commons and its committees from 2001 to 2009. Discussions cover a wide variety of topics, and speaking styles range from prepared speeches by a single speaker to more interactive discussions. It is significantly larger than Europarl, the common source of old domain data.

EMEA: Documents from the European Medicines Agency, made available with the OPUS corpora collection (Tiedemann, 2009). This corpus primarily consists of drug usage guidelines.

News: News commentary corpus made available for the WMT 2009 evaluation. It has been commonly used in the domain adaptation literature (Koehn and Schroeder, 2007; Foster and Kuhn, 2007; Haddow and Koehn, 2012, for instance).

Science: Parallel abstracts from scientific publications in many disciplines including physics, bio-

logy, and computer science. We collected data from two distinct sources: (1) Canadian Science Publishing made available translated abstracts from their journals which span many research disciplines; (2) parallel abstracts from PhD theses in Physics and Computer Science collected from the HAL public repository (Lambert et al., 2012).

Subs: Translated movie subtitles, available through the OPUS corpora collection (Tiedemann, 2009). In contrast to the other domains considered, subtitles consist of informal noisy text. [4]

In this study, we use the Hansard domain as the OLD domain, and we consider four possible NEW domains: EMEA, News, Science and Subs. Data sets for all domains were processed consistently. After tokenization, we paid particular attention to normalization in order to minimize artificial differences when combining data, such as American, British and Canadian spellings. This proved particularly important for the news domain; the impact of SEEN reduced by more than half after normalization.

6.2 MT systems

We build standard phrase-based SMT systems using the Moses toolkit (Koehn et al., 2007) for all experiments. Each system scores translation candidates using standard features: 5 phrase-table features, including phrasal translation probabilities and lexical weights in both translation directions, and a constant phrase penalty; 6 lexicalized reordering features, including bidirectional models built for monotone, swap, discontinuous reorderings; 1 distance-based reordering feature; and 2 language models, a 5-gram model learned on the OLD domain, and a 5-gram model learned on the NEW domain.

Features are combined using a log-linear model optimized for BLEU, using the n-best batch MIRA algorithm (Cherry and Foster, 2012). This results in a strong large-scale OLD system, which performs well on the old domain and is a good starting point for studying domain shifts. [5] The word alignments, language models and tuning sets are kept constant across all experiments per domain. For reference, we built systems using NEW domain data *only*; these achieved BLEU scores as follows: News = 21.70, EMEA = 34.63, Science = 30.72, Subs = 18.51.

7 Results

Before moving on to the interesting results, we show that SEARCH is not a major source of error. We analyzed search errors separately by computing BLEU scores for each domain with varying beam size from 10 to 1000, using the OLD system. We find that increasing the beam from 10 to 200 yields approximately a one BLEU point advantage across all domains. Increasing it further (to 500 or 1000) does not bestow any additional advantages. This suggests that for sufficiently wide beams, search is unlikely to contribute to adaptation errors. [6] This is consistent with previous results obtained in non-adapted settings using other measurement techniques: search errors account for less than 5% of the error in modern MT systems (Wisniewski et al., 2010), or 0.13% for small beam settings with a "gap constraint" (Chang and Collins, 2011). We use a beam value of 200 for all other experiments in this work.

7.1 Quantitative Results

Results are summarized in Tables 3 and 4. Table 3 gives an overview of our WADE analysis on test sets in each domain translated using OLD and MIXED models. Table 4 shows BLEU score results based on the TETRA analysis.

We first present general observations based on each set of results. WADE shows that for news, new domain data helps solve only a small number of SEEN issues, and SENSE and SCORE errors remain essentially unchanged. TETRA agrees that SENSE and SCORE are not issues in this domain. In general, the OLD system performs better on news than on the other three domains. For comparison, using the OLD system to translate a test set in the old (Hansard) domain yields a BLEU score of 37.41 and, according to our WADE analysis, 67.64% of all alignments are

4. Hansards, News and the Canadian Science Publishing are available, respectively, at: `http://www.parl.gc.ca`, `http://www.statmt.org/wmt09/translation-task.html`, and `http://www.nrcresearchpress.com`, preprocessed versions and data splits used in this paper can be downloaded from `http://hal3.name/damt`.

5. We use (unadapted) HMM word alignments (Vogel et al., 1996) in both directions, combined using grow-diag-final (Koehn et al., 2005). We estimate alignments jointly on all data sets. Thus, TETRA may have artificially good phrase tables.

6. This is likely dependent on language choice and the large amount of old domain parallel data.

Domain	% Correct		% Seen Errors			% Sense Errors			% Score Errors		
	OLD	MIXED	OLD	MIXED	%Δ	OLD	MIXED	%Δ	OLD	MIXED	%Δ
News	57.42	57.77	5.73	5.38	- 6%	11.02	11.13	+ 1%	25.84	25.72	+ 0%
EMEA	55.97	62.60	9.28	4.01	-56%	16.16	13.76	-15%	18.59	19.63	+ 6%
Science	56.20	61.50	10.22	5.63	-45%	13.58	13.11	- 3%	20.00	19.77	- 1%
Subs	55.76	59.58	5.25	1.67	-68%	13.93	9.71	-30%	25.06	29.04	+15%

TABLE 3: WADE: Percent correct, percent seen errors, percent sense errors, and percent score errors. The changes (%Δ) from OLD to MIXED are also given; here, negative changes are good (error reduction).

Domain	OLD	+SEEN		+SENSE		OLD vs MIXED SCORE			MIXED
News	22.81	23.87	+ 0%	23.95	+ 1%	23.72	23.86	+ 1%	24.15
EMEA	28.69	31.02	+ 8%	30.59	+ 7%	28.89	30.21	+ 5%	36.60
Science	26.13	27.72	+ 6%	27.29	+ 4%	26.09	28.68	+10%	32.23
Subs	15.10	15.96	+ 6%	16.41	+ 9%	14.99	16.25	+ 8%	18.49

TABLE 4: TETRA: Results on all new domains using OLD and MIXED models (first and last columns), OLD enhanced with seen translations (second), sense translations (third), and scores (fourth), together with percent improvements in terms of BLEU score. Here, positive improvements are good (higher BLEU scores).

correct. As in the news domain, most of the errors are SCORE followed by SENSE and then SEEN. For the other three domains, the two evaluation methods agree that SEEN is a fairly substantial problem. TETRA believes that SENSE is a fairly substantial issue, but WADE does not show this for Science. For SCORE, TETRA detects significant room for improvement, especially for Science.

The large changes in BLEU score found with TETRA are somewhat surprising given how little the phrase tables change in each of these experimental conditions. For News, EMEA, and Science, adding unseen words results in an increase in number of phrase pairs between 0.045% (News) and 0.3% (Science). The sense additions were similarly small: from 0.15% (EMEA) to 0.59% (News). For Subs the story was different: adding unseen words amounted to a growth of 4.2% in phrase table size; sense amounted to 25.1%. In all cases, the size of the score phrase tables was only 0.05% smaller than that of OLD.

At first glance, the WADE and TETRA analyses of the SCORE error type seem to contradict each other. The MIXED systems are *worse* in terms of SCORE (positive deltas, more errors than OLD), but have better BLEU scores. To understand this discrepancy, we must recognize that TETRA analyzes the score errors *in isolation*: by restricting the phrase tables to the *intersection* of the OLD and MIXED domain phrase tables, we remove all score and sense errors. In the WADE analysis however, many errors that "used to be" SEEN errors in the old domain *become* SCORE errors in the new domain.

	Correct		Incorrect			Total
	Cor	Seen	Score	Seen	Sense	
News						
Cor	53.7	0.0	**1.9**	0.0	0.0	55.6
Seen-C	**0.1**	1.7	0.0	0.0	0.0	1.8
Score	**2.2**	0.0	23.6	0.0	0.0	25.8
Seen-I	**0.1**	0.0	0.0	5.4	**0.3**	5.7
Sense	0.0	0.0	**0.2**	0.0	10.8	11.0
Total	56.1	1.7	25.7	5.4	11.1	100
EMEA						
Cor	48.3	0.0	**3.1**	0.0	0.0	51.5
Seen-C	**1.6**	2.8	**0.1**	0.0	0.0	4.5
Score	**5.3**	0.0	13.3	0.0	0.0	18.6
Seen-I	**2.3**	0.0	**0.5**	4.0	**2.5**	9.3
Sense	**2.3**	0.0	**2.6**	0.0	11.3	16.2
Total	59.8	2.8	19.6	4.0	13.8	100
Science						
Cor	49.8	0.0	**3.6**	0.0	0.0	53.3
Seen-C	**1.4**	1.4	0.0	0.0	0.0	2.9
Score	**5.8**	0.0	14.2	0.0	0.0	20.0
Seen-I	**1.8**	0.0	**0.3**	5.6	**2.5**	10.2
Sense	**1.4**	0.0	**1.6**	0.0	10.6	13.6
Total	60.1	1.4	19.8	5.6	13.1	100
Subtitles						
Cor	52.4	0.0	**2.5**	0.0	0.0	54.8
Seen-C	**0.6**	0.3	0.0	0.0	0.0	0.9
Score	**4.5**	0.0	20.6	0.0	0.0	25.1
Seen-I	**1.1**	0.0	**0.5**	1.7	**2.0**	5.3
Sense	**0.8**	0.0	**5.4**	0.0	7.7	13.9
Total	59.3	0.3	29.0	1.7	9.7	100

TABLE 5: Percent of WADE annotation changes moving from OLD (rows) to MIXED (columns) models, for each domain. Non-zero off-diagonals are bolded. Seen-C indicates Freebies, and Seen-I indicates unseen words that were mistranslated.

To see the full picture, we must look at how the different error categories *change* from the OLD system to the MIXED system in WADE. This is shown in Table 5. In this table, the rightmost column contains the total percentage of errors in the OLD systems; the rows labeled *Total* show the total percentage of errors in the MIXED systems; the remaining cells these errors changing from OLD to MIXED. For the news domain, the OLD system has 25.8% SCORE errors. Of those, 2.2% are fixed in the MIXED system.

For the three domains of interest (all except news), addressing SEEN errors can be substantially

helpful, in terms of both BLEU score and the fine-grained distinctions considered by WADE. The more interesting conclusion, however, is that simply bringing in new words isn't enough. Table 5 shows that in these three domains there are a substantial number of errors that transition from being SEEN-Incorrect to SENSE-Incorrect. This indicates that besides observing a new word, we must also observe it with all of its correct translations.

Likewise, there is a lot to be gained in BLEU by correcting new SENSE translation errors (essentially the same percentage as for SEEN). But this is harder to solve. We can see in Table 5 that from the SENSE errors of the OLD system, half become correct but the other half become SCORE errors. So giving appropriate scores to the new senses is a challenge. This makes sense: these new sense are now "competing" with old ones, and getting the interpolation right between old and new domain tables is difficult.

For SCORE, the situation is more complicated. Our TETRA analysis clearly indicates that there is room for improvement. But this is based on intersected phrase tables, from which we removed seen and sense distinctions, and in which there is no competition between phrases from the OLD and NEW systems. The WADE analysis shows a positive effect only for Science. The data in Table 5 shows that a lot ($5.8/20$) of the errors are corrected, but we also *introduce* a number of additional errors (3.6% that were correct, 0.3% that were SEEN and 1.6% that were SENSE). Similarly, in the EMEA domain, we fix 5% of 18% of SCORE errors but introduce 2.6% that were new sense errors before, 0.5% that were SEEN errors before, and make 3% additional error on words we got right before. Subs is similar: out of 25% SCORE errors we fix 4.5%, but introduce 0.5% from SEEN and 5.4% from SENSE, and suffer additional error on 2.5% of what we had correct before.

7.2 Qualitative Results

Table 7 shows examples of the French words that WADE frequently identified as incorrectly translated by the OLD system due to SCORE or SEEN but that were correctly translated under the MIXED system.[7] For example, in the Science domain, 'mesures' suffered from SCORE errors under the OLD system. While its correct translation was often 'measurements,' the OLD system preferred its most probable translations ('savings,' 'actions,' 'issues,' and 'provisions.'). Thirty of these error cases were correctly translated by the MIXED system. Similarly, in the Science domain, the French word 'finis,' when it should have been translated as 'finite,' was translated incorrectly due to a sense error 27 times. Its most frequent translations under the OLD system were 'finish,' 'finished,' and 'more.' The MIXED system corrected these sense errors. We omit examples of where seen errors made by OLD were frequently corrected by the MIXED system because they tend to be less interesting. Examples can be found in Daumé III and Jagarlamudi (2011).

We annotate the French test sentences using the Stanford part-of-speech (POS) tagger (Toutanova et al., 2003) and examine which POS categories correspond to the most errors of each type. Using the OLD system, new sense errors in the Subs domain are made on French nouns 40% of the time and on verbs 35% of the time. In EMEA, 51% are nouns and 23% are adjectives; in Science, 51% nouns and 20% adjectives; in News, 46% nouns and 23% verbs. Seen errors show a very similar trend: in the Subs domain 50% are nouns and 25% verbs; In EMEA, 48% are nouns and 37% adjectives; in Science, 46% are nouns and 40% adjectives; in News, 46% are nouns and 28% adjectives. Similarly, for all domains, more score errors are made on source nouns than any other POS category. In summary, we find that most errors correspond to source language nouns, followed by adjectives, except for Subs, where verbs are also commonly mistranslated due to all error types.

Table 6 (left) shows some examples of how TETRA can automatically estimate the errors due to unseen words when moving to a new domain. For example, the OCR error "miie" in the source sentence is correctly translated as "miss" by the enhanced system. The enhanced phrase tables of TETRA can also automatically estimate the errors due to poor lexical choice when moving to a new domain, and can select a more lucid translation term. For example, the enhanced system appropriately selected "shoot" instead of "growth" in the Science example in Table 6 (middle). When TETRA inserts the scores from the new domain into the transla-

7. Complete output lists are available at `http://hal3.name/damt`

Medical	
Inp	en ce qui concerne les indications thérapeutiques pour l'insuffisance cardiaque, le tamm a proposé le texte suivant: ≪ traitement de l'insuffisance cardiaque congestive. ≫
Ref	regarding the therapeutic indications for heart failure, the mah proposed the wording: "treatment of congestive heart failure."
Old	for therapeutic indications for heart failure, tamm proposed the following: treatment of congestive heart failure.
Fix	for therapeutic indications for heart failure, the mah has suggested the following: treatment of congestive heart failure.
Science	
Inp	les résultats à la base de cette hypothése sont révisés.
Ref	findings that form the basis of this hypothesis are reviewed.
Old	the results at the base of this assumption are reviewed.
Fix	the results of this hypothesis are reviewed.
Subtitles	
Inp	Bonne nuit, Mlle Kenton.
Ref	good night, miss kenton.
Old	good night, miie kenton.
Fix	good night, miss kenton.

Medical	
Inp	ce médicament est une solution limpide, incolore à jaune pâle.
Ref	this medicinal product is a clear, colorless to pale yellow solution.
Old	lantus is a clear, colorless to pale yellow.
Fix	this medicine is a solution is clear, colorless to pale yellow.
Science	
Inp	tous les traitements ont augmenté la production de pousses.
Ref	all treatments increased shoot production.
Old	all treatments have increased the production of growth.
Fix	all treatments increased shoot production.
Subtitles	
Inp	Le sexe c'est naturel.
Ref	sex is natural.
Old	the sex is natural.
Fix	sex is natural.

Medical	
Inp	les deux substances actives ont des effets inverses sur la kaliémie.
Ref	the two active substances have inverse effects on plasma potassium.
Old	the two active substances are reverse effects on the kaliémie.
Fix	the two active substances have side inverses on kaliémie.
Science	
Inp	par ailleurs, les constantes d'équilibre sont plus faibles.
Ref	in contrast, the equilibrium constants are lower.
Old	furthermore, the constant balance are lower.
Fix	moreover, the equilibrium constants are lower.
Subtitles	
Inp	Je bouge mieux.
Ref	i move better.
Old	i get better.
Fix	i move better.

TABLE 6: Example MT results obtained by fixing seen errors (left), sense errors (middle) and score errors (right). Includes source, a reference translation, the output of the OLD system and the output obtained via TETRA methodology.

D	#	French	Correct-E	Hansard-E
Score → Correct				
E	21	doit	should	has must needs shall requires
	8	association	combination	partnership association
	6	noms	names	names people nominee speakers
Sc	30	mesures	measurements	savings actions issues provisions
	27	courant	current	knowledge knew heads-up
	26	article	paper	standing clause order section
Su	9	comme	like	as because like akin how sort
	5	maison	house	home house homes head place
	4	fric	money	cash dough money fric bucks loot
Sense → Correct				
E	9	notice	leaflet	informed directions notice
	8	perfusion	infusion	perfusion intravenous
	8	molles	capsules	lax limpid soft weak
Sc	27	finis	finite	finish finished more
	18	jonctions	junctions	junction *(only once)*
	10	substrats	substrates	corn streams area substrata
Su	5	emmerde	fuck	annoying *(only once)*
	3	redites	say	repetitious tell covered again
	3	mec	man	cme guy mec

TABLE 7: For score/sense errors, in (E)MEA, (Sc)ience and (Su)bs, frequent French words that fall into that category (by WADE), as well as the corrected translations and the most frequent OLD translations.

tion tables, the system produces translations that take on the flavor of the new domain, yielding higher BLEU scores. This can be observed in Table 6 (right) where the TETRA-enhanced system used the science-specific word "equilibrium" rather than the political word "balance."

7.3 Results on an Adapted System

To show how WADE can be used on already adapted systems, we performed a simple experiment based on a standard adaptation technique. We used bilingual cross-entropy difference (Axelrod et al., 2011) to quantify the distance between each OLD domain sentence pair and each NEW domain. We selected the top K closest sentences for each domain. For EMEA and Science, we set K to the size of the NEW domain data. For Subs, this would select nearly all of Hansard, so we arbitrarily set $K = 1m$. (We excluded the news domain.) We took this data, concatenated it to the NEW domain data, trained full models, and ran the WADE analysis on their outputs.

The trends across the three domains were remarkably similar. In all, SCORE in the adapted system were lower by around 2% than even the MIXED baseline (as much as 4% for Subs). This is likely because by excluding parts of the OLD domain most unlike the relevant NEW domain, the correct sense is observed more often. However, this comes at a price: SENSE and SEEN errors go up about 1% or 2% each. This suggests that a more fine-grained adaptation approach might achieve the best of both worlds.

8 Limiting Assumptions

This paper represents a partial exploration of the space of possible assumptions about models and data. We cannot hope to explore the combinatorial explosion of possibilities, and therefore have restricted our analysis to the following settings:

Phrase-based models. All of our experiments are carried out using phrase-based translation, as implemented in the open-source Moses translation system (Koehn et al., 2007) to ensure that they are reproducible. Our methods are easily extended to hierarchical phrase-based models (Chiang, 2007). It is not clear whether the same conclusions would hold: on

the one hand, complex phrasal rules might overfit even more badly than phrases; on the other hand, hierarchical models might have more flexibility to generalize structures.

Translation languages. We only translate from French to English. This well-studied language pair presents several advantages; large quantities of data are publicly available in a wide variety of domains, and standard statistical machine translation architectures yield good performance. Unlike with more distant languages such as Chinese-English, or languages with radically different morphology or word order such as German or Czech, we know that the old-domain translation quality is high, and that translation failures during domain shift can be primarily attributed to domain issues rather than to problems with the SMT system.

Constant old domain. Our old domain is from Hansards, and we only vary our new domain. It would be interesting to consider other datasets as old domains. We deliberately only use the Hansard data: based on its size and scope, we assume that it yields the most general of our SMT systems.

Monolingual new-domain data. We assume that we always have access to monolingual English data in the new domain for learning a domain-specific language model. Our focus is on the effect of the translation model; the effect of adapting language models has been studied previously (see §3). Without access to a new domain language model, the effect of unseen words and words with new senses is likely to be dramatically underestimated, because their translations are likely to be "thrown out" by an old-domain LM. Moreover, since SCORE errors conflate language model and translation model scores, using a new-domain language model lets us mostly isolate the effect of the translation model.

Parallel new-domain data for tuning. We assume that we always have access to a *small* amount of parallel data in the new domain, essentially for the purpose of running parameter tuning. Without this, one would not even be able to evaluate the performance of one's system, typically a non-starter.

Automatic word alignments for WADE WADE is fundamentally based upon word alignments, so alignment errors may affect its accuracy. Such errors are obvious in manually inspecting sentence triples using the visualizer. When developing this tool, we checked that alignment noise does not invalidate conclusions drawn from WADE counts. In order to estimate how much alignment errors affect WADE, a French speaker manually corrected the word alignments for 955 EMEA test set sentences. The analyses based on manual experiments show fewer errors overall, but the erroneous annotations appear to be randomly distributed among all categories (details ommited for space). As a result, we believe that WADE yields results which are informative despite the inevitable automatic alignment errors. In particular, because alignments between a test and reference set are held constant in a system comparison, such errors should impact all analyses in the same way.

9 Discussion

Translation performance degrades dramatically when migrating an SMT system to a new and different domain. Our work demonstrates that the majority of this degradation in performance is due to SEEN and SENSE errors: namely, unknown source-language words and known source-language words with unknown translations. This result holds in all domains we studied, except for news, in which there appears to be little adaptation influence at all (especially after spelling normalization).

Our two analysis methods: WADE (Section 5.1) and TETRA analysis (Section 5.2), are both lenses on the adverse affects of domain mismatch. Using WADE, we are able to pinpoint precise translation errors and their sources. This could be extended to more nuanced, human-assisted, analysis of adaptation effects. WADE also "labels" translations with different error types, which could be used to train more complex models. Using TETRA, we are able to see how these errors affect overall translation performance. In principle, this performance could be *any* measure, including human assessment. We started with the BLEU metric since it is most widely used in the community. One point of possible improvement would be to replace exact string match in WADE, and BLEU in TETRA, with metrics that are more morphologically or semantically informed.

Error analysis opens the door to building adapted machine translation systems that directly target spe-

cific error categories. As we have seen, most existing domain adaptation techniques in MT aim to improve translation quality in general, and are accordingly evaluated using corpus-level metrics such as BLEU. Our intuitive finer-grained analysis suggests that finer-grained models might be better suited to understanding and comparing the errors made by adapted and unadapted systems. We have shown that considering the S^4 taxonomy is important: improving coverage, for example, does not necessarily improve translation quality. Translation candidates must also be complete and must be scored correctly. Our techniques provide an intuitive way to understand the effectiveness of new MT domain adaptation approaches.

Acknowledgments We gratefully acknowledge the support of the 2012 JHU Summer Workshop and NSF Grant No 1005411, as well as the NRC for Marine Carpuat, and DARPA CSSG Grant D11AP00279 for Hal Daumé III. We would like to thank the entire DAMT team (`http://hal3.name/damt/`) and Sanjeev Khudanpur for their invaluable help and suggestions, as well as all the reviewers for their insightful feedback.

References

Sankaranarayanan Ananthakrishnan, Rohit Prasad, and Prem Natarajan. 2011. On-line language model biasing for statistical machine translation. In *Proceedings of the Conference of the Association for Computational Linguistics (ACL)*.

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

Nguyen Bach, Fei Huang, and Yaser Al-Onaizan. 2011. Goodness: A method for measuring machine translation confidence. In *Proceedings of the Conference of the Association for Computational Linguistics (ACL)*.

Pratyush Banerjee, Sudip Kumar Naskar, Johann Roturier, Andy Way, and Josef van Genabith. 2012. Translation quality-based supplementary data selection by incremental update of translation models. In *Proceedings of the International Conference on Computational Linguistics (COLING)*.

Arianna Bisazza, Nick Ruiz, and Marcello Federico. 2011. Fill-up versus interpolation methods for phrase-based SMT adaptation. *International Workshop on Spoken Language Translation (IWSLT)*.

John Blatz, Erin Fitzgerald, George Foster, Simona Gandrabur, Cyril Goutte, Alex Kulesza, Alberto Sanchis, and Nicola Ueffing. 2004. Confidence estimation for machine translation. In *Proceedings of the International Conference on Computational Linguistics (COLING)*.

Marine Carpuat and Dekai Wu. 2007. Improving statistical machine translation using word sense disambiguation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Marine Carpuat, Hal Daumé III, Katharine Henry, Ann Irvine, Jagadeesh Jagarlamudi, and Rachel Rudinger. 2013. SenseSpotting: Never let your parallel data tie you to an old domain. In *Proceedings of the Conference of the Association for Computational Linguistics (ACL)*.

Yin-Wen Chang and Michael Collins. 2011. Exact decoding of phrase-based translation models through lagrangian relaxation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Colin Cherry and George Foster. 2012. Batch tuning strategies for statistical machine translation. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

David Chiang. 2007. Hierarchical phrase-based translation. *Computational Linguistics*, 33(2):201–228.

Hal Daumé III and Jagadeesh Jagarlamudi. 2011. Domain adaptation for machine translation by mining unseen words. In *Proceedings of the Conference of the Association for Computational Linguistics (ACL)*.

Kevin Duh, Katsuhito Sudoh, and Hajime Tsukada. 2010. Analysis of translation model adaptation in statistical machine translation. In *Proceedings of the International Workshop on Spoken Language Translation (IWSLT)*.

George Foster and Roland Kuhn. 2007. Mixture-model adaptation for SMT. In *Proceedings of the Workshop on Statistical Machine Translation (WMT)*.

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

Zhengxian Gong, Min Zhang, and Guodong Zhou. 2011. Cache-based document-level statistical machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Barry Haddow and Philipp Koehn. 2012. Analysing the effect of out-of-domain data on SMT systems. In *Proceedings of the Workshop on Statistical Machine Translation (WMT)*.

Almut Silja Hildebrand, Matthias Eck, Stephan Vogel, and Alex Waibel. 2005. Adaptation of the translation

model for statistical machine translation based on information retrieval. In *Proceedings of the Conference of the European Association for Computational Linguistics (EACL)*.

Ann Irvine, Chris Quirk, and Hal Daumé III. 2013. Monolingual marginal matching for translation model adaptation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Philipp Koehn and Josh Schroeder. 2007. Experiments in domain adaptation for statistical machine translation. In *Proceedings of the Workshop on Statistical Machine Translation (WMT)*.

Philipp Koehn, Amittai Axelrod, Alexandra Birch, Chris Callison-Burch, Miles Osborne, David Talbot, and Michael White. 2005. Edinburgh system description for the 2005 IWSLT speech translation evaluation. In *Proceedings of 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, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the Conference of the Association for Computational Linguistics (ACL)*.

Patrik Lambert, Holger Schwenk, and Frédéric Blain. 2012. Automatic translation of scientific documents in the HAL archive. In *Proceedings of the International Conference on Language Resourcesand Evaluation (LREC)*.

Thomas Lavergne, Alexandre Allauzen, Hai-Son Le, and François Yvon. 2011. LIMSI's experiments in domain adaptation for IWSLT11. In *Proceedings of the International Workshop on Spoken Language Translation (IWSLT)*.

Yajuan Lu, Jin Huang, and Qun Liu. 2007. Improving statistical machine translation performance by training data selection and optimization. In *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*.

Saab Mansour, Joern Wuebker, and Hermann Ney. 2011. Combining translation and language model scoring for domain-specific data filtering. In *Proceedings of the International Workshop on Spoken Language Translation (IWSLT)*.

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

Rada Mihalcea, Ravi Sinha, and Diana McCarthy. 2010. SemEval-2010 Task 2: Cross-Lingual Lexical Substitution. In *Proceedings of the 5th International Workshop on Semantic Evaluation*.

Jan Niehues and Alex Waibel. 2010. Domain adaptation in statistical machine translation using factored translation models. In *Proceedings of the European Association for Machine Translation (EAMT)*.

Maja Popović and Hermann Ney. 2011. Towards automatic error analysis of machine translation output. *Computational Linguistics*, 37(4).

Majid Razmara, George Foster, Baskaran Sankaran, and Anoop Sarkar. 2012. Mixing multiple translation models in statistical machine translation. In *Proceedings of the Conference of the Association for Computational Linguistics (ACL)*.

Rico Sennrich. 2012. Perplexity minimization for translation model domain adaptation in statistical machine translation. In *Proceedings of the Conference of the European Association for Computational Linguistics (EACL)*.

Jörg Tiedemann. 2009. News from OPUS - A collection of multilingual parallel corpora with tools and interfaces. In N. Nicolov, K. Bontcheva, G. Angelova, and R. Mitkov, editors, *Recent Advances in Natural Language Processing (RANLP)*.

Jörg Tiedemann. 2010. To cache or not to cache ? experiments with adaptive models in statistical machine translation. In *Proceedings of the ACL Workshop on Statistical Machine Translation and Metrics (MATR)*.

Kristina Toutanova, Dan Klein, Christopher Manning, and Yoram Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *NAACL*.

David Vilar, Jia Xu, Luis Fernando D'Haro, and Hermann Ney. 2006. Error analysis of statistical machine translation output. In *Proceedings of the International Conference on Language Resourcesand Evaluation (LREC)*.

Stephan Vogel, Hermann Ney, and Christoph Tillmann. 1996. HMM-based word alignment in statistical translation. In *Proceedings of the International Conference on Computational Linguistics (COLING)*.

Guillaume Wisniewski, Alexandre Allauzen, and François Yvon. 2010. Assessing phrase-based translation models with oracle decoding. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Bing Zhao, Matthias Eck, and Stephan Vogel. 2004. Language model adaptation for statistical machine translation with structured query models. In *Proceedings of the International Conference on Computational Linguistics (COLING)*.