Nordic Council of Ministers

Overview on Machine Translation services

Cross-border data exchange

Contents

This publication is also available online in a web-accessible version at:
https://pub.norden.org/nord2023-032

Executive summary

The objective of this report is to present an overview on the available machine translation engines and services. The report summarises machine translation methods and their key concepts, takes a quick overview on the machine translation service market and summarises some of the benefits and potential challenges of those services from the perspective of the Cross-Border Data Exchange (CBDE) Project.[1]

Machine translation is the process of using an engine to automatically translate human input text from one language to another. There are both cost-free and commercial machine translation tools available on the market. The engines' methodology varies: the translations are either rule-based, statistical-based or neural-based. The nature and characteristics of the needed translation defines what kind of machine translation engine suits the situation best.

This report views machine translation engines through the glasses of the above-mentioned CBDE project: using machine translation engines in cases like studying abroad, transferring health data and understanding legislation texts smoothly over the borders of Nordic and Baltic countries.

This overview on machine translation tools is based on literature and theory – the engines have not been tested for this report. The report takes a slightly closer view on three (3) machine translation services that provide a wide enough set of languages to be used across all Nordic and Baltic countries. Google Translate, Amazon Translate and EU eTranslation are capable of translating the whole set of Nordic and Baltic languages.[2] These three translation engines can all also be integrated to core systems via API interface – but before taking any machine translation service into use, not only the translation method but also information security and privacy issues are to be considered.

The report also suggests that there are no machine translator services that offer 100% accurate translations in every context; there is always a need for post-translation quality inspection done by a human.

To be able to make a more concrete and detailed analysis on the usability of machine translation services in different real-life environments, there is a need for more detailed information gathering, field testing and discussions with the service providers. In addition,

---

1. Cross-Border Data Exchange in Nordic and Baltic Countries is a 3-year project, funded by the Nordic Council of Ministers. More information on the project: https://wiki.dvv.fi/pages/viewpage.action?pageId=117377490
2. Nordic and Baltic languages: Danish, Estonian, Finnish, Icelandic, Latvian, Lithuanian, Norwegian and Swedish.

it would be interesting and reasonable to explore more on utilising existing specialised terminology and vocabularies[3] together with machine translation. These specialised vocabularies have been created by different administrative branches for varying needs. It would seem highly reasonable to evaluate whether these vocabularies could be used as an external source for broadening machine translators' vocabulary capabilities.

---

3.   e.g., https://sanastot.suomi.fi/

Motivation for the overview

Machine translation is the process of using an engine to automatically translate human input text from one language to another.

In recent years, machine translation has received more attention, as companies have started using it in their own operations. The benefits for companies are realised in the form of operational efficiency, time used on translating texts and translation costs. Machine translation is based on computational linguistics, where the machine translates source text into the desired target language. The size of the market is predicted to grow from 800 million USD in 2021 to 7.5 billion USD by 2030, with an annual growth rate of 30% from 2022 to 2030, according to Global Market Insight.[4] Popularity of the cloud-based application, and customer-driven company operations have been recognised as the key driver for this growth.

The European Commission has created the **eTranslation** service[5] for public administration and small and medium-sized enterprises (SMEs). The eTranslation service is intended to provide a quick, raw machine translation from and into any European language, including Icelandic and Norwegian, free of charge.

Also, several institutions have started their own machine translation projects, aiming to create machine translation engines or services to reduce existing language barriers. For example, the **University of Tartu** partnered with Baltic language technology and localisation company *Tilde* for an open-source machine translator engine development. The aim is to provide an engine that can translate Estonian to and from English, Russian and German. The development also aims at advancing the state of machine translation in Estonia and testing new methods for translating.[6]

Another interesting initiative for cross-border data exchange is **AuroraAI**[7], a Finnish national artificial intelligence programme started in 2020. The goal of AuroraAI is to connect public sector organisations through the AuroraAI network. The purpose is to create technical conditions that will enable information exchange and interoperability between different services and platforms. This could be a potentially interesting initiative for cross-border data exchange as well.

This report takes an overview on available machine translation engines and services. The

---

4. https://www.gminsights.com/industry-analysis/machine-translation-market-size
5. https://commission.europa.eu/resources-partners/machine-translation-public-administrations-etranslation_en
6. https://ut.ee/en/content/cooperation-between-university-tartu-and-tilde-takes-estonian-machine-translation-new-level
7. https://vm.fi/en/national-artificial-intelligence-programme-auroraai

report first summarises machine translation methods and their key concepts, then takes a quick overview on the machine translation service market, and finally summarises some of the benefits and potential challenges of those services from the perspective of the Cross-Border Data Exchange (CBDE) Project.[8]

Machine translation methods

Machine translation methods can be divided into three approaches:

- Rule-based machine translation (RBMT)
- Statistical machine translation (SMT)
- Neural machine translation (NMT)

Below are some notes from each of these methods.

Rule-based machine translation (RBMT)

The earliest machine translation systems were built with rule-based approaches. Rule-based machine translation engines require pre-coded linguistic rules.[9] Monolingual and bilingual dictionaries match input words to output words. In addition, rule-based machine translation engines require rules that present the structure of both the input language and the output language describing the grammatical structure of both languages.

Rule-based translating can be defined as an approach that includes a group of hardcoded linguistic rules that are used to analyse the grammatical input and create a representation into the output language structure. This approach requires knowledge of the source and target languages and the differences between them.

RBMT methods can be divided into three sub-logics: direct, transfer and interlingua approaches.

In the *direct approach,* the input text is translated word by word.

In the *transfer approach*, the input text is analysed sentence by sentence, after which the engine examines each word and sentence at a time and tries to understand its purpose. When the text input has been analysed and considered comprehensive, the engine translates the text using specific rules. Following the source sentence's structure, the system determines target words for each word and uses them to form target language sentences.

In the *interlingua approach,* the source text is transformed into semantic representation of the text, which will then form the basis for generating the target text.

---

9. https://machinetranslate.org/rule-based-machine-translation

Rule-based machine translation **can be customised to a specific industry or topic**. Challenges are, on the other hand, tightly connected with the nature of the method itself: all improvements in translation require either **manually updating** the built-in dictionaries or **new hardcoded rules**, which take a significant amount of time and human effort to create.

Statistical machine translation (SMT)

In the 2010s, the top method for machine translating was statistical machine translation.[10] Statistical machine translator engines use substantial amounts of bilingual data to make as accurate translation as possible.[11] These engines require storing of text corpus (parallel corpora), which serve as the basis for the translation. Translation engine analyses the text looking for statistical relationships between the original texts and their existing human translations.

The SMT engines can be divided into four approaches:

- Word-based model
- Phrase-based model
- Syntax-based model
- Hierarchical sentence-based model

All SMT engines follow the same logic. The input text is broken into sentences, after which the sentences are placed in their parallel counterparts, which are defined in the translation model. The language model then confirms that the translation is probable in the source language and translates the text.

SMT engines most commonly apply a *phrase-based model*, where words are translated in sentence-based sequences. The source text is segmented into sentences and compared to the targeted bilingual frame, after which statistical measurement is used to calculate the most likely target language segment based on the system's translation model and the collection of data of the target language. The translation model calculates the extent to which the source language word found in each sentence corresponds to target language words. It searches for this information from the text corpus input into the machine. The

10. https://machinetranslate.org/approaches
11. https://aws.amazon.com/what-is/machine-translation/

target language model calculates what is statistically the most probable translation.

In the *word-based model*, the translation is created word-by-word. The syntax-based model translates syntactic units. The *hierarchical sentence-based model* combines a phrase-based model with syntax-based methods.

Challenges with statistical machine translation are related to **creating the parallel data for each language pair** – it is both costly and time-consuming as statistical machine translation requires massive parallel data. Specific errors in the translation are also difficult to fix. SMT is considered a more challenging method for language pairs with differences in word orders.
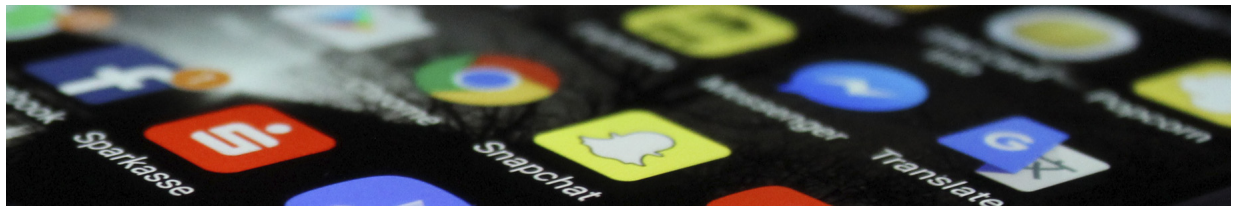
Neural machine translation (NMT)

During the 2020s, the machine translation methodology has turned to neural machine translating.[12] Neural machine translation (NMT) is the most used translation method in the commercial machine software industry, and it is widely seen as the most efficient and accurate method for machine translations.[13] Neural machine translation engines consider the entire sentences when creating the target sentence, where statistical translation engines calculate correspondence for individual words in sentences.

NMT engines use neural networks and teach themselves to recognise certain types of words and sentences. The process is time-consuming, but it is possible to train the translation process and thus make the processes more efficient. Neural machine translators are also improved as they are used – the more these machine translators are used, the better the translations become.

12. https://machinetranslate.org/approaches
13. https://aws.amazon.com/what-is/machine-translation/

Machine translation services

Overview

There are currently a large number of both publicly funded and commercial (cost-free and paid) machine translation services available. Cost-free commercial services usually collect the translated text for themselves and use them to improve the service. *Google Translate,* offered by Google, is an example of a cost-free commercial service that collects the translated data and uses it for improving future translations.[14]

The service providers also differ based on the translation methods used and the field of work the translator service is aimed at (healthcare, education, consulting, law, etc.). Large tech companies, such as Microsoft, Amazon, and Google, have produced their own translation services, but also smaller independent companies have started to emerge. Growing popularity of the cloud-based application strengthens the market, while at the same time, customer-driven company operations drive the need for machine translation services.

It is important to recognise the nature and characteristics of the content that needs to be translated, and the aimed use cases for the translated texts – whether close-enough translation is acceptable or is there a need for human inspection and quality control of the translated text. Many machine translator engines can technically be integrated into core systems through API interface, but it should be thoroughly assessed as to whether that is needed and reasonable.

Benefits of commercial machine translator services:

- Ability to quickly translate text into several languages
- Ability to save and store the original text
- Ability to organise stored translations for future use
- Ability to translate words, sentences, paragraphs and entire documents
- Possibility of add-on features
- Commercial machine translators are currently mostly based on neural networks

---

14. https://support.google.com/translate/answer/10400210?hl=en

Paid commercial translation services	Free commercial translation services
Amazon Translate	Bing Microsoft Translator
Crowd's	Google Translate
MemoQ Translator PRO	DeepL
Mesource	Reverso Translation
Systran Translate PRO	Systran
Smartling	
TextUnited	

Picture 1. Paid and free commercial translation services.

In this overview, Google Translate and Amazon Translate were selected to be more closely viewed, together with the European Commission's eTranslation service. Both Google Translate and Amazon Translate are based on neural networks, and their specific strength for this review was the scale of language pairs available, covering all Nordic and Baltic languages. The EC's eTranslation is aimed at European public authorities and is free-of-charge.

Google Translate

Google Translate is a cost-free online service that is based on neural networks. Google Translate can be used to translate words, texts and documents.[15] The engine is constantly learning from the inputs by other users, providing the users with relatively fast and accurate translations.

The engine works by scanning its own database that contains everything on Google and what other users have added to the Google Translate engine. The text is then analysed to find the most frequently used version of the target text. Google Translate is part of the commercial company Google, which means that in using this service, the input text can be stored and used for Google's business operations.

As an engine, it is flexible and versatile, because it can be used with speech, it recognises images, and it is fast and efficient to use via internet browser. It is an effective tool for translating short texts and words, but translating large documents is usually not as

---

15. https://support.google.com/translate#topic=7011659

accurate. The input text cannot be managed by the user.

The use of Google Translate requires an internet connection and a computing device, such as a computer, tablet or mobile phone. The source systems can be integrated with Google Translate via API interface.

Google Translate accepts the following document formats for translating: .doc, .docx, .odf, .pdf, .ppt, .pptx, .ps, .rtf, .txt, .xls and .xlsx.

Google Translate's languages				
Afrikaans	Albanian	Amharic	Arabic	Armenian
Assamese	Aymara	Azerbaijani	Bambara	Basque
Belarusian	Bengali	Bhojpuri	Bosnian	Bulgarian
Burmese (Myanmar)	Catalan	Cebuano	Chewa(Chichewa)	Chinese (Simplified)
Chinese (Traditional)	Corsican	Croatian	Czech	Danish
Dogri	Dutch	English	Esperanto	Estonian
Ewe	Finnish	French	Galician	Georgian
German	Greek	Guarani	Gujarati	Haitian
Creole	Hausa	Hawaiian	Hebrew	Hindi
Hmong	Hungarian	Icelandic	Igbollocano	Indonesian
Irish	Italian	Japanese	Javanese	Kannada
Kazakh	Khmer	Kinyarwanda	Konkani	Korean
Krio	Kurdish (Kurmanji)	Kurdish(Sorani)	Kyrgyz	Lao
Latin	Latvian	Lingala	Lithuanian	Luganda
Luxembourgish	Macedonian	Maithili	Malagasy	Malay
Malayalam	Maldivian (Dhivehi)	Maltese	Māori (Maori)	Marathi
Meitei (Manipuri)	Meitei (Meiteilon)	Mizo	Mongolian	Nepali
Northern Sotho (Sepedi)	Norwegian	Odia (Oriya)	Oromo	Pashto
Persian	Polish	Portuguese	Punjabi (Gurmukhi)	Quechua
Romanian	Russian	Samoan	Sanskrit	Scottish
Gaelic (Scots Gaelic)	Serbian	Sesotho	Shona	Sindhi
Sinhala	Slovak	Slovenian	Somali	Spanish
Sundanese	Swahili	Swedish	Tagalog (Filipino)	Tajik
Tamil	Tatar	Telugu	Thai	Tigrinya
Tsonga	Turkish	Turkmen	Twi	Ukrainian
Urdu	Uyghur	Uzbek	Vietnamese	Welsh
West Frisian (Frisian)	Xhosa	Yiddish	Yoruba	Zulu

Picture 2. Google Translate's set of languages. Nordic and Baltic languages highlighted.

Amazon Translate

Amazon Translate is a paid machine translation service based on neural networks that uses deep learning models to translate texts.[16] The machine is constantly learning, providing the user with fast and accurate translations. It is possible to customise the

---

16. https://aws.amazon.com/translate/

service according to users own needs, which makes the use of the service flexible. It can be integrated into various interfaces with the help of an API (Application Programming Interface).

Amazon Translate requires an AWS (Amazon Web Services) user account to use the service. AWS is a cloud services company, part of Amazon.com Inc. Creating an AWS user means a user is created in the AWS system, which allows the user to deploy and manage resources in AWS, including Amazon Translate. The data obtained from the machine translation is stored in the AWS cloud, which is protected by AWS, using the AWS Shared Responsibility Model.[17] The user is responsible for protecting data in the cloud (like IAM, Identity Access Management), while AWS is responsible for security of the cloud, such as the cloud infrastructure. Cloud data can be stored in the EU region, where it complies with the GDPR (General Data Protection Regulation).

For secure data storing, AWS recommends the following:

- Ensuring safe account management with multi-factor authentication (MFA)
- Use SSL/TLS to interact with the resources in AWS
- Ability to track activity in API and AWS with CloudTrail
- Security is high with AWS encryption possibilities, along with all default security controls of AWS and its services
- Possibility to use advanced managed security services, such as Amazon Macie, which assists in discovering and securing personal data that is stored in Amazon Simple Storage Service

The service includes "Amazon Comprehend Medical" U.S. Health Insurance Portability and Accountability Act (HIPAA)-compliant translation, for secure translation of medical domain texts – a service which is currently only available in English. AWS also contains a feature called "Active Custom Translation" (ACT), where the user can influence, in a certain way, what the machine translation output will be, giving users more control over translation output.

Using Amazon Translate requires the organisation to perform actions, such as creating users on AWS, understanding AWS and data security principles, implementing protocols and generally managing the AWS-based system. AWS offers its customers technical support and consultations related to using, maintaining, and managing AWS services. Amazon Translate can also be integrated with other IT systems via API interface.

Amazon Translate accepts the following document formats for translating: .docx, .xlsx

---

17. https://aws.amazon.com/blogs/security/the-aws-shared-responsibility-model-and-gdpr/

and .pptx.

Amazon Translate's languages				
Afrikaans	Albanian	Amharic	Arabic	Armenian
Azerbaijani	Bengali	Bosnian	Bulgarian	Catalan
Chinese	Croatian	Czech	Danish	Dari
Dutch	English	Estonian	Farsi (Persian)	Filipino
Tagalog	Finnish	French	French (Canada)	Georgian
German	Greek	Gujarati	Haitian	Creole
Hausa	Hebrew	Hindi	Hungarian	Icelandic
Indonesian	Irish	Italian	Japanese	Kannada
Kazakh	Korean	Latvian	Lithuanian	Macedonian
Malay	Malayalam	Maltese	Marathi	Mongolian
Norwegian	Pashto	Polish	Portuguese (Brazil)	Portuguese
Punjabi	Romanian	Russian	Serbian	Sinhala
Slovak	Slovenian	Somali	Spanish	Spanish (Mexico)
Swahili	Swedish	Tamil	Telugu	Thai
Turkish	Ukrainian	Urdu	Uzbek	Vietnamese
Welsh				

Picture 3. Amazon Translate's set of languages. Nordic and Baltic languages highlighted.

EU eTranslation

The European Commission has created the eTranslation service for European public administrations, small and medium-sized enterprises, and universities.[18] eTranslation is based on the neural machine translation method, and it works best with EU-related matters. The eTranslation service is compliant with GDPR legislation, and all translated data are processed within the data security protocols of the European Commission. The service can be used as part of the EC information systems and integrated into other IT systems via the API interface.

eTranslation can be used to translate sentences and documents. The input texts must be over 30 characters long for the engine to recognise the language. The engine translates and preserves the original text format in all except pdf files, where the document format changes to docx format. Users can select the domain of the source text (such as education, law or healthcare) to produce a translation that follows the possible linguistic peculiarities of the specified domain. Attached to this overview is an example of the EU Language Technology Resources[19] entered into the eTranslation service. The service provides a raw translation - not a perfect translation - and requires a skilled professional to validate the translated text and refine it. The service is currently free-of-charge but

18. https://commission.europa.eu/resources-partners/machine-translation-public-administrations-etranslation_en
19. https://joint-research-centre.ec.europa.eu/language-technology-resources_en

requires a license.

The use of eTranslation requires the creation of a user account, except if the user already has EU Login credentials (formerly ECAS).

eTranslate accepts the following document formats for translating: .txt, .doc, .docx, .odt, .ott, .rtf, .xls, .xlsx, .ods, .ots, .ppt, .pptx, .odp, .otp, .odg, . otg, .htm, .html, .xhtml, .h, .xml, .xlf, .xliff, .sdlxliff, .rdf, .tmx or .pdf.

eTranslate's languages				
Bulgarian	Croatian	Czech	Danish	Dutch
English	Estonian	Finnish	French	German
Greek	Hungarian	Icelandic	Irish	Italian
Latvian	Lithuanian	Maltese	Norwegian (Bokmål)	Polish
Portuguese	Romanian	Slovak	Slovenian	Spanish
Swedish				

Picture 4. EU eTranslation's set of languages. Nordic and Baltic languages highlighted.

Machine translation and GDPR

In machine translation, it is important to consider the requirements of the data protection legislation (GDPR) for sharing and managing the translated data.[20] There are translation services that use all the translated data for their own purposes or share that data to third parties. Should the translated text be of sensitive nature, such as patient data, regarding a person's health or ethnicity, GRPR legislation should be considered thoroughly.

One way to manage this challenge is to use a machine translator that does not share data with third parties, but instead stores the data with the users' own data warehouses. When procuring a service from a commercial operator, it should be ensured that the operator does not forward information to third parties, which is why having the translator engine on one's own servers is one possible option to be used. If any data is collected by any translator service, it is important to be aware of where the data is stored.

By ensuring a GDPR-compliant service provider, the user can inquire the following information:

- **Non-disclosure/confidentiality agreement** with contractors
- **Translation management system's security**
- **How, where and by who the translated texts are handled** before, during, and after

---

20. https://www.languagescientific.com/ensure-translation-services-provider-gdpr-compliant/

the actual translation operation

- **Standards and accreditations** of the service and the service provider

Use cases

To make this overview more concrete, the machine translating solutions are viewed in the context of the *Cross-Border Data Exchange (CBDE)* project's work packages.

The aim of the CBDE project is to support daily life in the Nordic and Baltic countries.[21] The project studies cross-border data exchange through three use cases, which translate into this overview as follows:

- In Work Package 1 (WP1), the potential machine translations are related to student data exchange and recognition of achievements, such as course credits. Translations are most often based on translating individual words and sentences, like course descriptions.

- In Work Package 2 (WP2), the subject of translations is related to health information. The information contains ePrescriptions and patient summaries, for which not just any machine translation can be used for information security issues based on the sensitive nature of the translated data. Also, this information is mainly in the form of code set. For WP2, this review focuses on potential future use cases.

- Work Package 3 (WP3) deals with legislative and regulatory information, where the use cases are divided into translation of legal text from legislative databases. The purpose is to translate texts into different languages than the original information is. There are limitations using translation tools when translating legislative data. Translated legislative information is usually unofficial and does not have legal force in court.

Please note that the report and this overview are based solely on theory and the available literature – the machine translation services have not been tested for accuracy with real-life sentences.

Using machine translation services in CBDE Work Packages

The translation services picked for this quick review all proved to have both advantages and some disadvantages. *Google Translate* is free-of-charge but not GDPR compliant. *Amazon Translate* keeps an organisation's data secured in the AWS cloud but requires AWS maintenance knowledge from the organisation utilising the service. *eTranslation* is free-of-charge and aimed at public administration, but the vocabulary is focused on

---

21. https://pub.norden.org/temanord2021-547/

EU-related matters. None of the systems mentioned should be used without post-translation quality inspection by a skilled language professional.

Below is a more detailed picture of the selected machine translation services based on the use cases studied in the CBDE project, first as a table and then a short analysis by service.

Service	Google Translate	Amazon Translate	eTranslate
Method	Neural network	Neural network	Neural network
Free-of-charge	Yes	No	Yes
Admin user required	No	Yes	Yes
GDPR compliance	No	Yes	Yes
Hybrid translation	Yes	Yes	Yes
Word	Yes	Yes	No
Sentence	Yes	Yes	Yes
Document	Yes	Yes	Yes
Code-to-text	No	Yes	No

Picture 5. Machine translation services comparison

WP1 (study abroad)	Google Translate	Amazon Translate	eTranslate
Used with single words	🟢	🟢	🟥
Used to translate whole sentences	🟢	🟢	🔺
Input whole document or text size of documents	🔺	🟢	🟢
Use it to transfer Code to Sentences	🟥	🟢	🟥

WP2 (health data)	Google Translate	Amazon Translate	eTranslate
Used with single words	🟢	🟢	🔴
Used to translate whole sentences	🟢	🟢	🔺
Input whole document or text size of documents	🔺	🟢	🟢
Use it to transfer Code to Sentences	🟥	🟢	🟥

WP3 (legislation texts)	Google Translate	Amazon Translate	eTranslate
Used with single words	🟢	🟢	🟥
Used to translate whole sentences	🟢	🟢	🔺
Input whole document or text size of documents	🔺	🟢	🟢
Use it to transfer Code to Sentences	🟥	🟢	🟥

Legend: 🟢 *Possible*   🔺 *Partly possible*   🟥 *Not possible*

Picture 6. Machine translation services and CBDE Work Packages

Google Translate

Google Translate is an easy-to-use, free-of-charge machine translation service to translate words, sentences and full documents as well as speech and images. The main

challenge is to realise that the content input to Google Translate will be utilised by Google to improve the future translations (and potentially also other Google services).[22] Any information translated with Google Translate must be non-sensitive data.

Google Translate does not understand codes (code-to-text), and the service cannot be trained to understand organisation-specific codes or rules. Other disadvantages of Google Translate include the public nature of the service, the lack of control over the entered data and the distribution of data ownership among third parties. Users cannot manage the stored data in any way, which can cause problems from a data management perspective.[23]

Google Translate will also require post-translation quality inspection and possible refining of the translated text.

Amazon Translate

Amazon Translate is a commercial service that can translate words, texts and documents. It provides an opportunity to manage the input data in a GDPR-compliant way, and uses the Amazon Web Services (AWS) cloud to store an organisation's data, so the organisation will have control over its own data. The AWS cloud requires an AWS admin from each organisation. The admin is also able to build and modify pre-defined rules to Amazon Translate, which makes it possible to also translate organisation-specific codes to text.

The use of Amazon Translate requires AWS know-how from the organisation. AWS usage will also create costs related to the service, such as AWS products, storage and maintenance.

Amazon Translate requires post-translation quality inspection and refining of the translated text, but the translated words and sentences can be influenced and taught in a certain way with AWS Active Custom Translator (ACT) by the organisation.

EU eTranslation

eTranslation is a translation service available free-of-charge to European public administration. It can be accessed with an EU login. The service can be used safely – the system is run by the European Commission, and it is compliant with GDPR, and can translate several documents into several languages in one session.

22. https://policies.google.com/terms?hl=en-US
23. https://proprivacy.com/blog/google-translate-privacy

eTranslation also gives the user the possibility to choose a domain of the input text. Domain function can be highly useful, as the system has analysed text and documents from various domains, including education, healthcare and legislation. The European Commission states, however, that the translations are most accurate on issues related to the EU.[24]

eTranslation is based on neural machine translation methodology, which means the service learns from the input data. It stores all the input data, but it gives the admin user an opportunity to delete the needed data. The eTranslation service contains all European languages.

The disadvantage of eTranslation could be seen in its limited use: the service requires the creation of EU Login credentials for each service user, and single words and texts under 30 words cannot be translated for eTranslate needs enough data to recognize the language. eTranslation is not capable of translating codes into text, and it also requires post-translation quality inspection and possible refining of the translated text.

24. https://commission.europa.eu/resources-partners/machine-translation-public-administrations-etranslation_en

Services comparison summary

Google Translate is comparatively the easiest to use for the end-user, as it is available via the internet. The engine has analysed a large number of texts over several years, during which the neural-networks-based machine has learned a huge amount of information. This makes Google Translate a very powerful and accurate machine translator. Its biggest benefits, from a service point of view, are cost savings and ease of use. However, one of the biggest disadvantages is the public nature of the service, as it does not enable GDPR-compliant data processing for sensitive information.

GDPR-compliant data management is emphasised in the Amazon Translate service, where data are stored in the AWS cloud, making GDPR-compliant data management and storage possible. The benefits of Amazon Translate are using the same neural networks method for translation as Google Translate and the additional possibility to store and manage data. The user can influence the development of the service itself, which makes the service better for certain domain areas compared to Google Translate. AWS knowledge in using Amazon Translate is vital, which can be seen as both a disadvantage and an advantage. The service can be expanded and taught, which makes the translation service more efficient, but this causes an increase in cost and requirements in AWS knowledge. The service costs are based on the use of Amazon Translate and its sub-services (e.g., ACT).

The European Commission's eTranslation is the safe and free option. Compared to Amazon Translate and Google Translate, eTranslation has large sets of information from specific domains input to it – securing rather high-quality translations in certain domain fields. It can translate documents into several languages with one translation request. The challenges of eTranslation are related to its limited use: the service requires creation of EU Login credentials for each service user, and single words and texts under 30 words cannot be translated. eTranslation is also not capable of translating codes into text.

In order to make a more concrete and detailed analysis on the usability of machine translation services in different real-life environments, there is a need for more detailed information gathering and discussions with service providers regarding what is possible. The machine translator services' capabilities and quality of translations in different fields of expertise also need to be tested with actual texts and documents, and with linguistic professionals to evaluate the accuracy of the translations. In addition, it would be of high interest to study the capabilities of machine translators in utilising specialised terminology and vocabularies. These specialised vocabularies have been created in recent years in

different administrative branches for divergent use. It would seem highly reasonable to examine further whether and how these vocabularies could be used as an external source for broadening machine translator services' vocabularies and utilise the excessive terminology work that has already been done.

Looking for a machine translator? Remember these!

1

It is important to **choose a translator service based on the type of text** that will be translated. First, clarify how complex the translated texts will be, how much text there will be and which input/output languages are required.

2

Machine translation **does not understand cultural differences**. The best translation result is obtained when the text is clearly structured, both grammatically and logically, and it conforms to the rules of the methods being used. Even high-quality machine translation does not produce 100% accurate text in every context.

3

Hybrid translation is a process where a machine translates the text, and it is then checked and refined by a human professional. This human-machine interaction is called human-in-the-loop (HITL).

4

Code-like texts are not automatically recognised by machine translators. Translating codes can only be used in certain types of machine translators that are based on codes. Machine cannot automatically translate a code "KHJ843287" to a certain course "Fundamentals of financial management". However, it is technically possible to build a translator that can transform code-to-text.

5

Certain services **share data with third parties**. which can cause a problem regarding **GDPR legislation**. If the machine translation service saves data, make sure you know who owns and manages the data. Open-source machine translation services might create similar challenges, as systems based on open-source code may give an open license for anyone using the service to also use the translated information worldwide.

6

The EU is currently updating legislation that guides the use of artificial intelligence in Europe. Considering the legislation, it may create both challenges and opportunities in the future. So, stay tuned!

Attachments

Attachment 1. EU Language Technology Resources[25]

EU Language Technology Resources is a translation memory that is a source for machine translation, as it provides a collection of small pieces of text and their translations. The following translation memories listed below are examples of EU translation memories. These translation memories and parallel texts can be important for several reasons:

- It can be used to train a machine translator using statistical machine translation (SMT)
- Training and testing of multilingual data extraction software
- Translations can be consistently checked automatically
- Ready-made data packages for many different languages

The EU-enabled resource is based on parallel text materials (parallel corpora) related to machine translation based on neural networks. Parallel language material is a large structured/controlled set of translated texts between two languages.

When using **EU Language Technology Resources**, the user should consider which languages and data the resource packages contain, and in which format. The packages contain different languages and purposes for machine translation memory. It must be considered that one package does not contain the necessary memory to serve the needs.

Here is a list of some of the EU Language Technology Resources to demonstrate what is in the resources:

JRC-Acquis

This collection of documents, and their manually produced translations, can be used for many purposes, including the training of statistical machine translation systems, the training and testing of text mining applications.

Languages: Bulgarian, Czech, Danish, Dutch, English, Estonian, German, Greek, Finnish, French, Hungarian, Italian, Latvian, Lithuanian, Maltese, Polish, Portuguese, Romanian, Slovak, Slovenian, Spanish, Swedish.

25.
   https://joint-research-centre.ec.europa.eu/language-technology-resources_en

Missing Nordic and Baltic languages: Norwegian, Islandic

DGT-Acquis

This collection of aligned full-text documents, and their manually produced translations, can be used for many purposes, including the training of statistical machine translation systems, the training and testing of text mining applications, and more.

Languages: Bulgarian, Czech, Danish, Dutch, English, Estonian, German, Greek, Finnish, French, Irish, Hungarian, Italian, Latvian, Lithuanian, Maltese, Polish, Portuguese, Romanian, Slovak, Slovenian, Spanish, Swedish.

Missing Nordic and Baltic languages: Norwegian, Islandic

DCEP-Digital Corpus of the European Parliament

The corpus includes a variety of different text types, including press releases, motions, minutes of plenary sessions, rules or procedure, reports and written questions to the parliament. This collection of sentence-aligned, full-text documents, and their manually produced translations, can be used for many purposes, including the training of statistical machine translation systems, the training and testing of text mining applications and more.

Languages: Bulgarian, Czech, Danish, Dutch, English, Estonian, German, Greek, Finnish, French, Irish, Hungarian, Italian, Latvian, Lithuanian, Maltese, Polish, Portuguese, Romanian, Slovak, Slovenian, Spanish, Swedish, Turkish.

Missing Nordic and Baltic languages: Norwegian, Islandic

DGT-Translation Memory (DGT-TM)

Translation memories are collections of small pieces of text and their manually produced translations. Translation memories are typically used to support human translators, but they can also be used to train statistical machine translation systems. DGT-TM consists of between 4 and 7 million units per language. It is distributed in the widely used TMX format.

Languages: Bulgarian, Croatian, Czech, Danish, Dutch, English, Estonian, German, Greek, Finnish, French, Irish, Hungarian, Italian, Latvian, Lithuanian, Maltese, Polish, Portuguese, Romanian, Slovak, Slovenian, Spanish, Swedish.

Missing Nordic and Baltic languages: Norwegian, Islandic

EAC-Translation Memory (EAC-TM)

The parallel corpus was provided by the European Commission's Directorate General for Education and Culture (EAC), and the data have been processed further by the JRC. The EAC-TM is smaller compared to the other parallel corpora available here, but it has the advantage that it focuses on a very different domain. EAC-TM consists of a total of over 32,000 units. It is distributed in the widely used TMX format.

Languages: Bulgarian, Czech, Danish, Dutch, English, Estonian, German, Greek, Finnish, French, Croatian, Hungarian, Icelandic, Italian, Latvian, Lithuanian, Maltese, Norwegian, Polish, Portuguese, Romanian, Slovak, Slovenian, Spanish, Swedish and Turkish.

ECDC-Translation Memory (ECDC-TM)

The majority of the documents talk about health-related topics (anthrax, botulism, cholera, dengue fever, hepatitis, etc.), but some of the web pages also describe the organisation ECDC (e.g., its organisation, job opportunities) and its activities (e.g., epidemic intelligence, surveillance). ECDC-TM consists of up to 2500 translation units per language. It is distributed in the widely used TMX format.

Languages: Bulgarian, Czech, Danish, Dutch, English, Estonian, German, Greek, Finnish, French, Hungarian, Icelandic, Irish, Italian, Latvian, Lithuanian, Maltese, Norwegian, Polish, Portuguese, Romanian, Slovak, Slovenian, Spanish, Swedish.

About this publication

Nordic co-operation

Nordic co-operation is one of the world's most extensive forms of regional collaboration, involving Denmark, Finland, Iceland, Norway, Sweden, and the Faroe Islands, Greenland and Åland.

Nordic co-operation has firm traditions in politics, economics and culture and plays an important role in European and international forums. The Nordic community strives for a strong Nordic Region in a strong Europe.

Nordic co-operation promotes regional interests and values in a global world. The values shared by the Nordic countries help make the region one of the most innovative and competitive in the world.

Read more Nordic publications on www.norden.org/publications