Overcoming barriers to NLP for clinical text: the role of shared tasks and the need for additional creative solutions

Wendy W Chapman,[1] Prakash M Nadkarni,[2] Lynette Hirschman,[3] Leonard W D'Avolio,[4,5,6] Guergana K Savova,[7] Ozlem Uzuner[8]

This issue of *JAMIA* focuses on natural language processing (NLP) techniques for clinical-text information extraction. Several articles are offshoots of the yearly 'Informatics for Integrating Biology and the Bedside' (i2b2) (http://www.i2b2.org) NLP shared-task challenge, introduced by Uzuner *et al* (***see page 552***)[1] and co-sponsored by the Veteran's Administration for the last 2 years. This shared task follows long-running challenge evaluations in other fields, such as the Message Understanding Conference (MUC) for information extraction,[2] TREC[3] for text information retrieval, and CASP[4] for protein structure prediction. Shared tasks in the clinical domain are recent and include annual i2b2 Challenges that began in 2006, a challenge for multi-label classification of radiology reports sponsored by Cincinnati Children's Hospital in 2007,[5] a 2011 Cincinnati Children's Hospital challenge on suicide notes,[6] and the 2011 TREC information retrieval shared task involving retrieval of clinical cases from narrative records.[7]

Although NLP research in the clinical domain has been active since the 1960s, progress in the development of NLP applications for clinical text has been slow and lags behind progress in the general NLP domain. There are several barriers to NLP development in the clinical domain, and shared tasks like the i2b2/VA Challenge address some of these barriers. Nevertheless, many barriers remain and unless the community takes a more active role in developing novel approaches for addressing the barriers, advancement and innovation will continue to be slow.

BARRIERS TO NLP DEVELOPMENT IN THE CLINICAL DOMAIN

Historically, there have been substantial barriers to NLP development in the clinical domain. These barriers are not unique to the clinical domain: they also occur in the fields of software engineering and general NLP.

Lack of access to shared data

Because of concerns regarding patient privacy and worry about revealing unfavorable institutional practices, hospitals and clinics have been extremely reluctant to allow access to clinical data for researchers from outside the associated institutions. The lack of reliable and inexpensive de-identification techniques for narrative reports has compounded the reluctance to share. Such restricted access to shared datasets has hindered collaboration and inhibited the ability to assess and adapt NLP technologies across institutions and among research groups. Several pioneering efforts[5 8–11] have made clinical data available for sharing—we need more of these grass-roots efforts.

Lack of annotated datasets for training and benchmarking

Closely related but not completely conditional on lack of shared datasets is the deficiency of annotated clinical data for training NLP applications and benchmarking performance. The sublanguage of clinical reports often necessitates domain-specific development and training, and, as a consequence, NLP modules developed for general text typically do not perform as well on clinical narratives. We need increased coordination to create annotation sets that can be merged to produce larger training and evaluation sets.

Insufficient common conventions and standards for annotations

Without the ability to share data, the community has lacked incentives for developing common data models for manual and automatic annotations. The result is that annotated datasets are usually unique to the laboratory that generated them and thus remain small and that NLP modules that perform the same tasks cannot be substituted and compared without considerable translational effort. At present, the clinical NLP community is leveraging existing standards and conventions and working together to develop shared data models and to map annotations across information extraction applications.

The formidability of reproducibility

Adopting an existing NLP application or module is complicated—source code and documentation may be unavailable, and published descriptions may lack sufficient detail for reproducibility. Open source releases of clinical information extraction and retrieval systems have improved the opportunity to reproduce performance.[12–15] Even with open source release, a tool may work less well in others' hands than in the hands of the original developers. Compounding the problem of reproducibility is the fact that proof-of-concept tools created in academic/research environments may not meet the highest software engineering quality, maintainability, scalability, or usability standards. And sometimes a tool may be over-fitted to a particular application, and modification to solve a similar problem may require wholesale changes. As Pedersen asserted,[16] the NLP community needs to invest more in assisting others in applying and reproducing our results.

Limited collaboration

In part due to previously listed barriers, collaboration within the clinical NLP community has been nominal. Development of NLP systems within the academic environment has centered around single institutions and single laboratories, and rather than building upon the foundations of previous work, the majority of clinical

[1]Department of Biomedical Informatics, University of California San Diego, La Jolla, California, USA; [2]Yale University, New Haven, Connecticut, USA; [3]The MITRE Corporation, Bedford, Massachusetts, USA; [4]Massachusetts Veterans Epidemiology Research and Information Center (MAVERIC), Boston, MA, USA; [5]Harvard Medical School, Division of Aging, Boston, MA, USA; [6]Center for Surgery and Public Health, Brigham and Women's Hospital, Boston, MA, USA; [7]Children's Hospital Boston Informatics Program, Harvard Medical School, Boston, Massachusetts, USA; [8]University at Albany-SUNY, Albany, New York, USA

Correspondence to Dr Wendy W Chapman, Department of Biomedical Informatics, University of California San Diego, 9500 Gilman Dr, Bldg 2 #0728, La Jolla, California, USA; wwchapman@ucsd.edu

NLP systems developed over the last four decades have been reinvented as silos that are neither expanded nor applied outside of the individual laboratory. Other factors limiting collaboration include insufficient infrastructure for facilitating cooperation and the reality that collaboration is inherently inefficient. Nevertheless, as with the biomedical research community at large, a surge in progression beyond the last half century of research can only come through enhanced teamwork. A recent trend for teamwork across NLP research laboratories is evident in funded initiatives such as the VA's Consortium for Healthcare Informatics Research (CHIR)[17] and in the ONC-funded SHARP Area 4 grant for Secondary Use of EHR data.[18] Also, advances in open source development and conformity to common frameworks has led to recent advances in NLP allowing one team to extend work by another (eg, HiTEX[12] built on GATE and cTAKES,[14] ODIE,[19] and Automated Retrieval Console (ARC) built on UIMA[16]).

Lack of user-centered development and scalability

Although we are improving incrementally the predictive performance of clinical NLP tools, clinical NLP applications are seldom deployed in clinical, public health, or health services research settings. Currently, the perceived cost of applying NLP outweighs the perceived benefit. Deploying an NLP system typically requires a substantial amount of time from an expert NLP developer—normally, applications do not generalize and must be rebuilt, retrained, enhanced, and re-evaluated for each new task; the output of an NLP system typically requires extensive mapping to the specific problem being addressed; and the ability to aid a user in customizing the application is generally inadequate (***see page 544***).[20] We need a shift of focus from accuracy in one task to generalizabiliy across many and from the production of papers as the sole output to production of usable software for medically relevant applications. We also need to understand where NLP tools fit into an overall user workflow so that the tools can be integrated into end-to-end applications for clinical, public health, and clinical research users.

SHARED TASKS—A PARTIAL SOLUTION FOR PROGRESS

Shared tasks like the i2b2/VA Challenge address several of these barriers in part. Shared tasks provide annotated datasets to participants and sometimes to non-participants (i2b2 datasets are available to others a year after the Challenge). The i2b2 shared task is standardizing its corpus as much as possible—the same records are used from one year to the next with layers of annotation that build on each other, and common input/output specifications are applied every year. Shared tasks partially address the barrier of reproducibility by providing an evaluation opportunity that minimizes the risk of over-fitting: participants have time to train their systems in supervised fashion with an annotated training dataset, but then evaluation must be performed against a separate non-annotated dataset within a stringent time limit that prevents non-trivial system modifications. Although shared tasks are not designed for this purpose, the i2b2 Challenge has been the impetus for some new collaborations across independent research groups.

Shared tasks have driven progress in related fields. For example, progress in speech understanding research was driven by a series of evaluations funded by DARPA from the late 1980s to the early 2000s.[21] The research community was able to consistently drive down the error rate by a factor of two every 2 years, on successively more challenging tasks, moving from recognition of small-vocabulary read speech to automated transcription of broadcast news in multiple languages. Associated with this progress was the incorporation of speech recognition products into applications, from dictation to speech interfaces.

Shared tasks provide value to the NLP community in several ways:

- Common evaluation metrics are developed.
- Annotated datasets are made available.
- Enticed by available annotated datasets, researchers in overlapping fields (both academic and corporate) participate in the tasks, bringing in new people and new approaches.
- Benchmarking evaluation on a shared dataset reveals the state-of-the-art performance for a given task.
- Students and post-docs receive excellent training opportunities.
- Preliminary results can be obtained by a new research group, which can potentially lead to funding opportunities.
- Pre-processed, standardized corpora with multiple layers of annotations on the same corpus pave the way for end-to-end evaluations in addition to evaluation on a single annotation layer.
- Conventions for standardizing annotations and input/output formats are developed, and despite other standardization efforts, shared task corpora often set de facto standards.

In spite of the value of shared tasks, the tasks have several shortcomings:

- Participants come mainly from teams with funded projects that overlap with the shared task.
- For academic participants, a significant motivation is the opportunity to publish; however, there is sometimes limited value for the larger community in publications resulting from a shared task. Because development time is limited during shared tasks, participants often build on applications that already exist and apply methods already described in the literature. This can result in many similar approaches being applied to the same task. Although publishing the high-performing systems can be interesting, the resulting publications may not be novel and therefore may not improve the general body of knowledge.
- If a particular challenge task is repeated over time, there is a tendency for system approaches to converge on the approach that showed most success in the previous evaluation—evaluations repeated over time tend to reduce the diversity of approaches.

Although shared tasks contribute to growth and progress, increased benefit to the community of clinical NLP developers and to potential users will require additional individual and community efforts that target existing barriers creatively.

THINKING CREATIVELY ABOUT THE FUTURE

Driving progress in a way that will increase the impact of NLP in the realm of individual and population health will require creativity at both the grass-roots and the community levels. No single activity can tackle all barriers. In addition to encouraging variations on the development of shared tasks and their incentives, we would like to see new types of shared activities that foster the outcomes described below.

Reproducibility of results

In exchange for access to the costly annotated dataset, shared task participation could be contingent on depositing code in a shared repository or creating a web service for prospective users. In this model, the organizers could send test data to the participants' servers and the servers return the results for evaluation (see Leitner *et al*[22] for a description of a meta-server used in evaluation of results from BioCreative II). The servers (and meta-server) could even persist, providing services to interested users beyond the initial shared task. Publication of computational methods in biomedical informatics journals like *JAMIA* could further encourage reproducibility of results through policies mandating simultaneous submission of code with a manuscript, as recommended by Pedersen.[16] The clinical NLP community could independently accelerate reproducibility (and lead by example) by depositing code and developing web services in a common repository.[23] This trend is occurring in settings restricted by affiliation, such as the VA VINCI framework[24] (available only to VA researchers) and a cloud environment being hosted by the SHARP Area 4 grant[18] (available to grant participants). The new National Center for Biomedical Computing iDASH[25] is developing a similar cyber-infrastructure that will be restricted not by affiliation but by adherence to privacy policies and agreements required by data contributors. The National Library of Medicine is currently hosting a registry developed by the AMIA NLP working group called ORBIT for listing and pointing to biomedical informatics and NLP resources.[26]

Collaboration

In a shared task, dozens of research groups duplicate the same task independently. Although a variety of techniques can emerge for the same task, given the relatively short time frame allowed for development and training, the features and approaches applied in the challenge are often very similar. Whereas similarity of approaches reveals agreement among teams on the best approaches and sets the stage for collaboration, the competitive nature of shared tasks provides a disincentive to collaboration; the reward system for shared tasks is not at all dependent on the ability to collaborate across teams but is solely geared toward competition in which a single winner arises. The open source development community has found inherent rewards in collaborative development through supportive environments like GitHub. Perhaps we can learn from collaborative development communities who participate in hackathons[27] and from the games industry in asking how a shared task can be designed so that collaboration is rewarded and becomes worthwhile, interesting, and attractive.[28]

User-centered design

Evaluation of a shared task is focused on accuracy, and existing challenges evaluate only predictive performance, not software engineering characteristics or usability. Imagine a shared task in which success is judged on usability of a system or direct portability of one technique to a new task or domain. Evaluating success of a system with this paradigm is inherently more complex, but we could learn from groupware evaluation, from the rich field of usability testing, and from incentivized competitions like those sponsored by the X-Prize Foundation.[29]

Scalability and tackling real problems

Because of the cost of creating annotated training data, shared tasks are often small scale, at least relative to real medical applications. We need new approaches to rapid adaptation of NLP systems to new applications, with less dependence on 'deeply annotated' data; such applications would present important opportunities for collaboration with the end user community, who might be motivated to provide domain expertise if they were likely to get a scalable, maintainable system out of the collaboration. Scalability will require more efficient techniques for manual annotation. And scalability will require an enriched ability to produce high quality software, which may necessitate better collaboration with industry[30] and funding models that include support for operational development.

CONCLUSION

The shared i2b2 evaluations have made a huge contribution to stimulating and vitalizing the field of clinical NLP; however, to ensure the transition into usable applications, the clinical NLP research community needs to address the critical issues of data access, development of shared infrastructure, and integration of software engineering methods to ensure the usability, maintainability, and availability of clinical NLP tools that are integrated into the workflow of real biomedical applications. This must be done in close collaboration with end users, software engineers, and clinical practitioners. We as a community need to think beyond the status quo of incremental improvement in the F score toward imaginative approaches that encourage collaboration, promote reproducibility, increase the scalability of NLP development, and provide value to end users.

Funding Authors are funded in part by U54HL108460, R01GM090187, SHARP ONC award 90TR0002, R01 CA127979, and U54LM008748.

Competing interests None.

Provenance and peer review Not commissioned; internally peer reviewed.

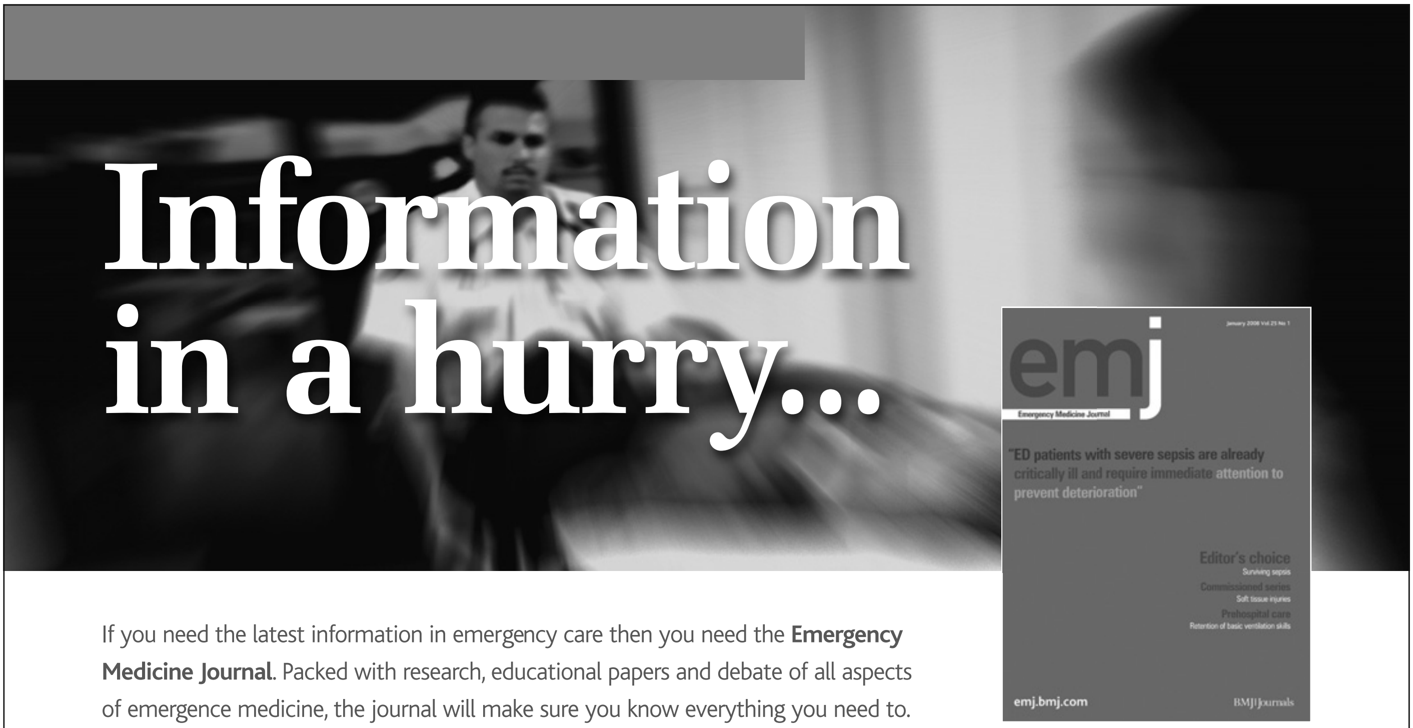
Accepted 6 July 2011

J Am Med Inform Assoc 2011;**18**:540—543. doi:10.1136/amiajnl-2011-000465

REFERENCES

1. **Uzuner O,** South B, Shen S, *et al*. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *J Am Med Inform Assoc* 2011;**18**:552—6.
2. **Grishman R,** Sundheim B. Message understanding conference - 6:a brief history. *16th International Conference on Computational Linguistics (COLING); 1996*. Copenhagen, Denmark: Center for Sprogteknologi, 1996:466—71.
3. **National Institute for Standards in Technology (NIST).** *Text Retrieval Conference (TREC)*. 2011. http://trec.nist.gov (accessed 1 Jun 2011).
4. **Moult J.** A decade of CASP: progress, bottlenecks and prognosis in protein structure prediction. *Curr Opin Struct Biol* 2005;**15**:285—9.
5. **Pestian JP,** Brew C, Matykiewicz P, *et al*. A shared task involving multi-label classification of clinical free text. *BioNLP Workshop of the Association for Computational Linguistics; 2007 June 29, 2007*. Prague, Czech Republic: Association for Computational Linguistics (ACL), 2007:97—104.
6. **Univ. of Cinncinnati Computational Medicine Center.** *NLP Challenge*. 2011. http://computationalmedicine.org/challenge/organizers (accessed 1 Jul 2011).
7. **National Institute for Standards in Technology (NIST).** *Text Retrieval Conference (TREC) 2011*. 2011. http://trec.nist.gov/pubs/call2011.html (accessed 1 Jun 2011).
8. **Uzuner O.** Recognizing obesity and comorbidities in sparse data. *J Am Med Inform Assoc* 2009;**16**:561—70.
9. **Uzuner O,** Goldstein I, Luo Y, *et al*. Identifying patient smoking status from medical discharge records. *J Am Med Inform Assoc* 2008;**15**:14—24.
10. **Uzuner O,** Luo Y, Szolovits P. Evaluating the state-of-the-art in automatic de-identification. *J Am Med Inform Assoc* 2007;**14**:550—63.
11. **Uzuner O,** Solti I, Cadag E. Extracting medication information from clinical text. *J Am Med Inform Assoc.* 2010;**17**:514—18.
12. **Zeng QT,** Goryachev S, Weiss S, *et al*. Extracting principal diagnosis, co-morbidity and smoking status for asthma research: evaluation of a natural language processing system. *BMC Med Inform Decis Mak* 2006;**6**:30.
13. **Christensen L,** Harkema H, Irwin J, *et al*: A system for the semantic analysis of clinical text. *Proceedings of the BioNLP2009 Workshop of the ACL Conference*. Boulder, CO: Association for Computational Linguistics (ACL), 2009.
14. **Savova GK,** Masanz JJ, Ogren PV, *et al*. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. *J Am Med Inform Assoc* 2010;**17**:507—13.
15. **D'Avolio LW,** Nguyen TM, Farwell WR, *et al*. Evaluation of a generalizable approach to clinical information retrieval using the automated retrieval console (ARC). *J Am Med Inform Assoc* 2010;**17**:375—82.

16. **Pedersen T.** Empiricism is not a matter of faith. *Comput Ling* 2007;**34**:465–70.
17. **Veterans Administration.** *Collaboration between VINCI and CHIR*. 2011. http://www.hsrd.research.va.gov/for_researchers/vinci/chir.cfm (accessed 1 Jul 2011).
18. **SharpN.org.** *Strategic Health IT Advanced Research Projects*. 2011. http://informatics.mayo.edu/sharp/index.php/Main_Page (accessed 1 Jul 2011).
19. **ODIE**. https://bmir-gforge.stanford.edu/gf/project/odie (accessed 22 Jul 2011).
20. **Nadkarni P,** Chapman W, Ohno-Machado L. Natural language processing: an introduction. *J Am Med Informat Assoc* 2011;**18**:544–51.
21. **Pallett DS.** The role of the National Institute of Standards and Technology in DARPA's Broadcast News continuous speech recognition research program. *Speech Comm* 2002;**37**:3–14.
22. **Leitner F,** Krallinger M, Rodriguez-Penagos C, *et al*. Introducing meta-services for biomedical information extraction. *Genome Biol* 2008; **9**(Suppl 2):S6.
23. **Chapman WW,** Saul M, Houston JD, *et al*. *Creation of a Repository of Automatically De-Identified Clinical Reports: Processes, People, and Permission. American Medical Informatics Association Clinical Research Informatics Summit; 2011*. San Francisco: American Medical Informatics Association (AMIA), 2011.
24. **Veterans Administration.** *VA Informatics and Computing Infrastructure (VINCI)*. 2011. http://www.hsrd.research.va.gov/for_researchers/vinci/default.cfm (accessed 1 Jul 2011).
25. **University of California San Diego.** *iDASH*. 2011. http://idash.ucsd.edu (accessed 2 Jul 2011).
26. **National Library of Medicine.** *ORBIT: Online Registry of Biomedical Informatics Tools*. 2011. http://orbit.nlm.nih.gov (accessed 1 Jul 2011).
27. **Wikipedia.** *Hackathon*. 2011. http://en.wikipedia.org/wiki/Hackathon (accessed 1 Jul 2011).
28. **Zagal JP,** Bruckman A. Designing online environments for expert/novice collaboration: Wikis to support legitimate peripheral participation. *Convergence* 2010;**16**:451–70.
29. **X-Prize Foundation.** *Heritage Health Prize*. 2011. http://www.heritagehealthprize.com/c/hhp (accessed 3 Jul 2011).
30. **Chapman WW.** Closing the gap between NLP research and clinical practice. *Methods Inf Med* 2010;**49**:317–19.


Information
in a hurry...
emj
Emergency Medicine Journal
"ED patients with severe sepsis are already critically ill and require immediate attention to prevent deterioration"
Editor's choice
Surviving sepsis
Commissioned series
Soft tissue injuries
Prehospital care
Retention of basic ventilation skills
emj.bmj.com
BMJ|Journals