The importance of sharing patient-generated clinical speech and language data

Kathleen C. Fraser
National Research Council Canada
Ottawa, Canada
kathleen.fraser@nrc-cnrc.gc.ca

Nicklas Linz
German Research Center
for Artificial Intelligence (DFKI)
Saarbrücken, Germany
nicklas.linz@dfki.de

Hali Lindsay
German Research Center
for Artificial Intelligence (DFKI)
Saarbrücken, Germany
hali.lindsay@dfki.de

Alexandra König
Memory Clinic at Nice University Hospital,
University of Côte d'Azur, and INRIA
Nice, France
alexandra.konig@inria.fr

Abstract

Increased access to large datasets has driven progress in NLP. However, most computational studies of clinically-validated, patient-generated speech and language involve very few datapoints, as such data are difficult (and expensive) to collect. In this position paper, we argue that we must find ways to promote data sharing across research groups, in order to build datasets of a more appropriate size for NLP and machine learning analysis. We review the benefits and challenges of sharing clinical language data, and suggest several concrete actions by both clinical and NLP researchers to encourage multi-site and multi-disciplinary data sharing. We also propose the creation of a collaborative data sharing platform, to allow NLP researchers to take a more active responsibility for data transcription, annotation, and curation.

1 Introduction

The Workshop on Computational Linguistics and Clinical Psychology (CLPsych) has brought together a strong community of NLP researchers and clinical experts, working on areas as diverse as the early detection of dementia through speech analysis, characterization of the properties of autistic children's language, identifying signs of depression and anxiety from written text, and many more. One theme that has emerged over time is the importance of clinically validated data, and at the same time, the difficulty in obtaining such data.

For example, and drawing only from the past proceedings of this workshop, numerous researchers have explicitly mentioned the small size of their dataset as a limitation of the work (Jarrold et al., 2014; Glasgow and Schouten, 2014; Fraser et al., 2014; Lamers et al., 2014; Bullard et al., 2016; Parish-Morris et al., 2016; Guo et al., 2017; Iter et al., 2018). These researchers point out that the consequences of such small datasets can include a lack of diversity in and representativeness of the training data, models which do not converge to a stable solution, unknown generalizability to other datasets, difficulty in interpreting the results, and limited clinical utility.

Other work has sought to overcome these limitations by using data scraped from social media or web forums (Coppersmith et al., 2014, 2015; Mitchell et al., 2015). While solving some problems, this approach introduces others, including uncertainty around the accuracy of the diagnosis and, crucially, the lack of a clinically-confirmed healthy control group (Coppersmith et al., 2014). Furthermore, such methods of data collection likely exclude many populations, including children and the elderly.

Here, we argue that large, clinically-validated datasets of patient-generated speech and langauge are imperative if we want to move the field forward, and that one way to create such datasets is to join together as a community and commit to finding better ways to share data.

2 Background

The issue of data sharing arises in many fields, including NLP more generally (where sharing corpora is strongly encouraged) and medical research (where data openness varies by domain). Clinical

Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology, pages 55–61
Minneapolis, Minnesota, June 6, 2019.

NLP sits at the intersection of these two fields, and thus faces its own unique challenges to data sharing (Chapman et al., 2011).

In NLP, data openness has long been recognized as the key to reproducible research and fair comparison between competing systems. One example of this is the popularity of the "shared task", in which systems from different research groups are trained, validated, and tested on the same data, allowing precise comparison across systems and leading to steady improvements in areas such as machine translation, speaker identification, parsing, information retrieval, etc. (Liberman and Cieri, 1998). In many areas of NLP, recent improvements in performance and generalizability have been reported due to the availability of larger and larger corpora (Jozefowicz et al., 2016; Koehn and Knowles, 2017).

The value of data sharing has been recognized in other scientific fields, where it has permitted the accumulation of massive data sets in areas such as astronomy and climatology. For example, while it is not possible for any one telescope to see all parts of the sky simultaneously, by sharing data with each other, astronomers can collectively build an accurate picture of the night sky (Borgman, 2012). The medical community has also identified important benefits to sharing data, as well as several critical practical and ethical challenges (Souhami, 2006; Hansson et al., 2016; Figueiredo, 2017).

In the following sections, we outline the benefits and challenges of data sharing as it applies specifically to patient-generated speech and text, within the context of NLP research.

3 Arguments for sharing data

Rationales for sharing data may vary for different stakeholders in the academic process (i.e., researchers, funding agencies, study participants).

When it comes to the computational study of clinical speech data, two broad groups of researchers are involved in the data sharing process: clinical researchers, who actively collect speech and language data, and computational linguistics researchers, who analyse and build models from the data. Both groups of researchers may be motivated by the fact that sharing data advances the state of research and innovation (Borgman, 2012; Figueiredo, 2017; Campbell et al., 2002; Fischer and Zigmond, 2010). Through the aggregation of multiple local studies, researchers are able to create a combined data set bigger than any single lab could reasonably collect (Borgman, 2012; Fischer and Zigmond, 2010), thus creating a more complete representation of reality. Proposals of innovative speech and language measures are more likely to attract the interest of the medical community when the conclusions are backed by a large study population. These large datasets can also support the application of complex computational modelling techniques, such as deep learning, that are not typically effective for small data.

Data sharing can also be used as a tool to reproduce and verify previous research (Borgman, 2012; Liberman and Cieri, 1998), which helps to validate findings for use in a clinical setting. Furthermore, data sharing can also have a professional benefit to researchers, as it fulfills the requirements of some granting agencies (e.g., NIH and NSF) (Borgman, 2012; Fischer and Zigmond, 2010), and can increase the citation rates and impact of researchers' studies (Piwowar et al., 2007; Figueiredo, 2017).

Societal interest in data sharing, and thereby that of funding agencies, is motivated differently. Since funding bodies often support research using tax revenue, there is interest in making results, including data, of publicly-funded research available to the public (Borgman, 2012; Figueiredo, 2017; Pennebaker, 2004). Additionally, data sharing has been found to increase the overall quality of the produced research. It maximizes the use of collected data, as it enables others to ask new questions of existing data (Borgman, 2012; Figueiredo, 2017; Fischer and Zigmond, 2010) and diversifies the perspective on these data (Fischer and Zigmond, 2010). Financially, sharing data leads to a greater return on public investment in research, since the production costs of data sets can be shared between different actors (Liberman and Cieri, 1998; Fischer and Zigmond, 2010) and it avoids the generation of duplicate data sets (Figueiredo, 2017; Liberman and Cieri, 1998; Fischer and Zigmond, 2010).

Participants in studies, including patient and healthy controls, might be motivated by the multiple benefits to society listed above. Participants are also often motivated by making a contribution to new, improved or safer medical treatments and want their participation to have the widest possible impact (Hansson et al., 2016). They are often willing to share de-identified personal data and do not necessarily see it as an invasion of their privacy (Hansson et al., 2016). The willingness to share data may be even greater in patient populations, since results from research may directly benefit themselves or other with the diseases (Souhami, 2006; Hansson et al., 2016).

4 Challenges to sharing data

Despite the many benefits, there are also challenges within scientific communities that can prevent the sharing of data, including ethical and legal considerations, practical barriers, and the desire for researchers to protect and manage access to the data that support their research programs.

A primary concern regarding the sharing of patient data is personal privacy and security (Souhami, 2006; Childs et al., 2011), which is magnified in the case of clinical speech and language data that will be linked by necessity to personal health data (e.g., medical diagnosis, cognitive test results). Audio and visual data may not be possible to fully anonymize, and are also considered personal information. Study participants in general are wary of being identified by insurance providers, employers or other third parties as the risk of exposure of personal information may result in social or psychological harm (Hansson et al., 2016). This can lead to inaccurate self-reporting or even the avoidance of medical care if a person believes that the disclosure of certain information (e.g. drug use) will be revealed to others, resulting in harm or persecution. Additionally, even if participants gave consent for the initial data collection, obtaining consent for the secondary use of data may be impossible, as patients may be deceased or have relocated (Souhami, 2006).

For these reasons, in some cases it may not be ethically or legally permissible to share clinical data, and legal measures are in place to protect the privacy of patients and research participants. For example, in the United States medical information is protected under the Health Insurance Portability and Accountability Act (HIPAA) and the Health Information Technology for Economic and Clinical Health Act (HITECH Act) (Annas, 2003; Blumenthal, 2010); similar regulations exist in countries around the world. These policies mean that data collected by clinicians acting in their clinicial capacities may be subject to stricter regulation than data in traditional academic research. Noncompliance with federal regulations can result in fines or loss of license. Additionally, many clinicians (including psychologists[1] and psychiatrists[2]) are bound by a professional code of ethics which may preclude the sharing of patient data.

Data sharing can be difficult on a practical level. Often, data collected at separate sites are not formatted for consistent and comparable sharing (Borgman, 2012). In some cases, audio or video data may not even exist as a digital file (MacWhinney, 2007). Limited financial and personnel resources may prevent the labour-intensive preparation and documentation of clinical speech and language data into convenient, transmittable formats (Campbell et al., 2002; Borgman, 2012). Different research projects may involve different speech/language tasks, different recording conditions, different diagnostic criteria, and different clinical populations, which may limit the extent to which datasets can be combined across projects.

In addition to these challenges are personal considerations within the research community itself. Allowing others to work on private datasets could expose errors within the data or in previous publications (Childs et al., 2011). A real example of this can be found in the social psychology literature, where the re-analysis of data from the implicit association test challenged the conclusions of the original study (Blanton et al., 2009; McConnell and Leibold, 2009). Data sharing efforts typically do not factor into tenure or promotional considerations (Borgman, 2012), and there is a perceived lack of reward or credit for the considerable time and effort required (Fischer and Zigmond, 2010; Borgman, 2012). This is compounded by the reality that one's research may be considered less novel or innovative, since allowing access to data resources would allow other researchers to publish similar work on the same data (Figueiredo, 2017; Childs et al., 2011; Campbell et al., 2002).

Other concerns relate to the inability to control the applications of the data and the possibility of misuse or misinterpretation (Campbell et al., 2002; Figueiredo, 2017). Research protocols describe the purpose of the data collection, e.g. improving care and providing timely intervention, and clinicians may be wary of outside parties using these data for more profit-oriented objectives.

[1]https://www.apa.org/ethics/code/principles.pdf

[2]https://www.psychiatry.org/psychiatrists/practice/ethics

5 Examples of successful data sharing

We now briefly discuss two case studies in successful data sharing, while acknowledging that many other models exist and may also be appropriate to our community (for example, shared tasks).

One successful example of a data repository in NLP is the Linguistic Data Consortium, or LDC (Liberman and Cieri, 1998). The LDC manages dozens of widely-used speech and language corpora, including TIMIT, Gigaword, the Penn Treebank, and many other foundational datasets in NLP. As of 2018, it has distributed more than 140,000 copies of datasets to over 4,000 organizations (Cieri et al., 2018). Originally supported by grants, the LDC has been sustained by membership fees and data sales since 2015. It also has a scholarship program to provide free data access to researchers who do not have the resources to pay for a membership (DiPersio and Cieri, 2016). Particularly relevant to our discussion here, the LDC has recently started to move in the direction of creating clinical databases, including for autism and neurodegenerative disorders (Cieri et al., 2018).

In the clinical speech research realm, one successful initiative has been the TalkBank Project, including AphasiaBank and DementiaBank (MacWhinney, 2007; Forbes et al., 2012). The project is supported by grants, and members of the TalkBank consortium are expected, wherever possible, to contribute data of their own. AphasiaBank has a standard protocol of tasks that facilitates comparison and aggregation of data across individual research projects. Furthermore, demographic and neuropsychological test data are also given for the participants, and all audio, video, and transcription files use a common format. Individual datasets in the database are protected according to the sensitivity of the data and the terms of the consent. The project has its own code of ethics, and provides guidelines for research ethics board applications and consent form templates. While AphasiaBank was started by and for researchers, it has become an important resource for clinicians and educators as well (Forbes et al., 2012).

Both platforms can be used as good examples for how sharing patient-generated clinical speech and language data can be realized. In particular, they create a separation between the work of creating the data from the work of maintaining and distributing the data (Cieri et al., 2018). They have also managed the issues of security and data privacy, and have created standards for data formatting and data collection.

However, contributions to TalkBank (and the limited clinical datasets on the LDC) appear to be made mostly by clinical researchers, which still places most of the burden of preparing, documenting, transcribing, and annotating the data on their shoulders. A more collaborative model of data sharing, which involves various contributions from both clinical and computational researchers, may encourage greater participation.

6 Recommendations

Based on the literature and examples above, we offer a preliminary (and surely incomplete) set of recommended best practices to promote collaboration and data sharing. Some actions that can be taken by *researchers who are collecting data* that will aid data sharing include:

- Having a long-term data management plan in place from the initial stages of a project, and including it in the funding proposal.
- Obtaining open and transparent consent from participants, that allows sharing and re-use of the data and realistically describes the benefits and harms of data sharing.
- Reviewing archival consent forms to determine if the original terms allow sharing to any degree.[3]
- Collecting data that can be anonymized to the greatest extent possible (e.g., eliciting speech on more general topics rather than personal histories, where appropriate).
- Where it is necessary to collect data of a more personal nature (as will be the case in many situations arising in couples and family therapy, or in relation to mental health conditions), considering automated or manual approaches to anonymizing the data, including offering participants the chance to anonymize their own data.
- Using file formats and transcription protocols that are common in the field, as well as a standardized protocol of tasks and meta-data (e.g. demographic information).

[3]For example, see https://talkbank.org/share/irb/ for some guidelines on this topic.

Some actions that can be taken by *researchers who intend to make use of shared data* that will encourage and support data sharing include:

- Making other kinds of contributions to shared repositories, including: digitized versions of archival data, transcriptions, scripts for data processing and feature extraction, spreadsheets of extracted information, etc.
- Incentivizing data sharing through citations, acknowledgements, collaborations, and respectful use of the data and adherence to the relevant codes of ethics.
- Creating resources/platforms to lower the technical barriers to data sharing, and improve security and privacy of data.
- Communicating openly with the data owners, both to promote trust and to increase awareness of the kinds of emerging technologies that can benefit research in the field.

7 Conclusion and next steps

Access to larger datasets would undoubtedly improve the accuracy, generalizability, and clinical utility of computer models of patient-generated speech and language. However, clinical data is expensive and time-consuming to collect. Therefore, we argue that increased data sharing across research groups may be the only way to collect datasets of the size needed for robust machine learning, and to establish the population norms and empirical validation that will be required to allow NLP technologies to be recognized and used in clinical practice.

Existing platforms like the LDC and TalkBank are one option, particularly for sharing existing data sets. However, other models of data sharing may also be appropriate. Specifically, we propose a collaborative platform to support the continuous aggregation of data in a multi-disciplinary setting, where different parties can contribute according to their expertise (e.g., clinicians collect data, NLP researches transcribe or curate data). This shifts some of the responsibility from the clinical researchers to the computational researchers, while increasing the total value of the resulting data resource for everyone.

As a first step towards this goal, we advocate for the creation of a multi-disciplinary working group, consisting of clinicians and clinical researchers, patient organizations, and NLP researchers. This group should carefully review the feasibility of the recommendations made in the previous section, gauge interest in such a project from the various stakeholders, define the concrete requirements of a platform that would enable multi-disciplinary data collection and sharing, and determine how it could be prototyped and sustained through funding, over a longer period of time. It is essential that clinicians take a leading role in defining the concrete objectives and orientation of this group, ensuring that clinical research goals and improved patient outcomes are the main focus.

References

George J Annas. 2003. HIPAA regulations—a new era of medical-record privacy? *New England Journal of Medicine*, 348(15):1486–1490.

Hart Blanton, James Jaccard, Jonathan Klick, Barbara Mellers, Gregory Mitchell, and Philip E Tetlock. 2009. Strong claims and weak evidence: Reassessing the predictive validity of the IAT. *Journal of Applied Psychology*, 94(3):567–582.

David Blumenthal. 2010. Launching HITECH. *New England Journal of Medicine*, 362(5):382–385.

Christine L Borgman. 2012. The conundrum of sharing research data. *Journal of the American Society for Information Science and Technology*, 63(6):1059–1078.

Joseph Bullard, Cecilia Ovesdotter Alm, Xumin Liu, Qi Yu, and Rubén Proano. 2016. Towards early dementia detection: Fusing linguistic and non-linguistic clinical data. In *Proceedings of the Third Workshop on Computational Linguistics and Clinical Psychology*, pages 12–22.

Eric G Campbell, Brian R Clarridge, Manjusha Gokhale, Lauren Birenbaum, Stephen Hilgartner, Neil A Holtzman, and David Blumenthal. 2002. Data withholding in academic genetics: Evidence from a national survey. *Journal of the American Medical Association*, 287(4):473–480.

Wendy W Chapman, Prakash M Nadkarni, Lynette Hirschman, Leonard W D'avolio, Guergana K Savova, and Ozlem Uzuner. 2011. Overcoming barriers to NLP for clinical text: The role of shared tasks and the need for additional creative solutions. *Journal of the American Medical Informatics Association*, 18(5):540–543.

Becky Childs, Gerard Van Herk, and Jennifer Thorburn. 2011. Safe Harbour: Ethics and accessibility in sociolinguistic corpus building. *Corpus Linguistics and Linguistic Theory*, 7(1):163–180.

Christopher Cieri, Mark Liberman, Stephanie Strassel, Denise DiPersio, Jonathan Wright, and Andrea

Mazzucchi. 2018. From 'solved problems' to new challenges: A report on LDC activities. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*, pages 3265–3269.

Glen Coppersmith, Mark Dredze, and Craig Harman. 2014. Quantifying mental health signals in Twitter. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 51–60.

Glen Coppersmith, Mark Dredze, Craig Harman, and Kristy Hollingshead. 2015. From ADHD to SAD: Analyzing the language of mental health on Twitter through self-reported diagnoses. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 1–10.

Denise DiPersio and Christopher Cieri. 2016. Trends in HLT research: A survey of LDC's data scholarship program. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, pages 1614–1618.

Ana Sofia Figueiredo. 2017. Data sharing: Convert challenges into opportunities. *Frontiers in Public Health*, 5:327.

Beth A Fischer and Michael J Zigmond. 2010. The essential nature of sharing in science. *Science and Engineering Ethics*, 16(4):783–799.

Margaret M Forbes, Davida Fromm, and Brian MacWhinney. 2012. AphasiaBank: A resource for clinicians. In *Seminars in Speech and Language*, volume 33, pages 217–222. Thieme Medical Publishers.

Kathleen C Fraser, Graeme Hirst, Naida L Graham, Jed A Meltzer, Sandra E Black, and Elizabeth Rochon. 2014. Comparison of different feature sets for identification of variants in progressive aphasia. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 17–26.

Kimberly Glasgow and Ronald Schouten. 2014. Assessing violence risk in threatening communications. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 38–45.

Jia-Wen Guo, Danielle L Mowery, Djin Lai, Katherine Sward, and Mike Conway. 2017. A corpus analysis of social connections and social isolation in adolescents suffering from depressive disorders. In *Proceedings of the Fourth Workshop on Computational Linguistics and Clinical Psychology — From Linguistic Signal to Clinical Reality*, pages 26–31.

Mats G Hansson, Hanns Lochmüller, Olaf Riess, Franz Schaefer, Michael Orth, Yaffa Rubinstein, Caron Molster, Hugh Dawkins, Domenica Taruscio, Manuel Posada, et al. 2016. The risk of re-identification versus the need to identify individuals in rare disease research. *European Journal of Human Genetics*, 24(11):1553.

Dan Iter, Jong Yoon, and Dan Jurafsky. 2018. Automatic detection of incoherent speech for diagnosing schizophrenia. In *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, pages 136–146.

William Jarrold, Bart Peintner, David Wilkins, Dimitra Vergryi, Colleen Richey, Maria Luisa Gorno-Tempini, and Jennifer Ogar. 2014. Aided diagnosis of dementia type through computer-based analysis of spontaneous speech. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 27–37.

Rafal Jozefowicz, Oriol Vinyals, Mike Schuster, Noam Shazeer, and Yonghui Wu. 2016. Exploring the limits of language modeling. *arXiv preprint arXiv:1602.02410*.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. *arXiv preprint arXiv:1706.03872*.

Sanne MA Lamers, Khiet P Truong, Bas Steunenberg, Franciska de Jong, and Gerben J Westerhof. 2014. Applying prosodic speech features in mental health care: An exploratory study in a life-review intervention for depression. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 61–68.

Mark Liberman and Christopher Cieri. 1998. The creation, distribution and use of linguistic data: The case of the Linguistic Data Consortium. In *Proceedings of the 1st International Conference on Language Resources and Evaluation (LREC)*, pages 159–164.

Brian MacWhinney. 2007. The TalkBank Project. In *Creating and Digitizing Language Corpora*, pages 163–180. Springer.

Allen R McConnell and Jill M Leibold. 2009. Weak criticisms and selective evidence: Reply to Blanton et al.(2009). *Journal of Applied Psychology*, 94(3):583–589.

Margaret Mitchell, Kristy Hollingshead, and Glen Coppersmith. 2015. Quantifying the language of schizophrenia in social media. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 11–20.

Julia Parish-Morris, Mark Liberman, Neville Ryant, Christopher Cieri, Leila Bateman, Emily Ferguson, and Robert Schultz. 2016. Exploring autism spectrum disorders using HLT. In *Proceedings of the*

Third Workshop on Computational Linguistics and Clinical Psychology, pages 74–84.

James W Pennebaker. 2004. Theories, therapies, and taxpayers: On the complexities of the expressive writing paradigm. *Clinical Psychology: Science and Practice*, 11(2):138–142.

Heather A Piwowar, Roger S Day, and Douglas B Fridsma. 2007. Sharing detailed research data is associated with increased citation rate. *PloS One*, 2(3):e308.

Robert Souhami. 2006. Governance of research that uses identifiable personal data. *The BMJ*, 333(7563):315–316.