Speech-to-Text and Evaluation of Multiple Machine Translation Systems

Evelyne Tzoukermann evelyne.tzoukermann@nvtc.gov
Steven Van Guilder sevanguilder@nvtc.gov
Jennifer Doyon jennifer.doyon@nvtc.gov
Ekaterina Harke eharke@nvtc.gov
National Virtual Translation Center, Washington, DC, USA

Abstract

The National Virtual Translation Center (NVTC) and the larger Federal Bureau of Investigation (FBI) seek to acquire tools that will facilitate its mission to provide English translations of non-English language audio and video files. In the text domain, NVTC has been using translation memory (TM) for some time and has reported on the incorporation of machine translation (MT) into that workflow. While we have explored the use of speech-to-text (STT) and speech translation (ST) in the past, we have now invested in the creation of a substantial human-created corpus to thoroughly evaluate alternatives in three languages: French, Russian, and Persian. We report on the results of multiple STT systems combined with four MT systems for these languages. We evaluated and scored the different systems in combination and analyzed results. This points the way to the most successful tool combination to deploy in this workflow.

1. Introduction

We report on the evaluation of multiple speech-to-text (STT) systems combined with four machine translation (MT) systems in order to perform comparisons on each combination. The goal of this project was to determine which combination of STT and MT systems would be optimal for a given language. That way, translators can benefit from this evaluation and use the optimal combination for any of these three languages. By combining and testing different configurations of STT and MT in a novel way, we have been able to determine strengths and weaknesses of these different workflows. In 2021, we reported on STT performance comparison and evaluation (see Miller et al. 2021). This year, we are presenting the results of the performance of multiple MT systems combined with the STT systems from last year. More specifically, we report on the evaluation of three to seven speech-to-text systems with four machine translation (MT) systems in order to perform comparisons for each combination. For French, we also compared the results with a speech translation system (all-in-one). The paper presents results and analysis of combinations for French, Russian, and Persian.

2. Evaluation Corpus

The corpus for this evaluation was comprised of two hours of audio for each language, which corresponded roughly to 2,000 sentences for each of these languages. French was based on a conversational document where a set of experts gathered in a panel, in person and virtually, to discuss state-of-the art in technological innovations in the space domain. Russian data collected was also conversational speech discussing technical innovations in the additive manufacturing and 3D printing domain. The Persian data was a broadcast interview on cybersecurity, cyber attacks and strategies for defense. Prior work (Miller et al. 2021) provides more detail on

the data selected, and the processes put in place for manual transcription, manual translation, as well as descriptions of the different STT systems that were used.

3. Systems and Scoring

We provide results for a multi-system comparison of French, Russian, and Persian. Table 1 lists all of the STT systems that were used[1]. STT-COTS are commercial-off-the-shelf STTs whereas STT-GOTS are government-off-the-shelf STT systems. As shown on the table, Russian is the only language on which seven STT systems were available to be run and analyzed. French was run on five available systems, which includes STT-COTS1 for European French and STT-COTS1-1 for Canadian French. Persian was run on the only three systems that were able to process Persian. For French only, we had access to a commercial, all-in-one speech system and the results are presented in the French Results section.

STT Systems	FRENCH	RUSSIAN	PERSIAN
STT-COTS1	**FRE** √	√	√
STT-COTS1-1	**CAN** √		
STT-COTS2	√	√	√
STT-COTS3	√	√	
STT-COTS4	√	√	
STT-COTS5		√	√
STT-GOTS1		√	
STT-GOTS2		√	

Table 1. Speech-to-text systems run with each of the three languages.

All but one of the four MT systems used in this project were commercial products. All three COTS products were multilingual, neural-based engines. The fourth system was a government product. This product was an integration of several, multilingual MT engines that employ a number of approaches for performing automatic translation, including direct and statistical methods[2]. This GOTS system is only available to government users. Table 2 lists the machine translation systems that were used in the combination.

MT Systems	FRENCH	RUSSIAN	PERSIAN
MT-COTS1	√	√	√
MT-COTS2	√	√	√
MT-COTS3	√	√	√
MT-GOTS1	√	√	√

Table 2. Machine translation systems used for the three languages.

Naturally, this yielded a very rich combination of STT systems and MT systems. The numbers of systems that were compared were as follows:

- 20 combinations for French
- 28 combinations for Russian
- 12 combinations for Persian

[1] For the sake of anonymity, we renamed all the systems by a number, differentiating them only by their commercial or government source.

[2] Direct machine translation systems are generally rule-based, whereas statistical machine translation systems learn how to translate by analyzing existing human translations based on bilingual text corpora.

This is the first time that we had the privilege of using such a large number of automatic tools in one of our evaluations. As a result, it generated a rich and comprehensive evaluation of multiple systems.

The BLEU (BiLingual Evaluation Understudy) metric was used to score the output of all the MT systems. BLEU is the commonly used metric for automatically evaluating machine-translated text (see Papineni et al. 2002 and NLTK[3]). The BLEU score is a number between zero and one that measures the similarity of the machine-translated text to a set of high-quality reference translations produced by human translators. It has been shown that BLEU scores tend to correlate with human judgment of translation quality (see Chen and Cherry 2014 and Banerjee and Lavie 2005)[4].

For the evaluation performed during this project, we had access to one human-produced reference translation available against which all the STT/MT workflow outputs were scored. The all-in-one pipeline was scored using this single human reference for comparison as well as using the output translations of all the STT/MT combinations to see if additional translation would yield different results when used as reference.

4. Speech-To-Text and Machine Translation Results

The following sections present the results of the different combinations for each of the four languages; the results are analyzed and discussed.

4.1. French Results

Table 3 shows the 20 system workflows along with the MT BLEU scores for each combination from the STT standpoint as opposed to Table 4 which shows results from the MT standpoint. The middle column displays the STT and MT specific workflow, and the numbers in the right column indicate the score for each. The higher the number, the better the score; red cells show lower scores than orange, yellow, and green cells. The green cells show the highest scores. The highest performing system pairs are STT-COTS2 + MT-COTS2 (line 19) followed by STT-COTS1-1 + MT-COTS2 (line 7), closely followed by STT-COTS2 + MT-COTS3 (line 20). The lowest performing system pair is STT-COTS3 + MT-GOTS1 (line 1), which is significantly lower. Although at first glance, the scores in Table 2 might look low, it is important to remember that these reflect a challenging pipeline from STT to MT.

[3] NLTK: nltk.translate.nist_score https://www.nltk.org/_modules/nltk/translate/nist_score.html

[4] Note that even human translators do not achieve a perfect BLEU score of 1.0. Several smoothing functions have been put forward for BLEU to deal with n-gram results of zero which often occur in higher-level n-grams. Since the overall BLEU score is the geometric mean of the different n-gram levels, a single n-gram result of zero would cause the overall score to be zero. The smoothing method used in this evaluation was a simple one put forward in the NIST Toolkit 'mteval-v13a.pl', where the first zero value encountered was assigned a value of ½, the second was assigned a value of ¼, the third a value of 1/8, and so on (See Papineni et al. 2002).

	French STT/MT Combinations	BLEU Scores
1	STT-COTS3 + MT-GOTS1	0.07779377
2	STT-COTS3 + MT-COTS1	0.12499899
3	STT-COTS3 + MT-COTS2	0.13734244
4	STT-COTS3 + MT-COTS3	0.12797222
5	STT-COTS1-1 + MT-GOTS1	0.15692103
6	STT-COTS1-1 + MT-COTS1	0.26834392
7	STT-COTS1-1 + MT-COTS2	0.30656324
8	STT-COTS1-1 + MT-COTS3	0.2770375
9	STT-COTS1 + MT-GOTS1	0.14525775
10	STT-COTS1 + MT-COTS1	0.25602079
11	STT-COTS1 + MT-COTS2	0.29186129
12	STT-COTS1 + MT-COTS3	0.26635345
13	STT-COTS4 + MT-GOTS1	0.1488969
14	STT-COTS4 + MT-COTS1	0.2354315
15	STT-COTS4 + MT-COTS2	0.26638751
16	STT-COTS4 + MT-COTS3	0.24955649
17	STT-COTS2 + MT-GOTS1	0.15188625
18	STT-COTS2 + MT-COTS1	0.27867294
19	STT-COTS2 + MT-COTS2	0.31683667
20	STT-COTS2 + MT-COTS3	0.29572112

Table 3. Twenty (20) STT and MT system combinations with BLEU scores for French

Table 4 synthesizes the system combinations showing the results from the MT system standpoint. Again, colors are very helpful to visualize the results. STT-COTS3 is clearly a low STT performer, and MT-GOTS1 is a low MT performer. In contrast, STT-COTS2 and MT-COTS2 show the highest performance, and MT-COTS2 remains a high performing system when paired with the other STT systems as well.

MT Systems	STT Systems				
	STT-COTS3	STT-COTS1-1	STT-COTS1	STT-COTS4	STT-COTS2
MT-GOTS1	0.077793768	0.156921027	0.145257748	0.148896898	0.151886249
MT-COTS1	0.124998991	0.268343922	0.25602079	0.235431496	0.278672936
MT-COTS2	0.137342442	0.306563239	0.291861286	0.266387507	0.316836666
MT-COTS3	0.127972218	0.277037504	0.266353446	0.249556491	0.295721118

Table 4. The 20 STT and MT systems from the MT standpoint for French

French was also evaluated on an all-in-one system consisting of integrated STT and MT to produce an English text translation of the French audio source. Note that we evaluated and compared two different versions for ST-COTS2[5], a January and February version, which returned slightly different results. Table 5 shows three sets of results. The results in A show the

[5] ST-COTS2 is based on STT-COTS2 and MT-COTS2, both high performing systems.

system scores after ingesting the audio, getting the English text, and scoring the results using the human translation as a reference. The B section results show the same process, but this time using the four machine translation outputs as reference translations, and not the human translation. Section C results show the same operation again, this time using the human translation as well as the other four machine translations as reference translations, for a total of five reference translations. For the BLEU score, the more reference translations, the better the system scores, because the addition of translation variants increases the system translation. The resulting scores in Table 5 clearly demonstrates this effect, and the five reference translations in C exhibit the highest translation scores. It is important to note that the results in sections B and C ranging from .78 to .85 are several orders of magnitude higher than the STT-MT combinations in Tables 3 and 4, varying from 0.316836666 to 0.316836666.

	All-in-One Speech Translation	**Bleu Scores**
A	**ST-COTS2 January (1 human reference translation)**	0.32152296
	ST-COTS2 February (1 human reference translation)	0.321454151
B	**ST-COTS2 January (4 MT translations)**	0.779563801
	ST-COTS2 February (4 MT translations)	0.797920901
C	**ST-COTS2 January (1 human + 4 MT translations)**	0.876123775
	ST-COTS2 February (1 human + 4 MT translations)	0.846498433

Table 5. Three ST systems with variable number of reference translations.

4.2. Russian Results

Russian, as mentioned earlier, was processed with the highest number of STT systems, that is seven (7). Table 6 shows the STT/MT combinations with BLEU scores. This time, the three dominant systems are STT-GOTS2 + MT-COTS2, as shown in line 27 of Table 6, STT-GOTS1 + MT-COTS2 in line 23, and STT-GOTS2 + MT-COTS3 (in line 28). Interestingly, these 3 combinations both have used a version of STT-GOTS, the government created systems. The 3 worst performing combinations are STT-COTS4 + MT-GOTS1 (line 13), STT-COTS1 + MT-COTS3 (line 4), and STT-COTS2 + MT-GOTS1 in line 5.

	Russian STT/MT Combinations	**BLEU Scores**
1	**STT-COTS1 + MT-GOTS1**	0.071936558
2	**STT-COTS1 + MT-COTS1**	0.128206719
3	**STT-COTS1 + MT-COTS2**	0.153879991
4	**STT-COTS1 + MT-COTS3**	0.064070521
5	**STT-COTS2 + MT-GOTS1**	0.064070521
6	**STT-COTS2 + MT-COTS1**	0.162268592
7	**STT-COTS2 + MT-COTS2**	0.148891882
8	**STT-COTS2 + MT-COTS3**	0.140978788
9	**STT-COTS3 + MT-GOTS1**	0.068786887
10	**STT-COTS3 + MT-COTS1**	0.107556239
11	**STT-COTS3 + MT-COTS2**	0.143115371
12	**STT-COTS3 + MT-COTS3**	0.131434129

13	**STT-COTS4 + MT-GOTS1**	0.045528381
14	**STT-COTS4 + MT-COTS1**	0.065932173
15	**STT-COTS4 + MT-COTS2**	0.093096203
16	**STT-COTS4 + MT-COTS3**	0.083497782
17	**STT-COTS5 + MT-GOTS1**	0.068218524
18	**STT-COTS5 + MT-COTS1**	0.101121023
19	**STT-COTS5 + MT-COTS2**	0.140348315
20	**STT-COTS5 + MT-COTS3**	0.129364632
21	**STT-GOTS1 + MT-GOTS1**	0.082513608
22	**STT-GOTS1 + MT-COTS1**	0.144663213
23	**STT-GOTS1 + MT-COTS2**	0.174511147
24	**STT-GOTS1 + MT-COTS3**	0.153340608
25	**STT-GOTS2 + MT-GOTS1**	0.089778715
26	**STT-GOTS2 + MT-COTS1**	0.157261373
27	**STT-GOTS2 + MT-COTS2**	0.190733115
28	**STT-GOTS2 + MT-COTS3**	0.168950871

Table 6. Twenty-eight (28) STT and MT system combinations with BLEU scores from the STT standpoint for Russian

Table 7 outlines MT scores. Indeed, MT-GOTS1 performs significantly lower than the other systems. The table shows how the combination with STT-COTS4 generates poor results. On the other hand, MT-COTS2 is the highest performing MT system except when in combination with STT-COTS4. This demonstrates how strong MT-COTS2 performs but that its performance is negatively impacted with the weaker STT-COTS4.

STT systems	**STT systems**			
	MT-GOTS1	**MT-COTS1**	**MT-COTS2**	**MT-COTS3**
STT-COTS1	0.071936558	0.128206719	0.153879991	0.064070521
STT-COTS2	0.064070521	0.162268592	0.148891882	0.140978788
STT-COTS3	0.068786887	0.107556239	0.143115371	0.131434129
STT-COTS4	0.045528381	0.065932173	0.093096203	0.083497782
STT-COTS5	0.068218524	0.101121023	0.140348315	0.129364632
STT-GOTS1	0.082513608	0.144663213	0.174511147	0.153340608
STT-GOTS2	0.089778715	0.157261373	0.190733115	0.168950871

Table 7. The twenty-eight (28) STT and MT systems from the MT standpoint for Russian

4.3. Persian Results

The Persian data was run on three STT systems and three MT systems. STT-COTS1 and STT-COTS2 performed well (see lines 2, 3, 4, and 6, 7, 8) except when combined with MT-GOTS1 (see lines 1 and 5). In contrast, STT-COTS5 did not perform well with any of the MT systems and shows the lowest results when associated with MT-GOTS1 (see line 9).

	Persian STT/MT Combinations	Bleu Scores
1	STT-COTS1 + MT-GOTS1	0.076480559
2	STT-COTS1 + MT-COTS1	0.155602641
3	STT-COTS1 + MT-COTS2	0.139287289
4	STT-COTS1 + MT-COTS3	0.131836115
5	STT-COTS2 + MT-GOTS1	0.092067026
6	STT-COTS2 + MT-COTS1	0.186981095
7	STT-COTS2 + MT-COTS2	0.186434702
8	STT-COTS2 + MT-COTS3	0.168386801
9	STT-COTS5 + MT-GOTS1	0.047054087
10	STT-COTS5 + MT-COTS1	0.095733738
11	STT-COTS5 + MT-COTS2	0.099165053
12	STT-COTS5 + MT-COTS3	0.081293357

Table 8. The twelve (12) STT and MT system combinations with BLEU scores from the STT standpoint for Persian

In Table 9, we clearly see that the combinations between STT-COTS1, STT-COTS2 and all of the MT-COTS systems is superior to the combinations with MT-GOTS1.

STT Systems	STT Systems			
	MT-GOTS1	MT-COTS1	MT-COTS2	MT-COTS3
STT-COTS1	0.076480559	0.155602641	0.13928729	0.131836115
STT-COTS2	0.092067026	0.186981095	0.1864347	0.168386801
STT-COTS5	0.047054087	0.095733738	0.09916505	0.081293357

Table 9. The twelve (12) STT and MT systems from the MT standpoint for Persian

4.4. Summary of Results across Languages

We compare here the highest combinations for the four languages in order to gain a deeper understanding of why each combination might be more (or less) effective. The first observation, as mentioned in the French results section, is that the scores of French ST (line 1 in Table 10) is several orders of magnitude higher than the French STT-MT combination. Our initial conjecture is that STT and MT working in tandem allow for the optimal translation. The system appears to be selecting the optimal hypothesis in the joint language models of French and English (see Matusov et al. 2005 and Lamel et al. 2011). These ideas will be further explored. The second observation is that the results of French STT-MT are almost two times better than the other languages. We believe that this may be due to the fact that French STT-MT pairs are more mature, thus more robust than they are for the other languages.

	Languages	STT/MT combinations	Bleu scores
1	French	ST-COTS2 with 1 human + 4 MT translation references	0.876123775
2	French	STT-COTS2 + MT-COTS2	0.31683667
3	Russian	STT-GOTS2 + MT-COTS2	0.190733115
4	Persian	STT-COTS2 + MT-COTS1	0.186981095

Table 10. Highest performing STT-MT Combinations

5. Conclusion[6]

The goal of this project is to determine which combination of STT and MT systems would be optimal for a given language in order to optimize an audio to translation workflow. We had access to STTs and MTs systems and analyzed 60 combinations. The results provided in the paper demonstrate very clearly the strong and poor STT and MT combinations. Our evaluation results show that for French, the speech translation (all-in-one system) along with multiple reference translations appears to be the best selection for integration into the NVTC workflow. Of course, this depends on the maturity of the systems, and we have observed these results since the selected French ST-COTS2 very robust. We are planning on exploring the use of ST-COTS2 on other languages to see if the conclusions appear to be the same as for French.

References

Banerjee, S., & Lavie, A. (2005, June). METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In Proceedings of the ACL workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization (pp. 65-72).

Chen, B., & Cherry, C. (2014, June). A systematic comparison of smoothing techniques for sentence-level bleu. In Proceedings of the ninth workshop on statistical machine translation (pp. 362-367).

Lamel, L., Courcinous, S., Despres, J., Gauvain, J. L., Josse, Y., Kilgour, K., ... & Woehrling, C. (2011). Speech recognition for machine translation in Quaero. In *Proceedings of the 8th International Workshop on Spoken Language Translation: Evaluation Campaign*.

Matusov, E., Kanthak, S., & Ney, H. (2005). On the integration of speech recognition and statistical machine translation. In *Ninth European Conference on Speech Communication and Technology*.

Miller, C., Tzoukermann, E., Doyon, J., & Mallard, E. (2021, August). Corpus Creation and Evaluation for Speech-to-Text and Speech Translation. In Proceedings of Machine Translation Summit XVIII: Users and Providers Track (pp. 44-53).

Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2002, July). Bleu: a method for automatic evaluation of machine translation. In Proceedings of the 40th annual meeting of the Association for Computational Linguistics (pp. 311-318).

Reiter, E. (2018). A structured review of the validity of BLEU. Computational Linguistics, 44(3), 393-401.

Tzoukermann, E. and Miller, C. (2018). Evaluating Automatic Speech Recognition in Translation. In *Proceedings of AMTA 2018*, vol. 2, pages 294-302, Boston.

[6] The authors would like to thank Dr. Corey Miller for his contributions.