Optimizing annotation efforts to build reliable annotated corpora for training statistical models

Cyril Grouin[1] **Thomas Lavergne**[1,2] **Aurélie Névéol**[1]
[1] LIMSI–CNRS, 91405 Orsay, France [2] Université Paris Sud 11, 91400 Orsay, France
firstname.lastname@limsi.fr

Abstract

Creating high-quality manual annotations on text corpus is time-consuming and often requires the work of experts. In order to explore methods for optimizing annotation efforts, we study three key time burdens of the annotation process: (i) multiple annotations, (ii) consensus annotations, and (iii) careful annotations. Through a series of experiments using a corpus of clinical documents annotated for personally identifiable information written in French, we address each of these aspects and draw conclusions on how to make the most of an annotation effort.

1 Introduction

Statistical and Machine Learning methods have become prevalent in Natural Language Processing (NLP) over the past decades. These methods sucessfully address NLP tasks such as part-of-speech tagging or named entity recognition by relying on large annotated text corpora. As a result, developping high-quality annotated corpora representing natural language phenomena that can be processed by statistical tools has become a major challenge for the scientific community. Several aspects of the annotation task have been studied in order to ensure corpus quality and affordable cost. Inter-annotator agreement (IAA) has been used as an indicator of annotation quality. Early work showed that the use of automatic pre-annotation tools improved annotation consistency (Marcus et al., 1993). Careful and detailed annotation guideline definition was also shown to have positive impact on IAA (Wilbur et al., 2006).

Efforts have investigated methods to reduce the human workload while annotating corpora. In particular, active learning (Settles et al., 2008) sucessfully selects portions of corpora that yield the most benefit when annotated. Alternatively, (Dligach and Palmer, 2011) investigated the need for double annotation and found that double annotation could be limited to carefully selected portions of a corpus. They produced an algorithm that automatically selects portions of a corpus for double annotation. Their approach allowed to reduce the amount of work by limiting the portion of doubly annotated data and maintained annotation quality to the standard of a fully doubly annotated corpus. The use of automatic pre-annotations was shown to increase annotation consistency and result in producing quality annotation with a time gain over annotating raw data (Fort and Sagot, 2010; Névéol et al., 2011; Rosset et al., 2013). With the increasing use of crowdsourcing for obtaining annotated data, (Fort et al., 2011) show that there are ethic aspects to consider in addition to technical and monetary cost when using a microworking platform for annotation. While selecting the adequate methods for computing IAA is important (Artstein and Poesio, 2008) for interpreting the IAA for a particular task, annotator disagreement is inherent to all annotation tasks. To address this situation (Rzhetsky et al., 2009) designed a method to estimate annotation confidence based on annotator modeling. Overall, past work shows that creating high-quality manual annotations is time-consuming and often requires the work of experts. The time burden is distributed between the sheer creation of the annotations, the act of producing multiple annotations for the same data and the subsequent analysis of multiple annotations to resolve conflicts, viz. the creation of a consensus. Research has addressed methods for reducing the time burden associated to these annotation

LAW VIII - The 8th Linguistic Annotation Workshop, pages 54–58,
Dublin, Ireland, August 23-24 2014.

activities (for example, adequate annotation tools such as automatic pre-annotations can reduce the time burden of annotation creation) with the final goal of producing the highest quality of annotations.

In contrast, our hypothesis in this work is that annotations are being developed for the purpose of training a machine learning model. Therefore, our experiments consist in training a named entity recognizer on a training set comprising annotations of varying quality to study the impact of training annotation quality on model performance. In order to explore methods for optimizing annotation efforts for the development of training corpora, we revisit the three key time burdens of the annotation process on textual corpora: (i) careful annotations, (ii) multiple annotations, and (iii) consensus annotations. Through a series of experiments using a corpus of French clinical documents annotated for personally identifiable information (PHI), we address each of these aspects and draw conclusions on how to make the most of an annotation effort.

2 Material and methods

2.1 Annotated corpus

Experiments were conducted with a corpus of clinical documents in French annotated for 10 categories of PHI. The distribution of the categories over the corpus varies with some categories being more prevalent than others. In addition, the performance of entity recognition for each type of PHI also varies (Grouin and Névéol, 2014). The datasets were split to obtain a training corpus (200 documents) and a test corpus (100 documents). For all documents in the training corpus, three types of human annotations are available: annotations performed independently by two human annotators and consensus annotations obtained after adjudication to resolve conflicts between the two annotators. Inter-annotator agreement on the training corpus was above 85% F-measure, which is considered high (Artstein and Poesio, 2008).

The distribution of annotations over all PHI categories on both corpora (train/dev) is: address (188/100), zip code (197/97), date (1025/498), e-mail (119/57), hospital (448/208), identifier (135/76), last name (1855/855), first name (1568/724), telephone (802/386) and city (450/217).

2.2 Automatic Annotation Methods

We directly applied the MEDINA rule-based de-identification tool (Grouin, 2013) to obtain baseline automatic annotations. We used the CRF toolkit Wapiti (Lavergne et al., 2010) to train a series of models on the various sets of annotations available for the training corpus.

Features set For each CRF experiment, we used the following set of features with a $l1$ regularization:

- Lexical features: unigram and bigram of tokens;
- Morphological features: (i) the token case *(all in upper/lower case, combination of both)*, (ii) the token is a digit, (iii) the token is a punctuation mark, (iv) the token belongs to a specific list *(first name, last name, city)*, (v) the token was not identified in a dictionary of inflected forms, (vi) the token is a trigger word for specific categories *(hospital, last name)*;
- Syntactic features: (i) the part-of-speech (POS) tag of the token, as provided by the Tree Tagger tool (Schmid, 1994), (ii) the syntactic chunk the token belongs to, from a home made chunker based upon the previouses POS tags;
- External features: (i) we created 320 classes of tokens using Liang's implementation (Liang, 2005) of the Brown clustering algorithm (Brown et al., 1992), (ii) the position of the token within the document *(begining, middle, end)*.

Design of experiments The models were built to assess three annotation time-saving strategies:

1. Careful annotation: (i) AR=based on automatic annotations from the rule-based system, (ii) AR∩H2=intersection of automatic annotations from the rule-based system with annotations from annotator 2. This model captures a situation where the human annotator would quickly revise the automatic annotations by removing errors: some annotations would be missing (average recall), but the annotations present in the set would be correct (very high precision), (iii) ARH2=automatic annotations from the rule-based system, with replacement of the three most difficult categories by

annotations from annotator 2. This model captures a situation where the human annotator would focus on revising targeted categories, and (iv) ARHC=automatic annotations from the rule-based system, with replacement of the three most difficult categories by consensus annotations;

2. Double annotation: (i) H1=annotations from annotator 1, (ii) H2=annotations from annotator 2, (iii) H12=first half of the annotations from annotator 1, second half from annotator 2, and (iv) H21=first half of the annotations from annotator 2, second half from annotator 1;
3. Consensus annotation: (i) H1∪H2=all annotations from annotator 1 and 2 (concatenation without adjudication), and (ii) HC=consensus annotations (after adjudication between annotator 1 and 2).

3 Results

Table 1 presents an overview of the global performance of each annotation run (H12 and H21 achieved similar results) across all PHI categories in terms of precision, recall and F_1-measure (Manning and Schütze, 2000). Table 2 presents the detailed performance of each annotation run for individual PHI categories in terms of F-measure.

	Baseline	**AR**	**AR∩H2**	**ARH2***	**ARHC**	**H1***	**H12**	**H1∪H2**	**H2**	**HC**
Precision	.820	.868	.920	.942	.943	.959	.962	.969	.974	.974
Recall	.806	.796	.763	.854	.854	.927	.934	.935	.936	.942
F-measure	.813	.830	.834	.896	.896	.943	.948	.951	.955	.958

Table 1: Overall performance for all automatic PHI detection. A star indicates statistically significant difference in F-measure over the previous model (Wilcoxon test, $p<0.05$)

Category	**Baseline**	**AR**	**AR∩H2**	**ARH2**	**ARHC**	**H1**	**H12**	**H1∪H2**	**H2**	**HC**
Address	.648	.560	.000	.800	.800	.716	.744	.789	.795	.791
Zip code	.950	.958	.947	.964	.958	.974	.984	.974	.984	.990
Date	.958	.968	.962	.963	.967	.965	.963	.963	.959	.970
E-mail	.937	.927	.927	.927	.927	1.000	1.000	1.000	1.000	1.000
Hospital	.201	.248	.039	.856	.868	.789	.809	.856	.861	.867
Identifier	.000	.000	.000	.762	.797	.870	.892	.823	.836	.876
Last name	.816	.810	.834	.832	.828	.953	.957	.954	.961	.963
First name	.849	.858	.900	.901	.902	.960	.956	.961	.965	.960
Telephone	1.000	.980	.978	.983	.980	.987	.994	.999	.999	1.000
City	.869	.874	.883	.887	.887	.948	.972	.962	.965	.972

Table 2: Performance per PHI category (F-measure)

4 Discussion

4.1 Model performance

Overall, the task of automatic PHI recognition has been well studied and the rule-based tool provides a strong baseline with .813 F-measure on the test set. Table 1 shows that there are three different types of models, in terms of performance: the lower-performing category corresponds to models with no human input. The next category corresponds to models with some human input, and the higher-performing models correspond to models with the most human input. This reflects the expectation that model performance increases with training corpus quality. However, it also shows that, within the two categories that include human input, there is no statistical difference in model performance with respect to the type of human input. We observed that the model trained on annotations from the H2 human annotator performed better (F=0.955) than the model trained on annotations from the H1 annotator (F=0.943). This observation reflects the agreement of the annotators with consensus annotations, where H2 had higher agreement than H1 (Grouin and Névéol, 2014). This is also true at the category level: H2 achieved

higher agreement with the consensus compared to H1 on categories "address" (F=0.985>0.767) and "hospital" (F=0.947>0.806) but H2 had lower agreement with the consensus on the category "identifier" (F=0.840<0.933).

4.2 Error Analysis

The performance of CRF models depends on the size of the training corpus and the level of diversity of the mentions. Error analysis on our test data shows that a few specific mentions are not tagged in the test corpus, even though they occur in the training corpus. For example, some hospital names occur in the clinical narratives either as acronyms or as full forms (e.g. "GWH" for "George Washington Hospital" in *transfer patient from GWH*). The acronyms are overall much less prevalent than the full forms and also happen to be difficult to identify for human annotators (depending on the context, a given acronym could refer to either a medical procedure, a physician or a hospital). We observed that the only hospital acronym present in the test corpus was not annotated by any of the CRF models. Nevertheless, only five occurrences of this acronym were found in the training corpus which is not enough for the CRF to learn.

Other errors occur in recognizing sequences of doctors' names that appear without separators in signatures lines at the end of documents (e.g. "*Jane BROWN John DOE Mary SMITH*"). In our test set we observed that models trained on automatic annotations correctly predicted the beginning of such sequences and then produced erroneous predictions for the rest of the sequence (models AR, AR∩H2, ARHC and ARH2). In contrast, models built on human annotations produced correct predictions on the entire sequence (models H1, H12, H1∪H2, H2 and HC). Similarly, for last names containing a nobiliary particle, the models trained on automatic annotations only identified part of the last name as a PHI. We also observed that spelling errors (e.g. "*Jhn DOE*") only resulted in correct predictions from the models trained on the human annotations. We did not find cases where the models built on the automatic annotations performed better than the models built on the human annotations.

4.3 Annotation strategy

Table 1 indicates that for the purpose of training a machine learning entity recognizer, all types of human input are equivalent. In practice, this means that double annotations or consensus annotations are not necessary. The high inter-annotator agreement on our dataset may be a contributing factor for this finding. Indeed, (Esuli et al., 2013) found that with low inter-annotator agreement, models are biased towards the annotation style of the annotator who produced the training data. Therefore, we believe that inter-annotator should be established on a small dataset before annotators work independently. Table 2 shows that using human annotations for selected categories results in strong improvement of the performance over these categories ("address", "hospital" and "identifier" categories in ARHC and ARH2 vs. AR) with little impact on the performance of the model on other categories. Therefore, careful human annotations are not necessarily needed for the entire corpus. Targeting "hard" categories for human annotations can be a good time-saving strategy. While the difference between the models using some human input vs. all human input is statistically significant, the performance gain is lower than between models without human input and some human input. Using data with partial human input for training statistical models can cut annotation cost.

5 Conclusion and future work

Herein we have shown that full double annotation of a corpus is not necessary for the purpose of training a competitive CRF-based model. Our results suggest that a three-step annotation strategy can optimize the annotation effort: (i) double annotate a small subset of the corpus to ensure human annotators understand the guidelines; (ii) have annotators work independently on different sections of the corpus to obtain wide coverage; and (iii) train a machine-learning based model on the human annotations and apply this model on a new dataset.

In future work, we plan to re-iterate these experiments on a different type of entity recognition task where inter-annotator agreement may be more difficult to achieve, and may vary more between categories in order to investigate the influence of inter-annotator-agreement on our conclusions.

Acknowledgements

This work was supported by the French National Agency for Research under grant CABeRneT[1] ANR-13-JS02-0009-01 and by the French National Agency for Medicines and Health Products Safety under grant Vigi4MED[2] ANSM-2013-S-060.

References

Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Peter F Brown, Vincent J Della Pietra, Peter V de Souza, Jenifer C Lai, and Robert L Mercer. 1992. Class-based n-gram models of natural language. *Computational Linguistics*, 18(4):467–79.

Dmitriy Dligach and Martha Palmer. 2011. Reducing the need for double annotation. In *Proc of Linguistic Annotation Workshop (LAW)*, pages 65–73. Association for Computational Linguistics.

Andrea Esuli, Diego Marcheggiani, and Fabrizio Sebastiani. 2013. An enhanced CRFs-based system for information extraction from radiology reports. *J Biomed Inform*, 46(3):425–35, Jun.

Karën Fort and Benoît Sagot. 2010. Influence of pre-annotation on POS-tagged corpus development. In *Proc of Linguistic Annotation Workshop (LAW)*, pages 56–63. Association for Computational Linguistics.

Karën Fort, Gilles Adda, and Kevin Bretonnel Cohen. 2011. Amazon Mechanical Turk: Gold Mine or Coal Mine? *Computational Linguistics*, pages 413–420.

Cyril Grouin and Aurélie Névéol. 2014. De-identification of clinical notes in french: towards a protocol for reference corpus development. *J Biomed Inform*.

Cyril Grouin. 2013. *Anonymisation de documents cliniques : performances et limites des méthodes symboliques et par apprentissage statistique*. Ph.D. thesis, University Pierre et Marie Curie, Paris, France.

Thomas Lavergne, Olivier Cappé, and François Yvon. 2010. Practical very large scale CRFs. In *Proceedings the 48th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 504–513. Association for Computational Linguistics.

Percy Liang. 2005. Semi-supervised learning for natural language. Master's thesis, MIT.

Chiristopher D. Manning and Hinrich Schütze. 2000. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, Massachusetts.

Mitchell P Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. 1993. Building a large annotated corpus of English: the Penn TreeBank. *Computational Linguistics*, 19(2):313–330.

Aurélie Névéol, Rezarta Islamaj Doğan, and Zhiyong Lu. 2011. Semi-automatic semantic annotation of PubMed queries: a study on quality, efficiency, satisfaction. *J Biomed Inform*, 44(2):310–8.

Sophie Rosset, Cyril Grouin, Thomas Lavergne, Mohamed Ben Jannet, Jérémy Leixa, Olivier Galibert, and Pierre Zweigenbaum. 2013. Automatic named entity pre-annotation for out-of-domain human annotation. In *Proc of Linguistic Annotation Workshop (LAW)*, pages 168–177. Association for Computational Linguistics.

Andrey Rzhetsky, Hagit Shatkay, and W John Wilbur. 2009. How to get the most out of your curation effort. *PLoS Comput Biol*, 5(5):e1000391.

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proc of International Conference on New Methods in Language*.

Burr Settles, Mark Craven, and Lewis Friedland. 2008. Active learning with real annotation costs. In *Proc of the NIPS Workshop on Cost-Sensitive Learning*.

W John Wilbur, Andrey Rzhetsky, and Hagit Shatkay. 2006. New directions in biomedical text annotation: definitions, guidelines and corpus construction. *BMC Bioinformatics*, 25(7):356.

[1]CABeRneT: Compréhension Automatique de Textes Biomédicaux pour la Recherche Translationnelle

[2]Vigi4MED: Vigilance dans les forums sur les Médicaments