"Hi, how can I help you?" Improving Machine Translation of Conversational Content in a Business Context

Bianka Buschbeck* **Jennifer Mell*** **Miriam Exel** **Matthias Huck**

SAP SE
Dietmar-Hopp-Allee 16, 69190 Walldorf, Germany
firstname.lastname@sap.com

Abstract

This paper addresses the automatic translation of conversational content in a business context, for example support chat dialogues. While such use cases share characteristics with other informal machine translation scenarios, translation requirements with respect to technical and business-related expressions are high. To succeed in such scenarios, we experimented with curating dedicated training and test data, injecting noise to improve robustness, and applying sentence weighting schemes to carefully manage the influence of the different corpora. We show that our approach improves the performance of our models on conversational content for all 18 investigated language pairs while preserving translation quality on other domains – an indispensable requirement to integrate these developments into our MT engines at SAP.

1 Introduction

At SAP we build machine translation systems to cope with a huge translation volume, covering product localization and translation of documentation, training materials or support instructions for up to 85 languages. We usually train mixed-domain neural machine translation (MT) engines, whose training input consists of a multitude of data sources including the contents of the company-internal translation memories from various domains. The resulting MT systems produce high-quality technical translations but have difficulties generating appropriate output for conversational content, required for multilingual chatbots or product support chats. For an enhanced user experience, it becomes more and more important that our customers can communicate in the language of their choice. Therefore, we started a project to improve translation quality for business-related conversations. This includes better translations, but also more robustness towards typos and poor input quality. We focus on methods that allow for improvements in one area without degrading quality in others, since we aim to deliver a single model per language pair.

The remainder of this paper is structured as follows. Section 2 introduces the particular challenges to be addressed when training an MT system for corporate conversational content. The methods we explored are presented in Section 3 and experimental results on selected language pairs can be found in Section 4. Section 5 contains results on 18 language pairs of major interest at SAP using the final experimental configuration. We point out open research questions in Section 6. Related problems and approaches are highlighted in Section 7 before we summarize our findings in Section 8.

2 Challenges

Our baseline MT system is built on large amounts of technical documents and UI-style training data with the consequence that it performs unsatisfactorily in business conversations, where the content is technical, but style is conversational. An example conversation is given in Table 1 (column *Original Conversation in English*). It contains a variety of features that are common in conversational content, with major challenges posed by incomplete or

*Equal contribution.

© 2022 The authors. This article is licensed under a Creative Commons 3.0 licence, no derivative works, attribution, CC-BY-ND.

Agent	Original Conversation in English	Baseline MT	Improved MT
User	hello anyone	Bonjour à tous	Bonjour à tous
User	why my gdm is not working	**pourquoi** mon gdm ne **fonctionne** pas	**Pourquoi** mon gdm ne **marche** pas **?**
User	it says .service file is not there	**cela indique** que le fichier .service **n'existe pas.**	**Ça dit** que le fichier .service **n'est pas là**
Support	what are you doing exactly	**que** faites-vous exactement	**Que** faites-vous exactement **?**
Support	where is it exactly	**où c'est** exactement	**Où est-ce** exactement **?**
Support	i didnt evn see the question..	**i didnt voir** la question...	**Je n'ai même pas vu** la question.
User	i have dsl connection my isp provides me and i got device not manaed so i edit.conf file of nm	**i a dsl connexion** mon isp me fournit et **i get device not manaed so i edit .conf file of** nm	**J'ai une connexion dsl que** mon isp me fournit et **je n'ai pas géré l'appareil donc je modifie le fichier .conf de** nm
User	yeah it is	**oui,**	**Oui, c'est ça**
Support	id leave it alone.	**et laissez-le** tranquille.	**Je le laisse** tranquille.
User	i thinking to filter that port	**i penser à** filtrer ce port	**Je pense** filtrer ce port.
Support	i never noticed it being open to the network. just to localhost	**i n'a** jamais remarqué qu'**elle** était **ouverte** au réseau. **Il s'agit simplement d'un hôte local**	**Je n'ai** jamais remarqué qu'**il** était **ouvert** au réseau. **Juste à localhost**
User	ok	OK	OK
Support	so you m ay be worried about nothing	**Vous vous inquiétez donc de ne** rien **voir**	**Donc tu es toujours inquiet pour** rien.
Support	seems its not really an security issue and it makes loookups quicker	**semble qu'il ne s'agit** pas vraiment **d'**un problème de sécurité et **qu'il** accélère les **loookups**	**n'est** pas vraiment un problème de sécurité et accélère les **recherches**
User	thanks	**remerciements**	**Merci**

Table 1: Excerpt of an English conversation (from the Ubuntu Dialogue Corpus (Lowe et al., 2015)) translated to French using the baseline and our improved MT model.

Phenomenon	Examples
Spelling	
Typos	thansk, tanks, thanx
Casing	cpu, i, aws
Spacing	ofcourse, any one, Id o
Lack of punctuation	Hi are you there
Conversational word forms	dunno, gotcha, doin'
Conversational variants	hey, hey hi, hiya, howdy
Abbreviations	
Word/phrase abbreviations	plz, thx, np, omg, ttyl
Letter/number homophones	u r, I c, c u, u 2, some1
Paralinguistic features	
Emoticons	:D ;-) :(
Emotional expressions	uh, hmm, oh, ah, whoa
Emphasis - duplication	no no no, oh noooo
Emphasis - typography	it's URGENT, It broke *EVERYTHING*!
Expletives	damn!, crap, sh*t

Table 2: Typical phenomena in conversational data.

ungrammatical sentences and high contextual dependency. Conversational expressions (*hello anyone*, *thanks*) and syntactic structures such as questions and utterances in first and second person singular are typical of conversational style. Technical documents do not provide a good coverage of these phenomena. Support chats, moreover, exhibit other challenging phenomena that are summarized in Table 2 based on initial exploration of in-domain data. While most of the listed linguistic issues could be corrected, paralinguistic phenomena that are a kind of textual equivalent to verbal prosodic features or facial expressions are more difficult. Emphasis expressed by word or letter duplication or typography are highly language-specific and cannot be easily transferred. Even emoticons are not used in the same way across languages.

3 Methods

In this section, we describe the methods we investigated to address some of these challenges.

3.1 High-quality Parallel Data

The most straightforward way to improve translation quality of conversational content would be adding appropriate training data. However, bilingual data in this domain is hard to find. Even largely conversational datasets, such as OpenSubtitles (Lison and Tiedemann, 2016) are not well suited for this purpose, as business conversations are highly technical.

Thus, we manually select and translate appropriate sentences to enrich our available training data with conversational style segments (Section 2). To collect suitable source segments, we draw on different resources such as support dialogues and expressions used for intents in our chatbots. But the most valuable resource is the Ubuntu Dialogue Corpus (UDC) (Lowe et al., 2015), a pub-

licly available dataset that contains almost one million two-person conversations extracted from Ubuntu technical support chat logs between 2004 and 2015. We create a list of utterances and their frequency from the UDC that helps us extract the following:

- Utterances that cover greetings, agreement, affirmations, refusal, uncertainty, wishes, regrets, hold-on expressions, thanks and responses to them, etc.
- Utterances starting with WH words and inverted questions (*Are you*, *Do you*, *Does that*, etc.), frequent in support dialogues but underrepresented in technical documentation.
- Utterances that contain the pronouns "*I*" and "*you*" to improve first- and second-person coverage.
- Frequent single word utterances, as they are especially problematic.

We mainly focus on short expressions that do not contain vocabulary specific to the UDC. The resulting list of approximately 10,000 English segments is then normalized, since it contains too many variants of the same expression, differing only in spelling, punctuation, and casing that would increase translation costs without resulting in more varied training data. The final corpus consists of 7,000 segments that we have manually translated by our professional translators into the required target languages. Source variations are later created using the methods described in Section 3.4.

3.2 Domain Adaptation

We define as *domain adaptation* the task of optimizing a natural language processing system's parameters towards improved quality on a specific text domain. A text domain typically exhibits particular charateristics wrt. aspects such as genre, topic, style, terminology, and so on. Domain adaptation for MT is an established field of study (Chu and Wang, 2018), with *fine-tuning* nowadays being one of the prevalent paradigms for neural MT models (Freitag and Al-Onaizan, 2016; Huck et al., 2017). In fine-tuning, training of a generic MT model is continued using in-domain data. The pitfalls of this method are overfitting and quality loss on out-of-domain data (Huck et al., 2015; Thompson et al., 2019). We found that *sentence weighting* (Chen et al., 2017; Rieß et al., 2021; Wang et al., 2017) suits our purpose of adapting towards conversational content better while at the same time not sacrificing translation quality on other text domains, thus keeping overall system performance stable. We apply a straightforward up-weighting technique by giving higher instance weights to subsections of the training set which contain conversational content. Experimental results on this will be reported in Section 4.3.

3.3 Error-sensitive Back-translation Scoring

The amount of conversational training data for MT models can be increased by employing synthetic bitext from back-translation (Huck et al., 2011; Schwenk, 2008; Sennrich et al., 2016a). We back-translate the UDC dataset with the aim of benefiting conversational style and vocabulary coverage without harming grammaticality and spelling of MT output. To that end, we first clean the dataset using in-house scripts, resulting in 4.6 million English sentences. We then machine-translate the English sentences into the source languages of the models which we intend to improve, using our existing engines for back-translation in the reverse direction. Experiments are thus only carried out on language directions with English target (Section 4.6).

We assume that grammatical and correctly spelled input sentences result in better back-translations, which in turn will lead to better performance of the final model. Furthermore, we require the final model to produce grammatical sentences despite the training references containing user-generated text. We therefore use Acrolinx[1] to measure the acceptability of a segment in terms of grammaticality and spelling. Acrolinx is AI-powered software that improves the quality and impact of enterprise content. Using a customized version of Acrolinx specialized for the technical support domain, we extract grammaticality, spelling, and clarity scores for every sentence and aggregate them into a sentence-level acceptability score. We further include sentence length into each sentence-level score since exploratory analysis has shown that longer sentences tend to achieve lower Acrolinx scores. The sentence-level scores will be used in Section 4.6 to either filter or weight the back-translated UDC training data.

[1] https://www.acrolinx.com/

3.4 Noise Injection

To improve and assess model robustness beyond the addition of conversational style segments, we inject noise into the in-domain subsets of training and test data. We replicate some typical chat phenomena (Table 2) by injecting noise in the form of (1.) typos, (2.) common chat variants and word forms, (3.) lowercasing and (4.) punctuation removal on the source side only. The required language data for typo injection and generation of chat variants (described below) is only available in English, restricting experiments to language directions with English source. Table 3 gives an overview of all generated variants. They are generated from the unmodified source data, except variants of conversational data (Section 3.1), which are based on the normalized dataset.

For typo generation we apply an approach similar to Shah and de Melo (2020) and compute a model of real-world typos based on a collection of character-level typos found in individual tokens. Typos are grouped into four categories: insertion (ex.: *th*r*ere*), deletion (ex.: *particu*_*ar*), substitution (ex.: *fa*v*ulous*) and transposition (ex.: *cor*c*ect*). For each error category and each character, we calculate probability distributions based on corpus occurrences. They constitute a statistical model of typos in the English language which we refer to as the *typo model*. For details on the computation of the probabilities, please see Shah and de Melo (2020).

For every token in a source sentence, we sample from a token corruption probability (c) to determine whether any noise will be injected. If a token is chosen for noise injection, we iterate over its characters and decide according to a typo probability (t) whether an error will be inserted at the current character. Using the typo model as a noise function, we sample from the calculated probability distributions to generate one of the four types of errors.

We inject spelling errors using two approaches. Simply applying the typo model and method as described above results in the *artificial* variants. Additionally, we inject typos and further filter the generated errors by checking corrupted tokens against token-level typo lists. This yields the *real* variants which are modified with real-world typos only.

Table 4 contains the hyperparameters used to generate three different misspelling levels for both approaches. They are based on preliminary experiments and settings reported by Shah and de Melo (2020). The parameters for the *real* approach were chosen such that, after the restrictive filtering step, the level of noise was comparable to that of the corresponding *artificial* variant. Comparability was assessed via the distribution of typos per sentence and manual checks of the resulting variants. We thus obtain a total of six variants from injecting typos for a single dataset (Table 3, rows 1–6).

Additionally, we create a variant of the dataset where we replace standard language with typical conversational expressions, abbreviations and homophones (Table 3, row 7) using an in-house expression mapping. For example, "*thanks*" is replaced with "*thx*", "*give me*" turns into "*gimme*", "*are you*" becomes "*r u*" etc.

Lastly, we generate three additional variants of the data by lowercasing it and/or removing punctuation (Table 3, rows 8–10).

	Variant
1	Low real typo injection
2	Medium real typo injection
3	High real typo injection
4	Low artificial typo injection
5	Medium artificial typo injection
6	High artificial typo injection
7	Colloquial replacements
8	Lowercasing
9	Punctuation removal
10	Lowercasing and punctuation removal

Table 3: List of generated source-side variants for a single dataset.

	artificial		real	
	c	t	c	t
Low	0.2	0.025	1.0	0.1
Medium	0.3	0.05	1.0	0.2
High	0.5	0.075	1.0	0.3

Table 4: Token corruption probability (c) and typo probability (t) for injecting noise using the typo model.

4 Experiments

We now empirically evaluate the methods introduced in Section 3, with the goal of improving MT quality on conversational content. We focus on conducting detailed experiments and presenting results for two language pairs per method, one being rather close languages, the other rather distant. These are English to French and Japanese (*en–fr*, *en–ja*) for up-weighting and noise injection, and Italian and Japanese to English (*it–en*, *ja–en*) for

back-translation. In Section 5 we will demonstrate that our main findings generalize to other language pairs.

4.1 Experimental Setup

For training we use large amounts of company-internal parallel data that mostly consists of documentation, training materials, UI strings and support instructions. We also utilize some publicly available datasets. The training data amounts to about 25 M parallel segments per language pair. The data is tokenized using a simple tokenization scheme based on whitespace and punctuation, then segmented into subwords using byte-pair encoding (Sennrich et al., 2016b).

We make use of the Marian toolkit (Junczys-Dowmunt et al., 2018) for this investigation. For all our experiments, we use a Transformer network in the standard base configuration (Vaswani et al., 2017) and train it on the training data of the corresponding language pair. The early stopping criterion is computed on a dedicated validation set of 4,000 parallel segments.

4.2 Test Corpora

Targeted changes to MT systems require meaningful test sets to guide experimentation and to measure improvement. As it is hard to find publicly available test data that reflects the technical support dialogue content we are interested in, we created new test sets consisting of customer support dialogues and some dialogues taken from the UDC. In contrast to the conversational training data, we kept the dialogue structure for the test data and selected a total of 21 dialogues, consisting of about 1,000 sentences, that were also translated by professional translators after normalization.

To measure performance on noisy input, we created ten variations of the normalized English source text of the support dialogues using the noise injection techniques introduced in Section 3.4, see Table 3. While we analyzed scores on the individual test set variants in the experimental phase, we will only present results on all variants combined here. Obviously, the impact of the methods on the individual test set variants differs but as we intend to cover different phenomena, the combined score also helps to select the best overall configuration.

We use three groups of test data for in-domain and out-of-domain testing in this study:

	en–fr		en–ja	
Weight	CHRF2	BLEU	CHRF2	BLEU
1	59.4	36.3	41.1	34.1
5	59.5	36.3	41.9	34.8
10	59.8	36.9	42.1	35.2
20	59.9	37.0	42.2	**35.6**
30	59.9	**37.2**	42.1	35.2
40	**60.0**	36.9	**42.3**	35.4
50	59.8	36.9	**42.3**	35.5

Table 5: CHRF2 and BLEU scores on the conversational test set with different weighting of the in-domain corpus. Best results are highlighted in bold.

Conversational comprises the original and normalized support dialogue test sets, their ten variants (Table 3) and two additional related publicly available test sets.

Corporate refers to a set of about 10 test sets with diverse SAP-internal content.

Generic groups together public test sets from news, Wikipedia, UN and EU sources.

Each of these groups contains about 10,000–15,000 test segments, amounting to a total of about 40,000 per language pair. We evaluate using case-sensitive CHRF2 (Popović, 2016) and BLEU (Papineni et al., 2002) and, in view of its better correlation with human judgment (Mathur et al., 2020), rely on CHRF2 for system choice. We report scores averaged over all test sets per group.

4.3 Sentence Weighting Experiments

The amount of conversational training data we have at our disposal is tiny compared to the rest of the training data. It corresponds to 0.02% for en–fr and to 0.06% for en–ja. Our first target is to effectively use the new in-domain training data described in Section 3.1 to adapt the model to the target domain of conversational content. We thus focus initially on conversational test sets, results on out-of-domain test data are reported in Section 4.5.

Instead of fine-tuning, we use sentence weighting, giving the in-domain training data more weight, see Section 3.2. We explore the up-weighting factor empirically (Table 5). A weight of 1 constitutes the baseline. Increasing the weight multiplier yields a small but steady improvement. A factor of 40 delivers the best performance for en–fr and is almost equal to the best CHRF2 for en–ja. For the purpose of applying a common weight setting across language pairs, we keep the factor of 40 fixed for subsequent experiments.

Level	Corpus	Typos		Lc.	Punct.	Colloq.
		real	art.			
0	None	–	–	–	–	–
1	Conv.	✓	–	✓	✓	✓
2	Conv.	✓	✓	✓	✓	✓
3	Conv.	✓	✓	✓	✓	✓
	Tatoeba	✓	✓ (low)	✓	✓	–

Table 6: Configurations of the different noise levels used in noise injection experiments. Conv. denotes the conversational corpus; Lc., Punct. and Colloq. refer to the lowercased, punctuation and colloquial variants; art. abbreviates artificial.

Level	en–fr		en–ja	
	CHRF2	BLEU	CHRF2	BLEU
0	60.0	36.9	42.3	35.4
1	60.7	37.9	42.5	36.3
2	60.8	38.3	42.8	36.8
3	61.4	38.6	43.4	36.9
3 + Tatoeba 3x	**61.5**	**39.1**	**43.5**	**37.4**

Table 7: Results of the noise injection experiments. The conversational corpus has a fixed weight multiplier of 40x. Tatoeba 3x indicates addition of the Tatoeba corpus with a 3x weight multiplier. Best results are highlighted in bold.

4.4 Noise Injection Experiments

As described in Section 3.4, noisy variants are injected into the training and test data on the English source only. The target remains in its original form so that the model learns to correct and translate at the same time. We categorize the noise injection experiments into three levels (Table 6) where we successively add more misspelled or wrongly cased data to the source of the training data. The additional noisy data is weighted with a factor of 1. Besides the newly created conversational dataset we also involve the Tatoeba corpus (Tiedemann, 2020) that was already part of our training data and is rich in conversational expressions.

The results on the conversational test sets combined are shown in Table 7. As the test sets cover different noise variants, we see a nice improvement with the highest noise level 3, and conclude that we gain in robustness of our MT system. Finally, we also up-weight the original Tatoeba corpus by a factor of 3. This gives an additional small, but consistent improvement on the conversational test data. Thus we select this configuration for further trainings and evaluations.

4.5 Out-of-domain Performance

As we want to integrate the selected configuration into a mixed-domain "one-size-fits-all" model, we need to make sure that the overall system quality remains stable. To check whether up-weighting or noise injection harms translation quality on non-conversational test data, we measure the performance of the systems that perform best on conversational test data on all other test sets, grouped into corporate and generic test sets, as explained in Section 4.2. The results are reported in Table 8. They show clear improvements on the conversational test sets of over 2.0 CHRF2 points and around 3.0 BLEU points for both en–fr and en–ja. Furthermore, the improvements do not lead to degradations on other test sets. These findings support the claim that the quality on all other test sets stayed quite stable.

4.6 Error-sensitive Back-translation Scoring Experiments

For language pairs targeting English, we experiment with adding different configurations of the UDC to the training data of the baseline systems:

Full adds the entire back-translated UDC to the training data of the baseline.

Filter adds only those pairs from the UDC where the source segment's acceptability score exceeds a set threshold.

Weight adds the entire UDC, but assigns a weight between 0.2 and 1 to all segments based on their acceptability score.

The filtering threshold was set based on manual exploration of resulting filtered corpora for a small development set of UDC sentences. The filtered UDC dataset contains roughly 840,000 parallel sentences. For the weighting approach, we decide to down-weight noisy segments rather than up-weight correct segments due to the user-generated nature of the dataset. Table 9 shows the number of UDC sentences per weight.

Table 10 contains the CHRF2 and BLEU scores on all test sets for it–en and ja–en. Adding the entire UDC data (*full*) improves performance for both language pairs on in-domain test data. This indicates that the back-translations are of sufficient quality to provide training signals despite the domain mismatch of the translation system used to obtain them. For generic test sets, performance remains stable, while there is a slight drop in quality on corporate test sets.

Comparing the filtering method (*filter*) with *full*, it performs similarly on generic and corporate test sets but does not achieve the same performance increase on the conversational test sets. It should be noted that filtering results in less than 20% of the

Language pair	Test domain	CHRF2		BLEU	
		Baseline	Final version	Baseline	Final version
en–fr	**conversational**	**59.4**	**61.5**	**36.3**	**39.1**
	generic	67.0	67.0	43.1	43.1
	corporate	81.5	81.4	63.8	63.7
en–ja	**conversational**	**41.1**	**43.5**	**34.1**	**37.4**
	generic	33.9	34.5	35.8	36.3
	corporate	67.8	68.0	69.8	70.0

Table 8: Results on all test sets when adding the noise-injected and up-weighted conversational training data to the baselines.

Weight	# segments
0.2	3,636
0.4	123,185
0.6	727,263
0.8	2,073,784
1.0	1,622,266

Table 9: Number of segments by weight for the *weight* experiment.

Language pair	Test domain	CHRF2				BLEU			
		Baseline	*full*	*filter*	*weight*	Baseline	*full*	*filter*	*weight*
it–en	conversational	64.3	**65.7**	65.1	65.5	41.9	**43.6**	42.8	43.4
	generic	65.9	**66.0**	65.9	65.9	43.1	**43.5**	43.4	43.4
	corporate	**80.8**	80.5	80.6	**80.8**	**63.2**	62.8	62.9	63.1
ja–en	conversational	45.6	**46.3**	45.9	**46.3**	20.5	**21.2**	20.7	21.1
	generic	51.8	**51.9**	**51.9**	**51.9**	22.3	22.3	**22.4**	22.1
	corporate	**74.9**	**74.9**	**74.9**	**74.9**	51.2	51.1	**51.4**	51.2

Table 10: Results on all test sets when adding back-translated UDC data to the training data of the baselines. Best results are highlighted in bold.

UDC being added to the training data. However, further experiments with larger subsets of UDC data have also not outperformed the *full* model.

Weighting the UDC data (*weight*) leads to in-domain improvements comparable to *full*. Additionally, adding the weighted UDC to the training data does not compromise performance in other domains. This may be on account of the down-weighting of ungrammatical segments, enabling the weighting model to learn from conversational data while preserving output quality.

5 From Experiments to Production

The experimental results from Section 4 motivated us to use the same data assembling techniques and configurations for other language pairs that had not been previously tested. For the translation directions with English source, Table 11 lists the language pairs and shows the gain in case-sensitive CHRF2 and BLEU for the three groups of test sets (see Section 4.2). *Base* constitutes the baseline, to which *New* adds up-weighted parallel data noise-injected using the best configuration found in Section 4. Note that the scores for en–fr and en–ja are slightly different from those in Table 8 as the overall setup and training data composition of the experimental and final systems are not exactly identical. Across all language pairs there is considerable improvement on the conversational test sets, while on the other domains (corporate and generic) the performance remains stable on average, according to both automatic metrics. Thus, our approach works similarly well for the other seven language pairs as for English to French and English to Japanese, showing that we can deliver high-quality business conversation MT broadly for many languages without compromising translation quality of other text types.

The results of adding the back-translated UDC data with error-sensitive weight factors for systems translating into English are shown in Table 12. Although the impact is less pronounced than for the other language direction, it is consistent and visible. It is quite surprising that the large amount of back-translated data is not harming the translation quality in other domains.

To illustrate the differences, we refer back to Table 1, comparing the French MT output after the quality improvements with the baseline engine's output on the English example dialogue. The example demonstrates that robustness to typos has improved, and that punctuation is placed correctly. Fewer words remain untranslated and the MT output is more fluent.

6 Outlook

Although we see nice improvements, the translation quality in technical business conversations could be further improved. We point out the main open issues in this section, leaving them for future work and calling for new methods to address them.

		CHRF2		BLEU	
	Test domain	Base	New	Base	New
en–de	**conversational**	**55.3**	**57.1**	**29.4**	**31.5**
	generic	66.2	66.4	40.4	40.7
	corporate	77.1	76.9	53.6	53.6
en–es	**conversational**	**65.4**	**68.0**	**44.0**	**47.3**
	generic	70.0	70.0	48.4	48.5
	corporate	81.6	81.6	64.4	64.3
en–fr	**conversational**	**58.8**	**61.7**	**35.7**	**39.0**
	generic	67.2	67.2	43.4	43.4
	corporate	81.8	81.8	64.2	64.3
en–it	**conversational**	**59.3**	**63.0**	**34.6**	**39.1**
	generic	67.2	67.4	42.0	42.1
	corporate	81.9	81.5	62.9	62.1
en–ja	**conversational**	**41.6**	**43.9**	**34.2**	**37.5**
	generic	33.8	34.2	35.3	36.1
	corporate	70.5	71.0	72.1	72.5
en–ko	**conversational**	**44.1**	**46.3**	**20.2**	**22.5**
	generic	65.9	65.2	44.0	43.1
	corporate	72.9	72.5	57.2	56.7
en–pt	**conversational**	**68.5**	**71.5**	**46.4**	**51.0**
	generic	69.6	69.9	45.5	46.1
	corporate	84.3	84.3	68.3	68.3
en–ru	**conversational**	**50.3**	**52.9**	**27.5**	**29.9**
	generic	64.9	65.0	38.8	38.9
	corporate	76.2	76.3	54.8	54.9
en–zh	**conversational**	**48.9**	**49.0**	**35.3**	**37.5**
	generic	42.6	43.3	45.6	46.2
	corporate	70.9	71.8	72.1	73.0

Table 11: CHRF2 and BLEU scores on test sets from all domains for the translation directions with English source.

		CHRF2		BLEU	
	Test domain	Base	New	Base	New
de–en	**conversational**	**60.1**	**60.7**	**36.1**	**36.6**
	generic	67.0	67.6	44.1	44.7
	corporate	81.7	81.5	65.4	65.0
es–en	**conversational**	**67.2**	**68.3**	**45.0**	**46.5**
	generic	69.2	69.8	46.3	47.2
	corporate	81.0	80.9	63.8	63.4
fr–en	**conversational**	**62.5**	**63.2**	**39.7**	**40.8**
	generic	67.7	67.4	44.8	44.5
	corporate	79.3	78.2	61.1	59.2
it–en	**conversational**	**63.5**	**65.1**	**40.8**	**43.1**
	generic	67.1	67.9	44.0	45.2
	corporate	82.6	82.5	66.2	65.9
ja–en	**conversational**	**44.1**	**45.7**	**19.1**	**20.5**
	generic	53.5	54.8	23.8	24.7
	corporate	74.5	75.2	50.9	51.9
ko–en	**conversational**	**50.8**	**52.8**	**24.2**	**26.3**
	generic	57.7	57.9	33.4	33.6
	corporate	75.8	76.1	52.9	53.8
pt–en	**conversational**	**69.5**	**70.6**	**47.8**	**49.4**
	generic	72.3	72.9	50.5	51.5
	corporate	84.6	84.7	69.6	69.6
ru–en	**conversational**	**56.5**	**57.5**	**32.8**	**33.8**
	generic	64.9	64.9	39.0	39.0
	corporate	75.9	75.8	55.7	55.2
zh–en	**conversational**	**52.4**	**53.6**	**27.0**	**28.5**
	generic	60.3	60.5	31.9	32.2
	corporate	78.9	79.1	57.5	57.7

Table 12: CHRF2 and BLEU scores on test sets from all domains for the translation directions with English target.

In order to enhance robustness with respect to misspellings, casing, chat-typical conversational forms, or abbreviations, a normalization step in preprocessing could be investigated (Chitrapriya et al., 2018; Clark and Araki, 2011). This would support subsequent MT. However, text normalization or automatic spelling correction (Peitz et al., 2013) is highly text-type specific and prone to overgeneration when applied to non-conversational text, especially for technical documentation with lots of acronyms and technical abbreviations. This is one of the reasons why we decided for the noise injection approach targeted at conversational content only.

Chat language includes other specific phenomena which we did not specifically address in this work, one of them being capitalization for emphasis, which could be tackled, e.g., using a factored representation for source and target (García-Martínez et al., 2016; Niehues et al., 2016; Wilken and Matusov, 2019). Another frequent phenomenon is emoticons, where one would need to decide whether they should just be copied over, or whether they also need to be localized to the target language. For expletives in conversations, applicable methods largely depend on the expectations in specific use cases, i.e., should a swearword be translated to its counterpart in the target language, should it be removed, or masked with asterisks?

Our MT model operates on the sentence level, and we treat each utterance as one sentence. However, in chat conversations, sentences are sometimes spread over multiple utterances, meaning the source is actually over-segmented, leading to poor translation quality. This could be improved by a different segmentation paradigm, and/or by an MT model that takes dialogue context beyond the sentence level into account (Liang et al., 2021). The latter should also improve the coherent use of pronouns and verbal forms within a dialogue.

Levels of politeness and their expression in conversations differ between cultures and languages. Accordingly, this also poses challenges for MT, especially when the target language has more fine-grained distinctions than the source language.

7 Related Work

Our work has focused on four methods: (1.) Integrating parallel high-quality conversational content into the training corpus, (2.) creating synthetic in-domain data via back-translation, (3.) data augmentation to make the model more robust to noisy input, and (4.) model adaptation towards the style of conversational content in the business domain. Prior work by other researchers has pursued aims related to ours while often employing slightly different techniques. For instance, high-quality parallel data is oftentimes identified by means of pseudo in-domain data selection (Axelrod et al., 2011); back-translation can be improved by sampling or noisy synthetic data (Edunov et al., 2018); better robustness towards noisy input may be achieved with a stochastically corrupted subword segmentation procedure (Provilkov et al., 2020); or domain adaptation might be feasible even in a semi-supervised or unsupervised manner in certain scenarios (Dou et al., 2019; Niu et al., 2018). We are confident that many of the existing related techniques are complementary to our work and will help further improve MT quality of conversational content in the business domain.

8 Conclusion

We have shown that an MT model specialized in the IT and business domains can be enhanced to also cover conversational content well. This balancing act is highly relevant in scenarios such as product support chats or multilingual chatbots. We have achieved that by curating high-quality parallel data to address phenomena where the model exhibited the most devastating shortcomings. We further add back-translated data from the dialogue domain, inject typos, punctuation and capitalization variants to make the model more robust, and carefully manage the influence of the different corpora using a sentence weighting scheme. We have demonstrated that promising results from experiments involving only a few language pairs generalize well to the main languages in our production scenario at SAP, achieving an improvement of 2.4 CHRF2 / 3.1 BLEU on average for language pairs from English and 1.2 CHRF2 / 1.5 BLEU for language pairs to English on our conversational test sets, while the performance on other domains and test sets remains stable.

Acknowledgments

We thank Vincent Asmuth, Nathaniel Berger, and Dominic Jehle for proofreading and valuable discussions, as well as the four anonymous reviewers for their feedback and helpful comments.

References

Axelrod, Amittai, Xiaodong He, and Jianfeng Gao. 2011. Domain Adaptation via Pseudo In-Domain Data Selection. In *Proc. of EMNLP*, pages 355–362, Edinburgh, Scotland, UK, July.

Chen, Boxing, Colin Cherry, George Foster, and Samuel Larkin. 2017. Cost Weighting for Neural Machine Translation Domain Adaptation. In *Proc. of the Workshop on Neural Machine Translation*, pages 40–46, Vancouver, Canada, August.

Chitrapriya, N., Md. Ruhul Islam, Minakshi Roy, and Sujala Pradhan. 2018. A Study on Different Normalization Approaches of Word. In *Advances in Electronics, Communication and Computing*, pages 239–251, Singapore.

Chu, Chenhui and Rui Wang. 2018. A Survey of Domain Adaptation for Neural Machine Translation. In *Proc. of COLING*, pages 1304–1319, Santa Fe, NM, USA, August.

Clark, Eleanor and Kenji Araki. 2011. Text Normalization in Social Media: Progress, Problems and Applications for a Pre-Processing System of Casual English. *Procedia - Social and Behavioral Sciences*, 27:2–11.

Dou, Zi-Yi, Junjie Hu, Antonios Anastasopoulos, and Graham Neubig. 2019. Unsupervised Domain Adaptation for Neural Machine Translation with Domain-Aware Feature Embeddings. In *Proc. of EMNLP-IJCNLP*, pages 1417–1422, Hong Kong, China, November.

Edunov, Sergey, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding Back-Translation at Scale. In *Proc. of EMNLP*, pages 489–500, Brussels, Belgium, October/November.

Freitag, Markus and Yaser Al-Onaizan. 2016. Fast Domain Adaptation for Neural Machine Translation. *CoRR*, abs/1612.06897.

García-Martínez, Mercedes, Loïc Barrault, and Fethi Bougares. 2016. Factored Neural Machine Translation Architectures. In *Proc. of IWSLT*, Seattle, WA, USA, December.

Huck, Matthias, David Vilar, Daniel Stein, and Hermann Ney. 2011. Lightly-Supervised Training for Hierarchical Phrase-Based Machine Translation. In *Proc. of the First Workshop on Unsupervised Learning in NLP*, pages 91–96, Edinburgh, Scotland, UK, July.

Huck, Matthias, Alexandra Birch, and Barry Haddow. 2015. Mixed-Domain vs. Multi-Domain Statistical Machine Translation. In *Proc. of MT Summit*, pages 240–255, Miami, FL, USA, October.

Huck, Matthias, Fabienne Braune, and Alexander Fraser. 2017. LMU Munich's Neural Machine Translation Systems for News Articles and Health Information Texts. In *Proc. of WMT, Vol. 2: Shared Task Papers*, pages 315–322, Copenhagen, Denmark, September.

Junczys-Dowmunt, Marcin, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018. Marian: Fast Neural Machine Translation in C++. In *Proc. of ACL, System Demonstrations*, pages 116–121, Melbourne, Australia, July.

Liang, Yunlong, Fandong Meng, Yufeng Chen, Jinan Xu, and Jie Zhou. 2021. Modeling Bilingual Conversational Characteristics for Neural Chat Translation. In *Proc. of ACL-IJCNLP (Vol. 1: Long Papers)*, pages 5711–5724, Online, August.

Lison, Pierre and Jörg Tiedemann. 2016. OpenSubtitles2016: Extracting Large Parallel Corpora from Movie and TV Subtitles. In *Proc. of LREC*, pages 923–929, Portorož, Slovenia, May.

Lowe, Ryan, Nissan Pow, Iulian Serban, and Joelle Pineau. 2015. The Ubuntu Dialogue Corpus: A Large Dataset for Research in Unstructured Multi-Turn Dialogue Systems. In *Proc. of SIGDIAL*, pages 285–294, Prague, Czech Republic, September.

Mathur, Nitika, Timothy Baldwin, and Trevor Cohn. 2020. Tangled up in BLEU: Reevaluating the Evaluation of Automatic Machine Translation Evaluation Metrics. In *Proc. of ACL*, pages 4984–4997, Online, July.

Niehues, Jan, Thanh-Le Ha, Eunah Cho, and Alex Waibel. 2016. Using Factored Word Representation in Neural Network Language Models. In *Proc. of WMT: Vol. 1, Research Papers*, pages 74–82, Berlin, Germany, August.

Niu, Xing, Sudha Rao, and Marine Carpuat. 2018. Multi-Task Neural Models for Translating Between Styles Within and Across Languages. In *Proc. of COLING*, pages 1008–1021, Santa Fe, NM, USA, August.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proc. of ACL*, pages 311–318, Philadelphia, PA, USA, July.

Peitz, Stephan, Saab Mansour, Matthias Huck, Markus Freitag, Hermann Ney, Eunah Cho, Teresa Herrmann, Mohammed Mediani, Jan Niehues, Alex Waibel, Alexander Allauzen, Quoc Khanh Do, Bianka Buschbeck, and Tonio Wandmacher. 2013. Joint WMT 2013 Submission of the QUAERO Project. In *Proc. of WMT*, pages 185–192, Sofia, Bulgaria, August.

Popović, Maja. 2016. chrF deconstructed: beta parameters and n-gram weights. In *Proc. of WMT: Vol. 2, Shared Task Papers*, pages 499–504, Berlin, Germany, August.

Provilkov, Ivan, Dmitrii Emelianenko, and Elena Voita. 2020. BPE-Dropout: Simple and Effective Subword Regularization. In *Proc. of ACL*, pages 1882–1892, Online, July.

Rieß, Simon, Matthias Huck, and Alex Fraser. 2021. A Comparison of Sentence-Weighting Techniques for NMT. In *Proc. of MT Summit*, pages 176–187, Virtual, August.

Schwenk, Holger. 2008. Investigations on Large-Scale Lightly-Supervised Training for Statistical Machine Translation. In *Proc. of IWSLT*, pages 182–189, Waikiki, HI, USA, October.

Sennrich, Rico, Barry Haddow, and Alexandra Birch. 2016a. Improving Neural Machine Translation Models with Monolingual Data. In *Proc. of ACL*, pages 86–96, Berlin, Germany, August.

Sennrich, Rico, Barry Haddow, and Alexandra Birch. 2016b. Neural Machine Translation of Rare Words with Subword Units. In *Proc. of ACL*, pages 1715–1725, Berlin, Germany, August.

Shah, Kshitij and Gerard de Melo. 2020. Correcting the Autocorrect: Context-Aware Typographical Error Correction via Training Data Augmentation. In *Proc. of LREC*, Marseille, France, May.

Thompson, Brian, Jeremy Gwinnup, Huda Khayrallah, Kevin Duh, and Philipp Koehn. 2019. Overcoming Catastrophic Forgetting During Domain Adaptation of Neural Machine Translation. In *Proc. of NAACL-HLT*, pages 2062–2068, Minneapolis, MN, USA, June.

Tiedemann, Jörg. 2020. The Tatoeba Translation Challenge – Realistic Data Sets for Low Resource and Multilingual MT. In *Proc. of WMT*, pages 1174–1182, Online, November.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Wang, Rui, Masao Utiyama, Lemao Liu, Kehai Chen, and Eiichiro Sumita. 2017. Instance Weighting for Neural Machine Translation Domain Adaptation. In *Proc. of EMNLP*, pages 1482–1488, Copenhagen, Denmark, September.

Wilken, Patrick and Evgeny Matusov. 2019. Novel Applications of Factored Neural Machine Translation. *CoRR*, abs/1910.03912.