Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 1032–1039
Marseille, 20-25 June 2022

Towards an Open-Source Dutch Speech Recognition System for the Healthcare Domain

Cristian Tejedor-García[*], Berrie van der Molen[†], Henk van den Heuvel[*], Arjan van Hessen[‡], Toine Pieters[†]
[*]CLS/CLST, Radboud University, Nijmegen, the Netherlands
{cristian.tejedorgarcia, henk.vandenheuvel}@ru.nl
[†]Freudenthal Institute, Utrecht University, Utrecht, the Netherlands
{b.j.vandermolen, t.pieters}@uu.nl
[‡]HMI, University of Twente, Enschede, the Netherlands
a.j.vanhessen@utwente.nl

Abstract

The current largest open-source generic automatic speech recognition (ASR) system for Dutch, Kaldi_NL, does not include a domain-specific healthcare jargon in the lexicon. Commercial alternatives (e.g., Google ASR system) are also not suitable for this purpose, not only because of the lexicon issue, but they do not safeguard privacy of sensitive data sufficiently and reliably. These reasons motivate that just a small amount of medical staff employs speech technology in the Netherlands. This paper proposes an innovative ASR training method developed within the Homo Medicinalis (HoMed) project. On the semantic level it specifically targets automatic transcription of doctor-patient consultation recordings with a focus on the use of medicines. In the first stage of HoMed, the Kaldi_NL language model (LM) is fine-tuned with lists of Dutch medical terms and transcriptions of Dutch online healthcare news bulletins. Despite the acoustic challenges and linguistic complexity of the domain, we reduced the word error rate (WER) by 5.2%. The proposed method could be employed for ASR domain adaptation to other domains with sensitive and special category data. These promising results allow us to apply this methodology on highly sensitive audiovisual recordings of patient consultations at the Netherlands Institute for Health Services Research (Nivel).

Keywords: speech recognition, language modeling, domain adaptation, healthcare

1. Introduction

1.1. Background

Every year we encounter more than 15,000 hospital admissions due to avoidable misuse of medicines in the Netherlands (Erasmus MC, Nivel, Radboud UMC, PHARMO, 2017). Often, this has to do with the patient's unintentional improper use or low levels of adherence. Practical limitations and barriers are in the foreground here, over which the patient has insufficient control (van Dijk et al., 2016). For instance, they are not good at reading, understanding (functionally illiterate), reproducing (forgetfulness, real cognition), and/or accurately performing the prescribed usage. This might result in rather diverse and inappropriate forms of use, low levels of adherence and waste of scarce financial resources (Maas et al., 2020).

In order to overcome these misunderstandings, we need to better understand the explicit and implicit attribution of meaning to medicines as part of the information processing. For that reason, effective and efficient transcriptions of doctor-patient interviews are indispensable. This, however, is a context with considerable privacy-sensitive constraints.

Research on and use of sensitive data involving audio/video (AV) recordings requires an infrastructure where both the data and the research environment are optimal in terms of General Data Protection Regulation (GDPR) safeguards. In the HoMed project[1] we have set out to establish a method and an infrastructure that has great potential for automatic transcription of AV-recordings at large with a special feature for employment in domains where sensitive AV-material needs to be transcribed and analyzed.

HoMed will develop an infrastructure in which an existing generic Dutch automatic speech recognition (ASR) system, named Kaldi_NL[2], used in the CLARIAH (Media Suite) infrastructure[3], is adapted to the medical/pharmaceutical domain on the semantic level, using a domain adaptation component (language model, LM). In the second stage, the ASR system will be adapted on both the semantic and the acoustic level using sensitive in-house data of the Netherlands Institute for Health Services Research (Nivel), whereby the AV-recordings themselves will not leave the Nivel building. The resulting ASR component will be made available at Nivel and in the CLARIAH (Media Suite) infrastructure.

1.2. Speech Recognition in Healthcare

Most doctors and nurses spend an estimated 30% of the working week on writing or typing patient notes and re-

[1] `https://homed.ruhosting.nl/`

[2] `https://github.com/opensource-spraakherkenning-nl/Kaldi_NL`

[3] `https://www.clariah.nl/` `https://mediasuite.clariah.nl/`

ports, which reduces their productivity (Luchies. et al., 2018; Friedberg et al., 2014). Current ASR technology constitutes a useful resource and offers a huge potential for revolutionizing the healthcare domain (Latif et al., 2021). Thus, through ASR technology, medical staff can save time, spent on data-entry by voice-typing their documentation on the go, and focus on patients more (Payne et al., 2018; Walker et al., 2011). Besides, when creating a medical record with the help of ASR, the possibility that important information is omitted during a patient visit is minimized (David et al., 2009; Fratzke et al., 2014). Finally, ASR-based dialogue systems or information retrieval systems, can increase the level of patient engagement in their treatment process outside the care facility (e.g., asking new questions or searching for more information after reading the complete consultation with the practitioner) while being monitored (Debnath and Roy, 2019; Hossain, 2016), or interacting with virtual avatars (O'Connor, 2019).

Nowadays, only 1% of the medical staff uses speech technology in the Netherlands (Luchies. et al., 2018), and in most cases, commercial ASR engines, such as Google ASR system, are not suitable for this infrastructure, since they neither contain the domain-specific jargon nor safeguard the privacy of sensitive data sufficiently and reliably (Vipperla et al., 2020). Besides, the current largest open-source generic ASR for Dutch, Kaldi_NL, does not include medical terms in its vocabulary set, which hampers the potential use of this ASR system in the Dutch healthcare environment. The aim of the present study is to discuss the results obtained after fine-tuning the LM of Kaldi_NL with healthcare-related Dutch words, as part of the HoMed project. In particular, we report on the word error rate (WER) improvement obtained after following a specific ASR training method in HoMed and the text error analysis comparison between Kaldi_NL and our proposed ASR system's output.

2. Related Work

Current publicly available and commercial generic ASR systems, which are trained on large amounts of speech data, perform well on their generic training domain. However, when applying ASR to domain-specific tasks, such as healthcare, the performance decreases substantially. It is reasonable to expect that better performing LMs for domain-specific tasks which include out-of-vocabulary (OOV) or infrequent words will result in better performing systems for these particular tasks (Xu et al., 2018) and ASR algorithms and clinical vocabulary will improve in the future, so that natural languages can be understood by ASR systems (Ajami, 2016).

Personalizing or fine-tuning LMs is a domain adaptation problem in which a LM that is trained on a large background corpus is adapted to text from a specific domain. In the literature, this technique has been applied to a variety of domains, such as adapting text from medical consultations (Renato et al., 2019; Liu et al., 2011) and automatically extract clinical meaning (Klann and Szolovits, 2009; Rajkomar et al., 2019). Renato et al. (2019) reduced the WER by 2.4% when using their fine-tuned ASR system for taking Spanish clinical notes in a mobile environment. Liu et al. (2011) evaluated the Generic and Medical version of the Nuance Dragon ASR system and the SRI Decipher system on spoken clinical questions, obtaining a WER of 68.1%, 67.4% and 41.5%, respectively. After fine-tuning the three ASR systems, the only system that improved significantly was the SRI one, with a WER of 26.7%. Klann and Szolovits (2009) reported on a WER of 26.4% of a simple proof-of-concept English ASR system for doctor-patient conversations. Rajkomar et al. (2019) reported on a sensitivity of 67.7% to identify symptoms and 80.6% to positive predict them with an English ASR system and machine learning in doctor-patient conversations.

Language modeling has also been applied for adapting text from news and the press (Mikolov et al., 2010), aspect-based sentiment analysis (Howard and Ruder, 2018; Rietzler et al., 2019), children books (Hill et al., 2016), and Wikipedia (Merity et al., 2016). Lin et al. (2017) allowed the LM that was trained on the source domain to continue training on the target domain to form personalized word vectors using specific tokens per user for user prediction and sentence completion tasks. Domain adaptation has been also applied for adapting one domain to another, such as adapting between English newspaper sections (Jaech and Ostendorf, 2018) or YouTube video topics (Irie et al., 2018).

Scarce research on language modeling and domain adaptation has been carried out in ASR for Dutch, even less in the healthcare domain. Maas et al. (2020) reported their vision for automated medical reporting of doctor-patient consultations in Dutch and the initial development of their state-of-the-art system for dialogue transcriptions using Google ASR system and its privacy limitations. Van der Klis et al. (2020) assessed the performance of an ASR system at extracting target words from infant-directed speech and adult-directed speech by using Kaldi_NL. The results revealed the necessity to build customized LMs for children speech due to the low level accuracy results obtained by using the generic vocabulary in the generic ASR system. Neerincx and Luijk (2020) claimed future efforts at implementing social-speaking robots in healthcare in order to avoid healthcare professionals' skepticism due to lack of knowledge about what robots could offer.

3. Method

3.1. Dutch Healthcare-Related AV-Data

3.1.1. Context and Problem Setting

The development of an ASR system for doctor-patient consultation recordings in Dutch is challenging due to two main factors. First, there is no established collection of medical terms specific to such audio recordings. Therefore, such a collection needs to be established. Second, the effectiveness of the ASR system in comparison with the effectiveness of general ASR systems needs to be tested, and therefore test material is required. The latter challenge is especially relevant given the privacy-sensitive nature of actual doctor-patient consultation recordings, which means such consultation recordings come with many use restrictions. In this section we describe how we addressed these challenges within the first stage of HoMed.

3.1.2. Data Collection

Doctor-patient consultation recordings have a number of specific properties: 1) they contain conversational speech between patients and medical experts; 2) the acoustic quality of the material varies greatly and is often low; 3) specific medical terms and medicines are mentioned in a particular way; 4) they are unscripted and relatively unstructured. Due to the many challenges in terms of privacy of this material, it was not yet possible to use it as test material: test transcripts would need to be shared freely between the different researchers and institutional settings of this research context.

Alternatively, we employed two other sources of Dutch healthcare data as training and testing material for our new ASR system. First, for the expansion of the standard lexicon with medical terms, we collected (licensed) lists of Dutch isolated medical terms from the following institutions: Dutch College of General Practitioners, Dutch Language Institute, Health Base, International Council for Harmonization of Technical Requirements for Pharmaceuticals for Human Use, Medicines Evaluation Board, National Health Care Institute, and Nictiz. These institutions provided us with one or more documents with Dutch medical terms. Many of the lists we received contained hierarchical and defining relationships between the included terms. These relations are seen as intellectual property of the different institutions, and in most of the license agreements we agreed that we would only extract the unique terms, not using the relationships between the terms.

Secondly, we used Medicijnjournaal[4] (MJ), the online periodic medicine news bulletin created by the Dutch Institute for Rational Use of Medicine (IVM). These around 10-minute online news bulletins consist of studio recordings, video footage with voice-over, and interviews with specialists and experts. Episodes generally contain around four longer items (approximately 2 minutes) and three to four short items (half a minute) and cover new medicines, recent medical research results, news on side effects etc. The bulletins are mostly scripted and primarily aimed at healthcare professionals. We transcribed 35 episodes of MJ (360 minutes) dated from 2017 to 2021 (Radboud University, 2022). When the MJ bulletins are compared to doctor-patient consultation recordings, the following differences are relevant: 1) MJ contains no conversational speech but merely scripted text or interview answers; 2) the acoustic quality of MJ is generally very good; and 3) there is a high frequency of mentioned medical terms, albeit not in the context of a doctor-patient dynamic. Although training the LM with these audio recordings enables us to work with a more domain-specific lexicon, we will only be able to test other properties specific to doctor-patient consultation recordings (the setting of the consultation room) when we use that material for training in the next stage of HoMed.

3.1.3. Data Preparation

Four native Dutch speakers manually corrected automatic speech transcripts of the MJ episodes. They employed the web service *Automatic Speech Transcription of Dutch Speech Recordings* [5] to create initial automatic speech transcripts, segmented in speaker turns for each MJ episode. The transcripts were manually corrected using the tool *ASRcorrector* using version 1.0 of the HoMed transcription protocol (see Section 3.1.4). The IVM provided us with the original production scripts of the news bulletins for this task. Each corrected transcript was exported as a text file and consequently processed with forced alignment in the *ForcedAlignment2* webservice[6].

Regarding the licensed lists of medical terms, we worked in a two phase selection process, creating two first versions of a medical Dutch lexicon: the full version and the downsized version. Compiling the extracted words of these relevant lists we received resulted in a full lexicon of approximately 400,000 words (tokens). In this full lexicon, we included all extracted unique words from lists actually containing medical terms. We noted that the very large amount of highly technical terms, such as pharmaceutical components and many variations of similar terms specific to this medical jargon, creates a potential challenge for ASR, as this could paradoxically reduce the chance that correct terms will eventually be recognized due to higher confusability. This might mean that a general ASR system for the Dutch medical domain seemed to be a bridge too far at this point: one can imagine that recordings of internal conversations between medical specialists, for instance, likely include more frequent use of specific technical medical terms. We therefore needed to make an informed selection of terms. The unique words from this second, informed selection of lists became our downsized lexicon. In this selection process the qualitative judgment was made on the basis of the likelihood that medical terms are used in doctor-patient consultations, as we work towards an ASR system of this specific type of medical discourse.

[4] https://www.medicijngebruik.nl/fto-voorbereiding/medicijnjournaals

[5] https://webservices.cls.ru.nl/oralhistory

[6] https://webservices.cls.ru.nl/forcedalignment2/

3.1.4. Transcription Protocol

For the transcription of the MJ material, we adapted the orthographic transcription protocol that was developed to transcribe automatic speech transcripts of lectures given in the context of the Universiteit van Nederland, an online platform on which prominent Dutch scientists give free lectures[7]. This protocol was largely based on the rules developed as part of the Corpus Gesproken Nederlands (CGN) project (Nederlandse Taalunie, 2004), in which around 900 hours of contemporary spoken Dutch were recorded and transcribed (Goedertier et al., 2000; Oostdijk et al., 2002). It is a database of Dutch as spoken by adults from the Netherlands and Flanders. The goal of this project was to create a corpus that would form a plausible sample of contemporary Dutch as spoken in the Netherlands and Flanders. In creating the corpus, the developers have attempted to assemble it to optimally suit the needs of diverse research disciplines and applications. Everything the speakers say is literally transcribed as long as it can be transcribed using existing Dutch words or deviations that are common in written Dutch (e.g., ”*’ns*” as short for ”*eens*” is accepted).

Transcribing the MJ bulletins highlighted a number of issues specific to medical terminology. First of all, medicines are often not pronounced fluently/correctly. The existing protocol demands that only existing terms are used in the transcription, which means the correct term should be noted in the transcript. Transcribing a mispronunciation would make it impossible to connect the actually intended medical term to the point where it was discussed in the news bulletin. Another issue we came across is the use of digits in medical terms, which is common in written Dutch. The protocol in general asks the transcriber to write out any number (e.g., the number 14 is transcribed as ”*veertien*”), but using unconventional spelling to denote medical terms that are already complex, would affect the searchability of the eventual transcript. Anyone searching medical recordings for ”*SGLT2-remmers*” (SGLT2 inhibitors), for instance, would need to know that they need to write ”*SGLT-twee-remmers*”. We therefore opted to create an exception for the transcription of digits in medical terms, where we leave the digits when they are used in written Dutch: ”*SGLT2-remmers*”. A last significant issue with transcribing the bulletins is that for many terms there are several common spellings. This can be illustrated with the same term: one might come across ”*SGLT-2-remmers*” and ”*SGLT2-remmers*” in written Dutch. We expanded the protocol accordingly, working towards a tailor-made protocol for transcribing Dutch medical domain audio recordings. In order to measure the inter-rater reliability between the four transcribers after following the protocol, we calculated the Cohen's Kappa score (Landis and Koch, 1977) and obtained an almost perfect inter-agreement (97%).

3.2. Speech Recognition Setup

Kaldi is one of the most well-known open-source and free toolkits for ASR research (Povey et al., 2011). This state-of-the-art ASR framework is currently utilized by most research groups in almost any language, including Dutch. One of its main features is that it is modular, i.e., it combines a LM, an acoustic model (AM), a lexicon and a word search algorithm (decoder). The LM is typically based on statistical n-grams. They consist in word sequences with some probabilities, taking into account the context of the words in sentences. The AM covers the different acoustic speech sounds in a particular language and associates these sounds with phonetic representations. The lexicon or dictionary relates the LM and the AM by mapping the word symbols to their respective pronunciations. Words that do not appear in the lexicon vocabulary (OOV words) cannot be recognized. Finally, when decoding, the most likely sequence of words according to the LM, AM and lexicon models are searched for. Nevertheless, searching for all the possible sequences sounds astonishingly inefficient. Even when the n-gram LM is huge, the amount of memory required to store the final models may be too large. Fortunately, we can use, for instance, the Viterbi algorithm, to find the best path in polynomial time (Arsadjaja and Kistijantoro, 2018).

Kaldi_NL is an open-source ASR project that integrates the CGN in a Kaldi ASR environment and reports WER values of 7% for daily conversation in Dutch. However, this general-purpose ASR system does not perform well when decoding domain-specific healthcare words. In this manuscript, we report on the results obtained after fine-tuning the LM and lexicon of Kaldi_NL including domain-specific healthcare words, as part of the initial proof-of-concept stage of HoMed.

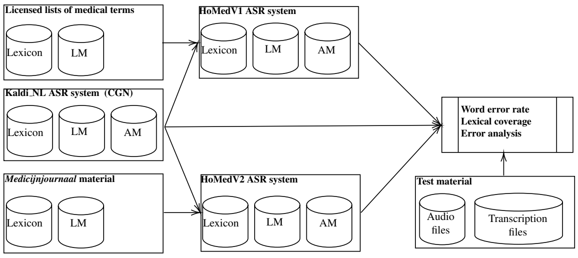

Figure 1: Steps of the HoMed ASR system development and evaluation.

Figure 1 shows the pipeline of the new HoMed ASR system development and evaluation. From left to right, we fine-tuned the CGN models (integrated in Kaldi_NL ASR system) by adapting the LM and lexicon with either just the licensed lists of medical terms (HoMedV1 ASR system) or the MJ material (HoMedV2 ASR system). Then, we carried out an ASR evaluation performance by comparing the decoding results offered by the generic Kaldi_NL ASR system with those obtained by our own ASR configuration.

[7] https://www.universiteitvannederland.nl

In particular, the HoMedV1 and HoMedV2 ASR systems AM was trained under a time-delay neural network (tDNN) (Georgescu et al., 2019) using CGN's speech data for Dutch[8], as in Kaldi_NL. We also explored different parameters for the LM training and fine-tuning in our experiments. The best configuration was to merge the generic 4-gram CGN's LM with a domain-specific 3-gram LM extracted from the training data by using interpolated Kneser-Ney smoothing (Heafield et al., 2013) and an interpolation weight of 0.25. The Dutch phonetic transcription of the words for the new lexicon was obtained by using a grapheme-to-phoneme (G2P) tool[9].

Three ASR systems are evaluated in this study. That is, the generic ASR system for Dutch, Kaldi_NL, and two fine-tuned versions of Kaldi_NL, HoMedV1 (with lists of Dutch isolated medical terms) and HoMedV2 (with MJ material) ASR systems. The Kaldi_NL lexicon contains approximately 255,000 tokens in the vocabulary, whereas HoMedV1 and HoMedV2 extend this lexicon by 13,934 and 5,342 words, respectively. Both fine-tuned ASR systems interpolate the Kaldi_NL 4-gram LM with a personalized 3-gram LM.

For testing the ASR systems, we employed the whole MJ AV material (360 minutes of speech recordings and 47,912 transcribed words in 35 transcribed bulletins) for Kaldi_NL and HoMedV1 ASR systems. However, for the evaluation of the HoMedV2 ASR system we employed a 12-fold cross validation scheme (Bengio and Grandvalet, 2004) to avoid overlapping between the train and test datasets of the MJ material. We opted to split the MJ dataset into 90%-10% for training and testing. In each one of the 12 experimental runs we tested 3 different bulletins (2 in the last test set) and the remaining bulletins were included in the fined-tuned LM.

3.3. Instruments and Evaluation Metrics

We employed two different instruments and metrics to evaluate the ASR system performance. First, we used *sclite*[10], a tool for scoring and evaluating the output of ASR systems, to obtain the WER, a measure of how accurate an ASR system performs. It is calculated by dividing the sum of word deletions, insertions and substitutions by the total number of words in the transcription. This value is expressed as a percentage. To standardize the texts, we changed the capitalization of text to lowercase, removed all punctuation and changed all numbers to their written form when comparing ASR output and reference transcriptions. And secondly, we calculated the lexical coverage (LC) by counting the number of tokens in the testing transcriptions that are included in the lexicon of the ASR system, by running a Python script[11]. This value is also expressed as a percentage. LC relates to WER in terms of the minimum WER possible value.

For the text error analysis between ASR output and reference transcriptions, we classified the words obtained in the final *sclite* report into five categories: 1. Regular spelling variant; 2. Compound word for which a white space was missed or inserted; 3. Morphological error (singular vs. plural in noun or verb tense); 4. Error within lexicon (the correct word was in lexicon but not recognized); and 5. OOV word.

4. Experimental Results

4.1. ASR System Performance

Table 1 reports on the ASR performance improvement achieved by the HoMedV2 ASR system over the other alternatives by testing the MJ material. In particular, our proposed HoMedV2 ASR system outperforms the generic Kaldi_NL ASR system substantially, on average, by 5.2% lower WER value, and the HoMedV1 ASR system by 4.1% (Kaldi_NL: WER = 25.8%, standard deviation (SD) = 4.0; HoMedV1: WER = 24.7%, SD = 3.1; HoMedV2: WER = 20.6%, SD = 3.0).

ASR system	WER	LC
Kaldi_NL	25.8	94.9
HoMedV1	24.7	96.1
HoMedV2	20.6	97.2

Table 1: Comparison of the ASR systems performance

By fine-tuning the LM with healthcare-related words, we also increased the LC, therefore, reduced the OVV rate, on average, by 2.3%, comparing HoMedV2 vs. Kaldi_NL ASR systems, and 1.1%, comparing HoMedV2 vs. HoMedV1 ASR systems (Kaldi_NL: LC = 94.9%, SD = 1.2; HoMedV1: LC = 96.1%, SD = 0.9; HoMedV2: LC = 97.2%, SD = 0.8). These results corroborate the importance of not only adding domain-related words to the lexicon, but the appropriate elaboration of the LM which such words.

4.2. Error Analysis

In order to obtain a better understanding of the remaining errors we looked at the most frequent confusions between the ASR output and reference transcriptions of our best HoMed ASR system (V2) and the generic Kaldi_NL one. For this, we looked at the confusions with a frequency of 5 and higher, and classified them into the five categories described in Section 3.3. The results are shown in Table 2.

Errors in categories 1-3 (Table 2) have no semantic effect on the resulting transcript. Disregarding these would lead to a WER of 24.5% for the Kaldi_NL ASR system and to a WER of 18.8% for the HoMedV2 ASR system. A second observation is that the category 5

[8] `https://github.com/cristiantg/kaldi_egs_CGN/tree/onPonyLand`

[9] `https://github.com/cristiantg/lexiconator`

[10] `https://github.com/usnistgov/SCTK`

[11] `https://github.com/cristiantg/ctmator`

Type of error	Kaldi_NL	HoMedV2
1. Spelling variant	457	798
2. Compound word	158	19
3. Morphological variant	21	31
4. Error within lexicon	598	872
5. OOV	286	78

Table 2: Categorization of main ASR output confusions

errors (OOV words) are substantially higher for the Kaldi_NL ASR system than for the HoMedV2 ASR system, (directly related to the LC values reported on Table 1), showing that the HoMedV2 ASR system is better suited for the medical domain, as intended. This is corroborated by the fact that the HoMedV2 ASR system deals much better with the compound words in the test material. As a final observation we note that the errors within the lexicon (category 4) have substantially increased for the HoMedV2 ASR system, demonstrating further room for improvement.

5. Conclusion

We have demonstrated a substantial improvement in recognition performance after fine-tuning a generic ASR for Dutch, Kaldi_NL, by expanding the lexicon with a limited generic set of healthcare medical terms and adapting the LM with these additional words and with limited transcribed material from the MJ. In principle, this method is domain-independent, enabling implementation in domains other than healthcare, e.g., for police reports, court recordings or the Dutch Immigration and Naturalization Service.

We can conclude that the resulting HoMedV2 ASR system WER scores substantially better (improvement of 5.2%) than the generic ASR system (Kaldi_NL) in reducing the errors for OOV words and compounds, and also than the other alternative, HoMedV1 ASR system (improvement of 4.1%), with isolated Dutch medical terms. This indicates the importance of contextual information for ASR in the healthcare domain. The resulting WER of around 20% is sufficient to use the recognition output and time stamps for document retrieval based on keyword spotting, topic modeling and sentiment mining, but the transcriptions are not yet of sufficient quality for subtitling purposes or detailed text analyses.

In our follow-up work we will focus on real doctor-patient consultation recordings, by making transcriptions of said recordings held at Nivel. In this process we will work on an updated version of the HoMed transcription protocol. The transcripts will again be used to expand and fine-tune the lexicon and the LM but also to train new AMs.

Finally, we will share some of the resources that we developed for building the HoMedV1 and HoMedV2 ASR systems. These include the manual transcriptions of the MJ bulletins.

6. Acknowledgements

The research funding for this study was supported by the Platform Digitale Infrastructuur Social Science and Humanities (PDI-SSH) 2020 (`https://pdi-ssh.nl/`). The authors would also like to thank Sandra van Dulmen for her valuable feedback.

7. Bibliographical References

Ajami, S. (2016). Use of speech-to-text technology for documentation by healthcare providers. *The National medical journal of India*, 29(3):148–152.

Arsadjaja, A. R. and Kistijantoro, A. I. (2018). Online Speech Decoding Optimization Strategy with Viterbi Algorithm on GPU. In *2018 5th International Conference on Advanced Informatics: Concept Theory and Applications (ICAICTA)*, pages 130–134.

Bengio, Y. and Grandvalet, Y. (2004). No Unbiased Estimator of the Variance of K-Fold Cross-Validation. *Journal of Machine Learning Resources*, 5:1089–1105.

David, G. C., Garcia, A. C., Rawls, A. W., and Chand, D. (2009). Listening to what is said – transcribing what is heard: the impact of speech recognition technology (SRT) on the practice of medical transcription (MT). *Sociology of Health & Illness*, 31(6):924–938.

Debnath, S. and Roy, P. (2019). Study of speech enabled healthcare technology. *International Journal of Medical Engineering and Informatics*, 11(1):71–85.

Erasmus MC, Nivel, Radboud UMC, PHARMO. (2017). Vervolgonderzoek medicatieveiligheid: eindrapport. Technical report, Rotterdam/Utrecht/Nijmegen. `https://www.nivel.nl/nl/publicatie/vervolgonderzoek-medicatieveiligheid-eindrapport`.

Fratzke, J., Tucker, S., Shedenhelm, H., Arnold, J., Belda, T., and Petera, M. (2014). Enhancing nursing practice by utilizing voice recognition for direct documentation. *JONA: The Journal of Nursing Administration*, 44(2):79–86.

Friedberg, M. W., Chen, P. G., Van Busum, K. R., Aunon, F., Pham, C., Caloyeras, J., Mattke, S., Pitchforth, E., Quigley, D. D., Brook, R. H., et al. (2014). Factors affecting physician professional satisfaction and their implications for patient care, health systems, and health policy. *Rand health quarterly*, 3(4).

Georgescu, A.-L., Cucu, H., and Burileanu, C. (2019). Kaldi-based DNN Architectures for Speech Recognition in Romanian. In *2019 International Conference on Speech Technology and Human-Computer Dialogue (SpeD)*, pages 1–6.

Goedertier, W., Goddijn, S., and Martens, J.-P. (2000). Orthographic transcription of the spoken Dutch corpus. In *Proceedings of the Second International Conference on Language Resources and Evaluation*

(LREC'00), Athens, Greece. European Language Resources Association (ELRA).

Heafield, K., Pouzyrevsky, I., Clark, J. H., and Koehn, P. (2013). Scalable modified Kneser-Ney language model estimation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, volume 2, pages 690–696.

Hill, F., Bordes, A., Chopra, S., and Weston, J. (2016). The Goldilocks Principle: Reading Children's Books with Explicit Memory Representations.

Hossain, M. S. (2016). Patient status monitoring for smart home healthcare. In *2016 IEEE International Conference on Multimedia Expo Workshops (ICMEW)*, pages 1–6.

Howard, J. and Ruder, S. (2018). Universal Language Model Fine-tuning for Text Classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339, Melbourne, Australia, July. Association for Computational Linguistics.

Irie, K., Kumar, S., Nirschl, M., and Liao, H. (2018). RADMM: Recurrent Adaptive Mixture Model with Applications to Domain Robust Language Modeling. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6079–6083.

Jaech, A. and Ostendorf, M. (2018). Personalized Language Model for Query Auto-Completion. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, volume 2, pages 700–705, Melbourne, Australia. Association for Computational Linguistics.

Klann, J. G. and Szolovits, P. (2009). An intelligent listening framework for capturing encounter notes from a doctor-patient dialog. *BMC medical informatics and decision making*, 9(1):1–10.

Landis, J. R. and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Latif, S., Qadir, J., Qayyum, A., Usama, M., and Younis, S. (2021). Speech Technology for Healthcare: Opportunities, Challenges, and State of the Art. *IEEE Reviews in Biomedical Engineering*, 14:342–356.

Lin, Z.-W., Sung, T.-W., Lee, H.-Y., and Lee, L.-S. (2017). Personalized word representations carrying personalized semantics learned from social network posts.

Liu, F., Tur, G., Hakkani-Tür, D., and Yu, H. (2011). Towards spoken clinical-question answering: evaluating and adapting automatic speech-recognition systems for spoken clinical questions. *Journal of the American Medical Informatics Association*, 18(5):625–630.

Luchies., E., Spruit., M., and Askari., M. (2018). Speech Technology in Dutch Health Care: A Qualitative Study. In *Proceedings of the 11th International Joint Conference on Biomedical Engineering Systems and Technologies (HEALTHINF)*, pages 339–348. INSTICC, SciTePress.

Maas, L., Geurtsen, M., Nouwt, F., Schouten, S., Water, R., Dulmen, A., Dalpiaz, F., Deemter, K., and Brinkkemper, S. (2020). The Care2Report System: Automated Medical Reporting as an Integrated Solution to Reduce Administrative Burden in Healthcare. In *IT Architectures and Implementations in Healthcare Environments*, pages 3608–3617.

Merity, S., Xiong, C., Bradbury, J., and Socher, R. (2016). Pointer Sentinel Mixture Models. *CoRR*, abs/1609.07843.

Mikolov, T., Karafiát, M., Burget, L., Černocký, J., and Khudanpur, S. (2010). Recurrent neural network based language model. In *Proc. Interspeech 2010*, pages 1045–1048.

Neerincx, A. and Luijk, A. (2020). Social Robot's Processing of Context-Sensitive Emotions in Child Care: A Dutch Use Case. In *Companion Publication of the 2020 International Conference on Multimodal Interaction*, ICMI '20 Companion, pages 503–505, New York, NY, USA. Association for Computing Machinery.

O'Connor, S. (2019). Virtual Reality and Avatars in Health care. *Clinical Nursing Research*, 28(5):523–528. PMID: 31064283.

Oostdijk, N., Goedertier, W., Eynde, F. v., Boves, L., Martens, J.-P., Moortgat, M., and Baayen, R. H. (2002). Experiences from the spoken Dutch corpus project. In *Proc. 3th Int. Conf. Lang. Resour. Eval. (LREC)*, pages 340–347, Las Palmas de Gran Canaria, Spain.

Payne, T. H., Alonso, W. D., Markiel, J. A., Lybarger, K., and White, A. A. (2018). Using voice to create hospital progress notes: Description of a mobile application and supporting system integrated with a commercial electronic health record. *Journal of Biomedical Informatics*, 77:91–96.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlíček, P., Qian, Y., Schwarz, P., Silovský, J., Stemmer, G., and Vesel, K. (2011). The Kaldi speech recognition toolkit. In *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*, pages 1–4, Waikoloa, Hawaii, HI, USA, Dec. 11–15.

Rajkomar, A., Kannan, A., Chen, K., Vardoulakis, L., Chou, K., Cui, C., and Dean, J. (2019). Automatically Charting Symptoms From Patient-Physician Conversations Using Machine Learning. *JAMA Internal Medicine*, 179(6):836–838.

Renato, A., Berinsky, H., Daus, M., Dachery, M. F., Jáuregui, O. I., Storani, F. D., Gambarte, M. L., Otero, C., and Luna, D. R. (2019). Design and Evaluation of an Automatic Speech Recognition Model for Clinical Notes in Spanish in a Mobile Online Environment. In *MedInfo*, pages 1761–1762.

Rietzler, A., Stabinger, S., Opitz, P., and Engl, S.

(2019). Adapt or Get Left Behind: Domain Adaptation through BERT Language Model Finetuning for Aspect-Target Sentiment Classification.

Van der Klis, A., Adriaans, F., Han, M., and Kager, R. (2020). Automatic Recognition of Target Words in Infant-Directed Speech. In *Companion Publication of the 2020 International Conference on Multimodal Interaction*, ICMI '20 Companion, page 522, New York, NY, USA. Association for Computing Machinery.

van Dijk, L., Hendriks, M., Zwikker, H., de Jong, J., and Vervloet, M. (2016). Informatiebehoeften van patiënten over geneesmiddelen. *Utrecht, NIVEL.*

Vipperla, R., Ishtiaq, S., Li, R., Bhattacharya, S., Leontiadis, I., and Lane, N. D. (2020). Learning to Listen... On-Device: Present and Future Perspectives of on-Device ASR. *GetMobile: Mobile Comp. and Comm.*, 23(4):5–9.

Walker, N. R., Trofimovich, P., Cedergren, H., and Gatbonton, E. (2011). Using ASR Technology in Language Training for Specific Purposes: A Perspective from Quebec, Canada. *CALICO Journal*, 28(3):721–743.

Xu, H., Li, K., Wang, Y., Wang, J., Kang, S., Chen, X., Povey, D., and Khudanpur, S. (2018). Neural Network Language Modeling with Letter-Based Features and Importance Sampling. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6109–6113.

8. Language Resource References

Nederlandse Taalunie. (2004). *Corpus Gesproken Nederlands. The Spoken Dutch Corpus project.* 2.0, ISLRN `https://taalmaterialen.ivdnt.org/download/tstc-corpus-gesproken-nederlands/`.

Radboud University. (2022). *HoMed Transcriptions Medicijnjournaal.* Radboud University via ELRA: ELRA-Id SXYZY, 1.0, ISLRN `https://doi.org/10.34973/dpjc-0v85`.