Task 2: ShARe/CLEF eHealth Evaluation Lab 2013

Danielle L. Mowery[1], Brett R. South[2], Lee Christensen[2], Laura-Maria Murtola[3], Sanna Salanterä[3], Hanna Suominen[4], David Martinez[5], Noemie Elhadad[6], Sameer Pradhan[7], Guergana Savova[7], and Wendy W. Chapman[8] *

[1] University of Pittsburgh, PA, USA, dlm31@pitt.edu
[2] University of Utah, UT, USA, brett.south@hsc.utah.edu, leenlp@q.com
[3] University of Turku, Finland, laura-maria.murtola@utu.fi, sansala@utu.fi
[4] NICTA and The Australian National University, ACT, Australia, hanna.suominen@nicta.com.au
[5] NICTA and The University of Melbourne, VIC, Australia, david.martinez@nicta.com.au
[6] Columbia University, NY, USA, noemie@dbmi.columbia.edu
[7] Harvard University, MA, USA, sameer.pradhan@childrens.harvard.edu, guergana.savova@childrens.harvard.edu
[8] University of California, San Diego, CA, USA, wwchapman@ucsd.edu

Abstract. In this pilot study, we aimed to generate a reference standard of clinical acronyms and abbreviations normalized to concepts from a standardized, medical vocabulary for the ShARe/CLEF eHealth 2013 challenge. In this paper, we review prior text normalization shared tasks, reference standard generation approaches, and recent clinical acronym and abbreviation normalization research. We report inter-annotator agreement for the reference standard and performance for participant systems.

Keywords: Natural Language Processing, Text Normalization, Reference Standard Generation

1 Introduction

Health care organizations are shifting towards a patient centered approach in care delivery. One aspect of this approach is patient access to personal health information (PHI) including their clinical reports. Allowing patients access to their PHI should increase patient knowledge of their own health status, enhance patient involvement in care related decision-making, and improve communication between the patients and care providers [1]. However, patients that have accessed their PHI experience worry and confusion due to use of medical jargon

* WWC, BRS, and DLM led the task, WWC, BRS, DLM, NE, SP, and GS defined the task, DLM, BRS, LMM and SS led the annotation effort, HS and SS co-chaired the lab, DLM, BRS, LC, and DM processed and distributed the dataset, and DLM, DM and WWC led result evaluations

such as unfamiliar concepts and abbreviations [2],[3]. Indeed, a lack of medical language understanding can contribute to poor post-encounter care adherence when patients can not understand their discharge summary instructions [4].

Natural Language Processing (NLP) can help patients understand their health status by enriching PHI with meta-data (presenting patient-specific words and definitions for unfamiliar concepts and abbreviations) that assists them in understanding the content of clinical reports.

2 Background

2.1 Shared Task Annotations

Annotated datasets are often used to train NLP systems to convert narrative texts into machine computable representations. The annotated datasets serve as a reference standard (also known as gold standard or ground truth) and supports both system development and evaluation [5]. The reference standard must be both reliable and valid to provide the most optimal training and evaluation data. Since 2006, various shared-task challenges have provided reference standards for the clinical NLP community, including the CCHMC Computational Medicine Challenge and the i2B2 Shared Tasks. Topics for these shared tasks include assigning discharge diagnosis codes to radiology reports [6], for identifying clinical events and their relations [7], for finding personally identifiable information [8], for classifying patient smoking status [9], for determining obesity comorbidities [10], and for extracting medication mentions [11]. Reference standards are usually created by annotations of multiple domain experts with separate adjudication for disagreements.

Previous large-scale annotation efforts include the Message Understanding Conference (MUC) [12], Text Retrieval Conference (TREC) [13], [14], Clinical E-Science Framework (CLEF) [15], [16], GENIA [17], [18], and Penn Treebank [19]. Work by Roberts [15], [16], Savova[20], [21], Chapman [22], [23], [24], Uzuner [25], and previous i2b2 challenges [7], [9], [10], [26] provide context and motivation for this year's first ShARe/CLEF eHealth shared task including the development of the annotation guidelines, annotation schema, and evaluation of participant systems against the resulting reference standard. Continuing the tradition of shared tasks providing annotated data for development and evaluation of NLP systems for potentially useful applications, this year's ShARe/CLEF eHealth shared task focused on facilitating understanding of information in narrative clinical reports, such as discharge summaries, by identifying and normalizing disease/disorders (Task 1), normalizing acronym/abbreviations (Task 2), and retrieving documents from the health and medicine websites for addressing questions patients may have about the disease/disorders in the clinical notes (Task 3). The ShARe/CLEF eHealth Evaluation Lab is the first step towards a shared task that evaluates our ability to help patients and family members understand their clinical records. In this paper, we discuss Task 2.

2.2 NLP for Acronym and Abbreviation

AAs occurring in clinical texts present unique challenges to patient readers due to genre-specific senses (e.g., in an echocardiogram, "BP" likely represents "blood pressure" rather than "Bell's Palsy"), lack of parenthetical definitions (e.g., "HTN" (Hypertension)), and ambiguous uses or word senses (e.g., "MS" can mean "mental status" or "multiple sclerosis" even in the same report genre) [27]. To potentially aid patient readers in understanding clinical reports, Task 2 involved normalizing pre-annotated Acronyms and Abbreviations (AAs) to the Unified Medical Language System (UMLS) [28].

Researchers have developed systems to normalize AAs in clinical texts for information extraction, information retrieval, and document summarization applications. Wu [29] compared the performance of current, existing clinical NLP tools - MetaMap, MedLEE, and cTAKES - for identifying boundaries of and normalizing AAs in discharge summaries, which performed with F-scores ranging from 0.21-0.71 (boundary detection) and 0.03-0.73 (normalization). The MedLEE system outperformed MetaMap and cTAKES for all tasks; poor performances by MetaMap and cTAKES were attributed to a lack of clinical sense inventories or disambiguation modules. Automated disambiguations methods developed by Moon [30] showed promise for developing effective acronym sense disambiguation solutions using minimal training data. Moon achieved accuracies greater than 0.90 using a support vector machine trained on 125 samples encoded with words, part of speech tags, MetaMap concept unique identifiers, and sections.

Our long-term goal is to facilitate development and evaluation of automated NLP tools for enriching clinical reports with meta-data that assists patients, providers, and family members in understanding the content of the reports. An important foundational step towards this goal is mapping acronyms and abbreviations to their definitions and potentially to consumer-oriented dictionaries like the Consumer Health Vocabulary [31]. Next, we describe the annotation schema, the dataset, the annotation process, and the evaluation methods used for the ShARe/CLEF eHealth Evaluation Lab Task 2.

3 Methods

3.1 Annotation Schema

We developed annotation schema guidelines based on the study by Xu [27], iterative annotation of 10 development reports, and discussions among the co-authors. Similar to Xu [27], we instructed annotators to only annotate clinically relevant AAs. For instance, annotators did not include general English terms such as salutations ("Mr.") or time ("am"), but could include services ("EMS"), locations ("ICU"), section headers ("HEENT"), and medications (see Ex. 1-3 below). Once an AA was annotated, we instructed annotators to select the closest UMLS concept sense.

Ex 1: He was given Vanco. "Vanco" is a mention of type Acronym/Abbreviation with CUI C0042313 (UMLS preferred term is "Vancomycin")
Ex 2: Patient has breast ca. "ca" is a mention of type Acronym/Abbreviation with CUI C0006826 (UMLS preferred term is "Malignant Neoplasms")
Ex 3: Mitral Valve: Trivial MR. "MR" is a mention of type Acronym/Abbreviation with CUI C0026266 (UMLS preferred term is "Mitral Valve Insufficiency")

3.2 Dataset

We annotated AAs on top of the ShARe (Shared Annotated Resources) dataset, a stratified subset of 300 de-identified clinical reports from over 30,000 ICU patients stored in the MIMIC (Multiparameter Intelligent Monitoring in Intensive Care) II database [32]. The ShARe corpus consists of discharge summary, electrocardiogram, echocardiogram, and radiology report types annotated for disease/disorders, corresponding SNOMED codes, and attributes such as negation and severity. For Task 1, disease/disorder named entities and their SNOMED codes were released for the Evaluation Lab. For Task 2, we maintained the training (n=200 reports) and test (n=100 reports) dataset splits from Task 1.

To characterize our dataset, we split and tokenized sentences in the reports using NLTK (Natural Language ToolKit) [33]. We measured average **report length** by *count of sentences in a report*, average **sentence length** by *count of tokens in a sentence*, and average **token count** by *count of tokens in a report.*

3.3 Annotation Access and Process

Due to the nature of sensitive, patient-oriented information stored in clinical reports, a data access procedure was implemented. After registration for annotation with the University of California NLP Annotation Registry [34], annotators were required to obtain permission to access the ShARe dataset, which included (1) a CITI [35] or NIH [36] Training certificate in Human Subjects Research, (2) registration on the Physionet.org site [37], (3) signing a Data Use Agreement to access the Mimic II data. See the ShARe website [38] for details.

For annotation training, annotators were provided an annotation kit consisting of 1) the eHOST (extensibleHuman Oracle Suite of Tools) [39] for annotating the text, 2) a quick start guide for using the tool, 3) a Camtasia video for training the annotators, and 4) an annotation guideline for learning the task. These materials were reviewed with each annotator through an interactive annotation training session using join.me.

We incorporated both clinical professionals and informatics experts to generate a reference standard reflecting both domain knowledge and NLP annotation expertise. We recruited a total of 15 annotators; 11 completed the data access procedure and attempted to annotate the dataset.

The reference standard was annotated in three steps:

Step 1) 9 Finnish nursing professionals, 1 Australian nurse, and 1 Australian biomedical informatician were provided pre-annotated disease/disorder annotations from Task 1. They were instructed to span each AAs, then map each

concept to one CUI (concept unique identifier) from the UMLS. If a CUI did not exist in the vocabulary for the AA, the annotator was instructed to assign the label "CUI-less".

Step 2) One US biomedical informatician reviewed and adjudicated the annotated spans from Step 1 annotators.

Step 3) One US respiratory therapist reviewed and adjudicated the annotated spans from Step 2.

Annotators for Steps 2 and 3 were instructed to delete spurious, modify existing, and add missing AA spans as well as correct their CUI mappings.

3.4 Participant Recruitment and Registration

To recruit participants, we sent emails to relevant listservs, including Corpora, SigIR, BioNLP, AMIA NLP Working Group, and CLEF. After registration for tasks through the CLEF Evaluation Lab, participants were required to take the same steps as annotators to obtain permission to access the ShARe/CLEF eHealth dataset.

3.5 Evaluation Metrics

We calculated inter-annotator agreement for the reference standard annotations and calculated accuracy of the participant systems when compared against the reference standard.

Annotator Agreement We determined inter-annotator agreement by comparing annotations resulting from Step 2 against those resulting from Step 3 using the Evaluation Workbench [40]. Since the number of strings not annotated as AAs (i.e., ***true negatives (TN)***) is very large, we followed [41] in calculating F1-score as a surrogate for kappa. F1-score is the harmonic mean of recall and precision, calculated from true positive, false positive, and false negative annotations, which were calculated as follows:

true positive (TP) = the annotation from Step 2 had overlapping character offsets with the annotation from Step 3 and was assigned the same CUI

false positive (FP) = an annotation from Step 2 did not exist in Step 3 annotations

false negative (FN) = an annotation from Step 3 did not exist in Step 2 annotations

$$\textbf{\textit{Recall}} = \frac{TP}{(TP + FN)} \tag{1}$$

$$\textbf{\textit{Precision}} = \frac{TP}{(TP + FP)} \tag{2}$$

$$\textbf{\textit{F1-score}} = 2\frac{(Recall * Precision)}{(Recall + Precision)} \tag{3}$$

System Performance We evaluated performance of participating systems by calculating the accuracy of system performance against the manual AA annotations as follows: *Accuracy = count of correct AAs* divided by *total count of AAs*. We calculated *Strict Accuracy* based on the AA annotations resulting from Step 3 review. Because there is sometimes more than one CUI that matches an annotated AA, we also calculated *Relaxed Accuracy* by defining a *correct AA* annotation as a match with the CUI assigned during Step 3 review or during Step 2 review.

We evaluated system performance with random shuffling [42], a non-parametric statistical significance test, to compare the Accuracy scores between participating systems.

4 Results

4.1 Dataset

Table 1 shows the distribution of report types in our dataset. In spite of random selection of reports for training and test sets, we observed a lower proportion of discharge summaries and higher proportions of all other report types in the training set compared to the test set.

Table 1: Distribution of Report Types

	Training	Test
Report Type	Report Ct (%)	Report Ct (%)
Discharge Summary	62 (31%)	76 (76%)
Electrocardiogram	54 (27%)	0 (0%)
Echocardiogram	42 (21%)	12 (12%)
Radiology	42 (21%)	12 (12%)
Total	200 (100%)	100 (100%)

The dataset showed an overall average report length of 39 sentences, sentence length of 20 tokens, and token count of 683 tokens. We also characterized the dataset by report type illustrated in Table 2. We observed similar distributions of report length, sentence length, and token counts for both training and test sets, in spite of the differing distributions of report types.

4.2 Annotator Agreement

Inter-annotator agreement scores between Step 2 and Step 3 annotations was 0.85 for the training set and 0.91 for the test set.

Table 2: Average Distributions for Report Types

	Training			Test		
Report Type	sentence per report	token per sentence	token per report	sentence per report	token per sentence	token per report
Discharge Summary	66.9	24.4	1584.8	65.4	25.4	1573.3
Electrocardiogram	2.3	12.7	38.1	0	0	0
Echocardiogram	32.4	12.9	417.2	35.0	12.1	430.8
Radiology	13.3	22.8	290.6	11.8	24.7	281.6

4.3 System Performance

We received a total of 56 data requests for individual researchers. Participating teams included between 3-7 people and were comprised of scientists, engineers, professors, post doctoral fellows, and graduate students. Our participants competed from the US, Australia, France, and China. Participants represented academic and industrial institutions including University of Texas, Vanderbilt University, Massachusetts Institute of Technology, West Virginia University, University of Sydney, Computer Sciences Laboratory for Mechanics and Engineering Sciences, Tsinghua University, Canon Information Technology, and M*Modal.

In total, five teams submitted systems for Task 2. Two system submissions used external annotations for training; three system submissions used no external annotations for training. As shown in Table 3, the UTHealthCCB team system had the highest performance, with accuracies of 0.72 for strict and 0.73 for relaxed accuracy. Using the majority class "CUI-less", a baseline system evaluated with strict accuracy only achieved 0.06 accuracy (not shown below).

Table 3: Ranking in Task 2. *added external annotations; **significantly better than one below**

Team.system	Strict Accuracy	Relaxed Accuracy
UTHealthCCB.B.1	**0.719**	0.725
UTHealthCCB.B.2	**0.683**	0.689
LIMSI.1	**0.664**	0.672
TeamHealthLanguageLABS.1	0.467	0.488
THCIB.B.1*	**0.657**	0.685
WVU.1*	0.426	0.448

5 Limitations

There are several limitations to this study. We only focused on clinical AAs from four report types using a convenience corpus. An NLP system may need to

disambiguate AAs from a variety of other report types to aid patients. Although, our annotators represented a variety of clinical and domain expertise, we did not evaluate inter-annotator agreement between Step 1 annotators, nor did we evaluate whether there were differences between annotators with clinical vs non-clinical training.

6 Discussion

Using a step-wise annotation process that incorporated clinical and NLP expertise, we were able to generate a reference standard with high inter-annotator agreement. By developing automated systems, participants demonstrated that an NLP system can interpret clinical AAs with reasonably high accuracy. We observed only between 4-6% of AAs were "CUI-less" in the strict training and test sets suggesting reasonable UMLS coverage of clinical AA terms. This finding demonstrates significant improvements since the 2007 study by Xu. Xu [27] evaluated coverage of the UMLS and Medline's ADAM vocabulary for abbreviations, acronyms, shortened words, and contractions annotated in admission notes. In particular, Xu observed low to moderate coverage of abbreviations (56-67%), senses (24-38%), and ambiguities (33-71%). However, the UMLS is primarily used to normalize words to a medical vocabulary. In order to improve understanding of clinical text by patients, we plan to evaluate the coverage of clinical AAs from the task dataset against the Consumer Health Vocabulary. Future tasks will build on the annotations in this task and include more user-based evaluation metrics, such as how well users understand the content of clinical reports with meta-data such as definitions of acronyms and abbreviations.

Acknowledgments

We greatly appreciate the hard work and feedback of our program committee members and annotators, including, but not limited to Qing Zeng, Tyler Forbush, Jianwei Leng, Maricel Angel, Eriikka Siirala, Heljä Lundgren-Laine, Jenni Lahdenmaa, Marita Ritmala-Castren, Riitta Danielsson-Ojala, Saija Heikkinen, and Sini Koivula.

This shared task was partially supported by NICTA, funded by the Australian Government as represented by the Department of Broadband, Communications and the Digital Economy and the Australian Research Council through the ICT Centre of Excellence program, the CLEF Initiative, the European Science Foundation (ESF) project ELIAS, the Khresmoi project, funded by the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement no 257528, the ShARe project funded by the United States National Institutes of Health (R01GM090187), the US Department of Veterans Affairs (VA) Consortium for Healthcare Informatics Research (CHIR), the US Office of the National Coordinator of Healthcare Technology, Strategic Health IT Advanced Research Projects (SHARP) 90TR0002, the Vårdal Foundation (Sweden), and the National Library of Medicine 5T15LM007059.

References

1. Medical Protection Society: Online medical records and the doctor-patient partnership. MPS research report (2013)
2. Ross, S., Lin, C.: The effects of promoting patient access to medical records: A review. J Am Med Inform Assoc **10**(2) (2003) 129–138
3. Delbanco, T., Walker, J., Bell, S., Darer, J., Elmore, J., Farag, N., Feldman, H., Mejilla, R., Ngo, L., Ralston, J., Ross, S., Trivedi, N., Vodicaka, E., Leveille, S.: Inviting patients to dread their doctors' notes: a quasi-experimental study and look ahead. Ann Intern Med **157**(7) (2012) 461–470
4. Engel, K., Buckley, B., Forth, V., McCarthy, D., Ellison, E., Schmidt, M., Adams, J.: Patient understanding of emergency department discharge summary instructions: Where are knowledge deficits greatest? Acad Emerg Med **19**(9) (2012) E1035–E1044
5. Meystre, S., Savova, G., Kipper-Schuler, K., Hurdle, J.: Extracting information from textual documents in the electronic health record: a review of recent research. Yearb Med Inform **35** (2008) 128–44
6. Pestian, J., Brew, C., Matykiewicz, P., Hovermale, D., Johnson, N., Cohen, K., Duch, W.: A shared task involving multi-label classification of clinical free text. BioNLP 2007: Biological, translational, and clinical language processing (2007) 97–104
7. Uzuner, Ö., Mailoa, J., Ryan, R., Sibanda, T.: Semantic relations for problem-oriented medical records. Artif Intell Med **50**(2) (October 2010) 63–73
8. Uzuner, Ö., Luo, Y., Szolovits, P.: Viewpoint paper: Evaluating the state-of-the-art in automatic de-identification. J Am Med Inform Assoc **14**(5) (2007) 550–563
9. Uzuner, O., Goldstein, I., Luo, Y., Kohane, I.: Identifying patient smoking status from medical discharge records. J Am Med Inform Assoc **15** (2008) 14–24.
10. Uzuner, O.: Recognizing obesity and co-morbidities in sparse data. J Am Med Inform Assoc **16**(4) (2009) 561–570
11. Uzuner, Ö., Solti, I., Cadag, E.: Extracting medication information from clinical text. J Am Med Inform Assoc **17**(5) (2010) 514–518
12. Grishman, R., Sundheim, B.: Message Understanding Conference-6: a brief history. In: Proceedings of the 16th conference on Computational linguistics - Volume 1. COLING '96, Stroudsburg, PA, USA, Association for Computational Linguistics (1996) 466–471
13. Hersh, W., Bhupatiraju, R., Corley, S.: Enhancing access to the Bibliome: the TREC Genomics Track. Stud Health Technol Inform **107**(Pt 2) (2004) 773–7
14. Jones, K.: Reflections on TREC. In: Information Processing & Management 31(3). (1995) 29131–4
15. Roberts, A., Gaizauskas, R., Hepple, M., Davis, N., Demetriou, G., Guo, Y., Kola, J., Roberts, I., Setzer, A., Tapuria, A., Wheeldin, B.: The CLEF Corpus: Semantic Annotation of Clinical Text. In: AMIA Annu Symp Proc. (2007) 625–629
16. Roberts, A., Gaizauskas, R., Hepple, M., Davis, N., Demetriou, G., Guo, Y., Roberts, I., Setzer, A.: Building a semantically annotated corpus of clinical texts. J Biomed Inform **42**(5) (2009) 950–66
17. Kim, J., Ohta, T., Tateisi, Y., Tsujii, J.: GENIA corpus - a semantically annotated corpus for bio-textmining. In: ISMB (Supplement of Bioinformatics). (2003) 180–182
18. Kim, J., Ohta, T., Tsujii, J.: Corpus annotation for mining biomedical events from literature. BMC Bioinformatics **9** (2008)

19. Marcus, M., Beatrice, S., Marcinkiewicz, M.: Building a large annotated corpus of english: The Penn Treebank. Computational Linguistics **19**(2) (1993) 313–330
20. Savova, G., Coden, A., Sominsky, I., Johnson, R., Ogren, P., Groen, P.d., Chute, C.: Word sense disambiguation across two domains: Biomedical literature and clinical notes. J Biomed Inform **41**(6) (December 2008) 1088–1100
21. Savova, G., Ogren, P., Duffy, P., Buntrock, J., Chute, C.: Technical brief: Mayo Clinic NLP system for patient smoking status identification. J Am Med Inform Assoc **15** (2008) 25–28
22. Chapman, W., Dowling, J., Hripcsak, G.: Evaluation of training with an annotation schema for manual annotation of clinical conditions from emergency department reports. Int J Med Inform **77**(2) (2008) 107–13
23. Chapman, W., Bridewell, W., Hanbury, P., Cooper, G., Buchanan, B.: A simple algorithm for identifying negated findings and diseases in discharge summaries. J Biomed Inform **2001** (2001) 34–301
24. Chapman, W., Haug, P.: Comparing expert systems for identifying chest x-ray reports that support pneumonia. In: AMIA Annu Symp Proc. (1999) 216–220
25. Uzuner, Ö., Solti, I., Xia, F., Cadag, E.: Community annotation experiment for ground truth generation for the i2B2 medication challenge. J Am Med Inform Assoc **17**(5) (2010) 519–523
26. Uzuner, Ö., South, B., Shen, S., DuVall, S.: 2010 i2B2/VA challenge on concepts, assertions, and relations in clinical text. J Am Med Inform Assoc **18**(5) (2011) 552–556
27. Xu, H., Stetson, P., Friedman, C.: A study of abbreviations in clinical notes. In: AMIA Annu Symp Proc. (2007) 821–825
28. Campbell, K., Oliver, D., Shortliffe, E.: The Unified Medical Language System: Towards a collaborative approach for solving terminologic problems. J Am Med Inform Assoc **5**(1) (1998) 12–16
29. Wu, Y., Denny, J.C., Rosenbloom, S.T., Miller, R.A., Giuse, D.A., Xu, H.: A comparative study of current clinical natural language processing systems on handling abbreviations in discharge summaries. AMIA Annu Symp Proc **2012** (2012) 997–1003
30. Moon, S., Pakhomov, S., Melton, G.: Automated disambiguation of acronyms and abbreviations in clinical texts: Window and training size considerations. In: AMIA Annu Symp Proc. (2012) 1310–1319
31. CHV: Consumer Health Vocabulary. http://consumerhealthvocab.org/ Accessed: 2013-06-30.
32. Saeed, M., Lieu, C., Raber, G., Mark, R.: MIMIC II: a massive temporal ICU patient database to support research in intelligent patient monitoring. Comput Cardiol **29** (2002)
33. NLTK: Natural Language ToolKit. http://nltk.org/ Accessed: 2013-06-30.
34. Fana, F.: NLP Ecosystem: Annotation Registry. http://nlp-ecosystem.ucsd.edu/annotators Accessed: 2013-06-30.
35. CITI: Collaborative Institutional Training Initiative. https://www.citiprogram.org/ Accessed: 2013-06-30.
36. NIH: National Institute of Health - ethics training module. http://ethics.od.nih.gov/Training/AET.htm Accessed: 2013-06-30.
37. Physionet: Physionet site. https:http://www.physionet.org/ Accessed: 2013-06-30.
38. ShARe: ShARe CLEF eHealth Website. https://sites.google.com/site/shareclefehealth/data#TOC-Obtaining-Datasets-Tasks-1-and-2-/ Accessed: 2013-06-30.

39. South, B., Shen, S., Leng, J., TB, F., DuVall, S., Chapman, W.: A prototype tool set to support machine-assisted annotation. In: Proceedings of the 2012 Workshop on Biomedical Natural Language Processing. BioNLP '12, Stroudsburg, PA, USA, Association for Computational Linguistics (2012) 130–139
40. Christensen, L.: Evaluation Workbench. http://nlp-ecosystem.ucsd.edu/content/documentation Accessed: 2013-06-30.
41. Hripcsak, G., Rothschild, A.: Agreement, the f-measure, and reliability in information retrieval. J Am Med Inform Assoc **12**(3) 296–8
42. Yeh, A.: More accurate tests for the statistical significance of result differences. In: Proceedings of the 18th Conference on Computational Linguistics (COLING), Saarbrücken, Germany (2000) 947–953