Tackling the ”We have no Data“ Challenge: Domain-Specific Machine Translation in SMEs

Frederik S. Bäumer & Sergej Denisov
Bielefeld University of Applied Sciences
Bielefeld, Germany
{fbaeumer1,sdenisov}@fh-bielefeld.de

Bastian Sirvend
Wonki GmbH
Bielefeld, Germany
bastian.sirvend@wonki.tech

Jens Weber
University of Applied Sciences Zwickau
Zwickau, Germany
jens.weber@fh-zwickau.de

***Abstract*—The use of translation software has decisive advantages for companies. For example, they facilitate communication and the editing and creation of multilingual documents. In contrast to the services of a translation agency, the results are immediately available and can be adapted flexibly. Nevertheless, concerns exist, especially regarding translation quality in case of specialized vocabulary, industry-specific phrases, and data security. Developing and deploying self-hosted business-specific translation models can address both problems by increasing speed and providing company-specific translations. However, this often leads to a situation where companies assume that they cannot contribute the necessary training data. In fact, many companies are sitting on a veritable treasure of data that needs to be lifted. This paper intends to show how we support enterprises with processes and software tools to create datasets for their translation solutions. For this purpose, we apply data acquisition techniques and data preparation methods, sentence alignment, and human-in-the-loop tools.**

***Index Terms*—machine learning, machine translation**

I. Introduction

Translation software is one of the tools needed daily in many companies, as it facilitates international cooperation immensely and simplifies working with multilingual documents and websites [1]. Compared to translation agencies, software has the advantage that companies can use them quickly, flexibly, and cost-effectively. However, many companies still have concerns about the quality of the translations, which is of crucial importance in corporate use since the technical vocabulary of a domain can differ considerably from that of the standard language [1]. In addition, data security concerns cause companies to be skeptical of cloud translation providers and drive the development of self-hosted translation models.

The possibility of training and operating own translation models is not a new concept. These machine translation solutions have been widely available for quite some time (e.g., [2]). The achievements of the last years, especially in the field of deep learning-based approaches, have brought the whole area of translation approaches a significant step forward [3]. In general, these machine translation systems generate translations by using statistical models that have been parameterized by analyzing documents available in the source and target languages. As a result, they can develop a basic sense of language and learn the specifics of domain-specific vocabulary. Especially low resource languages benefit from achievements such as transfer learning. However, today we face the situation that extensive resources for common languages exist, and the languages most requested can be served with existing models [1]. What remains is that these resources and models often do not reflect the peculiarities of enterprise- and domain-specific languages. However, this can be accomplished using the company's data.

While implementing in-house translation solutions, it often occurs that companies assume that they do not have their own data corpus that can be used to train the translation solutions at the very beginning. Often, companies are not aware that a parallel corpus does not necessarily have to be the starting point, but that web pages, instructions for products, advertising material, etc. can also be used to create parallel corpora – as long as they are multilingual, at least in parts. In this short paper, we report on our approach and developed software tool that makes it possible to generate necessary datasets in close cooperation with the companies.

The structure of this paper is as follows. Section II presents related work. Based on this, our approach is presented (Section III) and then discussed in Section IV. Finally, we conclude our work (Section V).

II. Related Work

For recent machine translation algorithms, underlying parallel corpora are essential, and the challenge of creating high-quality datasets is already well known. Thus, the questions of where such data comes from, who owns them, what their characteristics are and who is allowed to use them are not new either [4]. In the following, we describe existing work focusing on the mining of heterogeneous data sources to create parallel corpora for translation purposes.

Documents from parliaments or other public offices and institutions are a popular source for parallel corpora (esp. transcripts). Examples of this are *Europarl*, a corpus composed from texts from the proceedings of the European Parliament [5], and *The United Nations Parallel Corpus*, a corpus composed from United Nations documents [6]. These sources are very helpful because the documents are of very high quality, and the authors have a legitimate interest and often also the obligation to provide the identical content in several languages [6]. Furthermore, they are easy to process because they are homogeneous in format and structure [7]. This processing is different for unstructured or heterogeneous sources such

 ISBN: 978-1-61208-953-9

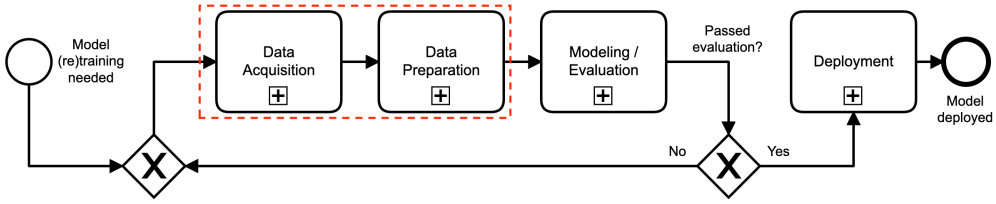

Fig. 1. Process overview for translation model development in close cooperation with the company

that more effort is required to prepare the data: An example for that is the Web. Due to the immense growth of publicly accessible multilingual websites, the Web has become an important data source for parallel corpora. A distinction must be made between approaches that consider the entire Web as a corpus [8] and approaches that use isolated pages [9] or related pages for defined subject areas [10]. Therefore, parallel corpora based on the Web exist and are widely used [11]. However, due to legal reasons not everything available online may also be used [5].

The general approach to developing parallel corpora on heterogeneous datasets is constantly evolving. It can be refined in detail depending on languages [7], sources [5], and data quality [4], but for the most part includes the following steps [5]: Obtain, Extract, Split, Prepare and Map/Alignment. A key step in this process is alignment. Alignment can be done at document, paragraph, and sentence levels. The final dataset is usually a parallel corpus at sentence level so that the sentence alignment is rarely omitted. However, alignment is not a new approach in machine translation [12]. Still, the underlying technologies have evolved immensely in recent years [13], making mass semantic alignment across hundreds of languages possible in a reasonable amount of time. Whether sentences are considered parallel depends on the alignment score. This score indicates the similarity of the source and target sentences. Existing parallel corpora often differ in the score from which sentences are considered parallel. The higher the average alignment score, the higher the quality of the training data. However, a high minimum score (e.g., 95%) leads to a significant reduction in sentences.

Less research exists in the area of business- and domain-specific parallel corpora. At the same time, commercial providers are aware of this need: The "*Microsoft Custom Translator*", for example, can be fine-tuned on datasets while using domain-specific base models (e.g., chemistry, art, agriculture). However, existing approaches can be well applied to internal company documents and domain-specific sources. A particular situation is that documents often have to be collected and merged from various platforms, data lakes, cloud services, and numerous providers and external suppliers in many different and partly proprietary formats. Ownership of the relevant documents may be subject to various departments, leading to conflicts. We will take a closer look at this situation in the following.

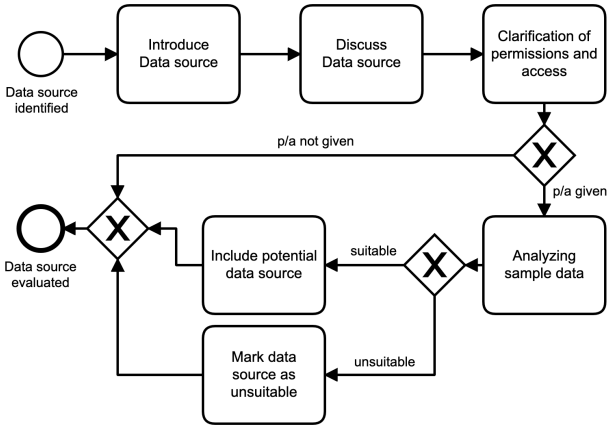

Fig. 2. Data acquisition process

III. Approach

Training an in-house translation solution requires data. The motivation of this approach is to obtain sufficient multilingual documents throughout the company, some of which are only accessible in the various functional departments. Here, a significant part is convincing people to participate, "demystifying" the process, and transparently communicating the individual steps of the necessary data acquisition and preparation.

A. Convince and involve people

Usually, individual departments place the order for a translation solution, often through internationally active units such as the "International Marketing" department. In these cases, the contact persons rarely have a technical background and often do not have permission to use the necessary data. To start the whole process, we recommend to "de-mystify" the process. This increases customer support and brings creativity to the data acquisition step and trust in the process and the results. It also makes it easier for contact persons within the company to communicate the process and obtain support.

From a process perspective, our approach (see Figure 1) at the top level is based on CRISP-DM [14]. This is a very well-known approach for data mining projects, which is already known in many companies. In the "Data Acquisition" step (see Figure 2), data sources are identified that contain potentially valuable data such as who owns this data, how it can be used, and how access can be established are examined. Once initial test data is available on the identified data source, it is

 ISBN: 978-1-61208-953-9

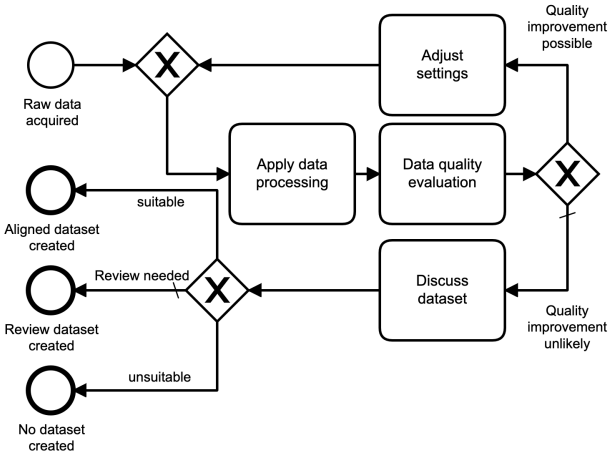

Fig. 3. Data preparation process

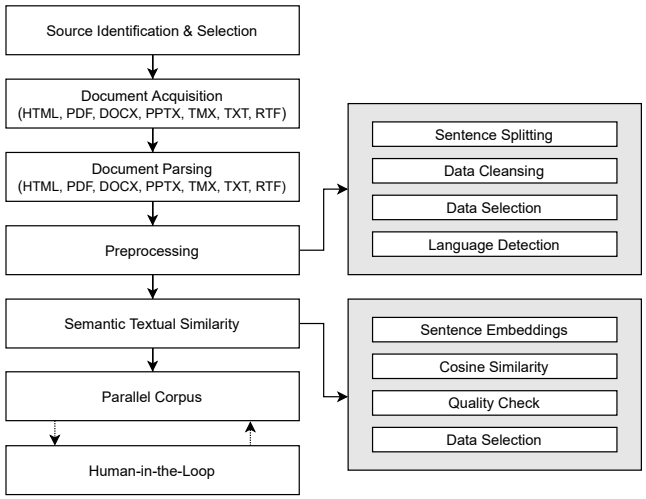

Fig. 4. Own corpus construction process

analyzed, and examples are used to discuss what the strengths and weaknesses are. If the strengths prevail, the data sources are included and considered in the data preparation step (see Figure 3). It is essential in this step to communicate and give examples of why data sources are suitable and why not. It is also advisable to ask for alternative formats that are easier to process in the case of proprieties or complex data formats. For example, the underlying DOCX files are often still available for PDF files.

Documents from accessed data sources are further processed in the "Data Preparation" step. This is a primarily automated step, which is discussed in Section III-B. In some cases, with high-quality datasets, these are of outstanding quality, so they can be adopted directly. In other cases, a "review" is necessary, since, e.g., faulty pagination due to OCR reduces the quality. These errors are quickly corrected but require manual reworking (cf. Section III-B0d). In addition, the evaluation process may also reveal that the resulting sentence pairs do not meet the requirements. If adjusting the settings (e.g., Alignment Score, Language Detection Confidence) cannot improve the quality, the dataset must be discarded. Here, it is essential to communicate the procedure and reasons to pay attention to any deficiencies in other data sources.

B. Automate repetitive, time-consuming tasks

Finding, extracting, preparing, and matching appropriate sentences across thousands of documents in numerous languages requires software assistance. In the following, we explain our approach (see Figure 4), using the processing steps from Koehn (2005) as mentioned in Section II.

a) Obtain & Extract: The origin of the data is diverse. In our real-world example, the information is located in two Content Management Systems (CMS), two Translation Management Systems (TMS), and a Product Database (PIM). There are also more than 5,000 PDFs containing general promotional material and stakeholder information. These PDFs can only be obtained from the company's website via web crawling (We use the free command-line tool *wget* to retrieve files via HTTP). The *Document Acquisition* step utilizes various Application Programming Interfaces (APIs), file operations, and database queries (e.g., MySQL) to get as many documents as possible. Often, this step is a company-specific adaptation of our standard processes since the existing systems always have peculiarities (VPN accesses, access restrictions, API modifications, limitations). The raw data varies strongly in format and structure. On the one hand, documents are available as plain text, standardized XML (this includes DOCX, TMX), or in Excel spreadsheets. On the other hand, documents can also appear as PDF files and InDesign files, which have to be converted to HTML in the *Document Parsing* step first.

b) Split & Prepare: The motivation behind the *Preprocessing* step is to split texts into sentences (Sentence splitting is done with *sentence-splitter*, v1.4) and detect the language (We use *langdetect*, v1.0.9 for language detection). It also removes sentences that contain incorrect characters, are incorrectly encoded, or consist of less than three or more than 50 words. Sentences in languages that are not needed for the translation system are also removed. As a result, sentences are stored in individual files per language.

c) Map: This step ensures similarity on the one hand but also checks how similar the sentences are in terms of distance in characters and words. The alignment of semantically highly similar sentences is done by using SentenceTransformers [15] together with the LaBSE model. The LaBSE model supports 109 languages and works well for finding translation pairs in multiple languages. As a result, semantically identical sentences are expected here, which exist in two different languages. Since errors in language detection can occur in a few cases, character distance and cosine distance are used to ensure that these are not identical sentences of one language that differ only in a few words or characters (e.g., OCR errors). If the semantic similarity is very high (90%+) and distance in characters and words is given, a very good match is assumed,

 ISBN: 978-1-61208-953-9

and the sentence pair is taken over into the parallel corpus.

d) Dataset Evaluation & Human-in-the-Loop: In cases where the data quality is supposedly not good enough, or inadequacies are detected automatically, a review process takes place. In this process, the sentence pairs are presented to a user for correction and approval (We use the Prodigy tool in this context). Users can edit both sentences and provide a comment, which can be helpful for further editing of the dataset. Once a sentence pair has been corrected, it is added to the training set and considered when the translation models are trained again. Furthermore, users of the translation software can specify that translations should be transferred to the review process if they are dissatisfied with the translation. In this way, incorrect translations are also corrected and taken into account in the next training of the models.

IV. Discussion

We believe that many companies have enough data to train translation solutions. One challenge is to access this data within a company. As shown, we use appropriate processes and technical tools for this purpose.

However, it is essential to understand and communicate that, ideally, this is a continuous process within the company. Language is subject to constant change, and domain-specific language changes and expands. It is necessary to continually train language models and also to constantly include new multilingual documents in this retraining process. Here, care must be taken to ensure that the rights to the documents remain with the company. Often, the idea is to crawl competitor websites or to use third-party product manuals as a data source. For legal reasons, we advise against this approach. We also check whether documents made available within the company meet the legal requirements for use.

A crucial role in improving the quality of in-house translation models is to involve employees in this process via the review process. We have had good experience enabling employees to report and correct "bad" translations. This improves the language model for everyone in the company and creates a sense of engagement and participation.

V. Conclusion

When it comes to self-hosted translation solutions tailored to a company's language, the question of suitable training data quickly arises. In this paper, we have shown how we support companies in finding relevant datasets and preparing them so that they can be used for fine-tuning translation models. When working with companies, it often becomes apparent that valuable data is spread across various platforms, cloud storage, and systems and that there is rarely an overview of the data. Furthermore, the information is available in various file formats, some of which must be converted into text.

Furthermore, there is often no understanding of how data can contribute to a good translation system. For this reason, we have developed processes for identifying and preparing the relevant data in the company and using it to train the translation models. Furthermore, we have established evaluation and data preparation loops in terms of a human-in-the-loop approach that help to keep data quality high. We want to build on this approach with our future activities by establishing the data preparation and training process as a continuous process, thus enabling companies to develop their translation models continuously. In this context, we are already working with major partners from research and industry.

References

[1] M. Druskoczi, "Final report on the SME panel consultation on eTranslation and language technologies," European Commission, Tech. Rep., 2020.

[2] M. Junczys-Dowmunt and et al., "Marian: Fast Neural Machine Translation in C++," pp. 1–6, 2018.

[3] A. F. Aji, N. Bogoychev, K. Heafield, and R. Sennrich, "In Neural Machine Translation, What Does Transfer Learning Transfer?" in *Proceedings of the 58th Annual Meeting of the ACL*. ACL, 2020, pp. 7701–7710.

[4] L. Tian and et al., "UM-Corpus: A Large English-Chinese Parallel Corpus for Statistical Machine Translation." in *LREC*. European Language Resources Association (ELRA), 2014, pp. 1837–1842.

[5] P. Koehn, "Europarl: A Parallel Corpus for Statistical Machine Translation," in *MT summit*, vol. 5, Citeseer. ACL, 2005, pp. 79–86.

[6] M. Ziemski, M. Junczys-Dowmunt, and B. Pouliquen, "The United Nations Parallel Corpus v1.0," in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. ACL, 2016, pp. 3530–3534.

[7] M. Morishita, J. Suzuki, and M. Nagata, "JParaCrawl: A large scale web-based English-Japanese parallel corpus," *arXiv preprint arXiv:1911.10668*, pp. 3603–3609, 2019.

[8] H. Schwenk, G. Wenzek, S. Edunov, E. Grave, and A. Joulin, "CCMatrix: Mining Billions of High-Quality Parallel Sentences on the WEB," 2019. [Online]. Available: https://arxiv.org/abs/1911.04944

[9] H. Schwenk, V. Chaudhary, S. Sun, H. Gong, and F. Guzmán, "WikiMatrix: Mining 135M Parallel Sentences in 1620 Language Pairs from Wikipedia," *arXiv preprint arXiv:1907.05791*, pp. 1351–1361, 2019.

[10] C. Christodouloupoulos and M. Steedman, "A massively parallel corpus: the Bible in 100 languages," *Language resources and evaluation*, vol. 49, no. 2, pp. 375–395, 2015.

[11] Y. Zhang, K. Wu, J. Gao, and P. Vines, "Automatic Acquisition of Chinese–English Parallel Corpus from the Web ," in *European Conference on Information Retrieval*. Springer, 2006, pp. 420–431.

[12] F. J. Och, C. Tillmann, and H. Ney, "Improved Alignment Models for Statistical Machine Translation," in *1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*. ACL, 1999, pp. 20–28.

[13] N. Reimers and I. Gurevych, "Making monolingual sentence embeddings multilingual using knowledge distillation," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. ACL, 11 2020, pp. 1–14, accessed: 01-04-2022. [Online]. Available: https://arxiv.org/abs/2004.09813

[14] R. Wirth and J. Hipp, "CRISP-DM: Towards a standard process model for data mining," in *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining*, vol. 1. Springer-Verlag London, UK, 2000, pp. 29–39.

[15] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. ACL, 11 2019, pp. 1–11, accessed: 01-04-2022. [Online]. Available: http://arxiv.org/abs/1908.10084

 ISBN: 978-1-61208-953-9