Evaluation of Konch's auto-transcription service for Aarhus University Arts

Maris Sala, Max R. Eckardt, & Kristoffer L. Nielbo,
Center for Humanities Computing Aarhus, Aarhus University, Jens Chr. Skous Vej 4, Building 1483, 3rd floor DK- 8000 Aarhus C.

Introduction

As of June 1, 2020, Aarhus University (AU) has had 154 users of the Konch service offered by DeiC at `deic.app.konch.ai`, predominantly from the Arts and Social Sciences faculties. For evaluation purposes Center for Humanities Computing Aarhus (CHCAA) has distributed a short software evaluation survey to AU users. Because CHCAA uses Google Cloud Platform's (GCP) Speech-to-Text API for auto-transcription, we have also included an illustrative comparison of sample texts.

Software Evaluation Survey

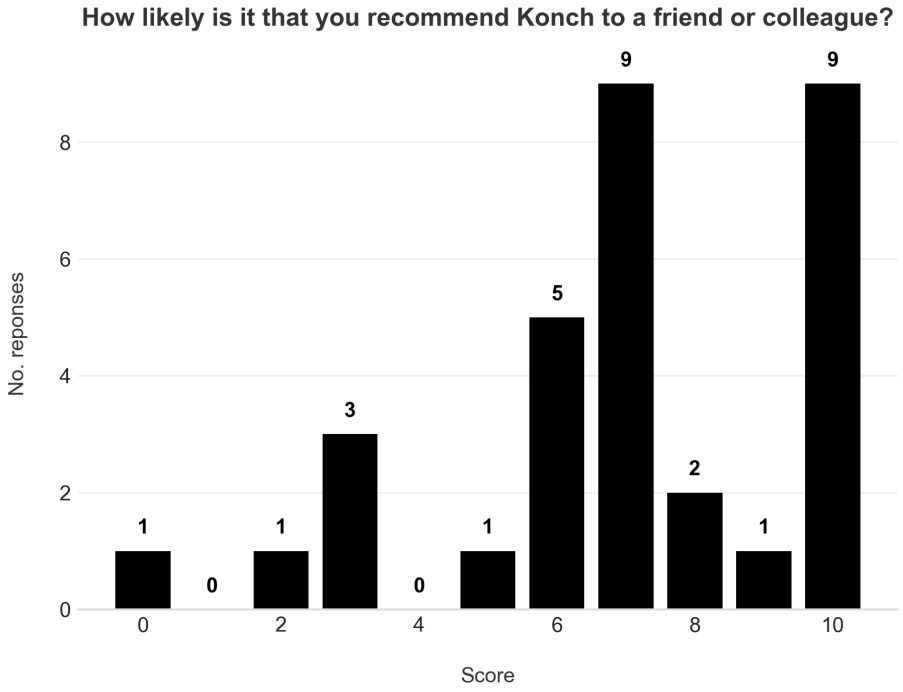
A short 12-item software evaluation survey was distributed to all users of Konch at AU of which 32 completed the survey. The items probed users' (self-reported) satisfaction with Konch's reliability, security, interoperability and support. Ten items were ordinal with five categories (*Not satisfied at all*, *Not so satisfied*, *Somewhat satisfied*, *Very satisfied*, *Extremely satisfied*), one item was interval (0-10) and the final item was open. Respondents were predominantly students (47%) or researchers (41%).

Results

More than half of the respondents were 0.70 or more likely to recommend Konch to friends or colleagues, see figure 1. 59.4% were only somewhat satisfied with Konch's reliability, while 12.5% were not satisfied, and only 28.2% were very or extremely satisfied. Respondents were generally satisfied with security and GDPR issues (78.1% very or extremely satisfied and 18.8% were somewhat satisfied). It should be noted that issues pertaining to GDPR were the second most frequent support request CHCAA received. Finally, 50% of the users were very or extremely satisfied with Konch ability to integrate with other software, 41% were only somewhat satisfied and 9% not satisfied. Several respondents used Konch together with computer-assisted qualitative data analysis software (e.g., NVivo or MAXQDA) and would have liked to use Konch as a plugin for their analysis software.

Respondents were generally satisfied with the UI experience, 69% were very or extremely satisfied with the look and feel and the remainder somewhat satisfied, 62.6% very or extremely satisfied with ease of use. The setup experience however is an issue that 12.5% are not satisfied with and 34.4% are only somewhat satisfied. Setup was by far the most frequent support request, where users either did not register or registered on `konch.ai`. CHCAA's default response was to provide users with the relevant links to Konch's documentation. For those that used the documentation, 40% were not satisfied or not satisfied at all, 30% were only somewhat satisfied and only 30% were very or extremely satisfied. 23% were not satisfied with Konch's user support, while 31% were somewhat satisfied, and 46% were very or extremely satisfied.

14 respondents contributed open responses to possible improvements of Konch. The most frequent suggestion was to improve the accuracy of Danish text-to-speech, because the service lacks conside-



Figur 1: Users' willingness to recommend Konch to friends or colleagues.

rably for Danish compared to English, and so resource requirements for post-processing were substantial.

GCP-Konch comparison

CHCAA has been providing a speech-to-text service for a select group of researchers at University of Southern Denmark (SDU) and AU using GCP's API. It is important to point out that neither AU or SDU have data-processing agreements with Google that allow us to use GCP services for personal data even without Google's data logging option. There are however no principal reason as to why such an agreement could not be obtained – several American universities have data processing agreements with Google. Another important issue is pricing, assumming Konch's pricing is 0.46 DKK/minute[1], GCP's service range from 0.11 to 0.24 DKK/minute for their most expensive service (i.e., with enhanced neural models and without data logging).

Method

For comparison we used a small sample of four ecologically valid interviews in English and Danish[2]. For each sample text, a gold standard was obtained from expert transcribers and Konch and GCP's performance were compared with reference to the gold standard on two metrics, BLEU and METEOR. BLEU (bilingual evaluation understudy) is an intuitive measure of distance between the word matches of the translation to the reference texts.[3] It's designed to work best on corpus-level, and on

[1]Estimate is based on previous report from Aalborg University.

[2]Performance was tested on a larger data set and the results follow the same result pattern. At the time of writing, we have not obtained sufficient rights to share these results.

[3]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).

sentence-level analysis it might give inflated scores.[4] The score range from 0 to 1, where 1 signifies a perfect match between the human-made translation and machine translation. METEOR (Metric for Evaluation of Translation with Explicit Ordering) is another evaluation tool used to assess quality of machine translation. It is more complex than BLEU and does approximate human judgment better since it also considers synonyms and uses a stemmer.[5] It weighs recall (matched unigrams divided by unigrams in reference) as more important than precision. METEOR is also adjusted to be used on sentence-level instead of document level - in contrast to BLEU.

Text-ID	Software	BLEU	METEOR
$Text_0$	GCP	0.95	0.58
$Text_0$	Konch	0.91	0.52
$Text_1$	GCP	0.94	0.53
$Text_1$	Konch	0.91	0.51
$Text_2$	GCP	0.9	0.5
$Text_2$	Konch	0.9	0.48
$Text_3$	GCP	0.11	0.17
$Text_3$	Konch	0.1	0.17

Tabel 1: BLEU and METEOR scores from GCP and Konch. BLEU scores are typically higher than METEOR, which is often seen in the literature.

Results

The original sample recording varied in quality and one ($Text_3$) was characterized by heavy Danish dialect. Conditions however were identical for both GCP and Konch (i.e., same audio files with no pre-processing) and results shown in table 1.

Overall, both scores agree on which service has the better performance. GCP's speech-to-text consistently outperforms Konch although only with a small margin. Even though the metrics differ in how they score the texts, they are consistent on which sample texts are translated better – dialect (Danish or otherwise) is a general problem in auto-transcription.

Conclusion

Survey results from AU showed that Konch users are generally satisfied with the tool usage and UI experience, but that proper support and documentation were lacking. This is important, because user support will require local investments – CHCAA provided additional support pro bono during the project evaluation. Several of the users were dissatisfied with the Konch performance for Danish transcription. On a side note, CHCAA observes that Konch does not provide an API or technical documentation, which makes the service irrelevant for larger research projects.

[4]Tatman, R. (2019, Jan). Evaluating Text Output in NLP: BLEU at your own risk. Medium: Towards Data Science. Retrieved from: `https://towardsdatascience.com/evaluating-text-output-in-nlp-bleu-at-your-own-risk-e8609665a213`

[5]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).

In comparison to GCP, which does provide an API, Konch is substantially more expensive, but has the important advantage of a data processing agreement with Danish universities. Konch is furthermore outperformed by GCP's speech-to-text, which does also integrate well with larger research projects. Regarding the DeiC inquiry of our continued usage of Konch after transitioning to a payed service, the answer is twofold. CHCAA will use cognitive services at GCP (or Azure) instead of Konch because they provide API, lower pricing, and improved performance. We would prefer that resources were spent on establishing a data processing agreement with these commercial cloud providers[6]. Some users at Arts may want to continue their use of Konch as a payed service, but this can be negotiated at an individual project basis – we have already been in dialogue with Konch.ai regarding this option.

[6] AU's data processing agreement with Azure does not include cognitive services at the moment.