# Building Machine Translation System for Software Product Descriptions Using Domain-specific Sub-corpora Extraction

**Pintu Lohar** pintu.lohar@adaptcentre.ie
ADAPT Centre, Dublin City University, Dublin, Ireland
**Maja Popović** maja.popovic@adaptcentre.ie
ADAPT Centre, Dublin City University, Dublin, Ireland
**Tanya Habruseva** thabruseva@linkedin.com
LinkedIn Corporation, Dublin, Ireland

**Abstract**

Building Machine Translation systems for a specific domain requires a sufficiently large and good quality parallel corpus in that domain. However, this is a bit challenging task due to the lack of parallel data in many domains such as economics, science and technology, sports etc. In this work, we build English-to-French translation systems for software product descriptions scraped from LinkedIn website. Moreover, we developed a first-ever test parallel data set of product descriptions. We conduct experiments by building a baseline translation system trained on general domain and then domain-adapted systems using sentence-embedding based corpus filtering and domain-specific sub-corpora extraction. All the systems are tested on our newly developed data set mentioned earlier. Our experimental evaluation reveals that the domain-adapted model based on our proposed approaches outperforms the baseline.

## 1 Introduction

The development of Machine Translation (MT) systems for a specific domain (e.g., science, politics, economics) is often a challenging task because of the lack of parallel corpus in these domains. It is impractical to develop large corpora in every domain as it requires a huge amount of time and cost even for a single domain. There can be following methods to build MT systems in this scenario: (i) training an MT system on the available data set from other domains while tuning the model parameters on in-domain development set, or (ii) extracting in-domain parallel texts from one or more corpora and then building an MT system on the concatenation of these extracted text pairs. The first method, although tuned on an in-domain development data, is not much useful because the training is done only on an out-of-domain data set. In contrast, the second method is better because it is aimed to extract the in-domain data which are then used for training. However, producing a sufficiently large in-domain data set is a difficult task. In this work, we mainly focus on the second method, i.e, we extract parallel texts similar to in-domain in order to build an MT system and also tune the parameters on the in-domain data set. This approach is useful for building MT system in a specific scenario, which is the domain of software product descriptions from LinkedIn[1] web pages in our case. LinkedIn is an American business and employment-oriented online service that operates via websites and mobile apps.

[1] https://linkedin.com

It is primarily used for professional networking and career development, and allows job seekers to post their curricula vitae (CVs) and employers to post jobs. It also contains product pages for brands to promote their products and grow their businesses, for product users to share their experiences and be recognised for their expertise, and for buyers to make confident decisions about products in a trusted environment. This work involves an initial analysis of domain-specific MT for public taxonomies on software product descriptions available in LinkedIn web pages using a novel approach of sub-corpora extraction. Our contributions in this work are as follows: (i) we develop a first ever parallel development and test corpus of software product descriptions which are originally written in English and then manually translated into French; we used this corpus to tune the system parameters and to test our MT systems, and (ii) we investigate methods for filtering a parallel corpus; first, we use LASER (Artetxe and Schwenk, 2019), the state-of-the-art tool for bitext mining with the help of measuring bilingual sentence similarity and then we use KeyBERT (Grootendorst, 2020) to extract key phrases from the in-domain data set developed by us, which is further used to extract relevant parallel texts from several corpora. More precisely, we exploit our development data set to extract parallel data which is similar to the domain of software production. We refer to this extracted data as sub-corpus. Afterwards, we build MT systems using these sub-corpora for training and the same development data set for tuning parameters. Finally, we evaluate all the systems on a separately held-out test data set. Our experimental results show that our approach of corpus filtering and keyphrase-based sub-corpus extraction improves the performance of the MT system even when trained on a much smaller data set.

## 2 Related Work

NMT has undergone huge evolution during the last few years. For example, in the shared tasks on News and biomedical translation in WMT 2019, it is found that several NMT systems perform at the same level of a human translator for some high-resource language pairs according to human judgement (Barrault et al., 2019a; Bawden et al., 2019). However, for many language pairs and for many domains, there is still no (sufficient) data available in order to build high-quality MT systems.

Domain adaptation is a well-explored research area in MT. The main objective is to facilitate adaptation of the MT system to a specific domain. For example, Hu et al. (2019) propose an approach of lexicon induction to extract an in-domain lexicon and then build a pseudo-parallel in-domain corpus with word-for-word back-translation of monolingual in-domain target sentences. Chu and Wang (2018) conduct a survey of the state-of-the-art domain adaptation techniques for neural machine translation (NMT). The work of (Poncelas et al., 2019) demonstrates the usefulness of Infrequent n-gram Recovery (INR) and Feature Decay Algorithms (FDA) for domain adaptation. Back-translation (or forward translation) is often used for domain adaptation, too (Hoang et al., 2018; Graça et al., 2019).

The vast majority of investigated MT systems covers only a limited set of domains, predominantly news (Akhbardeh et al., 2021; Barrault et al., 2020, 2019b). There is also work on biomedical domain, spoken language (Bérard et al., 2020; Duh, 2018) and some types of user-generated content (Lohar et al., 2019; Xu and Yvon, 2021). However, to the best of our knowledge, there is no previous work that involves the development of MT systems for software product descriptions. Moreover, no parallel corpus in this domain has been published so far.

## 3 Data Development

We develop the first ever corpus of software product descriptions in English and their manual translations into French. The corpus is suitable for development and testing of MT systems in this domain. The data set is collected from the LinkedIn webpage of software product de-

scriptions.[2] We scrape the contents of webpages and collect $1,395$ text segments in English on the description of software products. These texts are then manually translated by native French speakers who are also proficient in English. Product descriptions are usually different from natural texts and should be translated with special considerations. Bearing this in mind, the translators used the following guidelines to perform the translation task:

- some of the texts are not full sentences, which is perfectly acceptable in product-related texts and they need to be translated without considering the whole context,
- some of the texts contain URLs which should be left untranslated,
- names of the software products which contain valid English words should remain untranslated, and
- the translators should not delete any symbols or unwanted characters during the translation process

| **English text** | **French translation** |
|---|---|
| Kronologic is the world's first Calendar Monetization Platform. | Kronologic est la première plateforme mondiale de monétisation calendaire. |
| Cerebra is an Artificial Intelligence Platform powering connected operations, impacting Yield, Reliability, and Operational Excellence in a Sustainable way. | Cerebra est une plateform d'Intelligence d'Intelligence Artificielle alimentant des opérations connectées, impactant le rendement, la fiabilité et l'excellence opérationelle à travers une démarche durable. |
| - Multi-platform endpoint remote monitoring and management (RMM) | - Télésurveillance et télégestion de bout en bout multiplateforme |
| Prodoscore™ is a software solution that measures your most valuable asset: your people. | Prodoscore™ est une solution logicielle qui mesure vos actifs de plus grande valeur: votre personel. |

Table 1: Some example translations

Table 1 shows some example translations done by the native French speakers. As mentioned earlier, some texts in the data set are not full sentences. For instance, example 3 in the above table can be considered as an incomplete sentence or merely a text segment. Such segments are often seen in product descriptions and so the French translation is done accordingly.

The whole translation process was a challenging task and took a significant amount of time. One of the reasons was the presence of a large number of software or technical terms, some of which are not easy or straightforward to translate into French. People often use them as is, i.e, they keep them in English instead of translating into French. For example, the phrase "stacking-plans" was found to be very difficult to translate into French as its literal translation does not make much sense and therefore it should be left untranslated. In addition, the translators have to remember which terms should be translated and which should not, as they encounter many such terms. For example, the term "Cerebra Digital Assistants" should not be translated as it is the name of a software.

[2] `https://www.linkedin.com/products`

Once the translation of $1,395$ segments is complete, $696$ were held out as tuning data set and the rest $699$ as test data set. Therefore, all the MT models are tuned on these $696$ segment pairs and evaluated on the $699$ segment pairs. We refer to this data set as *SWP* corpus[3] which is now available online for free access.

## 4 Corpus Filtering for Domain Adaptation

In this section we describe our proposed approaches of corpus filtering and sub-corpora extraction for domain adaptation using LASER and KeyBERT. Here we use LASER for filtering and KeyBERT is used for extracting domain-specific sub-corpora, respectively.

### 4.1 Corpus Filtering

LASER is a state-of-the-art tool for calculating the Euclidean distance between a pair of bilingual sentences in order to measure their semantic similarity. This means that the smaller the distance, the more similar the sentence pairs. We use this score to filter out the pairs with low similarity score. To investigate its usefulness, we filter out the low-scoring sentence pairs from Europarl corpus (Koehn, 2005) and then the filtered corpus is used to train an MT model. On the other hand, the whole Europarl corpus is used to build a separate MT model. Table 2 shows the BLEU score produced by these three MT models, when evaluated on the test data set.

| Training corpus | #Sentence pairs | BLEU score |
|---|---|---|
| Europarl (all) | 2.05 million | 19.06 |
| Europarl (LASER-filtered) | 1.93 million | **19.82** |

Table 2: BLEU scores with and without corpus filtering

It can be seen from the table that the model trained on the LASER-filtered corpus produces better BLEU score than without filtering. We also used LABSE (Feng et al., 2022) with their optimal settings but it produced less BLEU score than that of LASER. We therefore decided to proceed with LASER filtering for the remaining experiments.[4]

### 4.2 Domain-specific Sub-corpora Extraction

Our approach of domain adaptation is different from the existing works. We compile the in-domain data by extracting sub-corpora (a part of the parallel corpus) using KeyBERT along with a tuning process. KeyBERT uses BERT-embeddings (Reimers and Gurevych, 2019) and the cosine similarity to find the key phrases in a document that are the most similar to the document itself. These key phrases can also be considered as key terms of a document. Usually, KeyBERT finds top n key terms from a document. Our goal is to find such terms from our development data set and then extract only those sentence pairs (from a corpus) that contain at least one of these key terms. However, it is not a good idea to simply extract an arbitrary number of key terms. We consider it as a tuning process and started with $n = 2,000$ and increase the value gradually. In order to identify the number of key term that should be extracted, we again use the Europarl corpus in the following steps: (i) top $n$ key terms are extracted from development data and then only those sentences that contain at lease one of these key terms are extracted from Europarl, (ii) an MT model is trained on these extracted sentence pairs and is then evaluated on the test data to calculate BLEU score, (iii) the value of $n$ is increased and we proceed from the first step, (iv) all the above steps are repeated until we obtain the highest BLEU score.

[3] https://github.com/loharp/SWP

[4] However, in future it will interesting to see how the combination of LASER and LABSE performs in corpus filtering

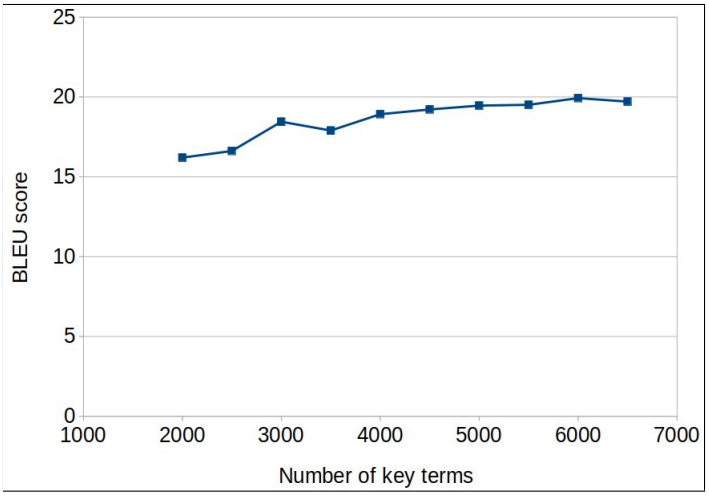


Figure 1: Tuning Europarl with different key term values

It is noticed that using a small value of $n$ results in extraction of very small number of sentence pairs from Europarl and so the MT model built from it produces a low BLEU score. Similarly, using a very large value of $n$ results in extraction of almost all sentence pairs from Europarl and therefore is not much helpful. Figure 1 shows the BLEU scores obtained when different values of $n$ is considered starting from $2,000$. This is also shown in Table 3 where we also mention the number of parallel sentences extracted for each key term values. Note that

| #Key terms used | #Parallel sentences extracted | BLEU score |
|---|---|---|
| N/A | 2.05M (all) | 19.06 |
| 2000 | 208K | 16.20 |
| 2500 | 292K | 16.62 |
| 3000 | 611K | 18.45 |
| 3500 | 1.04M | 17.90 |
| 4000 | 1.06M | 18.92 |
| 4500 | 1.21M | 19.22 |
| 5000 | 1.23M | 19.46 |
| 5500 | 1.34M | 19.51 |
| **6000** | **1.41M** | **19.93** |
| 6500 | 1.72M | 19.71 |

Table 3: Tuning Europarl with key term values

the first row in this table shows the scenario where no key terms are used and all the sentence pairs in the corpora are used to build the MT model. It produces BLEU score of $19.06$. In the second row $2,000$ key terms are used but they are capable of extracting only $208K$ sentence pairs which is insufficient for MT training and hence produces comparatively lower BLEU score of $16.20$. Afterwards, we continue to increase the number of key terms and notice that the BLEU score rises with the increase of key terms. It can be seen that the highest BLEU score is achieved when $6,000$ key terms are used. Using further higher value results in bringing the number extracted sentence pairs closer to the total number of sentence pairs in the whole

Europarl corpus but it cannot increase the BLEU score. Instead, the score decreases and hence shows that using the whole corpus may not be useful. Note that the optimal BLEU score of $19.93$ is obtained with $1.41$ million extracted sentence pairs which is much less than that of the original corpus ( $30\%$ less).

Once the tuning of Europarl with LASER and KeyBERT is done, we consider $6,000$ as the optimal value for key terms and use this optimal value for our further experiments on the larger data set.

## 5 Experiments

- **Data set:**

  There are a number of different parallel corpora available but not all of them are suitable for building a decent quality MT system. We explore a wide range of corpora available on the OPUS website.[5] We use Europarl corpus for tuning as mentioned earlier in Section 4. The optimal value of key terms is then applied to extract sub-corpora from a set of other corpora. Although we use several corpora in the initial stage of experiments, we consider using $12$ specific corpora in our later stage of experiments and filter them using our proposed approach as discussed earlier. The statistics of the data sets used are is shown in Table 4

| Corpus name | #Parallel sentences | Domain |
|---|---|---|
| Europarl | 2.05M | Mixed domain |
| XLEnt | 7.7M | Mixed domain |
| ELITR-ECA | 0.4M | European Court of Auditors |
| TED2020 | 0.4M | TED talks |
| GNOME | 0.9M | Software |
| QED | 1.0M | Educational |
| PHP | 45K | Software |
| GlobalVoices | 0.2M | News |
| TED2013 | 0.2M | TED talks |
| Tatoeba | 0.3M | Mixed domain |
| Ubuntu | 7K | Software |
| KDE | 0.2M | Software |

Table 4: Data sets used in our experiments

Following is a short description of the corpora we used in our experiments.

- **Europarl**: A parallel corpus extracted from the European Parliament web site. The main intended use is to aid statistical machine translation research.
- **XLEnt** (El-Kishky et al., 2021): This corpus was created by mining web data from Commoncrawl Snapshots and Wikipedia snapshots.
- **ELITR-ECA** (Williams and Haddow, 2021): This is a multilingual corpus derived from documents published by the European Court of Auditors.[6]
- **TED2020** (Reimers and Gurevych, 2020): This corpus contains a crawl of nearly $4,000$ TED and TED-X transcripts from July $2020$. The transcripts have been translated by a global community of volunteers to more than $100$ languages.

[5] https://opus.nlpl.eu/

[6] https://www.eca.europa.eu/

- **GNOME** (Tiedemann, 2012) : A parallel corpus of GNOME localization files.
- **QED** (Abdelali et al., 2014): The QCRI Educational Domain Corpus is an open multilingual collection of subtitles for educational videos and lectures collaboratively transcribed and translated over the AMARA web-based platform.
- **PHP** (Tiedemann, 2012): A parallel corpus originally extracted from a website containing documentation of PHP.[7] The original documents are written in English and have been partly translated into 21 languages.
- **Global Voices** (Tiedemann, 2012): A parallel corpus of news stories from the web site Global Voices compiled and provided by CASMACAT. [8]
- **TED2013** (Tiedemann, 2012): A parallel corpus of TED talk subtitles provided by CASMACAT.
- **Tatoeba** (Tiedemann, 2012): This is a collection of translated sentences from Tatoeba [9]
- **Ubuntu** (Tiedemann, 2012): A parallel corpus of Ubuntu localization files.
- **KDE** (Tiedemann, 2012): A parallel corpus of KDE4 localization files

Note that many of the above-mentioned corpora are published by Tiedemann (2012).

Although we could have used more corpora, we decided to use the above 12 corpora because of the following reasons: (i) after manually inspecting several corpora in a random manner, we found the above corpora to be good quality[10] (ii) many of them contain texts from multiple domains and thus are better to be used for domain adaptation using subcorpora extraction, and (iii) some of them are from software domain which is useful for our experiments. We also built a separate MT model using only those corpora that belong to software domain but we obtain low BLEU score of 10.41 as they are small in size. Due to this reason, we decided to combine other corpora as well.

- **Tools and Evaluation Metrics:**

  Initially we use LASER and KeyBERT (discussed in Section 4) for corpus filtering and sub-corpus extraction, respectively. To build MT models, we use OpenNMT[11] (Klein et al., 2017) with transformer architecture (Vaswani et al., 2017). Subword NMT[12] (Sennrich et al., 2016) is used to apply Byte-pair enconding (BPE) during the preprocessing. We use sacreBLEU (Post, 2018) for automatic evaluation of MT outputs.

- **Preprocessing:**

  We perform preprocessing in the following steps:

  (i) **Filtering out long sentences:** Extremely long sentences were deleted. If either side contains too many words (100 words is set as default limit), the sentence pair is discarded.

  (ii) **Removing blank lines:** Sentence pairs with no content on either side are removed.

  (iii) **Removing sentence pairs with odd length ratio:** Sentences with marginally longer or shorter translations when compared to their original sentences were removed because of the probability of their being incorrect translations. The filtering ratio is $1:3$ in our case.

[7] http://se.php.net/download-docs.php.

[8] http://casmacat.eu/corpus/global-voices.html

[9] https://tatoeba.org/en/

[10] However, we inspected only a tiny part of parallel corpus in a random manner. Inspecting whole corpora would be more useful but this is extremely impractical to achieve in a reasonable amount of time.

[11] https://github.com/OpenNMT/OpenNMT-py

[12] https://github.com/rsennrich/subword-nmt

(iv) **Removing duplicates:** All duplicate sentence pairs were discarded.

(v) **Tokenisation:** We break down the sentences into their most basic elements called tokens. Tokenisation is particularly relevant because it is the form in which MT models ingest sentences. In practice, most NMT models are fed with sub-words as tokens.

(vi) **Byte-Pair-Encoding (BPE):** In many cases, most out-of-vocabulary (OOV) words have similar morphemes to some of the words already in our vocabulary. With this in mind, the BPE technique was leveraged to resolve the OOV problem by helping the model infer the meaning of words through similarity. The BPE algorithm performs sub-word regularization by building a vocabulary using corpus statistics. Firstly it learns the most frequently occurring sequences of characters and then greedily merges them to obtain new text segments.

- **Building baseline MT systems:** In the first stage of our experiments, we explored several corpora and built different MT systems with different corpora individually and also with some combinations of them. We select one system from them that produces the best BLEU score and consider it as our baseline. Table 5 shows the BLEU scores obtained during building the baseline system.

| System name | Corpus used | BLEU score |
|---|---|---|
| Sys-1 | Europarl | 19.06 |
| Sys-2 | United Nations Parallel corpus (UNPC) | 12.81 |
| Sys-3 | Four different corpora:<br>Gnome, KDE, PHP and Ubuntu (GKPU) | 10.41 |
| **Sys-4**<br>**(Baseline)** | 12 different corpora:<br>ELITA, Europarl, PHP, TED2020,<br>XLEnt, Gnome, Global voice, KDE,<br>QED, TED2013, TATOEBA and Ubuntu | **30.17** |

Table 5: BLEU scores during building baseline

Table 5 shows that *Sys-1* and *Sys-2* are built from only one corpora. However, they do not produce the best BLEU score. Although the UNPC corpus is much larger than Europarl, it produces much less BLEU score because this corpus contains plenty of noise and so the MT system built from it is of low quality. We then explore some combined corpora to build *Sys-3* and *Sys-4*. These MT systems are trained from the combinations of 4 and 12 corpora, respectively. We initially consider the 4 corpora *GKPU* for MT training as it comprises of the corpora from software domain only. However, this combination still yields a small-sized corpus and therefore cannot produce a decent BLEU score. The best score is obtained by *Sys-4* which is trained from the concatenation of 12 different corpora. As explained earlier in this section we use this combination because all of them appeared to be good quality according to our random manual inspection and some of them are from software domain which is useful for us. We consider this as the baseline system for our further experiments. Note that there are numerous possible combinations of several corpora but it is impractical to try all of them. Our main focus is to select the combination that produces a decent BLEU score and proceed with the next stage of experiments on further improvement of MT systems with the same corpora combination.

- **Building domain-adapted MT system:** The domain-adapted MT system is the upgraded versions of the baseline system. The upgrade comes with our proposed approach of filtering and sub-corpus extraction. Firstly, we filter the concatenation of 12 different corpora

(shown in Table 5) using LASER and then extract a sub-corpora from the filtered corpus using the optimal key term value of $6,000$ as determined in Section 4. This results in a massive reduction in corpus size. Our domain-adapted model is built from this reduced corpus. Table 6 describes the baseline and the domain-adapted systems along with the corpus description.

| **MT system** | **Corpora used** | **#Training sentences** | **#Dev sentences** | **#Test sentences** |
|---|---|---|---|---|
| Sys-4 (Baseline) | 12 corpora (Original) | 15 million | 696 | 699 |
| Sys-5 (Domain-adapted) | 12 corpora (Filtered) | 4.73 million | | |

Table 6: Data distribution of Baseline and Domain-adapted systems

## 6 Results

Both the baseline and the domain-adapted systems are tuned on 696 and 699 texts from the data set of software product descriptions developed by us. The result is shown in Table 7 below.

| **MT system** | **#Sentence pairs** | **BLEU score** |
|---|---|---|
| Baseline | 15 million | 30.17 |
| Domain-adapted | 4.73 million | **31.47** |

Table 7: Baseline vs Domain-adapted system

We can notice from the table that the domain-adapted system outperforms the baseline system by $1.4$ BLEU points which is $4.3\%$ relative improvement. Another important observation is that both systems are trained on the same corpora but the domain-adapted system is the filtered version of it. Our proposed approach significantly reduces the corpus size by more than three times and at the same time produces the higher BLEU score.

## 7 Output Analysis

In this section we show some of the translation outputs produced by our MT system and compare them with the human translated references. Table 8 shows some examples translation outputs. In the first example of this table the output produced by our MT system is a very good translation but it misses the translation of *Instant* as compared to the reference translation. The second and the third outputs are the examples of excellent or perfect translation. The fourth example is an interesting and surprising one. Some native French speakers claimed that they would prefer this MT output than the reference translation. However, there are few cases where our MT system fails to produce good quality translations. For example, in the fifth output 'meet' and 'chat' still remain untranslated. The last example in the above table is a very good translation output except that the word 'empowers' has different ways to be translated into French. In fact, this word do not fit well in French with this context and so is difficult to translate. Moreover, although 'leurs' and 'les' are both correct but 'leurs' is better than 'les'.

## 8 Conclusions and Future work

In this work we proposed an approach of corpus filtering and sub-corpora extraction by using LASER and KeyBERT. In addition, we developed the first ever parallel corpus of software

| English text | French translation | Reference translation |
|---|---|---|
| Instant analysis of millions of real-time conversations happening online. | L'analyse de millions de conversations en temps réel se déroule en ligne. | Analyse instantanée de millions de conversations en ligne en temps réel. |
| Find out what your customers are doing, feeling, and thinking when it comes to your brand. | Trouvez ce que vos clients font, ressentent et pensent quand il s'agit de votre marque. | Découvrez ce que vos clients font, ressentent et pensent de votre marque. |
| No hidden fees or long-term commitments. | Pas de frais cachés ou d'engagements à long terme. | Pas de frais cachés ni d'engagements à long terme. |
| Know what matters, when it matters. | Savoir ce qui compte, quand c'est important. | Sachez ce qui est important et quand c'est important. |
| Meet, call, chat, share files, manage tasks. | Meet, appel, chat, partager des fichiers, gérer des tâches. | Réunions, appels, discussions, partage de fichiers, gestion des tâches. |
| Learn more about FICO Analytics Workbench, a next-generation analytics tool that empowers companies to improve business decisions across the customer lifecycle. | Apprenez-en plus sur FICO Analytics Workbench, un outil d'analyse de nouvelle génération qui permet aux entreprises d'améliorer les décisions commerciales à travers le cycle de vie du client. | En savoir plus sur FICO Analytics Workbench, un outil d'analyse de dernière génération qui permet aux entreprises d'améliorer leurs décisions commerciales tout au long du cycle de vie du client. |

Table 8: Some examples of translation outputs

product description suitable for development and testing, consisting of English product descriptions and their human translations into French. Our approach significantly reduces (more than three times) the corpus size and at the same time increases the BLEU score by $1.3$ points which is more than $4\%$ relative improvement over the baseline. This technique can easily and effectively be applied to any corpus in order to adapt or transform it into a refined corpus that is more similar to a specific domain. Moreover, the first ever corpus of software product descriptions developed by us can be beneficial for many researchers who are interested in building MT system in this domain. The data set is now freely available online. In future, our work can be extended by using the combination of multiple approaches such as LASER and LABSE together with KeyBERT. In addition, it is also possible to take the intersection of the filtered corpora obtained by applying LABSE and LASER separately. Afterwards, the intersection can be further refined by using KeyBERT. Moreover, it is to be noted that the developers of LABSE mentioned in their paper that they determine the optimal similarity threshold of $0.6$ after several trials on different corpora. However, it is possible to re-optimize the threshold for a specific domain such as ours and then select the threshold that exhibits the best performance. The overall translation quality produced by our MT system appeared to be very good after manual inspection in a random manner. However, it is better to perform detailed human evaluation on the translation outputs. Although it is a time consuming task, it is better to manually evaluate at least a small part of the translation outputs which will provide the clearer picture of translation quality to some extent.

## Acknowledgements

This work is funded by joint collaboration with Linkedin and the ADAPT Centre for Digital Content Technology which is funded under the Science Foundation Ireland (SFI) Research Centres Programme (Grant No. 13/RC/2106).

## References

Abdelali, A., Guzman, F., Sajjad, H., and Vogel, S. (2014). The AMARA corpus: Building parallel language resources for the educational domain. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1856–1862, Reykjavik, Iceland. European Language Resources Association (ELRA).

Akhbardeh, F., Arkhangorodsky, A., Biesialska, M., Bojar, O., Chatterjee, R., Chaudhary, V., Costa-jussa, M. R., España-Bonet, C., Fan, A., Federmann, C., Freitag, M., Graham, Y., Grundkiewicz, R., Haddow, B., Harter, L., Heafield, K., Homan, C., Huck, M., Amponsah-Kaakyire, K., Kasai, J., Khashabi, D., Knight, K., Kocmi, T., Koehn, P., Lourie, N., Monz, C., Morishita, M., Nagata, M., Nagesh, A., Nakazawa, T., Negri, M., Pal, S., Tapo, A. A., Turchi, M., Vydrin, V., and Zampieri, M. (2021). Findings of the 2021 conference on machine translation (WMT21). In *Proceedings of the Sixth Conference on Machine Translation*, pages 1–88, Online.

Artetxe, M. and Schwenk, H. (2019). Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Trans. Assoc. Comput. Linguistics*, 7:597–610.

Barrault, L., Biesialska, M., Bojar, O., Costa-jussà, M. R., Federmann, C., Graham, Y., Grundkiewicz, R., Haddow, B., Huck, M., Joanis, E., Kocmi, T., Koehn, P., Lo, C.-k., Ljubešić, N., Monz, C., Morishita, M., Nagata, M., Nakazawa, T., Pal, S., Post, M., and Zampieri, M. (2020). Findings of the 2020 conference on machine translation (WMT20). In *Proceedings of the Fifth Conference on Machine Translation*, pages 1–55, Online.

Barrault, L., Bojar, O., Costa-jussà, M. R., Federmann, C., Fishel, M., Graham, Y., Haddow, B., Huck, M., Koehn, P., Malmasi, S., Monz, C., Müller, M., Pal, S., Post, M., and Zampieri, M. (2019a). Findings of the 2019 conference on machine translation (WMT19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61, Florence, Italy.

Barrault, L., Bojar, O., Costa-jussà, M. R., Federmann, C., Fishel, M., Graham, Y., Haddow, B., Huck, M., Koehn, P., Malmasi, S., Monz, C., Müller, M., Pal, S., Post, M., and Zampieri, M. (2019b). Findings of the 2019 conference on machine translation (WMT19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61, Florence, Italy.

Bawden, R., Bretonnel Cohen, K., Grozea, C., Jimeno Yepes, A., Kittner, M., Krallinger, M., Mah, N., Neveol, A., Neves, M., Soares, F., Siu, A., Verspoor, K., and Vicente Navarro, M. (2019). Findings of the WMT 2019 biomedical translation shared task: Evaluation for MEDLINE abstracts and biomedical terminologies. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 29–53, Florence, Italy.

Bérard, A., Kim, Z. M., Nikoulina, V., Park, E. L., and Gallé, M. (2020). A multilingual neural machine translation model for biomedical data. In *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*, Online.

Chu, C. and Wang, R. (2018). A survey of domain adaptation for neural machine translation. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1304–1319, Santa Fe, New Mexico, USA.

Duh, K. (2018). The multitarget ted talks task. http://www.cs.jhu.edu/~kevinduh/a/multitarget-tedtalks/.

El-Kishky, A., Renduchintala, A., Cross, J., Guzmán, F., and Koehn, P. (2021). XLEnt: Mining cross-lingual entities with lexical-semantic-phonetic word alignment. In *Preprint*, Online.

Feng, F., Yang, Y., Cer, D., Arivazhagan, N., and Wang, W. (2022). Language-agnostic BERT sentence embedding. In Muresan, S., Nakov, P., and Villavicencio, A., editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, May 22-27, 2022*, pages 878–891, Dublin, Ireland.

Graça, M., Kim, Y., Schamper, J., Khadivi, S., and Ney, H. (2019). Generalizing back-translation in neural machine translation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 45–52, Florence, Italy.

Grootendorst, M. (2020). Keybert: Minimal keyword extraction with bert.

Hoang, V. C. D., Koehn, P., Haffari, G., and Cohn, T. (2018). Iterative back-translation for neural machine translation. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 18–24, Melbourne, Australia.

Hu, J., Xia, M., Neubig, G., and Carbonell, J. (2019). Domain adaptation of neural machine translation by lexicon induction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2989–3001, Florence, Italy.

Klein, G., Kim, Y., Deng, Y., Senellart, J., and Rush, A. (2017). OpenNMT: Open-source toolkit for neural machine translation. In *Proceedings of ACL 2017, System Demonstrations*, pages 67–72, Vancouver, Canada. Association for Computational Linguistics.

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

Lohar, P., Popović, M., and Way, A. (2019). Building English-to-Serbian machine translation system for IMDb movie reviews. In *Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing*, pages 105–113, Florence, Italy.

Poncelas, A., de Buy Wenniger, G. M., and Way, A. (2019). Adaptation of machine translation models with back-translated data using transductive data selection methods. *CoRR*, abs/1906.07808.

Post, M. (2018). A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels. Association for Computational Linguistics.

Reimers, N. and Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China.

Reimers, N. and Gurevych, I. (2020). Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Sennrich, R., Haddow, B., and Birch, A. (2016). Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany.

Tiedemann, J. (2012). Parallel data, tools and interfaces in opus. In Chair), N. C. C., Choukri, K., Declerck, T., Dogan, M. U., Maegaard, B., Mariani, J., Odijk, J., and Piperidis, S., editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, Long Beach, USA.

Williams, P. and Haddow, B. (2021). The ELITR ECA corpus. *CoRR*, abs/2109.07351.

Xu, J. and Yvon, F. (2021). Can you traducir this? machine translation for code-switched input. In *Proceedings of the Fifth Workshop on Computational Approaches to Linguistic Code-Switching*, pages 84–94, Online.