Evaluation of Machine Translation Performance Across Multiple Genres and Languages

Marlies van der Wees[1], Arianna Bisazza[2], Christof Monz[1]
[1]Informatics Institute, University of Amsterdam
[2]Leiden Institute of Advanced Computer Science, Leiden University

Abstract

In this paper, we present evaluation corpora covering four genres for four language pairs that we harvested from the web in an automated fashion. We use these multi-genre benchmarks to evaluate the impact of genre differences on machine translation (MT). We observe that BLEU score differences between genres can be large and that, for all genres and all language pairs, translation quality improves when using four genre-optimized systems rather than a single genre-agnostic system. Finally, we train and use genre classifiers to route test documents to the most appropriate genre systems. The results of these experiments show that our multi-genre benchmarks can serve to advance research on text genre adaptation for MT.

Keywords: Machine translation, parallel benchmarks, text genres, genre adaptation

1. Introduction

Text genre differences have shown to affect the output quality of statistical machine translation (SMT) systems: SMT systems trained on one genre often achieve poor performance when used for translating another genre (Foster and Kuhn, 2007; Matsoukas et al., 2009; Wang et al., 2012, among others). In addition, even if different genres in a test set are both present in equal amounts in the bilingual training data, performance differences between the test genres can be large, mostly due to poor model coverage for certain genres (van der Wees et al., 2015a; van der Wees et al., 2015b).

In this paper, we evaluate the impact of genre differences on phrase-based SMT for a diverse set of language pairs, covering both commonly and rarely studied language pairs. For common language pairs, parallel training data is abundant but limited to a few genres such as parliamentary and legal proceedings. For low-resource languages the situation is—by definition—much worse, with very few to no bilingual corpora available. To alleviate this problem, we present in this paper novel parallel training and evaluation corpora covering four genres for four language pairs that we automatically harvested from the web.

Next, we evaluate the usefulness of the newly collected bilingual resources by exploiting them for genre adaptation of SMT systems. Most existing adaptation approaches depend on the availability of provenance information and make the strong assumption that a translation task has known domain, genre or topic that is exploited to adapt the system (Matsoukas et al., 2009; Foster et al., 2010; Bisazza and Federico, 2012; Chen et al., 2013; Chen et al., 2014; Kobus et al., 2016; Sennrich et al., 2016; Freitag and Al-Onaizan, 2016; Chu et al., 2017, among others). While this is a fair assumption in a controlled research setting, it is less realistic in real world applications, such as general-purpose online MT services. In this paper, we provide the SMT system with a test document of unknown origin, and we show that we can use automatic genre classification to guide each test document to the most appropriate pre-trained system. While similar setups have been used in previous work (Xu et al., 2007; Banerjee et al., 2010; Pecina et al., 2011; Wang et al., 2012; Pecina et al., 2015), we are the first to extend this setup to four genres and four language pairs. Finally, we show that an adaptation method based on automatic classifiers also improves translation quality for genres with no parallel training data available.

2. Multi-genre benchmarks

In this section, we describe the construction of multi-genre corpora for four language pairs and four genres, which we obtained using an automated web-harvesting process.

2.1. Language pairs and genres

While most research in MT is evaluated on a small number of well resourced language pairs and domains or genres, we opt for a more balanced distribution of source languages that allows us to measure to what extent our findings for common language pairs generalize to languages with limited resources. We therefore evaluate our experiments in this paper on the following language pairs: *Arabic→English*, *Chinese→English*, *Bulgarian→English*, and *Persian→English*. For each of these language pairs we consider four different genres: *news*, as it can be found in (online) newspapers and in transcripts of broadcast news; *editorial*, covering Op-Ed pieces in (online) newspapers, that represent a subjective, and unlike news less matter-of-fact point of view; *colloquial*, covering informal conversation such as blog comments and Internet forum discussions; and *speech*, covering speeches for which transcripts are available such as TED talks and other public speeches.

2.2. Benchmark construction

For the language pairs and genres of interest, we collect parallel corpora from the web from twenty different websites, each covering at least one of our genres of interest.[1] All websites contain manual translations at the sentence level

[1]Sixteen websites contain news documents, six websites contain editorial documents, eight websites contain colloquial documents, and three websites contain speech transcripts.

	Train set		Dev set		Test set	
Genre	Lines	Tokens	Lines	Tokens	Lines	Tokens
Colloquial	273K	8.9M	1.5K	77.3K	1.5K	73.0K
Editorial	156K	4.7M	1.5K	45.6K	1.5K	47.3K
News	600K	18.0M	1.5K	50.4K	1.5K	48.1K
Speech	140K	3.4M	1.5K	35.7K	1.5K	38.7K
Total	1.2M	35.0M	6.0K	209K	6.0K	207K

(a) Arabic→English data.

	Train set		Dev set		Test set	
Genre	Lines	Tokens	Lines	Tokens	Lines	Tokens
Colloquial	55K	1.7M	1.5K	42.5K	1.4K	35.8K
Editorial	370K	10.2M	1.5K	43.1K	1.5K	42.6K
News	584K	16.4M	1.5K	39.2K	1.5K	35.8K
Speech	146K	3.3M	1.5K	42.6K	1.5K	37.5K
Total	1.2M	31.6M	6.0K	169K	5.9K	152K

(b) Chinese→English data.

	Train set		Dev set		Test set	
Genre	Lines	Tokens	Lines	Tokens	Lines	Tokens
Colloquial	–	–	–	–	1.4K	33.9K
Editorial	–	–	–	–	178	5.1K
News	215K	5.3M	1.2K	30.2K	2.0K	49.5K
Speech	206K	3.9M	1.2K	22.5K	2.0K	44.6K
Total	422K	9.2M	2.4K	52.7K	5.6K	133K

(c) Bulgarian→English data.

	Train set		Dev set		Test set	
Genre	Lines	Tokens	Lines	Tokens	Lines	Tokens
Colloquial	629K	16.4M	1.5K	40.3K	1.5K	37.7K
Editorial	–	–	–	–	600	19.4K
News	618K	16.8M	1.5K	44.5K	1.5K	47.4K
Speech	119K	2.5M	1.5K	31.2K	1.5K	35.6K
Total	1.4M	35.7M	4.5K	116K	5.1K	140K

(d) Persian→English data.

Table 1: Specifications of the harvested multi-genre training, development and test sets for four language pairs. Tokens are counted on the English side. We make the evaluation corpora available at `http://ilps.science.uva.nl/resources/genre-benchmarks`.

between English and one or more other languages. Unfortunately it is not always known which language is the original language and which language is a translation, especially for user comments, which may be written in any of a website's supported language.

Collection of the parallel data is done in an automated fashion. For each website, we provide URLs of one or more overview pages (e.g., a sitemap or an archive page) and then extract from it a list of URLs containing actual text documents. Since all websites support at least English, we start with collecting English documents, regardless of the original language of the websites' documents. Next, we identify translations of these English pages by (i) following direct links, such as 'read this article in Arabic,' or (ii) replacing the language abbreviation in the url, e.g., replacing '`en-US`' with '`ar`'.

To determine the genre of each text we use categories indicated on the respective website. For example, websites that support news articles and user comments (i.e., the genres news and colloquial), have clear website sections for these different genres.

While not all language-genre combinations are equally common, we can construct at least a translation test set for each of the four genres in each of the four language pairs. To do so, we first create sentence-parallel corpora using a combination of Moore's sentence alignment (Moore, 2002) and Champollion sentence alignment (Ma, 2006). We then organize the collected bilingual data into training, development, and test sets, such that each portion contains documents from non-overlapping time periods.[2] We tokenize all Arabic data using MADA (Habash and Rambow, 2005), segment the Chinese data following (Tseng et al., 2005), and use a simple in-house tokenizer for the other languages. The total numbers of foreign→English sentence pairs for the four genres and four language pairs are listed in Tables 1a–1d. In addition, Table 2 shows English example sentences for each of the four genres.

Genre	Example sentence(s)
Colloquial	Ministers should be sitting and attending the oath, like in Italy.
Editorial	This may sound like pie in the sky, but we have already tasted it in Africa, where Sierra Leone's agenda for prosperity 2013–2017 and the Liberia Vision 2030 exemplify the potential of such programs.
News	She is not only the first Saudi woman to ever attempt the climb but also the youngest Arab to make it to the top of the world's highest peak.
Speech	These are just a few of the milestones of recent progress. I have another reason to be optimistic. I know global health is guided by the right values.

Table 2: English example sentences for four genres in the web-harvested evaluation corpora.

3. Evaluating genre differences in SMT

In this section, we use our newly assembled resources to evaluate SMT performance across different genres and language pairs.

[2]The benchmarks are available for download at `http://ilps.science.uva.nl/resources/genre-benchmarks`.

Test genre	Baseline	SMT system optimized for				Combined best BLEU
		Coll.	Edit.	News	Speech	
Coll.	11.7	**13.8**	10.8	11.7	11.2	**17.9** (+1.1)
Edit.	22.6	19.6	**23.5**	21.6	21.0	
News	22.6	20.2	21.7	**23.2**	21.2	
Speech	11.5	11.5	11.1	11.0	**11.7**	
All	16.8	16.6	16.4	16.6	16.0	

(a) Arabic→English results.

Test genre	Baseline	SMT system optimized for				Combined best BLEU
		Coll.	Edit.	News	Speech	
Coll.	11.4	**11.6**	11.3	10.7	11.3	**13.9** (+0.5)
Edit.	15.5	14.9	**16.3**	14.6	14.3	
News	13.3	12.8	13.3	**13.5**	12.4	
Speech	12.8	12.5	12.5	12.1	**13.9**	
All	13.4	13.1	13.4	12.7	13.2	

(b) Chinese→English results.

Test genre	Baseline	SMT system optimized for				Combined best BLEU
		Coll.	Edit.	News	Speech	
Coll.	**29.1**	–	–	28.0	28.1	**33.4** (+0.6)
Edit.	24.7	–	–	**25.4**	21.3	
News	39.8	–	–	**40.4**	34.7	
Speech	27.4	–	–	25.8	**28.4**	
All	32.8	–	–	31.9	30.5	

(c) Bulgarian→English results.

Test genre	Baseline	SMT system optimized for				Combined best BLEU
		Coll.	Edit.	News	Speech	
Coll.	22.4	**22.5**	–	20.9	21.5	**22.3** (+0.4)
Edit.	**15.7**	15.2	–	15.6	15.1	
News	24.2	22.3	–	**24.3**	23.0	
Speech	21.3	19.5	–	20.7	**22.6**	
All	21.9	20.8	–	21.3	21.5	

(d) Persian→English results.

Table 3: Translation quality in BLEU of four test genres using genre-optimized systems and a genre-agnostic baseline. Best results for each test set genre are boldfaced. 'Combined best BLEU' indicates the overall BLEU score when combining the bold-faced results of all test genres in a single test set, followed by the difference with the genre-agnostic system.

3.1. Experimental setup

All SMT systems in this paper are trained using an in-house phrase-based SMT system similar to Moses (Koehn et al., 2007). To train our systems, we use our web-crawled corpora, supplemented with commonly used training data, if available: LDC corpora for Arabic→English and Chinese→English, and Europarl data (Koehn, 2005) for Bulgarian→English. In addition, we use a 5-gram language model that linearly interpolates various Gigaword subcorpora with the English sides of the bilingual training corpora. To evaluate the effect of our new bilingual resources, we do not vary the language model between experiments.

In order to create genre-specific SMT systems, we have to adequately use the available data. Simply concatenating the different corpora yields a general SMT system that performs reasonably well across a variety of genres, i.e., those covered in the training data, but is not optimal for each individual genre. Since we aim to create genre-specific systems, we use the fill-up technique proposed by Bisazza et al. (2011), in which we combine models trained on a particular genre with models trained on the remaining training corpora. Using this model combination technique, an additional feature is learned that favors genre-specific models, and 'backs off' to additional (out-of-genre) models for phrases that are unseen in the genre of interest. For instance, to train our news translation system, we train two phrase tables: one using all news data and one using all non-news data. We use the latter to complement the first with phrase pairs that are not covered in the first.

Following the above strategy, we can train genre-specific systems for all genres for which we have training data. Genres not covered in the training data have to be translated using a system trained on a mixture of genres or on one of the other genre-specific systems. For example, editorial Persian→English data is scarce, so for Persian editorial documents we have to resort to our colloquial, news, speech or mixed system. In addition to using the fill-up approach, we tune each genre-specific system on a development set covering only the genre of interest.

3.2. Results

Tables 3a–3d show the translation quality results for all language pairs. For each language pair, we measure case-insensitive BLEU (Papineni et al., 2002) for our four test genres with the available genre-specific systems as well as the genre-agnostic system. Note that some Arabic→English and Chinese→English BLEU scores might be lower than those reported in literature since our test data contains only a single reference translation.

The results confirm our expectation that the various test set genres benefit from being translated using a genre-optimized system rather than using a general system: generally, the highest BLEU scores are located on the diagonal of each table. In cases where no genre-specific system is available, we see that the best results are mostly obtained using the general system rather than a system optimized for a different genre.

4. Genre adaptation using automatic classifiers

We observed that translation quality is usually best when translating each genre using its respective genre-specific baseline system. This motivates the hypothesis that translation of a mixture-of-genre test set can be improved by using a genre classifier, which routes test sentences or documents to the most appropriate MT system. Adapting an MT system using this strategy involves two steps: training accurate

Genre	Arabic→English system		
	Genre-agnostic	Manual oracle	Genre-classified
Colloquial	11.7	13.8▲	13.8▲
Editorial	22.6	23.5▲	23.5▲
News	22.6	23.2▲	23.2▲
Speech	11.5	11.7	11.6
Overall	16.8	17.9▲	17.8▲

(a) Arabic→English results.

Genre	Chinese→English system		
	Genre-agnostic	Manual oracle	Genre-classified
Colloquial	11.4	11.6	11.5
Editorial	15.5	16.3▲	16.3▲
News	13.3	13.5	13.6
Speech	12.8	13.9▲	14.0▲
Overall	13.4	13.9▲	13.9▲

(b) Chinese→English results.

Genre	Bulgarian→English system		
	Genre-agnostic	Manual oracle	Genre-classified
Colloquial	29.1	29.1	28.6▼
Editorial	24.7	25.4△	25.4△
News	39.8	40.4▲	40.4▲
Speech	27.4	28.4▲	28.4▲
Overall	32.8	33.4▲	33.1▲

(c) Bulgarian→English results.

Genre	Persian→English system		
	Genre-agnostic	Manual oracle	Genre-classified
Colloquial	22.4	22.5	22.5
Editorial	15.7	15.7	15.6
News	24.2	24.3	24.2
Speech	21.3	22.6▲	22.6▲
Overall	21.9	22.3▲	22.1▲

(d) Persian→English results.

Table 4: Translation results in BLEU of baseline and genre-adapted systems. Manual oracle results are combined from several genre-optimized systems using manual genre labels of the test documents, see Tables 3a–3d. Statistically significant differences are indicated with △ or ▽ at the $p \leq 0.05$ level and with ▲ or ▼ at the $p \leq 0.01$ level.

genre classifiers (§4.1.) and incorporating these classifiers into an end-to-end MT pipeline (§4.2.).

Adaptation using automatic classifiers has been applied in a few previous efforts (Xu et al., 2007; Banerjee et al., 2010; Wang et al., 2012). However, all three methods limit their application to one language pair covering two provenance-based domains, which are typically very distinct, e.g., patents versus 'generic'. To the best of our knowledge, we are the first to extend this setup up to four genres and four language pairs, where documents within a genre originate from a variety of sources.

4.1. Training genre classifiers

Since we apply our genre classifiers to different languages, we aim at developing a single classification procedure that can be used on any source document regardless of the language it is written in. For this purpose, we apply our experiments to three languages: Arabic, Chinese, and English. To train the classifiers we randomly select documents from the training data listed in Tables 1a and 1b. The complete selection comprises 1,000 documents per genre, thus enforcing equal prior classification probabilities for all genres.

We train genre classifiers with Support Vector Machines (SVM) with linear kernels, using the WEKA data mining software (Hall et al., 2009). As our features, we use the union of the 500 most common words per genre. We do not remove stopwords since they have a high potential to distinguish between various genres, which is long known in text genre classification literature (Karlgren and Cutting, 1994; Kessler et al., 1997; Stamatatos et al., 2000; Dewdney et al., 2001). Using this classifier-feature combination, the classification accuracy on the documents in the test portion of our web-crawled corpora is 97.0%, 83.9%, and 88.1% for Arabic, Chinese, and English, respectively.

4.2. Genre adaptation experiments

Armed with accurate genre classifiers, we next classify for each document in the test set its genre, and guide it to the most appropriate SMT system. Note that while we do have access to the true genre labels in this controlled research scenario, we intentionally mimic a more realistic situation in which an incoming test document has unknown origin.

Figures 4a–4d show the translation quality in BLEU for all language pairs using (i) a genre-agnostic baseline system trained and tuned on a mixture of genres, (ii) several genre-specific systems which we combine manually and refer to as our 'oracle' system, and (iii) several genre-specific systems which we combine using automatic genre classifiers. We measure statistical significance with respect to the genre-agnostic baseline using approximate randomization (Riezler and Maxwell, 2005), reporting significant differences at the $p \leq 0.05$ (△/▽) or $p \leq 0.01$ (▲/▼) level.

For Arabic→English and Chinese→English (Tables 4a and 4b, respectively), we train our classifiers on four genres with a balanced prior distribution. Our Arabic genre classifier achieves near-perfect classification accuracy (97%), which is reflected by BLEU scores that are very similar to the oracle system. Our best Chinese genre classifier yields lower accuracy (84%), however BLEU scores of the genre-classified system do not suffer from this sub-optimal classification performance. On closer inspection we see that some documents actually benefit from being translated by a different genre-optimized system, for example the Chinese news documents classified as editorial improve with 0.4 BLEU if translated using the editorial system.

Next, we look at the languages for which not all genres are covered in the training data. Consequently, we can only train classifiers for two (Bulgarian) or three (Persian) of the

genres in the test set. For the remaining test genres, the predicted genre will be one of the genres in the training data, and translation is performed using the corresponding genre-specific system. Note that the genre-agnostic baseline system is never recommended based on classifier predictions, despite sometimes being the best option.

Table 4c shows the end-to-end results for Bulgarian→English translation. The Bulgarian genre classifier achieves 100% accuracy on news and speech. The editorial test documents are all classified as news, which is advantageous for the SMT output quality. Genre predictions for the colloquial test documents are distributed evenly over news and speech, achieving a BLEU score of 28.6. While the genre-agnostic system performs better (29.1), the result using an automatic classifier is superior to translating all colloquial documents with either the news (28.1) or the speech (28.0) system. This finding indicates that automatic genre classification can even be profitable if no training data for a given genre is available.

Finally, Table 4d shows the end-to-end results for Persian→English translation. The Persian classifier achieves 90% accuracy on the genres covered in the training data; colloquial, news, and speech. However, BLEU scores using the genre-agnostic and the genre-optimized systems are very similar for all genres except speech. Improvements using the genre-classified system are therefore small.

5. Conclusions and future work

In this paper, we have presented parallel evaluation corpora covering four genres for four language pairs. We used these multi-genre benchmarks to show that BLEU differences between genres can be large and that, for all genres and all language pairs, translation quality improves when using four genre-optimized systems rather than a single genre-agnostic system. Finally, we trained and used genre classifiers to route test documents to the most appropriate genre systems, and showed that this setup can be used to successfully adapt SMT systems to four different genres, even for genres with no available parallel training data.

While experiments in this paper are limited to phrase-based SMT, they can also be applied to neural MT, for which current research is still limited to a few language pairs and domains.

Acknowledgements

This research was funded in part by NWO under project numbers 639.022.213, 612.001.218, and 639.021.646.

6. Bibliographical References

Banerjee, P., Du, J., Li, B., Kumar Naskar, S., Way, A., and van Genabith, J. (2010). Combining multi-domain statistical machine translation models using automatic classifiers. In *Proceedings of the Ninth Conference of the Association for Machine Translation in the Americas*.

Bisazza, A. and Federico, M. (2012). Cutting the long tail: Hybrid language models for translation style adaptation. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 439–448.

Bisazza, A., Ruiz, N., and Federico, M. (2011). Fill-up versus interpolation methods for phrase-based SMT adaptation. In *Proceedings of the 8th International Workshop on Spoken Language Translation*, pages 136–143.

Chen, B., Kuhn, R., and Foster, G. (2013). Vector space model for adaptation in statistical machine translation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 1285–1293.

Chen, B., Kuhn, R., and Foster, G. (2014). A comparison of mixture and vector space techniques for translation model adaptation. In *Proceedings of the 11th Conference of the Association for Machine Translation in the Americas*, pages 124–138.

Chu, C., Dabre, R., and Kurohashi, S. (2017). An empirical comparison of domain adaptation methods for neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 385–391.

Dewdney, N., VanEss-Dykema, C., and MacMillan, R. (2001). The form is the substance: classification of genres in text. In *Proceedings of the Workshop on Human Language Technology and Knowledge Management*.

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

Foster, G., Goutte, C., and Kuhn, R. (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.

Freitag, M. and Al-Onaizan, Y. (2016). Fast domain adaptation for neural machine translation. *arXiv preprint arXiv:1612.06897*.

Habash, N. and Rambow, O. (2005). Arabic tokenization, part-of-speech tagging and morphological disambiguation in one fell swoop. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 573–580.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009). The weka data mining software: An update. *SIGKDD Explor. Newsl.*, 11(1):10–18.

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

Kessler, B., Numberg, G., and Schütze, H. (1997). Automatic detection of text genre. In *Proceedings of the eighth conference of the European chapter of the Association for Computational Linguistics (EACL)*, pages 32–38.

Kobus, C., Crego, J., and Senellart, J. (2016). Domain control for neural machine translation. *arXiv preprint arXiv:1612.06140*.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. (2007). Moses: Open source toolkit for statistical

machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, pages 177–180.

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. In *Proceedings of the X Machine Translation Summit*, pages 79–86.

Ma, X. (2006). Champollion: A robust parallel text sentence aligner. In *LREC 2006: Fifth International Conference on Language Resources and Evaluation*, pages 489–492.

Matsoukas, S., Rosti, A.-V. I., and Zhang, B. (2009). Discriminative corpus weight estimation for machine translation. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 708–717.

Moore, R. C. (2002). Fast and accurate sentence alignment of bilingual corpora. In *Proceedings of the 5th Conference of the Association for Machine Translation in the Americas*, pages 135–144.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318.

Pecina, P., Toral, A., Way, A., Papavassiliou, V., Prokopidis, P., and Giagkou, M. (2011). Towards using web-crawled data for domain adaptation in statistical machine translation. In *Proceedings of the 15th Conference of the European Association for Machine Translation*, pages 297–304.

Pecina, P., Toral, A., Papavassiliou, V., Prokopidis, P., Tamchyna, A., Way, A., and van Genabith, J. (2015). Domain adaptation of statistical machine translation with domain-focused web crawling. *Language resources and evaluation*, 49(1):147–193.

Riezler, S. and Maxwell, J. T. (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.

Sennrich, R., Haddow, B., and Birch, A. (2016). Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 86–96.

Stamatatos, E., Fakotakis, N., and Kokkinakis, G. (2000). Text genre detection using common word frequencies. In *Proceedings of the 18th conference on Computational linguistics*, pages 808–814.

Tseng, H., Chang, P., Andrew, G., Jurafsky, D., and Manning, C. (2005). A conditional random field word segmenter. In *Proceedings of the fourth SIGHAN workshop on Chinese language Processing*, pages 168–171.

van der Wees, M., Bisazza, A., and Monz, C. (2015a). Five shades of noise: Analyzing machine translation errors in user-generated text. In *Proceedings of the First Workshop on Noisy User-generated Text (WNUT)*, pages 28–37.

van der Wees, M., Bisazza, A., Weerkamp, W., and Monz, C. (2015b). What's in a domain? Analyzing genre and topic differences in statistical machine translation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Learning (ACL-CoNLL)*, pages 560–566.

Wang, W., Macherey, K., Macherey, W., Och, F., and Xu, P. (2012). Improved domain adaptation for statistical machine translation. In *Proceedings of the Tenth Conference of the Association for Machine Translation in the Americas*.

Xu, J., Deng, Y., Gao, Y., and Ney, H. (2007). Domain dependent statistical machine translation. In *Proceedings of the XI Machine Translation Summit*, pages 515–520.