123123

Project Proposal

Dineev Vladislav

Department of Software Engineering

Higher School of Economics

Moscow, Russia

vddineev@edu.hse.ru

***Abstract*—Translation of text documents will always be relevant. Nowadays we can solve that problem by creating software that will help people translate their documents more effectively rather than doing it by hand. This paper suggests ways of improving the existing collaborative translation software that allows users to upload files, split them in pieces (e.g. sentences), translate these pieces by a group of people in parallel and export the translated files. First, goals are stated and described, such as integrating with Yandex.API for basic translation, introducing user interactivity, news and support for Microsoft Word documents. Then, the market is analyzed for similar products and features that they already have to compare them with the current project, which outlines the need for the new features listed previously. Finally, a work plan is devised - it thoroughly describes the means by which new features will be implemented. Finally, expected results are stated, that suggest improving user experience and effectiveness of translating documents using the new version of current translation software.**

***Keywords—translation;collaborative;web-service; text; database***

# Introduction

Translation will always be a topic of current interest as long aspeoplespeak different languages. Books, articles and documentation are being written all the time, however there are a lot of individualswho have difficulties reading foreign literature due to their lack of language knowledge. The software engineering can ease the process of document translation by providing the instruments necessary to increase the quality and effectiveness of the said translations.

Existing approaches often suggest manual translation or basic translation with computer software that is later edited by qualified translators. The suggested approach combines both manual and software based translation which makes it new and unique - users split their text documents in parts (or segments) and other participants of the web service can translate these segments in parallel to achieve better quality and speed of the translation. In addition to that the application provides the basic machine translation to speed up the process so that the translators do not have to start from scratch.

The idea is to enhance the existing web application (implemented by the author) with new features in the suggested approach with some additions.

The expected results of the project is to introduce enhanced software that is capable of providing the end users with effective translation functionality.

The structure of the article is described as follows. First, the goals and tasks are described to provide the reader with a better understanding of what the existing solution is already capable of and how to build on it. Then a market review is conducted to compare that solution with other software products and outline their strong and weak points. After that a work plan is presented to formalize the details of the steps required to achieve the previously described goal. At last the conclusion summarizes the article and describes the expected results.

# Goals and tasks

For better understanding of the features to be implemented, it is crucial to provide a description of the current state of the web service. At this point the web application is capable of uploading a text file, splitting it in parts, allowing translation of each part into different languages and assembling the document back using the translated parts. Users can export the fully or partially translated document to download it on the PC.

To organize work, projects have been added to the web service. Projects allow users to attach several uploaded files to them. Users can later create a translation to a certain language inside the project. Each project can have an unlimited number of translations (e.g. Russian to English, English to Spanish), including translations to the same language (this could be useful if using different vocabulary is required - e.g. Russian formal to Russian informal).

The projectsare divided into two categories:

* public, meaning that all the users of the web service can access those projects and suggest the translations. The idea of public projects is strongly connected to the concept of open software - anybody participate in its development;
* private, where only a limited group of users can perform actions with the project.

For each translation inside a project users can translate parts of text documents, uploaded in the project earlier. The translation of these parts or text segments are tied to the overall translation while the translations are tied to the project.

The main goal is to improve on existing webapplicationwith additional features, providing a more complex but effective experience.

There are several tasks that are required to complete that goal.

## Integrating Yandex translate API

In order to help translators save their time on simple sentences and words, it is essential to provide them with basic machine translation. Yandex Translation API [1] solves this problem by providing access to Yandex translation services. This basic translation can be used as a basis for further polishing or accepted as is.

## Adding microsoft word format

Although plain text format is universal, a lot of documents and articles require more complex formatting, which the Word Documents (.docx file extension) [2] provide. However, additional parsing is required in order to preserve the initial structure of such a document and to reconstruct the document prior to exporting.

## Implementing User Interactions

Motivating users to engage in discussions and competing for best translation is important to the idea of open projects. Translators can simplify their job even further by reviewing and correcting variants of translations that have already become voted for by the majority of users.

## Integrating News feed

Self development and education is crucial if we allow the idea of common users translating the texts. Skilled translators can also try to write articles and share their knowledge. Thus, the news feed with article editor is introduced, allowing for creation and discussion of posts, containing useful information on certain subjects of languages, helping the users to broaden their horizons.

# Market review

There are several existing solutions on the market that have similar features to the proposed project in this article. Those solutions have a lot of useful features, however, they are not perfect and have their flaws.

## Notabenoid

Notabenoid [3] was a web application, allowing for the same functionality as the proposed application (splitting text in parts, collaborative translation, voting for translations, blog posts) with exception that only plain text formats are supported. This means that in order to use that application one would have to manually extract text from complex markdown file formats.

Another point to consider about that application is that it uses a fixed interface language - russian. Considering that this software’s main goal is to assist users with document translation it is hard to imagine non russian speakers learning the language simply to make use of Notabenoid.

It should also be noted that due to problems with copyright Notabenoid was closed and is no longer available for public use which is the biggest disadvantage about this product.

To compensate for its flaws it is planned to not limit the users to plain text format only and allow for word documents as both input and output. Moreover, different UI language is being implemented to allow users with different language backgrounds feel equally comfortable while using the product.

## Weblate

Weblate [4] is a solution for translating software based on the versioning control system (VCS) Git [5]. Git allows for complex versioning control using commits as snapshots of work and branches as series of commits. There are complex mechanisms in Git that allow to merge branches, resolve conflicts between them, etc.

Translation of simple text documents requires knowledge of Git. As a result it does not seem feasible for an ordinary user to utilize Weblate for purposes of translating simple text documents. That is the reason the proposed project aims to simplify the process and allow obvious interactions that do not require the additional knowledge of certain complex software.

That said, Weblate also includes one very important feature for collective translation - glossary. It represents a list of agreements on translation of certain words, names, etc. and aims to prevent conflicts in translations. That feature is already included in our product.

# Work plan

## Integrate Yandex.API

To provide the basic translations an API key needs to be created for Yandex translate API to bind to the application. This key will be used for all the api calls in order to retrieve the translations from the api.

After uploading the file and choosing the pattern for text splitting, a new phase of asynchronous translation will be created - the translation of the parts that uses Yandex.API will be running in the background allowing the translators to work on the project right away. As this asynchronous job progresses, the machine translation for each text segment will become available over time.

It should also be noted that an additional field for the project entity will be created in the database. It will be used to determine if the users want to enable machine translation for their project.

## Add microsoft word format

For Word Document parsing an already implemented library will be used. This library will allow for correct extraction of text and markdown from .docx format. To store the markdown a separate field will be added for each text segment entry in the database, which will contain only the corresponding styles for the text segment.

The styles will be required to reconstruct the document on demand in .docx format, preserving the original style and formatting.

## Interactions

User interactions while translating segments will require creating a separate database entity for comment and adding a field that denotes the number of votes for suggested translations. Comments are going to be displayed below the suggested translations, while number of votes will determine the position of the suggestion in the list - suggestion that is the most voted for will be seen on top of others.

## News feed

The news feed suggests having articles that are created by users displayed on the news page. A separate table in the database will be created for them. That table will contain both text and style for a post or an article and users will be allowed to create and edit their posts to be displayed on the main page.

To allow for easy creation and editing it is planned to use Editor.js - a popular Javascript markdown editor library. It already has all the basic features required to create styled texts (headings, quotations, pictures, etc.), so that most of the time will be spent developing the supporting features such as user rating and commentaries.

# Conclusion

To conclude, it must be said that right now the market has valuable products that provide rich features for collaborative text translation. However, those solutions have their advantages and disadvantages. Their strengths have been researched and will be implemented for theproduct in question, while their flaws will be used to build new features that will become the strong points of the product.

It is expected that the complexity of the product will increase, however the user experience and translation effectiveness will increase as well.

The support for new document format will introduce additional challenges of integrating and providing reliable methods of converting and storing the according data, although this feature gives the users more freedom, flexibility and removes the limit of being tied to plain text only.

The news feed and user interactions are going to motivate users to spend more time using the web service by engaging in activities and conversations with other users.

The basic machine translation will assist in constructing a draft of the translation that will require less time for creating a translation from scratch.

Overall, the web service will be heavily enhanced and appear more finalized and polished, thus the users will be more interested in using the product.

Word Count: 1748

##### References

1. *“API Переводчика - Технологии Яндекса” yandex.ru.(accessed Feb. 20, 2023)*[*https://yandex.ru/dev/translate/*](https://yandex.ru/dev/translate/)
2. “DOCX” docs.fileformat.com.*(accessed Feb. 20, 2023)*<https://docs.fileformat.com/word-processing/docx/>
3. “Notabenoid” notabenoid.com.*(accessed Feb. 20, 2023)*<https://web.archive.org/web/20121001043434/http://notabenoid.com/>
4. “Weblate Documentation” weblate.org.*(accessed Feb. 20, 2023)*<https://docs.weblate.org/en/latest/>
5. “Git - Documentation” git-scm.com.(accessed Feb. 20, 2023)<https://git-scm.com/doc>