Proceedings of the LREC2020 Industry Track , pages 31–35
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020

Industrial Machine Translation System for Automotive Domain

Maria Sukhareva*, Olgierd Grodzki†, Bernhard Pflugfelder*
*BMW Group
Bremer Straße 6, 80807 Munich
{maria.sukhareva, bernhard.pflugfelder}@bmwgroup.com

†Data Reply
Luise-Ullrich-Straße 14, 80636 Munich
o.grodzki@reply.de

Abstract

BMW Group, a large multinational company, inevitably faced the challenge of processing and creating large amounts of multilingual data. As manual translation fails to provide sufficient coverage and cost efficiency on a large scale, an acute need for a reliable machine translation service naturally arose. Translating highly technical automotive texts has proven to not be a trivial task and off-the-shelf commercial cloud solutions fail to deliver satisfactory translation quality. In this paper, we present a customized machine translation system tailored for automotive needs. Our system is well-suited for translating automotive texts and satisfies all data protection requirements. The use cases that we discuss in this paper have a projected business value between one and five million euros.

Keywords: machine translation, industrial system, domain adaptation

1. Introduction

BMW Group is a multinational company: it currently operates 30 production and assembly facilities in 14 countries and has a global sales network in more than 140 countries. As of December 2018, the BMW Group had a workforce of 134,682 employees from 124 nationalities. Having a fast and reliable machine translation infrastructure is vital for the company's value chain.

Assimilation of multilingual information in BMW Group The BMW Group receives daily thousands of customer comments in over 100 languages. These comments further undergo qualitative and quantitative analysis. BMW data scientists use machine learning methods to cluster, classify and extract information from the comments while customer relations specialists process the feedback individually addressing specific customer needs. Another source of multilingual data is dealer feedback. Currently car dealers operate in over 140 countries and BMW support service receives tickets in dozens of languages. Manual translation of such tickets is costly and causes unnecessary delay. Applying machine translation can significantly diminish the processing time.

Dissemination of multilingual information in BMW Group BMW Group creates a multitude of multilingual texts on a daily basis such as car manuals and training materials for dealers, marketers and customers. These kind of texts cannot tolerate any errors. Thus, the current approach is to automatically pre-translate the texts and involve human technical translators as post-editors.

Multilingual communication The company operates several support hotlines e.g. IT support, financial services and dealer support. Support agents are reachable by telephone as well as through an online chat. Machine translation here can be used for real time translation of low-resource languages for which it is not plausible to create a manned support service.

2. Automotive domain

The automotive domain includes a variety of texts, e.g. car manuals, promotional texts, error reports, protocols of production changes. All those texts pose a challenges of specialized terminology. While our machine translation encompasses a multitude of BMW domains and applications, this paper will focus on two use cases: (i) translation of car manuals and training materials and (ii) translation of production protocols. The savings from the usage of machine translation on these use cases are projected to be between one and five million euros annually.

Car manuals are instructions for car mechanics, dealers and customers on how to repair and maintain the vehicles. The original texts are written in German and are to be translated into multiple languages.

BMW Group has *prescriptive* terminological dictionaries for human translators and technical writers. The dictionaries contain lists of concepts and their lemmata in over 30 languages. The concepts are defined in German and, thus, all the dictionary entries have a German lemmata. An entry can have several non-German entries in other languages that should be used for translation. There is a list of synonyms for a given concept which shall not be used in technical documentation and translation (*negative* terms). Languages are not equally represented in the dictionary: English has most of the entries while Ukrainian, Bulgarian etc. have just a few thousands terms. Table 1 shows the statistics over the corporate lexicon. The lexicon has a total of 1241087 entries with 227590 being negative terms.

The dictionary is prescriptive i.e. the resulting translations have zero tolerance for synonyms i.e. using a term that is marked as negative invalidates the whole translation even if the meaning is preserved. This kind of restrictions pose several challenges, first of all, the challenge of integrating the lexica into the translation and the challenge of the automatic evaluation as commonly used measures such as TER or BLEU treat all the n-grams equally.

lang	entries	lang	entries	lang	entries	lang	entries	lang	entries
German	96032	Portuguese	54932	Turkish	48020	Swedish	13526	Bulgarian	6974
English (UK)	74404	Russian	53210	English (US)	47150	Slovene	12550	Portuguese (BR)	6632
French	73626	Greek	52832	Thai	46744	Hungarian	12174	Dutch (BE)	2284
Spanish	69050	Finnish	50932	Chinese	45946	Romanian	12096	French (BE)	2272
Italian	63466	Korean	50172	Czech	34888	Norwegian	9016	Ukrainian	2150
Dutch	60434	Danish	48332	Indonesian	34692	Chinese (TW)	8178	English (AU)	2104
Swedish	60358	Japanese	48090	Polish	28880	Arabic (SA)	7496	English (ZA)	1444

Table 1: The amount of entries per language in the prescriptive corporate lexicon

```
G34 ag upr/lwr seal/grommet opt. E34.234.2 assy line protection LU with ERWU.
If the electronic immobilizer (EWS) functionality is disabled,  the TEE shall
prevent disengagement of the parking lock. G83 = Non-return valve, cylinder head.
Close the trim grille CS/JCW/Cooper/CooperD/One/OneD.
```

Figure 1: An example of a production protocol

Translation of production protocols is another challenging MT use case. Daily, engineers protocol their actions e.g. ordering car parts, changing design decision etc. The protocols are highly technical but also contain a lot of abbreviations, acronyms etc. for brevity. Figure 1 shows a snippet of a production protocol. The texts are not easily decipherable by a person who lacks specialized training. Similarly, machine translation systems trained on out-of-domain data such as news corpora have a sub-par performance on this data. The original protocols are written either in English or in German. The German protocols are translated into English and communicated to the plants in non-German speaking countries. The English protocol translations are sent to the plants in Germany. Apart from the obvious challenges of non-standardized punctuation and domain specific lexica, additional challenges of copying numbers and translating abbreviation is added. Having erroneous translation of numbers has in fact proven to be worse than not having a translation at all as post-editors may fail to notice the error if the number is present in the text.

To sum up, the automotive domain is comprised of multiple genres of varying complexity: from easily readable customer reviews to terminology-rich production protocols. Implementing machine translation of automotive texts is not trivial and involves various degrees of domain adaptation depending on the use case.

3. Customized machine translation

Industrial MT also have to comply with restrictions imposed by the EU's General Data Protection Regulations[1] that states that no personal data can be disclosed to a third party without customer consent. This limits the options for cloud-based machine translation systems as data coming from customers and dealers may contain personal information. Another restriction is confidentiality: Production protocols are confidential and cannot be passed to a third party. Over the years, BMW Group has accumulated a large amount of translation memories for the languages listed in Table 1. The translation memories are parallel text fragments that have been translated to or from German. The fragments can be sentences and clauses, but are on average much shorter phrases.

We have implemented customized machine translation solutions for German and three high-resource languages: English, Italian and Spanish. English, unsurprisingly, has the largest amount of parallel data with over 5.5 million fragments. Spanish has 2.6 million parallel fragments and the Italian data has 2 million fragments. Training a machine translation system solely on this data is not possible as, for example, the average sentence length of a fragment is 9 words in German and 11 words in English for the production protocols and 7 words in German and 9 words in English for the car manuals. Thus, to ensure that the models learn to handle longer sequences, we used out-of-domain open-source data, which eventually doubled the training sets for all the language pairs: over 11 million parallel fragments for English and 5 million and 4 million for Spanish and Italian correspondingly.

3.0.1. Data filtering and preprocessing

We filtered the data with a language detector and eliminated all of the sentence pairs with a length discrepancy (the token ratio 0.6). We also deleted all the tags and non-ASCII characters. To prevent mistranslation of numeric data, we substituted all the tokens that contain digits (e.g. GB18, 10.02.2020, 24-241-123-123-432) with a placeholder and used lexical constraints (Post and Vilar, 2018) to make sure that all the digits are present in the translation. We also use placeholders to integrate a list of BMW abbreviations and untranslatables. Finally, we used byte-pair encoding to tokenize the data.

3.1. Model training

The models are on 8 GPUs with a toolkit for neural machine translation Sockeye (Hieber et al., 2017a). To avoid later bias towards shorter translations, we also learn the brevity penalty parameter during the training (Hieber et al., 2017b). The Spanish and Italian models are only used for car manuals while the English model is applied to production protocol as well. Despite terminological similarity, production

[1] https://gdpr-info.eu/

protocols are hastily written texts with multiple acronyms, abbreviations, digits and orthographic errors. Car manuals are, on the contrary, carefully crafted texts and void of errors. Thus, we have experimented with various set-ups: (i) training separate models for car manuals and protocols, (ii) training a multitask model by adding a tag to differentiate between use cases (Johnson et al., 2017) (iii) and training a joint model. The joint model has reached the highest BLEU score (Papineni et al., 2002), followed by the disjoint training for which the BLEU score dropped on average by 0.06. The multitask approach performed the worst mostly because the model would opt for short translations biased towards the length of the in-domain data.

3.2. Translation

We use the same preprocessing steps for translation as for training. The models are deployed on the BMW AWS cloud described in section 4. To accelerate inference, we also follow Post and Vilar (2018) and integrate the lexical translation probabilities learned from the training data with fast align (Dyer et al., 2013). The lexicon is learnt on the training data for each language pair and we set the top k candidates to 200.

3.3. Evaluation

We have compared our machine translation system to cloud-based commercial systems available on the market. The goal of the evaluation was to show that our system can achieve state-of-the-art performance and is more suitable for translating automotive domain texts than the off-the-shelf tools. While we are aware that some commercial systems allow in-domain data integration, our training data are confidential and cannot be shared with third parties. Therefore, our system has a clear advantage of having seen in-domain data but as the goal of the evaluation is not to compare algorithmic approaches but rather to show that an in-house machine translation system can deliver high quality translation while complying with data protection standards. 1000 reference sentences were translated by our CMT and by two cloud-based commercial translators (CS1 and CS2) Unsurprisingly, the cloud-based systems performed poorly on both datasets (see Table 2) with CS1 reaching the BLEU score of 0.55 and CS2 reaching the BLEU score of only 0.4. The CMT reached the BLEU score of 0.78 on the same test set. We have conducted qualitative evaluation of the results and observed that even for the sentences that have lower n-gram overlap the CMT produces better translations than commercial systems. Table 3 shows examples of sentences that were not translated perfectly by the CMT. The subsequent qualitative analysis by human translators concluded that CMT produces acceptable translations unlike off-the-shelf commercial systems that frequently fail to convey even the general idea of the sentence. As the main practical objective of translating production protocols was to facilitate information exchange between BMW engineers, we have concluded that CMT satisfies the quality requirements and the system has been launched into production.

Translation of car manuals posed an additional challenge as the end users are dealers and customers and, thus, the translation should not only adequately transfer the meaning but also be completely error-free as well as grammatically, orthographically and punctuation-wise correct. This could only be achieved by adding post-editors to the workflow. We have conducted three post-editing experiments for three customized models: German to English, German to Italian and German to Spanish. Human translators were asked to post-edit five documents (10259 sentences) translated from German into a corresponding language. The post-editors were instructed to only apply minimal edits in order to reach publishable quality. As the main pragmatic objective of integrating machine translation into the workflow was to accelerate the translation process, the usefulness of the system corresponds to the speed gains by post-editors as compared to translators. The best results were achieved for English with post-editing speed of 909 words per hour. Italian post-editor gained the speed of 650 words per hour and the Spanish post-editor was processing 641 words per hour. The better results for English can be easily explained by the fact that the English customized model was trained on three times as much data as compared to the other two languages. As the average translation speed of a BMW technical translator is 330 words per hour, we have shown that using our customized machine translation, BMW translators can process double or triple text volumes as compared to translation from scratch.

Domain	CS1	CS2	CMT
Protocols	0.56	0.4	0.73
Manuals	0.63	0.56	0.76

Table 2: BLEU Score evaluation of the CMT as compared to industrial systems

3.4. System updates and quality control

In order to improve the performance of the machine translation system and to keep the system up-to-date, the system is retrained each time when 100,000 new sentence pairs are available for a language pair. These data are then split into training, development and test sets with both test and development sets having 4,000 sentences. The evaluation is done on the newly acquired test set as well as on the test sets kept from previous (re-)trainings. In this way, we make sure that the new system is not overfitting the new data. If after retraining the model delivers satisfying performance on all the test data, the model is moved into production.

We also monitor the quality of the productive models. The post-editing results for each sentence are directly communicated back to the system and two measures are computed: the post-editing speed and the human-targeted translation edit rate (HTER) (Snover et al., 2006).

4. Scaling CMT to production

The solution is deployed in Amazon Web Services (AWS, 2020b), the world's most widely adopted cloud platform. It offers reliable, scalable, and inexpensive cloud computing services. The translation system was designed using a selection of the services offered by AWS. The architecture diagram is presented in Figure 2. All components were de-

Original	Human	CS1	CS2	BMW CMT
Neusausleitung Sensor Fussgaengerschutz G0x aufgrund der Max Boole	New export of sensor pedestrian protection G0x due to the Max Boole	Reuse sensor foot protector G0x due to the Max Boole	New version sensor football instrument protection G0x due to the Max Boole	New export sensor pedestrian protection G0x due to Max Boole
Rippe an Auflage LT entfernen, um Zugaenglichkeit an Sitzaussenlager zu gewaehrleisten	Remove rib at the support of the side member to ensure accessibility at the outer seat bearing.	Remove the rib on support LT to ensure traction on the outside bearing	Remove rib to edition LT to ensure access to seat outside warehouses	Remove rib on support LT to ensure accessibility to seat outer bearing
Fruehester moeglicher Wareneingangstermin BMW:	earliest possible incoming goods date BMW:	Early arrival date BMW:	Fruitful possible goods receipt date BMW:	Earliest possible receipts of goods BMW:
Freigabe der Bauteile die nicht Inhalt Huelle sind.	Release of components that are not part of the cover.	Release of components that are not content hulle.	Release of the components that are not content Huelle.	Release of the components that are not contained in the case.
Abloesung der BAW BET698	Replacement of deviation permit BET698	The removal of the BAW BET698	Abloation of BAW BET698	Release of the Closure of dev.perm BET698

Table 3: Comparison of selected translations from two commercial systems (CS1 and CS2) and BMW CMT

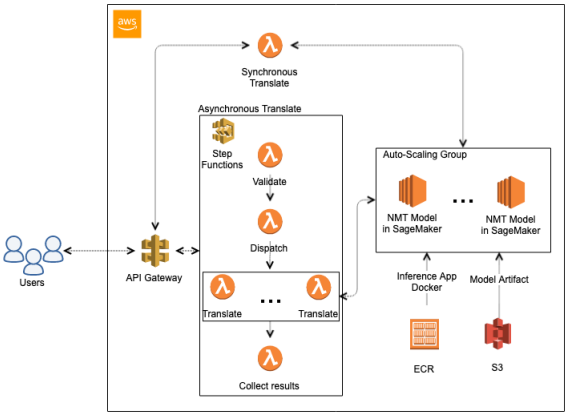

Figure 2: Implementation of the customized machine translation on AWS

ployed within an BMW AWS VPC with only private subnets and without any external connectivity.

The inference code was wrapped in a Flask application in order to expose a REST API to perform translations. It was then packaged in a Docker image and pushed to AWS Elastic Container Registry (AWS, 2020a). The model artifact is not a part of the image, instead it is packaged in an archive and uploaded to AWS Simple Storage Service (S3). This object storage service offers industry-leading scalability, data availability, security, and performance (AWS, 2016). This separation was done following engineering best practices - the model lifecycle should be managed separately from the inference code.

The application is deployed in AWS SageMaker, which is a fully managed service that provides every developer and data scientist with the ability to build, train, and deploy machine learning models quickly (AWS, 2019a). It provides many built in features needed for production, e.g. versioning, autoscaling, zero downtime redeployments, and canary testing. The SageMaker deployment makes uses of the inference application Docker image and the model artifact.

The translation system is exposed via an API deployed using AWS API Gateway (AWS, 2015). The API can be called in one of the following request-reply patterns:

- synchronous - the client submits a translation request and blocks execution until it receives a translation response,
- asynchronous - the client submits a translation start request, immediately receives a job ID as a response and proceeds to use it to poll for status and retrieve the results once the job is finished.

The asynchronous pattern is preferred because it leverages the potential parallelism of the translation service, reduces the risk of timing out, and accelerates the translation of large text packages. The system also includes a web frontend which allows users to engage with it from their web browsers.

Robustness of the service is ensured through the auxiliary code e.g. validation of the requests, choosing and invoking the correct SageMaker endpoint, etc. This code was deployed using AWS Lambda, which is a platform that enables code execution without provisioning or managing servers. The orchestration of the Lambda functions was done by AWS Step Functions (AWS, 2019b), which allow for the coordination of multiple AWS services within serverless workflows. In doing so it abstracts away the state and transformation management in order to focus on the business logic.

The production-readiness of the system is further enhanced via the application of a number of engineering best practices. The infrastructure is managed by an infrastructure as code (IaC) solution Terraform. Building and deploying the code for both the frontend and backend is managed by CI/CD pipelines (Chapman, 2014). Every component of the system provides logs and metrics to AWS CloudWatch. Alarms, based on the CloudWatch metrics, are sent to an SNS topic which delivers emails to subscribers in both the alarm and OK states. The observability of the service, meaning the distributed tracing of end-to-end requests through all layers of the solution, is performed via the use of AWS X-Ray to generate service maps and traces.

5. Conclusion

Machine translation has a wide spectrum of applications in BMW Group. This paper has focused on two challenging use cases: the translation of production protocols and the translation of car manuals. These use cases have a projected business value of several million euros. Our CMT systems have proven to deliver high quality translations of automotive texts and to accelerate human translation. Our future work includes introducing additional languages into the CMT system, integrating lexica and extending our evaluation methodology to assess terminological consistency according to corporate standards.

6. Bibliographical References

Dyer, C., Chahuneau, V., and Smith, N. A. (2013). A simple, fast, and effective reparameterization of IBM model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–648, Atlanta, Georgia, June. Association for Computational Linguistics.

Hieber, F., Domhan, T., Denkowski, M., Vilar, D., Sokolov, A., Clifton, A., and Post, M. (2017a). Sockeye: A Toolkit for Neural Machine Translation. *arXiv preprint arXiv:1712.05690*, December.

Hieber, F., Domhan, T., Denkowski, M., Vilar, D., Sokolov, A., Clifton, A., and Post, M. (2017b). Sockeye: A toolkit for neural machine translation. *CoRR*, abs/1712.05690.

Johnson, M., Schuster, M., Le, Q. V., Krikun, M., Wu, Y., Chen, Z., Thorat, N., Viégas, F., Wattenberg, M., Corrado, G., Hughes, M., and Dean, J. (2017). Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

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, Philadelphia, Pennsylvania, USA, July. Association for Computational Linguistics.

Post, M. and Vilar, D. (2018). Fast lexically constrained decoding with dynamic beam allocation for neural machine translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1314–1324, New Orleans, Louisiana, June. Association for Computational Linguistics.

Snover, M., Dorr, B., Schwartz, R., Micciulla, L., and Makhoul, J. (2006). A study of translation edit rate with targeted human annotation. In *In Proceedings of Association for Machine Translation in the Americas*, pages 223–231.

AWS, (2015). *AWS Serverless Multi-Tier Architectures*.

AWS, (2016). *AWS Storage Services Overview*.

AWS, (2019a). *Deep Learning on AWS*.

AWS, (2019b). *Implementing Microservices on AWS*.

AWS, (2020a). *Amazon ECR User Guide*.

AWS, (2020b). *Overview of Amazon Web Services*.

Chapman, D., (2014). *Introduction to DevOps on AWS*.