The Gold Standard in Corpus Annotation

Conference Paper · June 2014

DOI: 10.13140/2.1.4316.3523

4 authors:

Lars Wissler
Freie Universität Berlin
4 PUBLICATIONS **28** CITATIONS
SEE PROFILE

Mohammed Almashraee
Freie Universität Berlin
6 PUBLICATIONS **41** CITATIONS
SEE PROFILE

Dagmar Monett
Hochschule für Wirtschaft und Recht Berlin
74 PUBLICATIONS **173** CITATIONS
SEE PROFILE

Adrian Paschke
Freie Universität Berlin
248 PUBLICATIONS **2,176** CITATIONS
SEE PROFILE

The Gold Standard in Corpus Annotation

Lars Wissler and Mohammed Almashraee
Free University Berlin
Institute of Computer Science
Berlin, Germany
lars.wissler@googlemail.com
almashraee@inf.fu-berlin.de

Dagmar Monett
Berlin School of Economics and Law
Department of Computer Science
Berlin, Germany
Dagmar.Monett-Diaz@hwr-berlin.de

Adrian Paschke
Free University Berlin
Institute of Computer Science
Berlin, Germany
paschke@inf.fu-berlin.de

***Abstract*—Trustworthy corpora are necessary for training and meaningful evaluation of algorithms which use annotations. These standard collections are called *Gold Standard Corpora (GSC)*. However the construction of GSC is a laborious and time-consuming process and size, quality and most of all availability of task-specific GSC directly influence the development of machine learning based natural language processing algorithms. This paper provides an introduction to gold standard corpus construction in the context of natural language processing and gives an overview of alternative approaches.**

I. INTRODUCTION

Modern applications operating on natural language corpora automatically create summaries of articles, act as powerful semantic search engines [1] or extract trends in the stock market [2]. They often require information, which is more structured than plain text - annotated corpora. Trustworthy corpora are necessary to train algorithms which use annotations, because errors in the training corpus propagate to the final system. Quite similarly accepted standards are required for a meaningful evaluation of automatic annotation algorithms [3]. These standard collections are called *Gold Standard Corpora (GSC)*.

However, the construction of GSC is a laborious and time-consuming process, which is usually performed by experts. Size, quality and most of all availability of task-specific GSC directly influence the development of machine learning based natural language processing (NLP) algorithms. Therefore the costs of creating corpora, which are sufficient to reliably train and evaluate machine learning (ML) based algorithms need to be minimized. Recent approaches in this area attempt to determine the optimal size of a corpus [4], evaluate the influence of non-expert annotators on the quality of a GSC [5] or try to substitute GSC with automatically generated silver standard corpora (SSC) [6]. This paper provides an introduction to GSC as well as an overview and an evaluation of alternative construction approaches.

II. CORPUS ANNOTATION

Annotating means to tag documents, sentences or words with a choice from a predefined set of categories [7]. Interesting items, such as entities or grammatical structures, are identified and thus textual data is enriched with additional structured information. What information is added depends on the purpose of the annotations. It can range from syntactical information over lexical knowledge to semantic associations. Processing and interpreting text automatically in natural language are challenging tasks, because the meaning of terms is context dependent. As Sinclair points out in [8], *'... the machines may be required to make definite decisions about structure sentence by sentence. In contrast the ordinary user may be keeping a number of provisional points in an undecided state [...]'.*

Corpus annotation is often a preprocessing step to facilitate advanced automatic text processing. Annotated text contains additional *well defined* knowledge and thus it becomes possible to analyze corpora based on their feature set instead of the full text.

The annotation of corpora has become essential for NLP tasks that rely on ML techniques [9].

It is a time consuming process to add tags to terms and NLP algorithms usually need many thousand documents to work well. Whereas low-level syntactical tagging algorithms' reliability is generally high, reaching error rates as low as 2%, automatic high-level semantic tagging is still not accurate enough to be used for practical purposes in many areas [9]. These task specific differences are not only due to the difficulty of the task but also due to the (non-)existence of large, high quality GSC.

III. GOLD STANDARD CORPUS

The term "Gold Standard" was originally coined in the economical domain and denotes a monetary system, in which the value of currency units is based on a quantity of physical gold [10]. The system is not in use anymore, but it is still viewed as an extremely stable financial system. The meaning of the term has since been transformed to denote scientific procedures or collections which are accepted standards.

Gold standard corpora in NLP context are manually annotated collections of text. For high quality gold standard corpora multiple experts view the data independently and the inter-annotator agreement is computed to ensure quality. This makes the creation of gold standard corpora a very costly process. The agreement of the annotators depends on the ambiguity of the data, skill of the annotators and the task at hand [5]. Especially, both the ambiguity of data and the task influence the agreement of annotators, because these determine not only the difficulty when choosing a tag but also the number of (wrong) choices. The authors of [11] studied the inter-annotator agreement for word sense disambiguation. They report a Kappa value of 0.463 for 53 nouns with an average of 7.6 meanings per noun. After collapsing the set of nouns so that the average number of meanings drops to 4, the Kappa score rises to 0.86.

As NLP tasks largely depend on domain and intended results, proper gold standard corpora are required for each domain and task. An algorithm for the extraction of genes requires for training and evaluation a gold standard corpus, which is annotated genes. Together with the necessity for dependable quality and reasonable size, this leads to the issue, that no proper GSC exists for many NLP tasks.

The Penn Treebank[1] is a large corpus, which in 1993 consisted of over 4.5 million of words of American English, which are tagged with syntactic information. Although the Penn Treebank corpus is not a true GSC, since it was produced using a combination of automatic tagging and manual correction, its size, high quality and early development made it the *de facto* GSC for syntactical tagging. In named entity recognition (NER), the most used standards are MUC, CoNLL and BBN [12]. MUC and CoNLL were created in the context of challenges and mark location, person and organisation entities. BBN supplements the Penn Treebank Corpus with the same information and additionally numerical and time data.

Many highly specialised gold standard corpora exist for the biomedical domain, ranging from genes over species to diseases. For example, *PennBioIE-Oncology*[2] contains 18148 annotations of genes, the *Arizona Disease*[3] corpus has 3206 annotation of diseases and the *SCAI-Test*[4] corpus contains 1206 annotations of chemical mentions. The recognition of genes, chemicals and diseases are among the most common applications of biomedical NER. It is apparent, that specialised and manually compiled GSC cannot compare to the semi-automatically created corpora like the Penn Treebank in terms of size. In many cases no GSC exists at all.

IV. REDUCING THE GSC COSTS

Not only does the lack of proper GSC hinder the development of ML based NLP algorithms but, more important, it prohibits a meaningful evaluation of such algorithms, since no reliable test sets are available. To reduce the cost of the creation of GSC, several approaches are possible: (1) Reducing the amount of annotations. (2) Reducing the number of

[1] http://www.cis.upenn.edu/~treebank/cdrom2.html

[2] http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2008T21

[3] http://diego.asu.edu/downloads/AZDC

[4] http://www.scai.fraunhofer.de/chem-corpora.html

reviews. (3) Using non-expert annotators via crowd-sourcing. (4) Replacing or supplementing GSC with automatically created SSC.

Reducing the GSC size is a valid approach if it does not significantly impair classifiers that are trained. The influence of the training corpus size on accuracy and precision is evaluated in [13], resulting in minimal changes for large sizes. This points to a threshold in GSC size, that limits the usefulness of large GSC. A method to determine that threshold depending on the number of features and their class size is proposed in [4]. In [5], the authors compare the inter-annotator agreement of experts with non-experts from Amazon Mechanical Turk[5]. Their results show high agreement of over 90% for experts, leading to the conclusion that multiple reviews, as it is common practice, are not necessary for expert annotators. The same study shows a similarly high agreement for experts and non-experts of 92%; however, a low Kappa measure of 0.62 compared to 0.76 for experts. A similar study reports significantly lower agreements of 51% for non-experts and 61% for experts in correlation to a pool of experts and non-experts. However, it shows that four non-expert labeler reach on average the same correlation as an expert [14].

The idea of SSC is fairly new. In the context of the CALBC[6] project, an effort to create an SSC is made, which uses harmonized annotations produced by multiple automatic annotation systems [6]. The study [13] evaluates the usefulness of the CALBC-SSC-I corpus as GSC. When evaluating, it trains the ML-based biological NER system BioEnEx on the SSC and compares the results to the entities in the gene GSC BioCreAtIve II GM corpus. It shows that classifiers trained on the SSC reached slightly lower precision than GSC trained classifiers (80% vs. 87%); however, it also reports a significant drop in recall resulting in a F-score of 54% for the SSC compared to 86% for the GSC. By removing sentences without annotations, the F-score rises to 62%. By adding an GSC of equal size to the SSC and using the combined corpus for training, the F-score was raised to 77%.

[5]http://mturk.com

[6]http://www.ebi.ac.uk/Rebholz-srv/CALBC/project.html

V. DISCUSSION

Using SSC that are automatically created by multiple annotation algorithms is a promising approach. However, previous studies suggest, that the different potentially disagreeing annotation systems lead to many omitted genes in CALBC-SCC, which in turn leads to low recall. The performance of the SSC can be increased, as shown in the previously discussed study. But without the combination with a GSC, the resulting F-score is more than 20% lower as opposed to using the GSC. This is why the substitution of an GSC with an SSC, as described above, is only sensible for applications that do not require high recall. The approach to minimize the size of the GSC is of course valid, but adequate sizes still require thousands of tokens [13]. The contrasting inter-annotator agreement scores of 92% reported in the first and 51% to 61% reported in the second study show the risks of creating a GSC without multiple reviews, because the agreement is highly dependent on the corpus and the task.

This leaves the use of crowd-sourcing with a large number of non-expert annotators as the most viable option to reduce the cost of GSC creation while retaining the high quality necessary for training and evaluation purposes. Both mentioned studies show good results when using non-experts as annotators and the reported factor of four for non-experts to reach satisfactory agreement will still be cheaper than acquiring one domain expert for the task. The annotation system KAFNotator enables through its role-system of 'blind annotator' and 'annotator, who resolves conflicts' the controlled annotation of a corpus where non-experts are assigned as blind annotators and experts resolve conflicts if there is disagreement [15].

Active learning (AL) seems to be predestined for GSC creation, which is a semi-supervised machine learning technique used to train classifiers. Active learning requires manual annotation, reduces the necessary amount of tokens for equal performance and preselects documents for annotation based on a small initial set. It selects the most promising data-points from a corpus for human annotation in such a way, that it maximizes the knowledge gain per tag, thereby decreasing the overall number of tags

necessary to reach a certain accuracy. Theoretically an exponential cost reduction using AL is possible; for applications in corpus annotation reduction rates of 48% to 72% have been reported [9]. With the use of AL for document selection and collaborative crowd-sourcing for annotation, the cost of creating GSC can be reduced while retaining high quality through experts reviewing and resolving conflicts.

VI. CONCLUSION

Annotation is a powerful mechanism for storing, reusing and analyzing information. The possibilities and techniques to create and use annotated text have evolved from scribbles in a book to digital and structured data collections. While annotation techniques work well in some areas, i.e. grammatical mark up, in other areas further advancement is necessary for reliable results. Faster and more accurate algorithms can and will revolutionize many tasks, in which processing large data sets is crucial. However, in order to develop and to evaluate these algorithms, reliable corpora for testing and training are necessary, the gold standard corpora. In many areas such corpora are unavailable and despite numerous efforts, the creation of GSC is still a costly process due to the manual labour involved. Crowdsourcing, expert review and active selection schemes can significantly reduce that cost while retaining high quality.

REFERENCES

[1] M. Almashraee, "Feature extraction based on semantic sentiment analysis," in *Business Information Systems Workshops*. Springer, 2013, pp. 270–277.

[2] O. Streibel, L. Wißler, R. Tolksdorf, and D. Montesi, "Trend template: mining trends with a semi-formal trend model," in *Workshop on Ubiquitous Data Mining*, Beijing, Peking, 2013, p. 49.

[3] A. Kilgarriff, "Gold standard datasets for evaluating word sense disambiguation programs," *Computer Speech & Language*, vol. 12, no. 4, pp. 453–472, 1998.

[4] D. Juckett, "A method for determining the number of documents needed for a gold standard corpus," *Journal of biomedical informatics*, vol. 45, no. 3, pp. 460–470, 2012.

[5] S. Nowak and S. Rüger, "How reliable are annotations via crowdsourcing: a study about inter-annotator agreement for multi-label image annotation," in *Proceedings of the international conference on Multimedia information retrieval*. ACM, 2010, pp. 557–566.

[6] D. Rebholz-Schuhmann, A. J. Jimeno-Yepes, E. M. van Mulligen, N. Kang, J. A. Kors, D. Milward, P. Corbett, E. Buyko, K. Tomanek, E. Beisswanger *et al.*, "The calbc silver standard corpus for biomedical named entities-a study in harmonizing the contributions from four independent named entity taggers." in *LREC*, 2010.

[7] A. Hinze, R. Heese, M. Luczak-Rösch, and A. Paschke, "Semantic enrichment by non-experts: Usability of manual annotation tools," in *International Semantic Web Conference (1)*, 2012, pp. 165–181.

[8] J. Sinclair, *The automatic analysis of corpora: proceedings of Nobel Symposium 82 Stockholm, 4-8 Aug*. Walter de Gruyter, 1992, vol. 65.

[9] K. Tomanek, J. Wermter, and U. Hahn, "An approach to text corpus construction which cuts annotation costs and maintains reusability of annotated data," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*. ACL Press, 2007, pp. 486–495.

[10] M. D. Bordo and H. Rockoff, "The gold standard as a good housekeeping seal of approval," *The Journal of Economic History*, vol. 56, no. 02, pp. 389–428, 1996.

[11] C. Yong and S. K. Foo, "A case study on inter-annotator agreement for word sense disambiguation," *SIGLEX*, vol. 99, pp. 9–13, 1999.

[12] J. Nothman, T. Murphy, and J. R. Curran, "Analysing wikipedia and gold-standard corpora for ner training," in *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, ser. EACL '09. Stroudsburg, PA, USA: Association for Computational Linguistics, 2009, pp. 612–620. [Online]. Available: http://dl.acm.org/citation.cfm?id=1609067.1609135

[13] F. M. Chowdhury and A. Lavelli, "Assessing the practical usability of an automatically annotated corpus," in *Proceedings of the 5th Linguistic Annotation Workshop*, ser. LAW V '11. Stroudsburg, PA, USA: Association for Computational Linguistics, 2011, pp. 101–109. [Online]. Available: http://dl.acm.org/citation.cfm?id=2018966.2018978

[14] R. Snow, B. O'Connor, D. Jurafsky, and A. Y. Ng, "Cheap and fast—but is it good?: Evaluating non-expert annotations for natural language tasks," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, ser. EMNLP '08. Stroudsburg, PA, USA: Association for Computational Linguistics, 2008, pp. 254–263. [Online]. Available: http://dl.acm.org/citation.cfm?id=1613715.1613751

[15] M. Tesconi, F. Ronzano, S. Minutoli, C. Aliprandi, and A. Marchetti, "KAFnotator: a multilingual semantic text annotation tool," in *Proceedings of the Second International Conference on Global Interoperability for Language Resources*, 2010.