Proceedings of the 1st International Workshop on Language Technology Platforms (IWLTP 2020), pages 50–53
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020

Eco.pangeamt: Industrializing neural MT

Mercedes García-Martínez, Manuel Herranz, Amando Estela, Ángela Franco, and Laurent Bié
Pangeanic / B.I Europa - PangeaMT Technologies Division
{m.garcia, m.herranz, a.estela, l.bie, a.franco}@pangeanic.com

Abstract

Eco is Pangeanic's customer portal for generic or specialized translation services (machine translation and post-editing, generic API MT and custom API MT). Users can request the processing (translation) of files in different formats. Moreover, a client user can manage the engines and models allowing their cloning and retraining.

Keywords: neural machine translation, customize translation, adaptive machine translation, NLP ecosystem

1. Introduction

Pangeanic is a language service provider (LSP) and language processing tool developer specialised in natural language processing and machine translation. It provides solutions to cognitive companies, institutions, translation professionals, and corporations. Pangeanic was the first company in the world to make use of the Moses statistical machine translation models in the translation industry (Yuste et al., 2010; Yuste et al., 2012). To this purpose, a platform to build models by the user was developed (PangeaMT's first platform[1]).

Eco.pangeamt[2] is a platform managing translation engines and an NLP ecosystem. It allows the access of three types of user profiles:

- Super Admin, is a reserved profile with which the translation infrastructure can be monitored and managed.
- Client, is an admin profile that allows the management of users and their access rights and statistics. Clients can check metrics and usage of their users, manage the access of the users to the different engines and process files. Moreover, Clients can also manage their models, they can clone models and train them from a baseline with new bilingual material, thus automating the task of engine specialization.
- User, this profile allows the processing of files and checking of information about usage and metrics of the API calls and processed files.

After logging in, the home page of the website shows the Dashboard with the charts about statistics and usage (see Figure 1).

The dashboard will be shown with information about the processes (translations) that have been carried out (processes per week, per month, total expenses, weekly, last processes, etc.).

The options (appearing in the left-side menu) are:

1. New Process: in this page Clients can process files and check their processes.
2. Services/Processes: in this page, Clients can check the files that are being processed and the ones already finished. Here, they will find all the information about them.
3. Profile: here, Clients can change their name, email, password and billing information.
4. Stats: here Clients can check their API stats, File stats and in the Details tab they can check the number of characters, words, segments, files and pages processed by their Users. The Range Date can be set to check the statistics of a particular period of time.
5. Corporate: in this tab Clients can manage their models and engines.
6. Users: where the list of created users is displayed. For a user it is possible to check which engines can be accessed and data about the usage. New users can be created with credentials for their access and with an APIKey that can be used in API or other applications access.
7. Subscriptions: in this page, Clients can check the assigned subscription and manage it.

User and Client profiles can directly translate text or send a file to be translated via the Eco platform. The system saves the files privately, only the file's owner has access to those files handling GDPR compliance. After processing, the translated file in its original format will be available to download. These features are described in services and processes (see Section 2.).

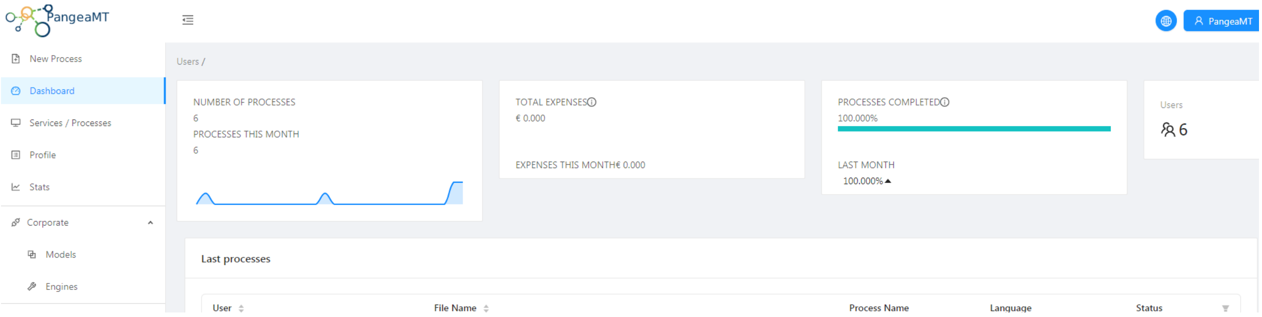
One of the main features of Eco.pangeamt is the possibility of adapting a neural machine translation (NMT) model to the user's own data in a friendly user interface. This feature is presented in Section 3. Finally, conclusions are explained in Section 4.

2. Services/Processes

The services and processes option allows Clients to process new files or translate paragraphs or sentences directly. In order to start a process, Clients have to choose the *Upload file* or *Translate text* option.

[1] https://www.gala-global.org/ondemand/pangeamt-platform-user-empowering-and-data-driven-domain-machine-translation

[2] https://eco.pangeamt.com/main



NUMBER OF PROCESSES
6
PROCESSES THIS MONTH
6

TOTAL EXPENSES
€ 0.000
EXPENSES THIS MONTH€ 0.000

PROCESSES COMPLETED
100.000%
LAST MONTH
100.000%▲

Users
6

Last processes

User	File Name	Process Name	Language	Status

Figure 1: Example of dashboard for a client.

2.1. Processing a file

For processing a file the *Upload file* option is selected, the user selects the source language of the document or documents and the target language (into which language it is translated). To upload the files, the user clicks on the gray box or drags the files to the box (see Figure 2).

Once the files have been selected (the name of the selected files appears below the box) the user can click *Start upload* to upload them. A confirmation message will appear. A process must be carried out per language combination, i.e. if two files need to be translated from English to French, they can be uploaded together, if another file needs to be translated, for example from Japanese to Korean, another process must be carried out by pressing Send another.

Clicking on *List of processes* will display the processes that are being carried out and those that have already been completed. In the Finished tab (see Figure 3), the details of the process are displayed: file name, process type, language combination and status.

In the Actions column, the option to download the translated file appears. Once it has been downloaded, next to the download button, the Trash icon appears; pressing it deletes the selected file from the list of completed files. In the Dashboard page, Clients can check that the process has been added to the list of last processes.

2.2. Translation of text

If *Translate text* is selected, users can enter text to translate in the box, the source and the target language have to be selected and pressing *Translate* the translated text will appear (see Figure 4) as the output of the selected engine.

3. Adapting an NMT model via Eco

One of Eco's most popular features is its model adaptation feature (Client role). A model can be trained with generic data (no specific domain). Usually, a generic model has been trained with a lot of data from several general domains. Users can use their clean data to adapt this generic model to a specific domain using specific Machine Learning routines. User material quickly specializes engines into for example technical, legal or science domains (see Figure 5).

A Client can clone or adapt models copying or cloning a model and specializing the model into a domain with the data it has previously acquired. Therefore, you can have a structure with father models (more generic models) and child models (specific models). Once the new model is created by cloning a father model, a Client can retrain the model with their own specific data.

By clicking on the Clone model icon, Clients can clone a model by entering the name and description and selecting the different options (see Figure 6).

Clients can also manage their models and engines. By clicking models on the corporate menu, Clients find all their models and data. Here, they can check all the information about these models: which models they can clone and train, the language pairs, description, the model's father, updates, when it was last saved, etc.

In the Engines section, Clients can verify their engines and check which ones are granted to their users. If the Grant all option is activated all users will have access to the engine.

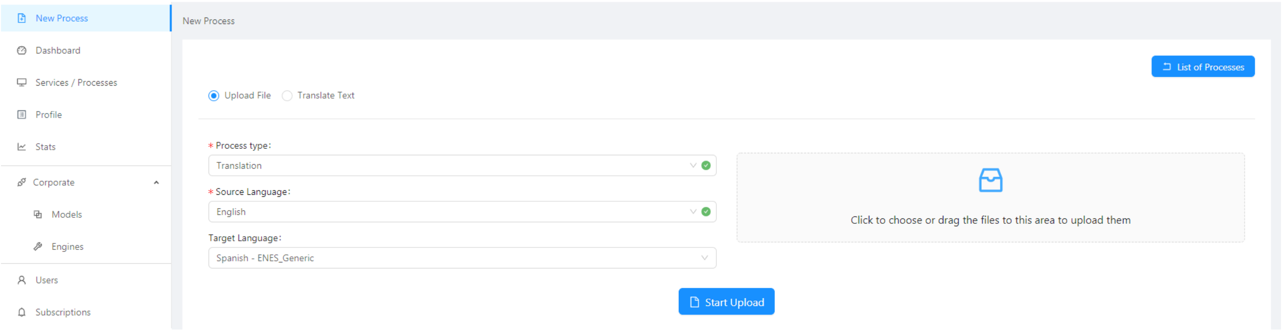
3.1. Training models

Eco makes training models easy thanks to its user friendly interface. Clients just have to click on the *To Train* icon and upload a bilingual file with language declaration or ID. The allowed file formats are preferably .tmx although .csv and .af (aligned format) are also accepted. Training files must contain perfectly aligned and recognisable source and target segments. Clients can decide the weight or aggressivity of the training. This affects how data will be incorporated into the model and its impact on the engine. A series of ML techniques weigh the data, its length, its vocabulary, etc. Effects on the model are to train it heavily on specific data to ultra-specialise it on the field of application or to just add domain data without changing severely whilst keeping its more generic features. Eco has 3 selectable levels of aggressivity from less to more weighing, shallower or deeper learning: Conservative, Normal or Aggressive (see Figure 7). The time needed for training depends on the size of the file and the level of aggressivity. Training is available with GPU making it much faster.

After a training file has been sent, Clients can access the training page by clicking on the *Trainings* icon. This page shows the completed trainings, the requested ones and the failed trainings. If a training fails, the system notifies where the error is.

The effectiveness of model retraining allowing its specialization in specific data is well known and it has been shown (Domingo et al., 2019a; Domingo et al., 2019b).

Pangeanic has run many trials with the training feature. For that, we used a generic English to Spanish transla-



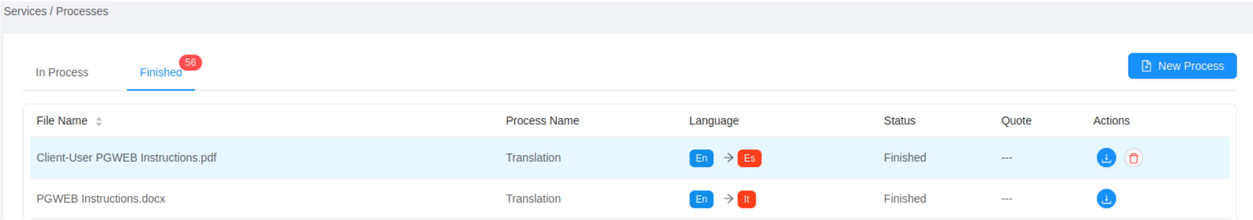
● Upload File ○ Translate Text

* Process type: Translation

* Source Language: English

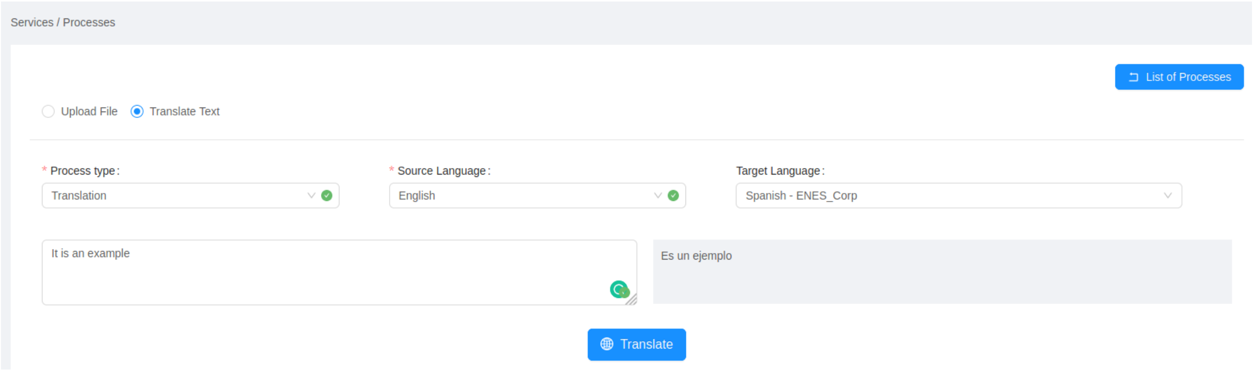
Target Language: Spanish - ENES_Generic

Figure 2: New process view for translating a file.



File Name	Process Name	Language	Status	Quote	Actions
Client-User PGWEB Instructions.pdf	Translation	En → Es	Finished	---	
PGWEB Instructions.docx	Translation	En → It	Finished	---	

Figure 3: Finished process tab view.



○ Upload File ● Translate Text

* Process type: Translation

* Source Language: English

Target Language: Spanish - ENES_Corp

It is an example

Es un ejemplo

Figure 4: Translate text view.

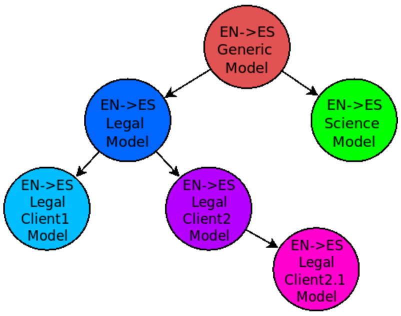
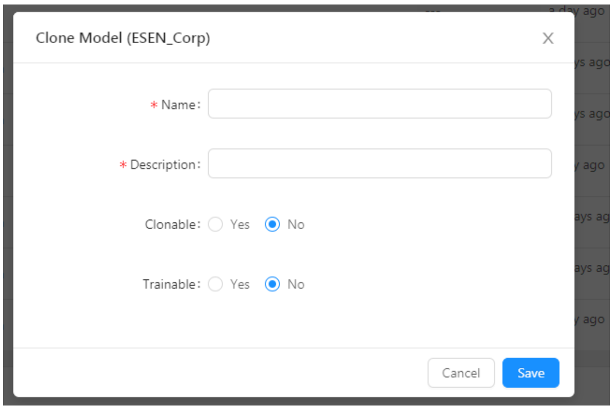

EN->ES Generic Model
EN->ES Legal Model
EN->ES Science Model
EN->ES Legal Client1 Model
EN->ES Legal Client2 Model
EN->ES Legal Client2.1 Model

Figure 5: Example model cloning - each child specializes in an area with its own specific data.



* Name:

* Description:

Clonable: ○ Yes ● No

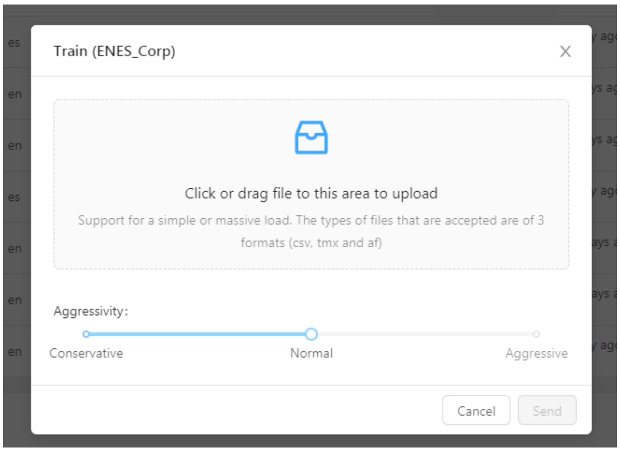
Trainable: ○ Yes ● No

Figure 6: Clone a model view.

tion model trained with public corpora (filtered Paracrawl dataset[3]). We have retrained it with 2 different test files of 500 sentences from the DGT dataset[4].

[3] https://paracrawl.eu/

[4] https://ec.europa.eu/jrc/en/language-technologies/dgt-



Train (ENES_Corp)

Click or drag file to this area to upload

Support for a simple or massive load. The types of files that are accepted are of 3 formats (csv, tmx and af)

Aggressivity:

Conservative — Normal — Aggressive

Figure 7: Training a model view where Client can chose the level of aggressivity.

The first test file (DGT test1) has been used to retrain the generic model. We used the 3 options of aggressivity and compared them with no training. We translated the 2 DGT test files (test1 and test2) and a generic test and we compared the results using the standard automatic translation metric BLEU score (Papineni et al., 2002). The results are shown in Table 3.1.

Training	Generic test	DGT test1	DGT test2
No train	66.29	38.25	38.01
Conservative	59.29	45.26	39.18
Normal	56.89	46.54	40.52
Aggressive	55.24	46.83	38.73

Table 1: Results in BLEU score using different types of trainings for the generic test and DGT test1 and test2 files.

Generic test results show a decrease in BLEU score when specializing in DGT domain, this is a normal behaviour because the model will translate better within the same domain. By contrast, the translation of DGT test1 file results show how BLEU score increases with the number of trainings as expected. Furthermore, when translating DGT test2 file, BLEU score improves using retraining. We expect this due to the fact that DGT test1 and test2 files are from the same domain. However, when translating DGT test2 file using aggressive training we obtain lower BLEU score than using conservative and normal training. This can be the case if the model has been adapted too much (overfitting) to data from DGT test1 file and translations to other files do not obtain the best results.

4. Conclusion

We have introduced Eco, Pangeanic's commercial translation platform describing its usage and different options. Eco incorporates a user friendly option for model adaptation. We have shown its effectiveness in a small set of experiments. This platform allows the translation of text and documents as well as APIKey machine translation. The platform is hosted by Pangeanic but can be hosted by clients. Moreover, users are able to build their own models by cloning a generic model and can retrain those models with their own data and as many times as they wish to obtain specific results. Engines can be stored and recalled at a later date. These adapted models will adjust to their domain and generate translations with more quality for their purposes. As a result, machine translation output will be more accurate and productivity will increase due to a decrease in machine translation manual corrections.

For future work, in addition to machine translation more tasks will be added to this platform such as anonymization, summarization or sentiment analysis.

5. Acknowledgements

The research leading to these results has received funding from the Spanish Centre for Technological and Industrial Development (Centro para el Desarrollo Tecnológico Industrial) (CDTI) and the European Union through Programa Operativo de Crecimiento Inteligente (Project IDI-20170964).

6. References

Domingo, M., García-Martínez, M., Estela Pastor, A., Bié, L., Helle, A., Peris, Á., Casacuberta, F., and Herranz Pérez, M. (2019a). Demonstration of a neural machine translation system with online learning for translators. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 70–74, Florence, Italy, July. Association for Computational Linguistics.

Domingo, M., García-Martínez, M., Peris, Á., Helle, A., Estela, A., Bié, L., Casacuberta, F., and Herranz, M. (2019b). Incremental adaptation of NMT for professional post-editors: A user study. In *Proceedings of the Machine Translation Summit*, pages 219–227.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 311–318.

Yuste, E., Herranz, M., Lagarda, A.-L., Tarazón, L., Sánchez-Cortina, I., and Casacuberta, F. (2010). Pangeamt - putting open standards to work... well. In *Proceedings of the Association for Machine Translation in the Americas*, Denver, Colorado, USA.

Yuste, E., Herranz, M., Helle, A., Lagarda, A.-L., García-Martínez, M., Pla-Civera, J., Blasco, M., Morella, A., and Mallach, J. (2012). Pangeanic's do-it-yourself machine translation: User empowerment and user-driven mt processing. *Journal of the Asia-Pacific Association for Machine Translation*.

translation-memory