A Workbench for Rapid Generation of Cross-Lingual Summaries

Nisarg Jhaveri, Manish Gupta[†], Vasudeva Varma
International Institute of Information Technology, Gachibowli, Hyderabad, Telangana - 500 032, India.
nisarg.jhaveri@research.iiit.ac.in, {manish.gupta, vv}@iiit.ac.in

Abstract

The need for cross-lingual information access is more than ever with the easy accessibility to the Internet, especially in vastly multilingual societies like India. Cross-lingual summarization can help minimize human effort needed for achieving publishable articles in multiple languages, while making the most important information available in target language in the form of summaries. We describe a flexible, web-based tool for human editing of cross-lingual summaries to rapidly generate publishable summaries in a number of Indian Languages for news articles originally published in English, and simultaneously collect detailed logs about the process, at both article and sentence level. Similar to translation post-editing logs, such logs can be used to evaluate the automated cross-lingual summaries, in terms of effort needed to make them publishable. The generated summaries along with the logs can be used to train and improve the automatic system over time.

Keywords: information access, cross-lingual summarization, machine translation, human evaluation of summaries

1. Introduction

In Indian news media, most of the content gets published in English first and then in regional languages, especially for news categories such as national or international news, technology or lifestyle news. The delay can be just a few hours, days or sometimes the news does not appear in the regional languages at all. On the other hand, some content gets generated and consumed in regional languages alone. With the Internet becoming easily accessible and the rise of digital journalism, it is now crucial to make the large amount of news published in English or other popular languages on the Internet available to the readers of other languages having fewer native publications.

Advances in Machine Translation (MT) and other fields of Computational Linguistics in recent years make it possible to automate cross-lingual news access. However, the current state of Machine Translation is not able to generate publishable articles in most Indian Languages from English. Although, post-editing MT output has been shown to increase translator's productivity over translating from scratch (Aziz et al., 2012), it still requires a significant amount of human effort to produce end-user consumable articles.

Making the highlights or summaries of articles originally published in English accessible to non-English speaking users helps in making a large amount of *critical* information accessible as fast as possible with minimal human effort. We aim to make the process completely or partially automatic so that the gist of the articles can be published in regional languages with minimal delay.

Working in this direction, we've developed a pluggable system, implementing a pipeline for cross-lingual summarization of news articles. While summarization and translation are two major modules in the automatic cross-lingual summarization pipeline, a number of other modules can be included, for example, preprocessing, automatic post-editing, etc. After automatic processing, the articles are sent to humans for post-editing the summary and the automatically translated text to produce publishable cross-lingual summaries.

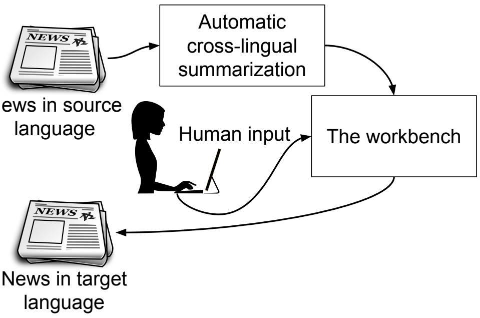


Figure 1: Overall flow of information

In this paper, we present a web-based tool for post-editing cross-lingual summaries, and briefly describe a pluggable pipeline for cross-lingual summarization. The source code of the tool, which we refer to as the workbench, is available on GitHub[1].

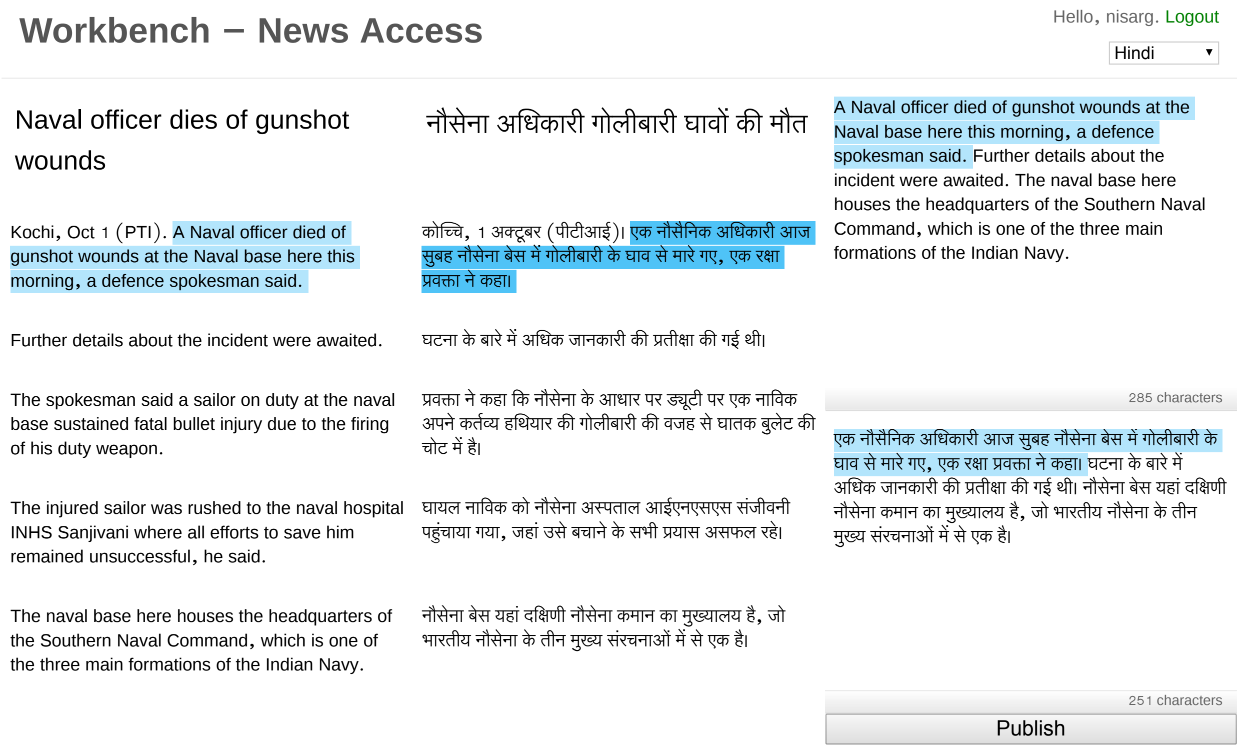
The workbench is pluggable by design, which enables it to work with a number of different MT systems, summarization systems and other tools available for different languages.

The workbench also records edit logs and other parameters while editing the automatic summary and translation. These logs can give meaningful insights for the task or can also be used as continuous feedback to the system.

The rest of the paper is organized as follows: Section 2 presents some related works. Section 3 describes the workbench in detail, including its main functionalities, interface and architecture. In Section 4 we show some examples of data collected and result of a pilot study. In Section 5 we discuss other possible use-cases for the workbench and possible future works. At the end, Section 6 contains conclusions.

[†] The author is also a Principal Applied Scientist at Microsoft.

[1] `https://github.com/nisargjhaveri/news-access`



Naval officer dies of gunshot wounds

Kochi, Oct 1 (PTI). A Naval officer died of gunshot wounds at the Naval base here this morning, a defence spokesman said.

Further details about the incident were awaited.

The spokesman said a sailor on duty at the naval base sustained fatal bullet injury due to the firing of his duty weapon.

The injured sailor was rushed to the naval hospital INHS Sanjivani where all efforts to save him remained unsuccessful, he said.

The naval base here houses the headquarters of the Southern Naval Command, which is one of the three main formations of the Indian Navy.

नौसेना अधिकारी गोलीबारी घावों की मौत

कोच्चि, 1 अक्टूबर (पीटीआई)। एक नौसैनिक अधिकारी आज सुबह नौसेना बेस में गोलीबारी के घाव से मारे गए, एक रक्षा प्रवक्ता ने कहा।

घटना के बारे में अधिक जानकारी की प्रतीक्षा की गई थी।

प्रवक्ता ने कहा कि नौसेना के आधार पर ड्यूटी पर एक नाविक अपने कर्तव्य हथियार की गोलीबारी की वजह से घातक बुलेट की चोट में है।

घायल नाविक को नौसेना अस्पताल आईएनएसएस संजीवनी पहुंचाया गया, जहां उसे बचाने के सभी प्रयास असफल रहे।

नौसेना बेस यहां दक्षिणी नौसेना कमान का मुख्यालय है, जो भारतीय नौसेना के तीन मुख्य संरचनाओं में से एक है।

A Naval officer died of gunshot wounds at the Naval base here this morning, a defence spokesman said. Further details about the incident were awaited. The naval base here houses the headquarters of the Southern Naval Command, which is one of the three main formations of the Indian Navy.

एक नौसैनिक अधिकारी आज सुबह नौसेना बेस में गोलीबारी के घाव से मारे गए, एक रक्षा प्रवक्ता ने कहा। घटना के बारे में अधिक जानकारी की प्रतीक्षा की गई थी। नौसेना बेस यहां दक्षिणी नौसेना कमान का मुख्यालय है, जो भारतीय नौसेना के तीन मुख्य संरचनाओं में से एक है।

Figure 2: Screenshot of the workbench, with a sentence highlighted.

2. Related Work

To the best of our knowledge, there is no available end-to-end system that allows post-editing of automated cross-lingual summaries to generate publishable summaries and at the same time can collect useful data about the process.

Computer Aided Translation (CAT) tools or translation post-editing tools like *SDL Trados Studio*[2], *MateCat*[3], *OmegaT*[4], *PET* (Aziz et al., 2012) and *CATaLog online* (Pal et al., 2016) are available. They compare with our system in following ways. *a*) While they support translation post-editing, we support editing of cross-lingual summaries as well. *b*) A few of them allow the recording of various kinds of logs about the translation post-editing process, while we allow recording comprehensive logs about the human editing of summary and translations.

Some work exists on cross-lingual summarization. Most recently, Zhang et al. (2016) proposed abstractive cross-lingual summarization. Yao et al. (2015) proposed compressive cross-lingual summarization inspired by phrase-based translation models. Wan (2011) proposed summarization using information from both source and translated article, while Wan et al. (2010) proposed to summarize considering the translation quality prediction.

Most extractive cross-lingual summarization systems have a sequential pipeline architecture. Additionally, most of them output a proposed mono-lingual summary and its translation at the end (Litvak et al., 2010; Orasan and Chiorean, 2008; Pingali et al., 2007; Wan et al., 2010; Wan, 2011; Yao et al., 2015).

This motivates the design of the workbench where the annotator can edit the mono-lingual summaries and its translation easily to get publishable cross-lingual summaries, and which can also collect various logs.

3. The Workbench

The workbench is a flexible, language independent tool for editing automatically generated cross-lingual summaries. The main features of the workbench are:

- The workbench provides a unique user-friendly environment for annotators to edit summaries in the source language, the cross-lingual summaries, and optionally, translation of the original article, in a seamless way.
- The pluggable and generic architecture provides possibility of using the workbench for almost any language pair, and with any set of external tools to plug into the pipeline.
- The workbench collects a wide range of logs from the editing jobs, which can be used as feedback by any module in the pipeline to improve the automatic process over time, and can also provide useful insights for the task in question.

3.1. Interface

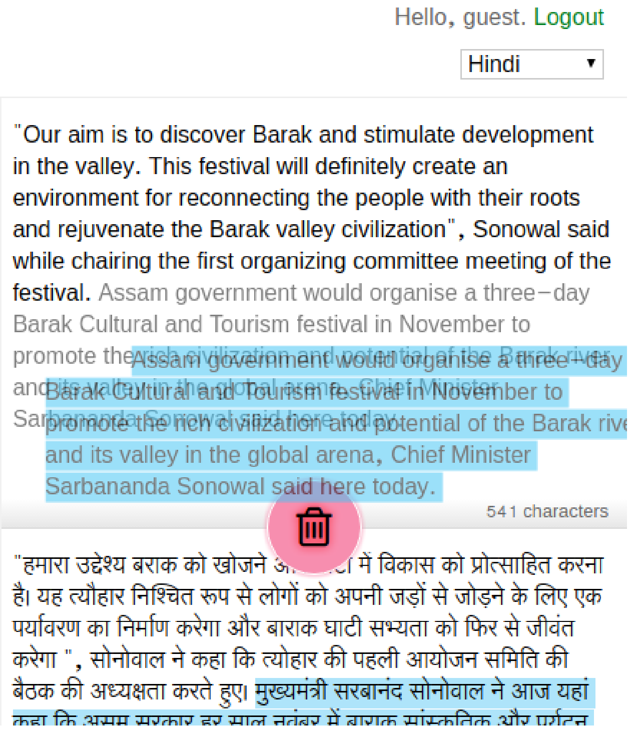
Figure 2 shows the default interface of the workbench with source and target languages being English and Hindi respectively.

The default interface for the workbench splits the working area in three vertical columns. The first column contains the source article. The second column contains the translation of the article in the target language. Paragraphs are aligned

[2] http://www.sdltrados.com/
[3] https://www.matecat.com/
[4] http://omegat.org/

in both these columns and the scrolling is synced. The third column contains the mono-lingual extractive summary in source language and the translation of the summary in the target language.

The annotator can edit the sentences in the mono-lingual summary and the translations of the sentences in summary, with the source and automatically translated article as the context. Optionally, the sentences in the translation of the complete article can also be edited.


Hello, guest. Logout

Hindi ▾

"Our aim is to discover Barak and stimulate development in the valley. This festival will definitely create an environment for reconnecting the people with their roots and rejuvenate the Barak valley civilization", Sonowal said while chairing the first organizing committee meeting of the festival. Assam government would organise a three−day Barak Cultural and Tourism festival in November to promote the rich civilization and potential of the Barak river and its valley in the global arena, Chief Minister Sarbananda Sonowal said here today.

Assam government would organise a three−day Barak Cultural and Tourism festival in November to promote the rich civilization and potential of the Barak rive and its valley in the global arena, Chief Minister Sarbananda Sonowal said here today.

541 characters

"हमारा उद्देश्य बराक को खोजने अ… में विकास को प्रोत्साहित करना है। यह त्यौहार निश्चित रूप से लोगों को अपनी जड़ों से जोड़ने के लिए एक पर्यावरण का निर्माण करेगा और बाराक घाटी सभ्यता को फिर से जीवंत करेगा ", सोनोवाल ने कहा कि त्योहार की पहली आयोजन समिति की बैठक की अध्यक्षता करते हुए। मुख्यमंत्री सरबानंद सोनोवाल ने आज यहां कहा कि असम सरकार हर साल नवंबर में बाराक सांस्कृतिक और पर्यटन

Figure 3: Screenshot of a summary sentence being dragged for deletion.

To edit the mono-lingual summary, a simple drag-drop interface is provided to make the usage intuitive. The annotator can add or remove sentences from the mono-lingual summary with an easy drag-drop interface. When the annotators starts dragging a sentence from the mono-lingual summary, a bin is shown near the end of the summary box, see Figure 3. The annotator can either drop the sentence at some position inside the box to reorder or they can drop it on the bin to remove the sentence from the summary. To add a sentence from original article into the summary, the annotator has to pick a sentence from the source article and drop into the summary box at the desired position. The position of the sentence is previewed live while the sentence is being dragged. Additionally, the content of the sentences in the summary can be changed in-line by clicking on it. Any changes in mono-lingual summary reflects in the cross-lingual summary immediately. The number of characters in the summary is also shown below the summary.

Contrary to the translation post-editing tools, since the structure of the article is important for summarization, we cannot show the original article broken into segments, or even sentences. The view shows paragraphs of the source and target article aligned with synced scrolling for easy navigation. To distinguish between multiple sentences in the paragraph, the sentence which is hovered or is active for editing is highlighted along with all the linked sentences, such as the source sentence, the corresponding sentence in mono-lingual summary if included, and the corresponding sentence in cross-lingual summary.

3.2. Architecture

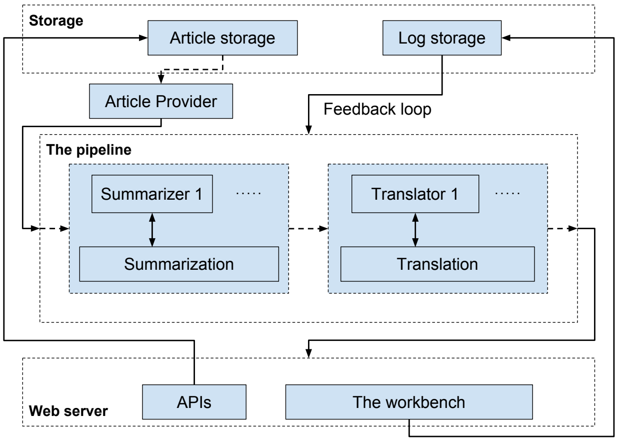


Figure 4: Overall architecture diagram

Figure 4 shows overall architecture of the system. The dotted arrows in the diagram indicate a possible flow, which is quite flexible.

The modular architecture of the system results in high flexibility in terms of use of external resources. The article provider module retrieves news article from either a corpus or other external sources.

The article is then passed through the automated pipeline, which depending on specifics, populates the article object with translation of the article, mono-lingual summary and the cross-lingual summary obtained based on the mono-lingual summary.

Each module in the pipeline can potentially access external resources, make API calls or access previous edit logs in order to generate the output. The modules themselves can be composite of other components.

In our system, summarization and translation are the two major modules along with pre-processing stage. Multiple summarizers and translators are integrated in summarization and translation module respectively. For example, in our case, the translation module may use Google Translate[5] or Microsoft Translator[6] depending on the language pair or a configuration option.

3.3. Log Collection

One of the primary goal of the workbench is to generate a large amount of cross-lingual news summaries in multiple languages, which can be used to build new systems in the area. Along with that, various kinds of user-activity logs are collected by the workbench, which can also be used to evaluate or improve new systems.

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

[6] https://www.microsoft.com/translator

Log level	Log type	Short description
Sentence-level	Focus/Blur Keystrokes Text selection Text input	When the sentence is activated for editing or de-activated Key presses, along with IME compositions Text selection by any means, e.g. with mouse or shift-arrow keys When the text is changed, including copy/cut/paste events
Summary-level	Add/Remove sentence Sentence reordering	When a sentence is added or removed from the summary When the order of the sentences in the summary is changed
Article-level	Page events	Events about page load, article load, and completion of editing

Table 1: Summary of various kinds of logs collected by the workbench

Table 1 summarises the different kinds of logs collected by the workbench. All events are logged with precise time in milliseconds.

The following sentence-level editing events are logged for all editable sentences in source language or target language.

- Focus/Blur: Focus and blur events are logged for each editable item when the item is activated for editing and when leaving the focus from the item respectively. A single item can be focused multiple times. The time spent on the item can be estimated by adding difference between all focus-blur pairs for the item.

- Keystrokes: All keystroke information available are logged, including IME (Input Method Editor) composition information.

- Text selection: Text selection by any means (mouse, shift-arrow) is logged.

- Text input: For all items, all the changes in text are logged, along with all copy/cut/paste events. This is specifically important as keystroke logging doesn't provide accurate and wholesome information in case of complex scripts and use of IMEs.

For translation post-editing, Human-targeted Translation Edit Rate (HTER) (Snover et al., 2006), along with information about insertion, deletion, and substitution of single words as well as shifts of word sequences can be calculated and stored when the translation is finalized.

For editing of mono-lingual summary, along with the sentence level editing logs, following summary-level events are also logged by the workbench.

- Add/Remove sentence: An event is logged when a sentence is added to the summary or removed from the summary, along with its position in the summary before adding or after removing.

- Sentence reordering: When the ordering of the sentences in the summary is changed, an event is logged with the information about previous and new ordering of sentences in the summary.

Additionally, some article-level events described below are also logged.

- Page events: The events about page load, article load, and completion of editing of an article are logged.

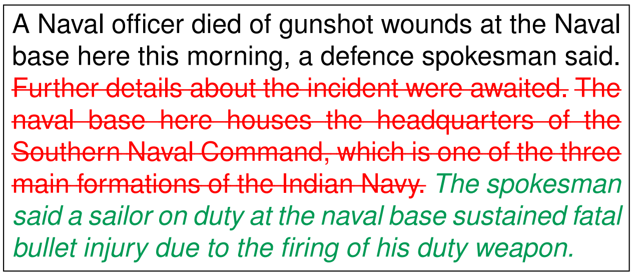

A Naval officer died of gunshot wounds at the Naval base here this morning, a defence spokesman said. ~~Further details about the incident were awaited. The naval base here houses the headquarters of the Southern Naval Command, which is one of the three main formations of the Indian Navy.~~ *The spokesman said a sailor on duty at the naval base sustained fatal bullet injury due to the firing of his duty weapon.*

Figure 5: Edits made to the mono-lingual summary. ~~Removed sentences~~, *Added sentences*

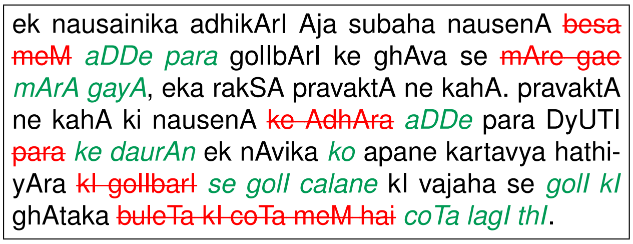

ek nausainika adhikArI Aja subaha nausenA ~~besa meM~~ *aDDe para* golIbArI ke ghAva se ~~mAre gae~~ *mArA gayA*, eka rakSA pravaktA ne kahA. pravaktA ne kahA ki nausenA ~~ke AdhAra~~ *aDDe* para DyUTI ~~para~~ *ke daurAn* ek nAvika *ko* apane kartavya hathiyAra ~~kI golIbArI~~ *se golI calane* kI vajaha se *golI kI* ghAtaka ~~buleTa kI coTa meM hai~~ *coTa lagI thI*.

Figure 6: Edits made to the Hindi translation of the summary. ~~Removed~~, *Added*

The total time taken for editing can be calculated as the time difference between completion of editing and article load time, and reducing the difference by the amount of time the annotator was marked away.

In addition to these logs, annotator's browser and platform information is also collected. This information is important to give us a better idea of client's environment, and helps interpreting the logs with a better context.

With all these logs, we can virtually replay the complete editing process for any article.

4. Examples

Figure 2 shows an example article without editing. The title of the article is wrongly translated to "nausenA adhikArI golIbArI ghAvoM kI mauta", which is neither syntactically nor semantically correct. Using the workbench, an annotator can fix the translation to "nausenA adhikArI kI golIbArI ghAvoM se mauta".

Additionally, we can see that the summary shown in Figure 2 is not very informative, as it is not completely coherent with the title. Figure 5 shows the edits made to the mono-lingual summary using the workbench, which is also reflected in cross-lingual summary. Once the sentences in mono-lingual summary are fixed, with a few corrections in

Article	Total Time	Summary Editing			Final Summary	
		Sentences Added	Sentences Removed	Sentences Reordered	Total Sentences	Total Characters
1	320s	0	0	0	3	371
2	251s	1	1	0	2	364
3	389s	1	1	0	2	310

Table 2: Statistics for summary level editing of example articles

Article	Source Article		Summary					Title			
				Source		Target		Source		Target	
	Sent.	Words	Sent.	Words	Chars	Words	Chars	Words	Chars	Words	Chars
1	8	139	3	72	430	77	371	6	30	9	36
2	6	163	2	38	276	46	364	5	40	9	49
3	5	139	2	53	337	60	310	8	53	10	56

Table 3: Statistics of example articles edited

Article	Total Time	Estimated Translation Editing Time	Total Items	Number of Items Corrected	HTER (Summary and Title)				
					Ins	Del	Sub	Shft	Total
1	320s	223s	4	3	6	2	14	0	25.58 (22/86)
2	251s	49s	3	1	3	0	1	0	7.27 (4/55)
3	389s	277s	3	3	4	0	9	2	22.06 (15/68)

Table 4: Translation statistics for example articles

the translation of the summary, similar to what we have shown for the title, we can generate a publishable cross-lingual summary. Figure 6 shows an example of edits made to the translation of the summary in Hindi.

For demonstration, we used the workbench to generate cross-lingual summaries of three randomly selected articles of similar sizes. The articles were originally in English and we set the target language to Hindi. Table 2 shows the total time taken to generate human edited cross-lingual summaries and the summary-level edits made the the articles. Table 3 shows the number of sentences, words and characters in source articles, as well as mono-lingual and edited cross-lingual summaries. Table 4 shows the estimated time taken for editing erroneous translations and HTER along with number of insertions, deletions, substitutions and shifts performed.

4.1. Pilot Study

We conducted a pilot study to understand the usability and effectiveness of the workbench. One language expert was hired to generate translations of the articles and cross-lingual summaries for English-Gujarati language pair using the workbench.

For the pilot study, the workbench was configured to allow editing of the translation of the entire article along with the mono-lingual and the cross-lingual summary. The mono-lingual extractive summaries were provided by Veooz[7] and Google Translate was used for automatic translations.

The language expert was told to follow the following sequence to correct the article.

- First, correct the translation of all the sentences in the source article.
- Once the translation is corrected, fix the mono-lingual summary.
- The cross-lingual summary automatically picks up the sentences and their translation included in the mono-lingual summary from the article. Fix the translation errors in cross-lingual summary if required.

In this setting, we observed that the mono-lingual summary was never getting changed. On investigating, the feedback from the language expert was that the summaries are “good enough”. Although, the summaries were not always good, we observed that in this setting, due to multiple tasks (translation of the article, mono-lingual and cross-lingual summary), the “good enough” summaries were not getting enough attention.

Another feedback from the language expert was that “it would be easier and faster to translate by hand instead of post-editing the machine translation”. To verify this claim, we did a small experiment to compare time and effort taken in both the cases. The results and statistics of the experiment are shown in Table 5.

We can clearly see that the post-editing approach is faster. We see that post-editing machine translations takes about 33% lesser time compared to translating by hand. Though, post-editing machine translation takes lesser time, we notice that it might be more difficult to correct an erroneous translation compared to translation by hand and the difference noticed in time taken might be simply due to the fact that all sentences do not need to be corrected.

[7] https://www.veooz.com/

	Translation by hand	Post-editing machine translation
Avg. time per article	14.9 min	9.85 min
Avg. number of sentences per article	13.70	13.17
Avg. number of sentences edited per article	100%	57.8%
Total number of articles included in study	112	395

Table 5: Comparison between human translation and Post-Editing machine translation

5. Discussion and Future Work

The workbench can be used to collect human edited cross-lingual summaries along with mono-lingual summaries and translations of the original article for English to a number of Indian Languages. Such a dataset can be used for training statistical mono-lingual or cross-lingual summarizers as well as can help research efforts on machine translation. We also aim to collect comprehensive logs and use that as continuous feedback to some of the modules in our pipeline.

As the workbench and the architecture is not limited to a specific set of languages, the same can be used with a number of other language pairs too.

The flexibility of the workbench and the pipeline makes it possible to use the system for a number of other related tasks. The workbench can be used for extractive or abstractive mono-lingual summary generation or post-editing. It can also be used just as another translation post-editing tool, or can be used to prepare paraphrasing datasets.

Apart from the human edited data collected by the workbench, the logs collected about the process can also be important. Keystroke logs, along with translation time and HTER is a common measure of Translation Quality Estimation (Aziz et al., 2012). Automated Post-Editing (APE) systems also use similar information to automatically post-edit and try and remove systematic errors made by a particular MT system. The workbench can also be used to compare different settings of the pipeline such as different MT systems or different approaches to summarization, etc., by comparing time taken to edit or other relevant measures.

In future, following possible modules could be integrated with the workbench to improve the usability and the effectiveness of the workbench.

- Cross-lingual dictionaries to refer words and get possible translations. The dictionary can be triggered by double-clicking or selecting a word or phrase and can be shown as a pop-up to ease the work-flow.
- A Translation Quality Estimation module, that adapts and learn from previous edits, and can highlight sentences or part of sentences that need attention.
- An Automatic Post-Editing module, to automatically remove common errors made by the MT system, based on previous usage of the workbench.

6. Conclusion

We presented the workbench, a flexible, language independent, web-based tool for human editing of cross-lingual summaries along with a pluggable pipeline for cross-lingual summarization. We also explained some of the core features of the tool and possible usage scenarios for the tool, and discussed briefly about the possible uses of the data collected by the tool.

7. Acknowledgements

This work is done in collaboration and with partial support from Veooz Labs (now NewsPlus)[8].

8. Bibliographical References

Aziz, W., Castilho, S., and Specia, L. (2012). Pet: a tool for post-editing and assessing machine translation. In *LREC*, pages 3982–3987.

Litvak, M., Last, M., and Friedman, M. (2010). A new approach to improving multilingual summarization using a genetic algorithm. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, pages 927–936. Association for Computational Linguistics.

Orasan, C. and Chiorean, O. A. (2008). Evaluation of a cross-lingual romanian-english multi-document summariser. In *LREC*.

Pal, S., Zampieri, M., Naskar, S. K., Nayak, T., Vela, M., and van Genabith, J. (2016). Catalog online: Porting a post-editing tool to the web. In *LREC*.

Pingali, P., Jagarlamudi, J., and Varma, V. (2007). Experiments in cross language query focused multi-document summarization. In *Workshop on Cross Lingual Information Access Addressing the Information Need of Multilingual Societies in IJCAI2007*.

Snover, M., Dorr, B., Schwartz, R., Micciulla, L., and Makhoul, J. (2006). A study of translation edit rate with targeted human annotation. In *Proceedings of association for machine translation in the Americas*, volume 200.

Wan, X., Li, H., and Xiao, J. (2010). Cross-language document summarization based on machine translation quality prediction. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 917–926. Association for Computational Linguistics.

Wan, X. (2011). Using bilingual information for cross-language document summarization. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 1546–1555. Association for Computational Linguistics.

[8] https://www.veooz.com/

Yao, J.-g., Wan, X., and Xiao, J. (2015). Phrase-based compressive cross-language summarization. In *EMNLP*, pages 118–127.

Zhang, J., Zhou, Y., and Zong, C. (2016). Abstractive cross-language summarization via translation model enhanced predicate argument structure fusing. *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, 24(10):1842–1853.