Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 3808–3815
Marseille, 11–16 May 2020

The FISKMÖ Project: Resources and Tools for Finnish-Swedish Machine Translation and Cross-Linguistic Research

Jörg Tiedemann*, Tommi Nieminen, Mikko Aulamo*,
Jenna Kanerva†, Akseli Leino†, Filip Ginter†, Niko Papula◇
*Department of Digital Humanities, †Department of Future Technologies, ◇Multilizer
University of Helsinki, University of Turku, Espoo
{jorg.tiedemann, mikko.aulamo}@helsinki.fi, {jmnybl,akeele,figint}@utu.fi, niko.papula@multilizer.com

Abstract

This paper presents FISKMÖ, a project that focuses on the development of resources and tools for cross-linguistic research and machine translation between Finnish and Swedish. The goal of the project is the compilation of a massive parallel corpus out of translated material collected from web sources, public and private organisations and language service providers in Finland with its two official languages. The project also aims at the development of open and freely accessible translation services for those two languages for the general purpose and for domain-specific use. We have released new data sets with over 3 million translation units, a benchmark test set for MT development, pre-trained neural MT models with high coverage and competitive performance and a self-contained MT plugin for a popular CAT tool. The latter enables offline translation without dependencies on external services making it possible to work with highly sensitive data without compromising security concerns.

Keywords: parallel corpora, machine translation

1. Introduction

Finland is a country with two official languages, Finnish and Swedish. The demand for translation between the two languages is huge and translation efforts produce a lot of overhead and costs in public administration and organisations as well as any service providers in the country. Yet there is no systematic collection of translated data nor any large scale effort to develop translation services and tools for the public and administrative use in the country. For this reason, we started FISKMÖ[1] (finsk-svensk korpus och maskinoversättning) a joint project of the universities of Helsinki and Turku and Kites (an umbrella organization for the language sector in Finland), currently funded by the Swedish Culture Foundation.[2]

The primary goal of the project is to create a massive parallel corpus of Finnish and Swedish to support linguistic research and machine translation (MT) development. The secondary goal of the project is to establish a public machine translation service for translation between Finnish and Swedish. As the quantity and quality of training data is the most important factor in machine translation quality, the FISKMÖ translation service can be expected to provide translations of higher quality than other public machine translation services, for which less training data is available.

Our mission is to systematically collect material that has been translated from Finnish to Swedish and vice versa in the public sector, in private organisations and at language service providers in Finland. We intend to publish the data with permissive licenses to increase the impact on on-going research and development. Nevertheless, we will also include restricted data sets that can be used internally for improvements of the coverage of machine translation models and for the optimisation of domain-specific translation engines.

This paper reports the achievements of the project including the data sets we have collected so far, machine translation models we have developed and translation tools that we have released. In particular, we describe

- data provided by governmental organisations and academic institutions,
- methods for extracting parallel segments from web-crawled data,
- tools for pre-processing and alignment of parallel data,
- the implementation of a general-purpose on-line translation service and
- the development of computer-assisted translation (CAT) tools with fully integrated machine translation.

The latter is especially interesting for real-world applications in which translation efforts need to be done off-line and with strict security constraints. We released a plugin to the popular SDL Trados Studio CAT tool that integrates the MT engine inside the plugin itself making it obsolete to send (potentially sensitive) data to on-line services. More details about this tool will be provided in Section 4.2.

2. Data Collection

The primary goal of FISKMÖ is the systematic collection of Finnish-Swedish parallel data from public as well as private sources in order to improve the development of machine translation between the two languages. Furthermore, our project intends to support cross-linguistic research by making those data sets available to researchers in translation studies and general linguistics. Computer-aided language learning is another area where translation data is becoming increasingly useful. Our efforts in collecting data relate to two strategies:

[1] https://blogs.helsinki.fi/fiskmo-project/
[2] https://www.kulturfonden.fi

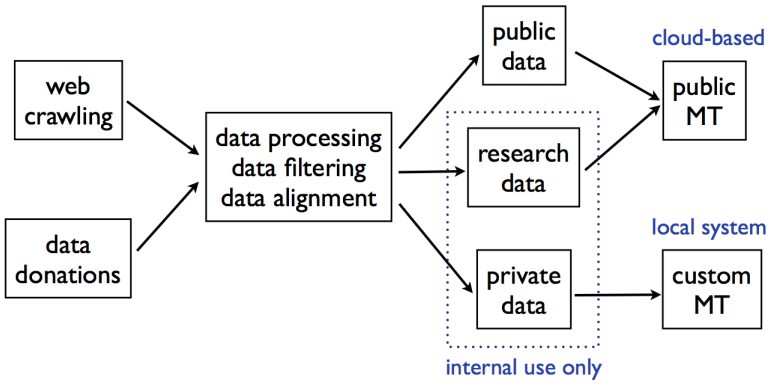

Figure 1: The basic workflow for data collection and machine translation development in the FISKMÖ project.

1. Collecting data from public and private organisations and language service providers. We also ask for general data donations from the public.

2. Web crawling and parallel segment extraction. We develop and apply methods for the identification of translated sentences and documents from unrestricted web crawls.

The workflow for data collection and its connection to machine translation development is illustrated in Figure 1. Below, we give more details about the efforts and procedures that we apply.

2.1. Collecting data from organizations

So far the majority of the data received through direct contributions has come from public organizations. Public organizations in Finland are obliged to give their data for public use if there are no good grounds to do otherwise. However, collecting the data is not always straightforward as their publication can differ greatly. Mostly, organisations refer to their websites to point to their public data sets. Unfortunately, mining parallel data from the web comes with various caveats as website structures and functionalities vary. Hence, our goal in FISKMÖ is to obtain the data directly from the original source in suitable electronic formats and with explicit correspondences between translations and their original documents. Furthermore, we expected translation memories to be in wide use internally to make it easy to incorporate collected translation knowledge in our project.

Unfortunately, the situation is more complicated than expected for numerous reasons. In some cases, translation units work independently from the publishing unit, which makes it difficult to identify translation data that is published openly without further editing and modification. Additionally, public data and sensitive data are often not systematically separated and translation memories incorporate mixed content. Furthermore, copyrights and intellectual property rights (IPRs) may not always be clear for all data files in larger collections. Those reasons and the fear of compromising privacy information make it difficult to convince even public organisations to share their data.

Fortunately, recent developments in machine translation lead to greater interest in computer-assisted translation tools in the workflow of professional translators. The appearance of the "EU Presidency Translator"[3] developed in collaboration with the Finnish Prime Minister's Office and their translation unit shed some new light on the integration of MT in human translation demonstrating the benefits of such tools in the translator's daily work.

The private sector has been an even more challenging segment despite initial promises to donate data. Finnish language service providers (LSPs) are generally not using MT in large scale and thus they are not very experienced in that technology. It seems that there is a common believe that the development of machine translation would benefit their competitors more than themselves and, therefore, it is better to hold back with handing out data. Another issue is also that language service providers usually do not own the translations they have done for their clients. It will be necessary to change the way IPRs are regulated between LSPs and their customers but permissions could easily be asked from the clients to improve services that they order from LSPs. In general , convincing LSPs to see the benefits of translation services that they can incorporate in their own workflow requires the successful demonstration of practical tools. Therefore, it is crucial to develop proper integrations into CAT tools similar to the one we introduce in Sections 4.2..

To help potential providers to donate their data, we have created different procedures they can follow to support our efforts.

Donating translation memories: In this method a translation memory is donated as is. The translation memory can be an existing translation memory but sometimes organizations prefer to create a specific one for this purpose to better control its contents. In return we offer automatically cleaned translation memories and customized MT engines based on the data if there is sufficient training material and resources available.

Crawling and filtering with translation memories: This method effectively overcomes many privacy concerns. We crawl e.g. organization websites and thus create a monolingual corpus that contains only openly published data. The organization then uses

[3] https://presidencymt.eu

sentences in those data sets to extract the translation equivalents from their internal translation memory, which they can offer to us. In this way, the filtered memory only contains public material that can be found on-line. Using this procedure we can skip complex alignment methods of partially translated open web content and the organisation avoids privacy concerns. We can also assist with the internal filtering process if necessary.

Alignment of data: A third option is to provide documents and their translations and we extract and align the data directly from those documents. The benefit for the provider is that we will return an aligned corpus in TMX format that can directly be useful in translation workflows and that we apply data pre-procesing and cleaning pipelines that help further data curation. We accept a variety of formats including websites, PDFs, Microsoft Word document to name a few. We also ask the general public to suggest sources like translated websites or to send translated documents through our public translation tool (see Figure 2). This is connected to a data crunching backend that runs data conversion, text extraction and alignment jobs. More details about that system are given in (Aulamo and Tiedemann, 2019).

One common concern of potential data providers is related to privacy issues as discussed earlier. Some data is too sensitive to be published for general use. However, such data may still be used in model development and internal research. We offer different levels of privacy according to the choice of the provider. The data can be public, used internally in training public MT and language models or it can be used for training customized MT only for the provider itself. The main principles are illustrated in Figure 1.

2.2. Sentence-level Bitext Mining

Recently, several methods have been proposed for bitext mining from incomparable, monolingual corpora (see for example (Espana-Bonet et al., 2017; Guo et al., 2018; Schwenk, 2018)). Such corpora, most typically resulting from web crawls, contain no metadata information that would allow direct linking of sentences, or at least documents across the languages. Further, typically only a small fraction of sentences can be assumed to have a translation counterpart in such data. These methods allow us to utilize also text sources inaccessible to methods that assume a prior document-level alignment.

Bitext mining from incomparable corpora relies on cross-lingual sentence embeddings constructed such that sentences from different languages are embedded into a single vector space, allowing their subsequent comparison using, for instance, the cosine similarity measure. Here we apply the LASER embeddings (Language Agnostic SEntence Representations) of Artetxe and Schwenk (2019b). This approach is based on an encoder-decoder architecture, with a shared sentence encoder, and a set of language-specific decoders. The encoder and decoders are trained using existing parallel data and the shared encoder can subsequently be used to produce embeddings of previously unseen sentences in a vector space shared across the languages.

	size		
data set	gold	silver	mono (fi/sv)
Web crawl	500K	2M	54M/90M
FISKMÖ crawl	85K	130K	4.3M/1.3M
YLE News Archive	140K	300K	13.9M/4.5M
YLE News RSS	25K	100K	1.4M/1.1M
Total	**750K**	**2.5M**	**74M/97M**

Table 1: The number perfect or near-perfect (Gold) and partial translations or highly related (Silver) translation pairs extracted from different data collections. Sentence counts for the monolingual collections (*mono*) are given in term of unique (deduplicated) sentences.

The embeddings of all sentences in one language are then compared to the embeddings of all sentences in the other language, and the most similar pairs are likely translations. The cosine similarity scores of sentence embeddings are shown not to be numerically comparable across different source sentences, which is accounted for by the margin-based ranking method of Artetxe and Schwenk (2019a). As calculating the all-vs-all sentence embedding comparison for large corpora is computationally demanding, the highly optimized FAISS library of Johnson et al. (2019) is typically applied to carry the comparison out efficiently. Recently, this method was applied to Wikipedia, obtaining 135M parallel sentences for 85 languages (Schwenk et al., 2019a) and to CommonCrawl web crawl data, obtaining 3.5B parallel sentences for 38 languages (Schwenk et al., 2019b).

Here, we use the LASER+FAISS method to extract Finnish-Swedish parallel data from several monolingual corpus pairs, whose sizes are summarized in Table 1. For web crawl data, we use the Finnish Internet Parsebank (Luotolahti et al., 2015), a large-scale dedicated Finnish web crawl corpus, and the Swedish section of the CoNLL-17 Shared Task raw data (Ginter et al., 2017), based on CommonCrawl. The news data are sourced from the Finnish national broadcast organization YLE 2011–2018 archive available through the Language Bank of Finland (*YLE news archive*) (Yleisradio, 2019b; Yleisradio, 2019a), and from the RSS feed of the Finnish and Swedish language YLE news (*YLE News RSS*), gathered during 2018–2019. Finally, we crawled web pages of various government organizations (*FISKMÖ crawl*).

For each corpus pair, we manually evaluated a sample of sentence pairs at various points of the list of sentence pairs ranked by their similarity. We estimate how many sentence pairs can be judged perfect or near-perfect translations (*Gold*) and how many can be judged as partial translations or highly related (*Silver*). The results of the evaluation are summarized in Table 1, showing that in total we estimate having extracted 750K perfect or near-perfect translation pairs from 74M sentence Finnish and 97M sentence Swedish corpora.

2.3. Document-level Bitext Mining

As an initial experiment, we also tested whether document-level alignments could be obtained with similar techniques as described above if the document-level structure is known in the raw data. Here, we used the filtered translation pairs obtained from above mentioned sentence-level comparison to find candidate document pairs from dedicated web crawls of over 60 Finnish domains.

When considering all possible Finnish-Swedish document pairs from the raw data, the document pair was considered as a translation candidate if it shared at least one sentence. These candidate document pairs are then scored using a document-level margin-based ranking method. In the original sentence-level margin-based ranking score, defined by Artetxe and Schwenk (2019a), a global order of candidate sentence pairs is determined by comparing the similarity of the highest ranking candidate translation with the average similarity of its k nearest neighbours. Similarly, given a document-level similarity metric between two documents, a document-level margin metric can be defined by comparing the similarity of the highest ranking candidate document with the average similarity of its k nearest neighbours. The document-level similarity metric is defined as an average cosine similarity of each source and target sentence to its most similar sentence in the candidate document of opposite language, and the value of k is determined to cover the full candidate document space for a given document.

The document level alignment was tested on two of the above mentioned datasets, *YLE news archive* and the *FISKMÖ crawl*. *YLE news archive* includes 700K Finnish and 230K Swedish documents, of which 20K parallel documents were retrieved and manually determined to be at least partial translations of each other. *FISKMÖ crawl* has 260K Finnish and 230K Swedish document of which we were able to identify 7,500 as good quality, at least partially parallel documents.

The extracted documents have then automatically been sentence-aligned using hunalign[4] (Varga et al., 2005) with its two-pass realignment strategy and pre-extracted bilingual dictionaries as integrated in Uplug.[5] In order to increase the quality of the extracted bitexts even further, we also filtered the resulting sentence alignment with rather strict alignment score thresholds. As a result we released a small, medium-sized and large bitext collection from the *FISKMÖ crawl* using 0.8, 0.5 and no score threshold, respectively. Table 2 summarizes the size of each of the parallel data sets. Unfortunately, the sentence alignment of the *YLE news archive* did not work very well and we decided not to release a bitext even after heavy filtering. However, the raw document pairs are still available from our git repository.

released data set	size
FISKMÖ crawl – small	199K
FISKMÖ crawl – medium	322K
FISKMÖ crawl – large	473K

Table 2: The number of aligned translation units extracted from document pairs identified in the FISKMÖ crawl.

3. Data Processing and Distribution

One reason for collecting data is to make the translation data available to other researchers for studying cross-linguistic effects in human translations as well for training multilingual models for natural language processing. The data sets that we collect from organisations and public web content is processed in a way that makes them conform to the standards in OPUS and the parallel corpora that are distributed in that data collection. This means that we encode data sets in XML with document information and sentence boundary tags containing IDs that are linked with each other using standoff alignment information. The data files are also tokenized and automatically annotated with universal dependencies using UDPipe (Straka and Straková, 2017). The parallel data are available from OPUS[6], for example, in the sub-corpora Finlex and fiskmö and can also be searched on-line.[7]

Furthermore, we create packages that are distributed in aligned plain text files and also files in translation memory exchange (TMX) format, which can easily be incorporated in professional translation workflows. Those files are also distributed via OPUS and additionally from our FISKMÖ public git repository.[8]

4. Machine Translation

The main downstream application of our data collection is the development of machine translation between Finnish and Swedish in both directions with a high quality and coverage to be used in the general public or domain-specific use. Below, we describe the current state of our development, focusing on the general-purpose models that we train without specific domain adaptation. For this, we apply the data we have collected in OPUS and the FISKMÖ project and train state-of-the-art neural machine translation (NMT) models using the popular transformer model (Vaswani et al., 2017) as implemented in Marian-NMT (Junczys-Dowmunt et al., 2018). The training data comprises roughly 33 million training examples with about 1 billion tokens (counting both languages together). The data is derived from a wide mix of sources ranging from legislative texts to translated movie subtitles, software localization and general web content. We pre-process the data with the common Moses tools (Koehn et al., 2007) applying Unicode character normalization, corpus cleaning tools and the Moses tokenizer for Finnish and Swedish. Furthermore, we perform BPE-based subword segmentation using the Subword NMT package[9] (Sennrich et al., 2016) with BPE models trained separately for each language and setting the number of merging operations to 32,000.

For training the MT models we apply standard settings of a 6-layer transformer model (in both, encoder and decoder) with 8 self attention heads per layer, tied embed-

[4]https://github.com/danielvarga/hunalign

[5]https://github.com/Helsinki-NLP/uplug

[6]http://opus.nlpl.eu

[7]http://opus.nlpl.eu/bin/opuscqp.pl

[8]https://version.helsinki.fi/Helsinki-NLP/fiskmo

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

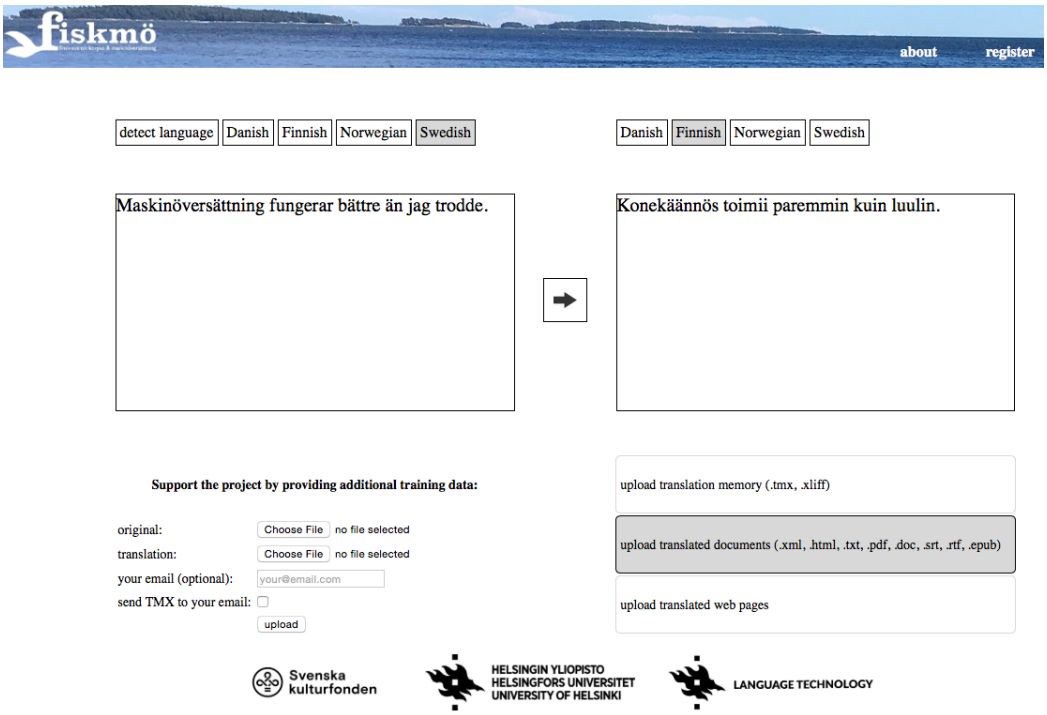

Figure 2: The public FISKMÖ translation interface.

dings and a shared vocabulary of 65,000 tokens. We follow the recommendation from the documentation[10] and use label smoothing with a factor of 0.1, transformer dropout of 0.1, a learning rate of 0.0003 with inverted squared decay, optimizer parameters of 0.9, 0.98 and 1e-09, learning rate warmup until 16,000 updates and a clip norm of 5. For development and testing, we use data from Tatoeba (5,000 randomly selected sentences in each set), a collection of user contributed translations, which is part of OPUS. We run training until convergence with a stopping criterion of 10 subsequent validation steps without further perplexity improvements on validation data. Validation is done after 10,000 training batches with dynamic batch fitting on four NVIDIA V100 GPUs. For decoding we apply a beam of 12 and the scores are computed using sacreBLEU (Post, 2018).

4.1. On-line Services

The pre-trained models from our project are publicly available from our project website. They can be downloaded and deployed in any appropriate computing environment using the open-source tools that are used for developing them. Besides of the models themselves, we also offer an on-line service that runs translation engines in a virtual environment. For that purpose, we have developed a server architecture that provides websocket-based connections and a simple API to call the service with plain text input in JSON format. The services seamlessly integrate all pre- and post-processing steps that are necessary to run the actual translation decoder from the MarianNMT server. Our server software also incorporates language detection and caching to enhance the functionality and performance of the system. The implementation of the server backend and a simple client script are open source and available from github.[11]

The translation API can be accessed via a simple web frontend developed in the Python web framework flask.[12] The interface is shown in Figure 2 and can be accessed from https://translate.ling.helsinki.fi. The internal service can easily be extended with additional language pairs and the demonstrator already now supports other Nordic languages such as Norwegian and Danish through the same interface. We are currently also integrating a highlighting function of linked subword units based on cross-lingual attention. This is trained using the guided alignment feature of MarianNMT and word alignment of the training data produced by eflomal[13] (Östling and Tiedemann, 2016).

Another useful tool in this interface is the integration of data uploads as mentioned earlier. Figure 2 shows the upload buttons at the bottom that can be used to upload new data sets to the data processing backend, which then can be used for further enhancements of the system. Translation memories can be uploaded in TMX format as well as translated documents in various formats to be converted and aligned by the system. The data provider can also specify an e-mail address to receive the resulting bitext in TMX format, an important incentive to convince users to provide new data sets.

4.2. CAT Tool Integration

Professional translators translating from Finnish into Swedish or vice versa are one of the most important potential user groups for FISKMÖ's MT systems. Most of such

[10] https://github.com/marian-nmt/marian-examples/tree/master/transformer

[11] https://github.com/Helsinki-NLP/Opus-MT

[12] https://palletsprojects.com/p/flask/

[13] https://github.com/robertostling/eflomal

professional translation is performed in CAT tool environments, such as SDL Trados Studio, Memsource, memoQ and Wordfast. While FISKMÖ's online translation service can provide translations also for professional use, efficient use of MT requires that the MT is directly integrated into a CAT tool. This is usually achieved by developing a plugin that provides translations on demand for each sentence to be translated.

Based on correspondence with translators, organisations that produce translations, and translation agencies, many different CAT tools are used in Finland. Of these CAT tools, SDL Trados Studio appears to have most users. SDL Trados Studio also has more extensive plugin development features than the other CAT tools, so due to the size of the potential user base and ease of development, FISKMÖ MT was initially integrated with SDL Trados Studio. A plugin has also been created for memoQ at the request of a government organization interested in participating in FISKMÖ, and the feasibility of other CAT tool integrations is being investigated.

The FISKMÖ MT plugin for CAT tools is a conventional MT plugin in all but one very important aspect: the plugin is almost entirely self-contained and performs NMT decoding on the local machine. The plugin package contains a full NMT decoding pipeline, based on a MarianNMT decoder executable built for Windows with support for CPU decoding (available as part of the MarianNMT repository). The speed of CPU decoding varies depending on hardware, and it is approximately 1 second for a sentence of 10-20 words on a middle-tier Intel i5-8250U CPU. The speed is clearly not sufficient for real-time decoding on all modern computer hardware. However, since translators work on documents sequentially segment by segment, later segments in the document can be translated ahead of time and cached. This means that the FISKMÖ plugin can display translations instantaneously to the translator, with the exception of the very first segments in a document, where there may be a slight delay.

The FISKMÖ plugin can be used offline, since the decoding is performed locally. A network connection is only required for downloading the Marian NMT models from the FISKMÖ object storage. Once a model has been downloaded, the plugin is entirely self-contained and can be used without any dependencies on external services. This inherent offline-capability of the plugin has many significant benefits. The foremost benefit is that the confidentiality and security of the MT system can be absolutely guaranteed. The risk posed by outages or discontinuation of external services is also eliminated. The deployment of the plugin is also simple, and can be performed by the end user. FISKMÖ MT is also free to use, unlike most online MT services intended for professional translation.

From a technical point of view, the offline implementation of the FISKMÖ plugin offers a promising basis for future development of interactive MT capabilities that require low latency, such as interactive translation prediction (Knowles and Koehn, 2016). In an online MT system, these features are constrained by the latency of the two-way communication between the server and the client. Of course, an offline system is constrained by the available hardware, but the hardware constraint is not fixed like network latency, as it can be alleviated by optimization and design.

A major motivation for creating the plugin was to generate interest in the data collection part of the project by offering potential partners a clear way to benefit from the project. This goal has been achieved, since the FISKMÖ MT plugin for SDL Trados Studio has already generated publicity for the whole project, and it has been tested by organisations interested in participating in the FISKMÖ project. It is available from our project website at https://blogs.helsinki.fi/fiskmo-project/resources/ as well from github.[14]

4.3. The FISKMÖ Benchmark Test Set

For development purposes we also created a small test set to evaluate the translation tools we develop. For this, we sampled nine documents from three different on-line sources and manually aligned them to create a clean test set for automatic evaluation. The documents come from the webpages of the municipality of Espoo, the Fennia investment and insurance company and the University of Helsinki. The genres include informative texts, minutes from public meetings, text from F.A.Q. pages, terms and conditions, feature stories, historical remarks and a text about codes and principles. With this, we cover a variety of styles and domains making it possible to judge the generality and coverage of a translation system. So far, the data set is small but will be extended in the future. Each document consists of 28 sentences each on average. Most sentences are quite long with an average of 34 tokens per sentence in Swedish 26 tokens in Finnish. The test set is available from https://version.helsinki.fi/Helsinki-NLP/fiskmo.

We applied the test set for a quick automatic evaluation of our current translation models in both directions also in comparison to two on-line systems that support the translation between Finnish and Swedish, namely Google Translate[15] and Presidency MT[16]. The latter has been released in connection with the EU presidency of Finland and it has been heavily optimized for the translation from Finnish into English and Swedish. It is developed by Tilde, a Latvian language service provider, in collaboration with the Prime Minister's Office of Finland using large data sets collected from European resources and data provided by the Finnish authorities. Both on-line systems have been accessed on October 29, 2019 and the results in terms of BLEU (Papineni et al., 2002) and chr-F2 (Popović, 2015) scores are shown in Table 3.

From the results, we can see that our system fairs quite well in comparison to both on-line translation engines. Note that it is not optimised for the data set in any way and trained without any fine-tuning for any specific domain. The advantage over Google Translate is astonishing showing that uncommon language pairs are still not well supported by general-purpose engines. In comparison to the Presidency MT engine, our system performs on a similar scale, slightly better for the translation into Finnish but worse in the other directions. However, the results need to be taken with a

[14] https://github.com/Helsinki-NLP/fiskmo-trados

[15] https://translate.google.com

[16] https://presidencymt.eu/

Finnish to Swedish		
MT engine	BLEU	chr-F2
Google Translate	17.7	0.513
Presidency MT	29.5	0.627
FISKMÖ	26.6	0.600
Swedish to Finnish		
MT engine	BLEU	chr-F2
Google Translate	19.2	0.564
Presidency MT	25.6	0.612
FISKMÖ	26.1	0.623

Table 3: Translation performance on the FISKMÖ benchmark test set.

grain of salt as the test set is small and the training data used in all three MT engines are not directly comparable. Nevertheless, we can see the competitiveness of our model proving its applicability in professional workflows.

5. Conclusions and Future Work

The FISKMÖ project will continue with further enhancements of translation tools for Swedish and Finnish. In the sections above, we have demonstrated that the project already delivered novel large scale data sets crawled from public sources and prepared from contributed data sets. We have developed competitive and freely available translation engines that can be easily deployed as on-line services and also be used in a self-contained plugin for computer-assisted translation. We have created a network of collaborators and strive for increased coverage of the two official language of Finland in order to improve translation capacity and information accessibility across language barriers. Our tools are public and data sets are distributed with permissive licenses when possible. Furthermore, we support offline translation in our CAT tools to remove the dependency on online services that may compromise security concerns when working with sensitive data. With this, we want to support translation in the public as well as the private sector to improve the bilingual services in Finland.

In the future, we would like to integrate our services and tools in additional platforms. We would like to develop streamlined procedures for the creation of domain-specific and customized translation models and we plan to investigate the use of interactive translation tools to provide a natural interface between MT and human translators. Finally, we would also like to extend language coverage by integrating additional language pairs and multilingual translation models.

Acknowledgments

The work reported in this paper has been funded and supported by the Swedish Culture Foundation in Finland (project-ID: 139592).

6. Bibliographical References

Artetxe, M. and Schwenk, H. (2019a). Margin-based parallel corpus mining with multilingual sentence embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3197–3203, Florence, Italy, July. Association for Computational Linguistics.

Artetxe, M. and Schwenk, H. (2019b). Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.

Aulamo, M. and Tiedemann, J. (2019). The OPUS resource repository: An open package for creating parallel corpora and machine translation services. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, pages 389–394, Turku, Finland, 30 September – 2 October. Linköping University Electronic Press.

Espana-Bonet, C., Varga, A. C., Barrón-Cedeno, A., and van Genabith, J. (2017). An empirical analysis of nmt-derived interlingual embeddings and their use in parallel sentence identification. *IEEE Journal of Selected Topics in Signal Processing*, 11(8):1340–1350.

Ginter, F., Hajič, J., Luotolahti, J., Straka, M., and Zeman, D. (2017). CoNLL 2017 shared task - automatically annotated raw texts and word embeddings. http://hdl.handle.net/11234/1-1989, LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Guo, M., Shen, Q., Yang, Y., Ge, H., Cer, D., Hernandez Abrego, G., Stevens, K., Constant, N., Sung, Y.-H., Strope, B., and Kurzweil, R. (2018). Effective parallel corpus mining using bilingual sentence embeddings. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 165–176, Brussels, Belgium, October. Association for Computational Linguistics.

Johnson, J., Douze, M., and Jégou, H. (2019). Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*.

Junczys-Dowmunt, M., Grundkiewicz, R., Dwojak, T., Hoang, H., Heafield, K., Neckermann, T., Seide, F., Germann, U., Fikri Aji, A., Bogoychev, N., Martins, A. F. T., and Birch, A. (2018). Marian: Fast neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, Melbourne, Australia, July. Association for Computational Linguistics.

Knowles, R. and Koehn, P. (2016). Neural interactive translation prediction. In *Proceedings of the Conference of the Association for Machine Translation in the Americas (AMTA)*.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. (2007). Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic, June. Association for Computational Linguistics.

Luotolahti, J., Kanerva, J., Laippala, V., Pyysalo, S., and Ginter, F. (2015). Towards universal web parsebanks. In *Proceedings of the International Conference on Depen-*

dency Linguistics (Depling'15), pages 211–220. Uppsala University.

Östling, R. and Tiedemann, J. (2016). Efficient word alignment with Markov Chain Monte Carlo. *Prague Bulletin of Mathematical Linguistics*, 106:125–146, October.

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.

Popović, M. (2015). chrF: character n-gram f-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal, September. Association for Computational Linguistics.

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, October. Association for Computational Linguistics.

Schwenk, H., Chaudhary, V., Sun, S., Gong, H., and Guzmán, F. (2019a). Wikimatrix: Mining 135m parallel sentences in 1620 language pairs from wikipedia. *arXiv preprint arXiv:1907.05791*.

Schwenk, H., Wenzek, G., Edunov, S., Grave, E., and Joulin, A. (2019b). Ccmatrix: Mining billions of high-quality parallel sentences on the web.

Schwenk, H. (2018). Filtering and mining parallel data in a joint multilingual space. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 228–234, Melbourne, Australia, July. Association for Computational Linguistics.

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

Straka, M. and Straková, J. (2017). Tokenizing, pos tagging, lemmatizing and parsing ud 2.0 with udpipe. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 88–99, Vancouver, Canada, August. Association for Computational Linguistics.

Varga, D., Halácsy, P., Kornai, A., Viktor, N., László, N., László, N., and Viktor, T. (2005). Parallel corpora for medium density languages. In *Proceedings of the RANLP 2005*, pages 590–596.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pages 5998–6008.

Yleisradio. (2019a). Yle svenska webbartiklar 2012-2018, källmaterial. http://urn.fi/urn:nbn:fi:lb-2016111401.

Yleisradio. (2019b). Ylen suomenkielinen uutisarkisto 2011-2018, lähdeaineisto. http://urn.fi/urn:nbn:fi:lb-2017070501.