Replicability of Research in Biomedical Natural Language Processing: a pilot evaluation for a coding task

Aurélie Névéol and **Cyril Grouin**
LIMSI, CNRS, Université Paris-Saclay, F-91405 Orsay
firstname.lastname@limsi.fr

Kevin Bretonnel Cohen
University of Colorado, USA
kevin.cohen@gmail.com

Aude Robert
INSERM-CépiDC, Paris, France
aude.robert@inserm.fr

Abstract

The scientific community is facing raising concerns about the reproducibility of research in many fields. To address this issue in Natural Language Processing, the CLEF eHealth 2016 lab offered a replication track together with the Clinical Information Extraction task. Herein, we report detailed results of the replication experiments carried out with the three systems submitted to the track. While all results were ultimately replicated, we found that the systems were poorly rated by analysts on documentation aspects such as "ease of understanding system requirements" (33%) and "provision of information while system is running" (33%). As a result, simple steps could be taken by system authors to increase the ease of replicability of their work, thereby increasing the ease of re-using the systems. Our experiments aim to raise the awareness of the community towards the challenges of replication and community sharing of NLP systems.

1 Introduction

Reproducibility, or replicability, is the quality of a scientific experiment that can be performed independently several times and yield the exact same results on each iteration.

1.1 Why should research strive for reproducibility and methods to achieve it

The advantages of reproducibility notably include increased work productivity and recognition in the community (Piwowar et al., 2007; Schultheiss et al., 2011; Markowetz, 2015). However, in practice, reproducibility is not always achieved or maintained over time (Davis and Walters, 2011). The scientific community is facing raising concerns about the reproducibility of research in many fields (Baker, 2016), including Natural Language Processing (Fokkens et al., 2013).

Is there really a reproducibility problem in natural language processing that needs to be dealt with? Different observations support different conclusions regarding this question. On the one hand, the relative paucity of attention to the question until recently suggests that the community does not seem to think that there is one. On the other hand, recent activity in the area suggests that the community might not be quite so sanguine about the situation: an editorial in a major journal in our field (Pedersen, 2008) and the healthy level of participation in a workshop on the topic associated with a major conference[1] suggest that in fact, reproducibility is an issue—not just reproducibility of work outside of one's own lab, but even reproducibility of work *within* one's own lab.

Can we investigate empirically the extent of reproducibility issues in natural language processing? Previous work has pointed out that in computer science in general, it is difficult to assess reproducibility even at very superficial levels and even with very unambitious definitions of "reproducibility" (Goodman et al., 2016). If the null hypothesis is that it is not any more difficult to assess reproducibility in natural language processing than it is in other areas of computer science, then there is reason to suspect that the null hypothesis does not hold, and that in

[1] Workshop on Research Results Reproducibility and Resources Citation in Science and Technology of Language; http://4real.di.fc.ul.pt/

Proceedings of the Seventh International Workshop on Health Text Mining and Information Analysis (LOUHI), pages 78–84,
Austin, TX, November 5, 2016.

fact it is *more* difficult to assess reproducibility in natural language processing due to the nature of the data our discipline studies: large corpora of natural language texts that are updated on a regular basis (e.g. PubMed, hospital information systems) and subject to being processed in a myriad of different yet similar ways by every researcher (e.g. for the purpose of word segmentation, part of speach tagging).

1.2 The shared task model in evaluation of natural language processing

Early in the history of natural language processing, it was quite difficult for researchers to learn from comparisons of systems because they generally differed on the most basic issues of goals and metrics. Answering questions that are commonplace today, such as *what are the advantages and disadvantages of purely rule-based methods and purely learning-based methods for information extraction?,* was not possible when the differences between projects included not only different methods, but also different extraction targets, data, and figures of merit. In that context, the idea developed that one could learn more from research by standardizing some of those basic aspects of the work. The resulting *shared task model* of evaluation consists of multiple groups agreeing on a shared task definition, a shared data set, and a shared evaluation metric.

Thus, shared tasks provide an opportunity to overcome some of the challenges to replication in natural language processing—in particular, the definitional, data, and scoring issues. The work reported here explores the question of whether the evaluation of replicability in natural language processing can be pushed forward to the highest level of the replicability hierarchy by taking advantage of these aspects of the shared task model. The rationale behind this approach is that one can capitalize on the fact that the systems that are used to address a challenge task all accommodate the same input and output formats, as specified in the challenge. And, the scoring code is open and freely available. Therefore, system results on a challenge dataset should be very easy to replicate without incurring significant training and effort—or, at least, they should be *possible* to replicate, if given access to the original system.

1.3 Leveraging the shared task model to achieve reproducibility

The project reported here is an attempt to pursue the issue of reproducibility in the light of the language-processing-specific problems in research methodology. Discussing reproducibility in computer science in general, Collberg et al. (Collberg et al., 2014) suggest that in the context of computer science research, the notion of reproducibility—defined by them as *the independent confirmation of a scientific hypothesis through reproduction by an independent researcher/lab*—can usefully be replaced by the concept of *repeatability.* In particular, they define three types of what they call *weak repeatability.* The highest level is the ability of a system to be acquired, and then built in 30 minutes or fewer. The next level is the ability of a system to be acquired, and then built, regardless of the time required to do so. The lowest level is the ability of a system to be acquired, and then either built, regardless of the time required to do so, *or* the original author's insistence that the code would build, if only enough of an effort were made.

Previous work has reached only as far as the 3rd level (Anda et al., 2009). However, the shared task environment (defined below) gives us the opportunity to come quite close to the fourth and highest level: the ability of a system to be acquired, built, and used to produce results consistent with published reports. In particular, the facts that the shared task model gives one access to the same data on the one hand, and the same scoring script on the other, provide a rather unique opportunity to evaluate reproducibility at the fourth level. In fact, this paper reports the only work that we are aware of that travels this high up the computer science reproducibility hierarchy.

Reproducibility is a real challenge because of the complexity of scientific experiments and experimental set-up. When describing experiments, researchers are often encouraged to focus on the novelty and interest of their research while devoting less time (and report space) to describe steps that might appear as easy routine. This situation leaves researchers (the authors themselves, or colleagues) trying to reproduce experiments described in a paper with a series of minute technical questions. Without

answers to these questions, the experimental set-up may or may not be reproduced exactly, and it becomes difficult to interpret differences in results.

Beyond the observation that reproducibility can be hard to achieve, the scientific community is also trying to understand the specific challenges associated with reproducibility in order to devise strategies to overcome them (Nosek et al., 2015; Cohen et al., 2016). The work we present here follows this direction and aims to study the ease of reproducing experiments in the highly constrained setting of a community shared task, and to yield first-hand actionable knowledge of what makes an experiment easy or difficult to reproduce.

2 Replication track at CLEF eHealth 2016

The CLEF eHealth 2016 lab (Kelly et al., 2016) offered three tasks to promote information extraction and information retrieval in the clinical domain. Task 2 (Neveol et al., 2016) focused on clinical information extraction in languages other than English. It challenged participants with the task of extracting UMLS (Unified Medical Language System) concepts from biomedical text French in the form of anchored normalized entities or ICD10 (International Classification of Diseases, 10th revision) codes.

2.1 Description of the replication task and system requirements

Participation in the replication track was open to all teams who submitted results to the task. After submitting their result files, participating teams had one extra week to submit the system used to produce them, or a remote access to the system, along with instructions on how to install and operate the system.

The "replication track" consisted in attempting to replicate a team's results by running the system supplied on the test data sets, using the team's instructions.

2.2 System analysts and evaluation environment

Four system analysts committed to spend a maximum of one working day (8 hours) with each system. The analysts attempted to install and configure the systems according to the instructions supplied. Participants were also allowed to supply a contact address to make themselves available to address any additional questions.

Two analysts had a Computer Science background with experience developping research systems in the field of bioNLP, and represented the use-case of a colleague trying to reproduce experiments in their field (research-oriented role). Another analyst had a computer science background, and the fourth analyst had a mixed linguistics/computational linguistics background. Both had experience using bioNLP applications and represented the use-case of a user trying to leverage an existing tool for a task of interest (user-oriented role).

In contrast with (Zheng et al., 2015), we did not foster a controlled environment (e.g. using a virtual machine with standard configuration for all analysts) for installing the systems evaluated because we wanted the analysts to work in an experimental setting that would be similar to the one they would use for reproducing experiments. For the same reason, we did not rely on the use of containers.

2.3 Evaluation of the replication experience

The analysts independently ran the systems on the appropriate CLEF eHealth task 2 test sets. The results obtained were be compared to those submitted by the teams using the same system. During this process, the analysts took notes on the various aspects of working with the systems (ease of installing and using, ease of understanding supplied instructions, success of the replication attempt), using a specific score sheet developed by the analysts, following some of the criteria evaluated by (Zheng et al., 2015). The score sheet comprised 10 questions addressing the experience of analysts at each stage of the experiment: system configuration, system installation, running the system, obtaining results, and overall impressions. Table 1 shows the specific questions and answer scales. The analysts were also encouraged to complete their answer to questions with free text comments.

3 Results

A total of seven teams participated in CLEF eHealth 2016 task 2. Three teams submitted systems to the replication track. One team submitted a system that addressed the subtask of named entity extraction and

Question	Scoring Scale
Part 1. System configuration	
Q.1 Is it easy to understand which are the system prerequisites, to check whether they are already installed?	Yes/No
Q.2 Is it easy to follow the installation instructions to install the prerequisites that may be missing?	5-point scale
Part 2. Installing the System	
Q.3 Is it easy to follow the installation instructions to install the system itself?	5-point scale
Q.4 Did you need to contact the system authors to install any part of the system?	Yes/No
Part 3. Running the System on the CLEF eHealth 2016 datasets	
Q.5 Is it easy to follow the instructions in the user manual to use the system to process the challenge dataset(s)?	5-point scale
Q.6 Are there sufficient information to assess whether the system is running as expected, e.g. progress visualization, running time, information messages	Yes/No
Part 4. Obtaining Results	
Q.7 Are the results produced directly in the challenge format?	Yes/No
Q.8 Did applying the challenge evaluation tool yield the exact same results as the participant submitted run?	4-point scale
Part 5. Overall Impression	
Q.9 Do you have any suggestions on what the authors of the system can do to make it more usable? For example: Additional information on where to find prerequisites; Examples of installation or run commands; Screenshots, videos, or tutorials of the installation process or using the system.	free text
Q10. Would you feel comfortable using the system outside the challenge?	Yes/No

Table 1: Score sheet presented to analysts when working with the systems. The 5-point scale comprised the following options: 5-Effortless or nearly effortless, 4-Somewhat easy but there are challenges, 3-Somewhat difficult, 2-Extremely difficult, nearly impossible, 1-I was not able to perform the task. The 4-point scale used for question Q.8 comprised the following options: 4-Yes, exactly the same results, 3- Results are slightly different (less than .01 difference in F-measure), 2- Results are quite different (more than .01 difference in F-measure), 1- Evaluation tool error.

the subtask of ICD10 coding. However, for named entity extraction, the system submitted relied on preprocessed intermediate results obtained by applying an indexing tool on the test corpus. From the perspective of replication, we considered that we had adequate material for reproducing only the ICD10 coding task with this system. The other two systems submitted also addressed the ICD10 coding subtask.

3.1 Characteristics of systems submitted and experimental set-up

Table 2 presents the characteristics of the systems submitted by participants to the replication track. To our knowledge, none of these systems are made available by the authors outside of the CLEF eHealth replication track.

All systems were research prototypes used with terminal-based command-line.

Four analysts (the authors of this paper) participated in the replication experiments. One analyst had access to both Windows and Linux OS and worked will all three systems. One analyst had access to a Windows OS and worked with System 1. Two analysts had access to a Mac OS and worked with System 2 and 3.

Participant	Operating System	Language
System 1	Windows	java
System 2	Linux	python
System 3	Linux	python

Table 2: Characteristics of the systems submitted to the replication track.

Table 3 presents the configuration of the machines used by the analysts to reproduce experiments.

3.2 Assessment of the replication process

Table 4 presents the time spent by each analyst working with the three systems to attempt reproducing results.

Table 5 presents the aggregated scoring of sys-

Analyst	Configuration
1	Windows 7 16Go ram, 9470Mb cache Intel Core i5-3437U CPU 1.90 GHz (2.40GHz)
2	Windows and Ubuntu 4Go ram, 3MB cache Intel Core i5-3210M with dual-core (2.50GHz)
3	Ubuntu 14.04.3 LTS 62Go ram, 42G cache Intel Xeon, CPU L5520 (2.27GHz)
4	Mac OS X 8Go ram, 3Mb cache Core i5-3427U CPU (1.8GHz)

Table 3: Configuration of the machines used by the analysts to reproduce experiments.

Participant	Analyst	Human Time	Run Time
System 1	User	47	150
System 1	Developer	180	510
System 2	User	204	720
System 2	Developer	45	96
System 3	User	55	240
System 3	Developer	10	93

Table 4: Time (in minutes) spent by each analyst reproducing results with the participant systems. For analysts with the *User profile, human time is averaged between the two analysts, while run time only reflect the run time of the analyst who succeeded in obtaining results from the systems.*

tems performed by analysts while reproducing results.

Phase	Question	Score
Configuration	Q1(*) Easy to understand?	33%
	Q2 Easy to configure?	55%
Installation	Q3(+) Easy to install?	93%
	Q4(*) Contact Author?	0%
Running	Q5(+) Easy to run?	55%
	Q6(*) Info while running?	33%
Results	Q7(*) Challenge format?	100%
	Q8(*) Reproduced?	71%
Overall	Q10(*) Use outside challenge?	33%

Table 5: Aggregated scoring of systems. A star symbol * indicates binary scales (yes/no) and a plus symbol + indicates a 4 or 5 level scale as detailed in table 1. For questions Q7 and Q8, data is averaged over analysts who did succeed in obtaining results.

3.3 Reproducibility of the results

Between them, the analysts were able to replicate results exactly for System 1 and System 3: the precision, recall and F-measure obtained from running the systems were identical to that of the runs submitted by participants for two analysts, while one analyst did not suceed in obtaining results. For System 2, only one analyst was able obtain results (one analyst obtained a memory error before obtaining results and one analyst was not able to run the system), and the results obtained showed a 0.02 difference in F-measure, which was statistically significant. For System 3, one analyst obtained results that showed less than 0.01 difference in F-measure with the results submitted by the participants. For System 3, it can also be noted that the system came with two configuration options and the analysts were only able to implement one of the configuration options each (not the same one). These difficulties are reflected in the score of 71% for the overall reproducibility (Table 5, Q8).

4 Discussion

Almost half of the participants to the CLEF eHealth 2016 task 2 submitted a system to the replication track, and an additional two teams expressed interest in submitting but did not do so due to lack of time and resources to prepare a system suitable for sharing. It can also be noted that the three systems submitted addressed the task of ICD10 coding viewed as a classification task - a relatively simpler task compared to named entity recognition and normalization also offered in task 2, which did not receive any system submission. This confirms that there is a strong interest from the community in the production of reproducible research. It also confirms that the time and resources required to ensure reproducibility are not readily available.

Table 5 shows that the weakest aspects of the systems were the ease of understanding the system requirements (Q1, overall score of 33%) and the quality of information supplied when the systems are actually running (Q6, overall score of 33%). The analysts experienced varying degrees of difficulty to install and run the systems. Differences were mainly due to the technical set-up of the computers used to replicate the experiments. For example, for System

1, one of the analysts had a version of java compatible with the system installed by default, so that running the system was effortless and the question of the java version required never came up. In contrast, the other analyst had an older version of java installed. Running the system produced errors that had to be interpreted to understand that the problem came from the incompatibility between java versions. The analyst then had to look into the system code files to find the java version requirement for the system and then update their work environment accordingly. In our opinion, this highlights the fact that reproducibility needs to be thought through preferably at the time of system development and in any case before a system can be shared or re-used.

Analysts also report that additional information on system requirements, installation procedure and practical use would be useful for all the systems submitted. For system 3, one analyst reported they stopped the experiment because they feared that installing the python configuration required would interfere with the current setting they had and would prevent them from using tools they had set-up. Additional explainations of the system requirement would have helped provide a better understanding of whether the system was compatible with an existing configuration. Free-text comments elicited specific recommandations for each of the systems.

Interestingly, one analyst in the user-oriented role reported that they would feel confident using all of the systems outside the challenge, while the other analysts did not.

5 Concluding remarks

In Section 1, we pointed out that prior to the development of the shared task model, there was no way to explore questions such as *what are the advantages and disadvantages of purely rule-based methods and purely learning-based methods for information extraction?,* due to gross differences in task definition, data, and figures of merit. Despite having developed and matured the shared task model in natural language processing, we still cannot answer questions like that. The shared task model controls three very important variables: task definition, data and evaluation metrics. However, it leaves an *enormous* number of variables unexplored, and those variables can have a large number of values. Suppose that everyone always used the default settings on every out-of-the-box machine learning package: even in that case, one only knows what the default settings are if one knows which version of the package was used, and that is often not recorded in published papers—we looked at 11 of our own machine learning papers, and found that we had given version numbers only 9% of the time.

Nonetheless, the approach that is described in this paper moves the study of replicability in natural language processing forward quite a bit. Replicating the CLEF eHealth challenge results was feasible, and this is the first paper that we know of that has demonstrated that in computer science in general, and in natural language processing in particular. For each of the three systems studied, we were able to replicate the results exactly or closely.

Not only does this work show that the approach is feasible, but it also shows that the approach is able to find problems—a very different kind of value from validating the lack of problems, and in some ways a more valuable one. The ease of replicating results varied. In particular, it generally was based on the analysts' work environment set-up. Moreover, the work reveals something about replicability in natural language processing that is "actionable," something that can be done to improve the situation: most of the difficulties encountered could be alleviated by additional documentation from system authors.

There is some reason to think that the reproducibility situation in natural language processing may be changing, and for the better. The Association for Computational Linguistics is now allowing extra pages in conference papers for documenting the fine details of system configurations. Meetings like the recent workshop at a major conference in the field—and the CLEF eHealth meeting—are exploring the issues and the opportunities for their empirical investigation. In the context of that change, work such as that reported here moves the conversation further along, to higher levels of reproducibility, and it does uncover issues in that respect. The problem of the difficulty of asking the interesting big questions—*what are the advantages and disadvantages of purely rule-based methods and purely learning-based methods for information extraction?*—due to inability to answer the lit-

tle questions—*which tokenizer did we use, did they use a linear kernel or a radial kernel, do our run times reflect performance before or after we fixed that bug*—may be closer to being resolved.

References

Bente CD Anda, Dag IK Sjoberg, and Audris Mockus. 2009. Variability and reproducibility in software engineering: A study of four companies that developed the same system. *IEEE Trans. Softw. Eng.*, 35(3):407–429.

Monya Baker. 2016. 1,500 scientists lift the lid on reproducibility. *Nature*, 533(7604):452–4.

Kevin B Cohen, Jingbo Xia, Christophe Roeder, and Lawrence Hunter. 2016. Reproducibility in natural language processing: A case study of two R libraries for mining PubMed/MEDLINE. In *LREC 4REAL Workshop:Workshop on Research Results Reproducibility and Resources Citation in Science and Technology of Language*, pages 6–12. European Language Resources Association (ELRA).

Christian Collberg, Todd Proebsting, Gina Moraila, Akash Sankaran, Zuoming Shi, and Alex M. Warren. 2014. Measuring reproducibility in computer systems research. Technical report, Department of Computer Science, University of Arizona.

Philip M Davis and William H Walters. 2011. The impact of free access to the scientific literature: a review of recent research. *J Med Libr Assoc*, 99(3):208–17.

Antske Fokkens, Marieke van Erp, Marten Postma, Ted Pedersen, Piek Vossen, and Nuno Freire. 2013. Offspring from reproduction problems: What replication failure teaches us. In *Proc of ACL*, pages 1691–1701.

Steven N Goodman, Daniele Fanelli, and John PA Ioannidis. 2016. What does research reproducibility mean? *Science Translational Medicine*, 8(341):ps12.

Liadh Kelly, Lorraine Goeuriot, Hanna Suominen, Aurelie Neveol, Joao Palotti, and Guiddo Zuccon. 2016. Overview of the CLEF eHealth evaluation lab 2016. In *Lecture Notes in Computer Science (LNCS), CLEF 2016 7th Conference and Labs of the Evaluation Forum*, Berlin, Heidelberg. Springer.

Florian Markowetz. 2015. Five selfish reasons to work reproducibly. *Genome Biol*, 8(16):274.

Aurelie Neveol, Kevin B Cohen, Cyril Grouin, Thierry Hamon, Thomas Lavergne, Liadh Kelly, Lorraine Goeuriot, Gregoire Rey, Aude Robert, Xavier Tannier, and Pierre Zweigenbaum. 2016. Clinical information extraction at the CLEF eHealth evaluation lab 2016. In *CLEF 2016 Working Notes*, number 609 in CEUR-WS, pages 28–42.

BA Nosek, G Alter, GC Banks, D Borsboom, SD Bowman, SJ Breckler, S Buck, CD Chambers, G Chin, G Christensen, M Contestabile, A Dafoe, E Eich, J Freese, R Glennerster, D Goroff, DP Green, B Hesse, M Humphreys, J Ishiyama, D Karlan, A Kraut, A Lupia, P Mabry, TA Madon, N Malhotra, E Mayo-Wilson, M McNutt, E Miguel, EL Paluck, U Simonsohn, C Soderberg, BA Spellman, J Turitto, G VandenBos, S Vazire, EJ Wagenmakers, R Wilson, and T Yarkoni. 2015. Scientific standards. promoting an open research culture. *Science*, 348(6242):1422–5.

Ted Pedersen. 2008. Empiricism is not a matter of faith. *Computational Linguistics*, 34(3):465–470.

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.

Sebastian J Schultheiss, Marc-Christian Münch, Gergana D Andreeva, and Gunnar Rätsch. 2011. Persistence and availability of web services in computational biology. *PLoS One*, 6(9):e24914.

Kai Zheng, VG Vinod Vydiswaran, Yang Liu, Yue Wang, Amber Stubbs, Özlem Uzuner, Anupama E Gururaj, Samuel Bayer, John Aberdeen, Anna Rumshisky, Serguei Pakhomov, Hongfang Liu, and Hua Xu. 2015. Ease of adoption of clinical natural language processing software: An evaluation of five systems. *J Biomed Inform*, 58:Suppl:S189–96.