Enhancement of deep learning based semantic representations with acoustic-prosodic features for automatic spoken document summarization and retrieval

NKFIH FK-214413

Final report

György Szaszák

June 2023

Introduction

In our research we had a general focus on automatic spoken document summarization and retrieval in a modular architecture, which allowed us to inspect and analyse each module either separately or in connection. Beside optimizing overall spoken documentation summarization performance, we had planned the research in a way that we could get insights into different aspects of human language and speech in particular. We have also addressed aspects that often get passed on, namely the exploitation of speech prosody, the combination of audio and text processing NLP techniques, and we have also benchmarked automatic processing w.r.t. human perception, that is which errors, losses of information cause perceivable degradation on the human side.

We started our research at the beginning of the revolution Artificial Intelligence (AI) brought in many fields of the technology, either in research, engineering or everyday life. Since these changes triggered by the development of AI were really revolutionary, from the 3rd year of the project we decided to adopt our research plan as some tasks we had found reasonable to deal with at project planning became less relevant, whereas others became more crucial. This essentially means that according to the paradigm shift caused mostly by transfer learning, we built a bit more than planned on top of third party resources and models and fine-tuned rather those then forcing the use of our own implementations, if these latter proved to be less efficient or outdated. Indeed, also in the industry, development cycles became shorter and showed large deviations from original plans because of the fast progression of the underlying technology.

Although we were not capable of exploiting full scaled big data practices because of the lack of resources and the limited budget, we believe that even if the industrial and research context has changed much and rapidly, the results we obtained are still important and contribute to industrially exploitable solutions, especially regarding the Hungarian language, where due to its agglutinating nature and the related relative data scarsity, it is still not trivial to adopt approaches elaborated essentially for English or fine-tune models with limited amount of in domain data.

Baselines

Before moving on to items in our research plan, we created our baselines. In its vanilla form, speech summarization was implemented as a cascade of an Automatic Speech Recognizer (ASR) and text based extractive summarization. As ASR baseline, an in house Kaldi based implementation was created consisting of recordings of several datasets, including but not limited to MRBA and BEA. For several experiments we could use the ASR of SpeechTex Ltd, which is already in production use. The text summarization baseline was relying on Gensim, which provides graph based extractive summarization. In this form of summarization, we rank the sentences and extract the most relevant ones from them to create the summary. In order to use a baseline which allows for a wide comparison, we chose Gensim which is available as a Python module and is well documented and well-known. We refactored this module to speed up processing and performed the intitial performance analysis [1].

We separately created a prosodic modelling baseline, emerging from our previous project NKFIH PD–112598, based on phonological phrase modelling composed of a Hidden Markov model (HMM) for time alignment, and a Gaussian Mixture Model (GMM) for phonological phrase entity modelling. Regarding prosody modelling, we also explored a new robust approach based on formal atom decomposition of the pitch and energy tracks of speech utterances (WCAD, Weighted Correlation based Atom Decomposition). The hybrid approach combining the phonological phrase modelling approach and the WCAD components yielded significant improvement over our baseline. We tested these for Hungarian and French languages and also published our method and the obtained results in a high impact international journal [2].

From the 2nd year of research we switched to these improved baselines and compared further experiments to them.

Experiments on prosody based tokenization

The main interest in prosody based tokenization is that it is thought to be robust to Automatic Speech Recognition (ASR) errors. Moreover, it is able to operate at the sentence level, whereas ASR typically operate at word or character level. From there two use-cases arise: (1) sentence level tokenization for summarization of spoken documents and (2) rich transcription of ASR output by adding punctuation marks.

We conducted a preliminary analysis on broadcast news data, what ratio of phonological phrases match real sentences boundaries. An encouraging 80% accuracy was measured. Spontaneous speech cannot be tokenized for sentences, as complete grammatical sentences are often not present in spontaneous speech. Therefore prosody based tokenization yields intonational phrase tokens for spontaneous speech. Subjective tests had confirmed that such a tokenization improves mean opinion scores when evaluating summarization.

Punctuation restoration

In ASR, two main approaches can be applied or combined for punctuation insertion. Text based punctuation approaches exploit word context dependency of the punctuation marks, whereas audio based punctuation approaches exploit acoustic markers which correlate with clause or sentence boundaries. Audio based approaches have the advantage of being independent of ASR errors and hence can block the word errors propagating further in the processing pipeline.

We first investigated an audio only punctuation model, which exploits the correlation between phonological phrasing and punctuation marks. As the phonological phrasing represents the building blocks of sentence level intonation, we model them as a sequence and map this sequence to the sequence of the punctuation marks. The most suitable machine learning framework for such tasks is using recurrent neural networks with Long-Short Term Memory cells (LSTM). Inserting punctuation marks into the word chain hypothesis produced by ASR has long been a neglected task. In several application domains of ASR, real-time punctuation is however vital to improve human readability. We proposed and evaluated a prosody inspired approach in the form of a phrase sequence model implemented as a recurrent neural network to predict the punctuation marks from the audio. In a very basic and lightweight modeling framework, we showed that punctuation was possible by state-of-the-art performance, solely based on the audio signal for speech close to read quality. We tested the approach on more spontaneous speaking styles and on ASR transcripts which may contain word errors. A subjective evaluation was also carried out to quantify the benefits of the punctuation on human readability, and we also showed that when a critical punctuation accuracy is reached, humans were not able to distinguish automatic and human produced punctuation, even if the former may contain punctuation errors [3].

Later we added text based features to the model. We paid particular attention to model both syntactically correct, and syntactically impaired (ASR errors) text, and also to build a compatible framework with the audio based model for later combination. The text models were based on word and/or character embeddings, followed by a neural model exploiting contextual information through recurrent layers. Although these models have a larger footprint than the audio only model, they work at a better accuracy. Combining audio and text based approaches yielded further improvement [4]. We performed a throughout evaluation of the proposed models and their combinations. We used the F1-score computed from precision and recall for evaluation, as well as the Slot Error Rate (SER) corresponding to state-of-the-art practice. We trained and tested models for Hungarian and English.

Summarizing the results on punctuation, for the highly agglutinating and relatively free word order Hungarian, we could obtain significant improvement in overall punctuation over the word based baseline by adding the character based and audio based models. Considering the most important use-case, where ASR transcripts are input, adding these helps by rel. 18% in Hungarian over the word only baseline. For English, adding prosody lead to an improvement of 4.4% over the word baseline on ASR transcripts, still significant by $p<0.05$.

Analysis of word and audio embeddings

This research direction was spanning over the first three years of our project, we started the work by collecting a large representative Hungarian webcorpus consisting of news, magazines, etc. covering everyday topics. We cleaned the corpus and built a dictionary from it, based on term frequency. In the corpus 1.5 million different Hungarian words occur (45% of them more than 5 times), the total number of word tokens is over 65 million.

We trained skip-gram embeddings with various hyperparatemer settings using Tensorflow and the Google word2vec recipe, adapted to our needs. We tested several approaches which allow for reducing high context variability and large dictionary sizes, resulting from the highly agglutinating nature of Hungarian and its less constrained word order. This is especially important when using an ASR , itself of a limited vocabulary too. Due to these characteristics of Hungarian, the semantic robustness and coherence of the word embeddings is lower compared to English. Indeed, semantic accuracies measured on word analogy tasks are known to decrease in parallel with increasing agglutination (i. e. on Wikipedia 78% for English, 62% for German, 52% for Italian and 27% for Czech). We measured 15% semantic accuracy for Hungarian, which, taking into account the extensive agglutination is inline with the reported state-of-the-art (hence it is a fair baseline). We also prepared the Hungarian version of the Google analogy test set used for the evaluation of embeddings syntactic and semantic accuracies.

We concentrated our efforts on limiting vocabulary size and reducing contextual variation by eventually using larger context windows in order to improve the semantic capabilities of our embeddings. This again was motivated based on the constraint an ASR brings in when processing spoken language. We did not consider stemming (chunking the tokens to a fixed length) as a real alternative as it obviously impacts semantics by sometimes preserving, sometimes removing word endings leading to incoherence. Lemmatization on the other hand was the primary choice, we performed it based on a rule based dependency parser (magyarlánc) and by using an unsupervised data driven approach (morfessor). Using lemmatization with magyarlánc increased semantic accuracy from 15% to over 28%, but obviously made syntactic testing senseless. Using morfessor for lemmatization yielded 18% semantic accuracy, and by incorporating all morph units, syntactic accuracy was as high as 40% (baseline: 27%) [5].

Finally we extended our research to cover speech embeddings, as evidence suggests that complementary information can be extracted from speech and added to text features. We have analysed acoustic bag-of-words and Gábor Gosztolya lead an experiment on using Fisher representation for audio related tasks [6].

Audio embeddings used during the project were further developed and used for projects targeting classification of emotions, cognitive and physiological states in a number of connected research items [8].

Analysis of ASR error propagation into downstream summarization

When word level features are used, and transcripts are generated by ASR, word errors may appear which propagate further into the tokenization and summarization pipeline. Therefore the effect of this error propagation on sumarization performance was also assessed.

We were primarily interested in the assessment of semantic bias introduced by the presence of ASR and/or punctuation errors. We created therefore 4 kinds of transcripts to be compared: Manual Transcript with Manual Punctuation (MT-MP); ASR Transcript with Manual Punctuation (AT-MP); Manual Transcript with Automatic Punctuation (MT-AP); ASR Transcript with Automatic Punctuation (AT-AP).

Then we prepared summaries for the MT-MP, AT-MP, MT-AP and AT-AP scenarios and compared them based on the Recall-Oriented Understudy for Gisting Evaluation (ROUGE) metric family. As reference, we use human made summaries prepared by 3 independent annotators based on the MT-MP scenario transcripts. We performed the summarization with Python’s Gensim and the BM25 scoring function (our baseline).

The most relevant cases to compare are the MT-MP (gold) and the real use-case AT-AP, where both transcription and punctuation are derived automatically. Briefly summarizing the results for 3 different genres of audio in Hungarian, we observed that summarization on the MT-MP transcripts is highly correlated with ASR accuracies, most likely because the language model in the ASR and the semantic ranking module in the summary module have to face a task of similar linguistic complexity. Punctuation errors were found to be more crucial for summarization. This also correlates well with the sentencewise results: word errors caused limited bias in the semantic space at the sentence level, provided that the true sentence level was known (AT-MP). In the AT-AP scenario, word errors already propagate into the AP phase as well. This also confirms that sentence level tokenization errors (that is, punctuation errors) influence summarization at a larger scale than word errors [7].

Analysis of human readability and understandability in automatic summaries

Beside the objective and automatic evaluation of ASR and punctuation error propagation we run subjective tests too, since summaries are mostly read by human, who have subjective criteria when getting impression about summary quality. Therefore we have also performed subjective tests in Hungarian and English (including reading automatic summaries or watching a video with automatic subtitles), following a similar experimental setup than the one presented in the previous section. Here we focussed on subjective evaluation, i.e. whether punctuation adds a subjectively confirmed benefit to the captions, and what can we say about the relation between the used objective and subjective measures. We also involved Deaf and Hard of Hearing (DHH) subjects in order to inspect the primary end-user audience of closed captioning. The subjective evaluation process was designed such that it makes the assessment possible on word error-free transcripts (to evaluate clearly the share of punctuation in understanding a text) and ASR-produced transcripts (to test for realistic use-cases and to see whether punctuation keeps to be useful when word errors already degrade text quality). For punctuation, we compared three strategies: missing punctuation, error-free punctuation and machine produced punctuation. Our results, obtained from a big sample survey, demonstrated clearly that users prefer punctuated text, even if punctuation is prone to some errors. MOS were significantly higher when using a recurrent model for punctuation, with the condition, that word errors occur up to a 20-25% WER (in Hungarian broadcast tasks). Indeed, it is easy to agree that once word errors trespass a critical amount, the punctuation task itself becomes hard to define, as the word chain to be punctuated is grammatically incorrect. A similar problem arises with spontaneous speech, where punctuation is not defined in the sense it is used in written language. Beside significance tests on the obtained MOS for the 6 examined caption strategies, a Generalized Additive Model (GAM) approach also confirmed that punctuation errors account for approx. 1/3 of the variance measured in the user scores. Experiments with DHH subjects showed a more pronounced benefit in favour of punctuated captions, punctuation of ASR transcript was preferred over missing punctuation marks [4].

Language modelling (ASR)

As a connected research on ASR using subword units and / or approximated language models (low footprint models approximating high footprint performance), and especially as we also experimented with subword embeddings earlier, we took part in experiments focusing on ASR only organized by Balázs Tarján and related to his project on language modelling for ASR [9]. We also found this relevant as the ASR errors may be of different nature, hence cause different bias in summaries.

Abstractive summarization

At the time of planning the project, abstractive summarization was of poor performance. With the emergence of BERT however, abstractive summarization became a reality. These models are very data hungry, therefore often trained through transfer learning, i. e. a generic model is trained and made available, which is fine-tuned for a specific task (summarization in our case) in a specific domain (news) in a specific language (Hungarian). Unfortunately the framework of such models does not allow easy exploitation and integration of audio, moreover, when adding audio information to them with shallow fusion, no improvement was observed.

In recent years, abstract content extraction has undergone huge development at the international level, and it has become possible to generate extracts that reach or exceed the quality of extractive summaries. Due to the high data and resource requirements of the experiments, we established a close cooperation with ELTE's Department of Digital Arts (data) and SZTAKI (servers) and built on the Transformer-based models (BERT, HUBERT) which got also published in Hungarian in the meantime. In our work, we examined the extraction of news data, for which we used articles on various topics available from popular Internet news portals from previous years. As an extract, we considered the leads of the articles as a reference, since it is not realistic to produce the reference with tens of thousands of human annotations, and human references can also contain distorting, subjective elements. We solved the separation of the training and test sets in time to avoid news known from different news portals being included in both sets. The HUBERT model available in Hungarian was fine-tuned, optimized, and then applied to generate abstract summaries. The obtained results can be compared with the state-of-the-art performance available in other languages, although the performance values described for the English language are not reached - however, this would not be a realistic goal due to the different organization of the languages and the amount of available data. We mainly used the ROUGE metrics for evaluation [10].

The spoken documents are first converted using ASR, then automatic punctuation restoration is performed. When using speech recognition, compared to the written text, we have to expect that word and punctuation errors will appear in the text. We examined the impact of these errors as in the case of extractive summarization. In the extractive case propagation was detectable for both types of errors, which made the end performance somewhat worse compared to pure text analysis. The errors typically affected less important terms from the point of view of extracting, presumably because they are more carefully articulated in the document, as they carry important information. The punctuation errors sometimes caused a sentence boundary shift, so in the extractive case, the ranking of the sentences also changed, which typically caused a small but detectable performance deterioration in the ROUGE metrics. In abstract summarization, neither word errors nor punctuation errors cause a significant difference in the quality of the summary. The latter is understandable, since punctuation is often not included in the sequence-to-sequence model, so this information can also be omitted. With regard to word errors, we think that they would probably degrade the quality of the summary, if we could generate better summaries with more accurate decoding than is currently possible. However, the impact of word errors could not be measured at the current performance level. We supposed that the language model of the ASR has a smoother effect on the text input to the summarization module, so that it better fits the input of the encoder, which can lead to improved performance.

The abstractive summarization module ‘HunSumm’ was publicly made available [10].

References

[1] Valér, Kaszás ; Máté, Ákos Tündik ; György, Szaszák: ***A semantic space approach for automatic summarization of documents***, In: Sallai, Gyula (szerk.) 9th IEEE International Conference on Cognitive Infocommunications, pp. 153-158., 2018

[2] György, Szaszák ; Máté, Ákos Tündik ; Branislav, Gerazov: ***Prosodic stress detection for fixed stress languages using formal atom decomposition and a statistical hidden Markov hybrid***, SPEECH COMMUNICATION 102 pp. 14-26., 2018

[3] Máté, Ákos Tündik ; György, Szaszák ; Gábor, Gosztolya ; András, Beke: ***User-centric Evaluation of Automatic Punctuation in ASR Closed Captioning***, Proc. Interspeech, pp. 2628-2632., 2018

[4] Máté Ákos Tündik, Balázs Tarján, György Szaszák: ***A low latency sequential model and its user -focused evaluation for automatic punctuation of ASR closed captions***, Computer Speech & Language 63 (2020): 101076., 2020

[5] Döbrössy Bálint, Makrai Márton, Tarján Balázs, Szaszák György: ***Investigating Sub-Word Embedding Strategies for the Morphologically Rich and Free Phrase-Order Hungarian***, In: Isabelle, Augenstein; Spandana, Gella; Sebastian, Ruder; Katharina, Kann; Burcu, Can; Johannes, Welbl; Alexis, Conneau; Xiang, Ren; Marek, Rei (szerk.) Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019), Association for Computational Linguistics (2019) pp. 187-193., 2019

[6] Gábor Gosztolya: ***Using the Fisher Vector Representation for Audio-based Emotion Recognition***, Acta Polytechnica Hungarica, Vol. 17, No. 6, pp. 7-23, 2020

[7] Tündik Máté Ákos, Kaszás Valér, Szaszák György: ***Assessing the Semantic Space Bias Caused by ASR Error Propagation and its Effect on Spoken Document Summarization***, In: Gernot, Kubin; Zdravko, Kačič (szerk.) The 20th Annual Conference of the International Speech Communication Association (2019) pp. 1333-1337., 2019

[8] Egas-López, J.V., Kiss, G., Sztahó, D., Gosztolya, G.: ***Automatic Assessment of the Degree of Clinical Depression from Speech Using X-Vectors***, Proceedings of ICASSP, pp. 8502-8506, Singapore, 2022, 2022

[9 Balázs Tarján, György Szaszák, Tibr Fegyó, Péter Mihajlik: ***N-gram Approximation of LSTM RecurrentLanguage Models for Single-pass Recognition ofHungarian Call Center Conversations***, CoginfoCom 2019, 2019]

[10] Márton Makrai, Ákos Máté Tündik, Balázs Indig, György Szaszák: ***Towards abstractive summarization in Hungarian***, Berend Gábor. XVIII. Magyar Számítógépes Nyelvészeti Konferencia : MSZNY 2022. (2022) ISBN:9789633068489 pp. 505-519, 2022