Controlled propagation of concept annotations in textual corpora

Cyril Grouin
LIMSI, CNRS, Université Paris-Saclay, F-91405 Orsay
cyril.grouin@limsi.fr

Abstract

In this paper, we presented the annotation propagation tool we designed to be used in conjunction with the BRAT rapid annotation tool. We designed two experiments to annotate a corpus of 60 files, first not using our tool, second using our propagation tool. We evaluated the annotation time and the quality of annotations. We shown that using the annotation propagation tool reduces by 31.7% the time spent to annotate the corpus with a better quality of results.

Keywords: Annotation propagation; Corpus annotation; Texts

1. Introduction

1.1. Corpus annotation

Corpus annotation is a crucial step to develop suitable natural language processing (NLP) systems, to carry out evaluations of system outputs, or to train statistical models while using supervised machine-learning approaches (e.g., conditional random fields (Lafferty et al., 2001) for sequence labelling). Nevertheless, corpus annotation is a really time-consuming task.

A useful way to reduce the time spent to annotate corpora consists in providing the annotators a pre-annotated version of the corpus to annotate. Automatic pre-annotations can be made through a lexicon mapping (i.e., all existing entities found in a lexicon would be automatically pre-annotated) or a system designed to annotate entities, either using a rule-based system or a machine-learning approach. The choice of the method used to pre-annotate corpora depends on the type of entity to process: regular entities such as numeric values can be formalized using rules while more complex entities or contextual annotations would be processed using statistical approaches. Annotators working on automatic pre-annotations have to check those annotations, in order to remove non relevant annotations and to complete missing annotations. In a previous study, we demonstrated that using automatic pre-annotation based on CRF system both reduces the time spent by humans (annotators spent about 10% less time) and improves the quality of the final annotations (we computed a gain of 6 points in κ inter-annotator agreements), in comparison with annotation task made on similar raw corpora (Grouin and Névéol, 2014).

Another solution to reduce annotation time consists in selecting documents from a corpus, parts of documents, or parts of text (e.g., a few sentences), using a sampling process (Patton and Potok, 2006; Kantner et al., 2011), in order to annotate only a few samples of corpora. Those samples are considered by the corpus manager as representative enough of phenomenon to annotate and study.

1.2. Annotation propagation

The basic principle of annotation propagation relies on existing annotations that will be associated with new documents, for which parts—either a single word or a whole part depending on the type of annotation to be made—are found to be similar with previously annotated documents. Two main objectives are expected while using annotation propagation systems: first, the reduction of time spent by humans to annotate corpora, and second, an improvement of the final quality of annotations made. As a consequence, human annotators can focus on unseen annotations.

Existing systems designed to propagate and enrich human annotations—either semantic annotations or POS tagging—use external ressources such as deep parser (Swift et al., 2004), meta-data and ontologies (Zonta Pastorello Jr et al., 2010), as well as transformation rules and graph (Lansdall-Welfare et al., 2012). Existing annotations can be proposed to the user through an interactive system, as done by Voutilainen (2012) for a POS tagging task.

Some existing systems take advantage of several sources of distinct type to enrich and propagate annotations made by humans. Chevallet et al. (2006) designed a propagation annotation system for medical image annotation based on visual similarity. Their approach relies on concept extraction from texts in order to duplicate those concepts for images which share visual similarity. Budnik et al. (2014) proposed a multimodal system (speaker diarization and face clustering) in order to manually annotate persons in TV shows.

In this paper, we present the tool[1] we designed to automatically propagate annotations on textual corpora and the experiments we made to evaluate such propagations. Our experiments rely on the BRAT Rapid Annotation Tool, a system designed by Stenetorp et al. (2012) to annotate text corpora through a browser. Although a few functionalities are provided with this tool (e.g., keyboard shortcuts), no propagation annotation plugin exists. Since sophisticated propagation annotation tools still exist, our motivation was to produce a tool with basic functionalities so as to rapidly propagate relatively slight ambiguous annotations (e.g., named entity vs. POS tagging). The aim of this tool consists to improve both annotation quality and time processing.

We originally designed this tool to manually annotate a corpus of 13,500 clinical records so as to produce a fully de-identified corpus. In this paper, we applied this tool on a corpus composed of messages from a pharmacovigilance forum. We draw similar conclusions on both corpora.

[1]The tool is available at: `https://perso.limsi.fr/grouin/propagation-BRAT-annotation.tar.gz`

2. Material and methods

2.1. Corpus

2.1.1. Presentation

Our corpus is composed of 60 files corresponding to messages written in French and posted on the `meamedica.fr` website. This website allows the users to report adverse drug reactions they experienced.

2.1.2. Annotations

Guidelines The annotation work we focus on relies on 16 categories of concepts relying on medical treatments, clinical information, and additional information. Those annotations are then used to produce systems to automatically identify drug names and adverse drug reactions as reported by patients in messages from health forums (Morlane-Hondère et al., 2016).

We used the following categories of concepts in our annotation task, following the guidelines we defined (Grouin, 2015), mainly based on the semantic types from the UMLS (Lindberg et al., 1993) and completed by useful categories for an adverse drug reaction identification task: *Chemical or drug; Dosage; Concentration; Mode of administration; Anatomical part; Gene or protein; Biological process or function; Disorders; Sign or symptom; Medical procedure; Date; Duration; Frequency; Time; Weight; Job.*

Despite this annotation framework, our propagation annotation tool can be used for every annotation task on text data using the BRAT stand-off annotation schema.

Statistics Table 1 presents the numbers of annotations for each category from our corpus, for a total number of 651 annotations. We observe that *Sign or symptom* and *Anatomical part* constitute the two main categories of information to annotate (i.e., 53.3% of all annotations). Entities from those categories are found in all documents from the corpus, since adverse drug reactions mainly involve a problem *(Sign or symptom)* and a location in the body *(Anatomical part)*. As an example, the sentence *I'm suffering from back pain* combines the anatomical part "back" with the symptom "pain". Additional information can be found such as intensity marker (e.g., *severe back pain*) or frequency marker (e.g., *chronic back pain*).

Category	#	Category	#
Chemical or drug	76	Sign or symptom	254
Dosage	32	Medical procedure	37
Concentration	1	Date	0
Mode of administration	19	Duration	18
Anatomical part	93	Frequency	16
Gene or protein	0	Time	13
Biological process	53	Weight	4
Disorders	19	Job	16

Table 1 – Number of annotations for each category

Table 2 shows the number of annotations depending on the number of tokens in each annotated part in the corpus. While the main number of annotations concerns parts of text composed of single words (583 annotations, i.e., 89.6% of all annotations), a few annotations imply longer sequences (up to 7 tokens).

Number of tokens	1	2	3	4	5	7
Number of annotations	583	43	16	5	4	1

Table 2 – Number of annotations for each size (number of tokens) of annotated parts

Longer annotations (more than 2 tokens) only concern the category *Duration*, composed of temporal marker, number, and unit: *pendant un peu plus d' 1 mois* ("for slightly more than one month"), *cela fait déjà 7 ans* ("it has been almost 7 years"), *depuis plus d' un an* ("for over a year now"), etc.

2.2. Method

2.2.1. Annotation tool

The BRAT rapid annotation tool relies on stand-off annotations: each text file is associated with its annotation file. Annotation files are composed of three columns separated by a tabulation: (i) annotation ID, (ii) entity type, beginning and ending offset of characters for the annotated phrase, and (iii) the annotated phrase.

2.2.2. Propagation annotation tool

The tool we designed to propagate annotations of concepts is a PERL script which relies on two main steps:

1. First, all existing annotations are saved in a hash table, in order to keep the correspondence between entities and category;
2. Second, for each remaining file to be annotated:
   - existing annotations for this file are saved (in case of automatic pre-annotations done on the whole corpus),
   - annotations saved from the already annotated files (first step) are searched within the file. Then, beginning and ending offsets of characters are computed for each occurrence found in the file,
   - and a new stand-off annotation file is produced, combining existing annotations (pre-annotation step) with new annotations (propagation step).

The user can configure two features in this tool:

- The minimum size for annotations to be saved, in terms of number of characters. We defined a minimum size of 3 characters per annotation (i.e., all existing annotations of tokens composed of at least three characters will be propagated). This feature depends on the type of annotations to be propagated (namely, annotations propagation for tokens composed of only one character will result in annotating each similar character found in the corpus);
- The starting file from which annotation propagation will be performed. This feature allows the user to do not propagate annotations on files already annotated, reducing the risk of over-annotation.

Additionally, the user can define if propagations occur on full tokens (existing tokenization is kept), or if propagations can be found within portion of text (embedded annotations,

useful for inconsistent tokenization). We did not defined any confidence score to determine whether an annotation must be propagated or not. We considered that only annotations which are not ambiguous can be propagated. Since we did not want to add noisy annotations, the user has to process ambiguous annotations (e.g., annotations depending on the context). In its current version, contrary to active learning approaches, our tool relies on iterative actions from the user (i.e., the user decides when he wants to perform the propagation annotation process: either at the end of the annotation of each file, or at the end of a set of files). Figure 1 presents the general framework of the propagation annotation tool.

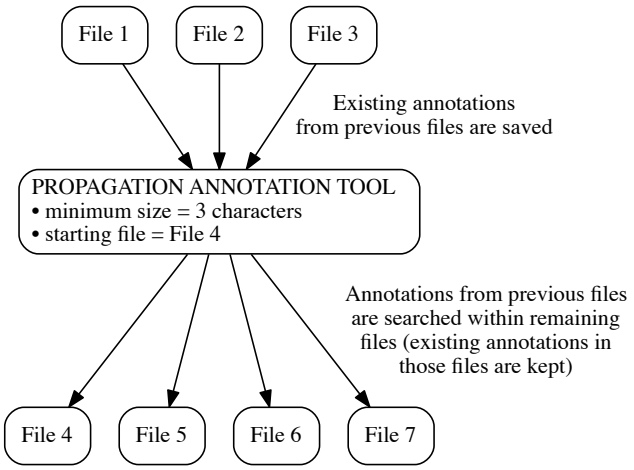

Figure 1 – General framework of the propagation annotation tool

3. Evaluation

3.1. Design of experiments

We defined two situations of annotation performed on the corpus of 60 files (cf. section 2.1.1.) for a concepts annotation task (cf. section 2.1.2.). Since no pre-annotation step has been done, human annotations are done on a raw version of the corpus:

- First, human annotations are done on the whole corpus without any annotation propagation step;
- Second, human annotations are done on the whole corpus using the annotation propagation tool. In this configuration, each time we completed a file, we launched the tool on the remaining files in order to optimize the human annotation work.

Both annotation situations rely on the same set of files. Nevertheless, annotations done in the first situation were not reused in the second situation. The same human annotator annotated files from the two situations during two distinct stages.

3.2. Results

Evolution of number of annotations Figure 2 presents the evolution of the total number of annotations along the human annotation process, depending on whether the annotation propagation tool was used (green line) or not (red line).

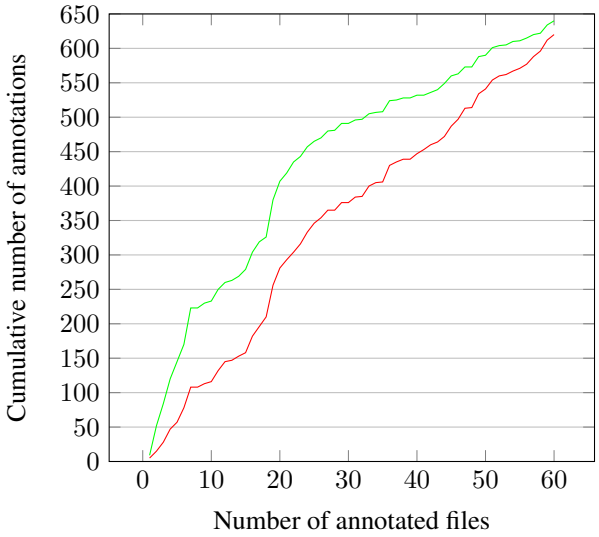

Figure 2 – Evolution of total number of annotations along the annotation process, without annotation propagation (red), with annotation propagation (green)

Annotation time Table 3 presents the time spent to annotate the corpus, the average number of files processed in one minute and the average number of annotations done in one minute, whether the propagation annotation tool was used or not.

Experiment	No propagation	Propagation
Annotation time	41 minutes	28 minutes
Average number of file/minute	1.5	2.1
Average number of annotations/minute	15.1	22.9

Table 3 – Human annotation time and statistics

Figure 3 presents the cumulative minutes spent to annotate all files, whether the annotation propagation tool was used (green line) or not (red line).

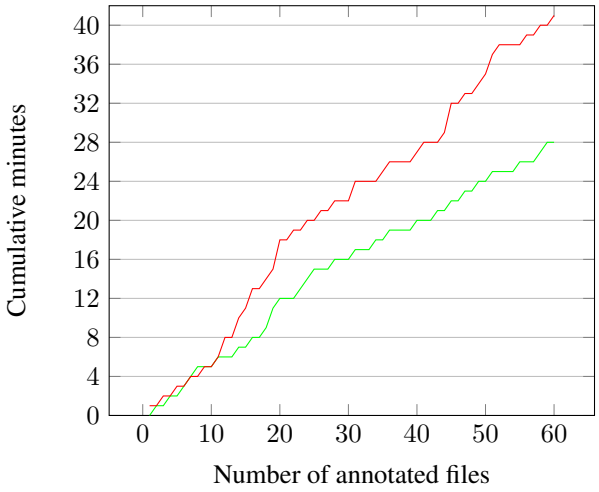

Figure 3 – Time spent to annotate all files, without annotation propagation (red), with annotation propagation (green)

Annotation quality We manually built a gold standard by revising one set of annotations. We then computed precision, recall and F-measure for both situations of annotation, based on this gold standard, using the BRATeval evaluation script. Table 4 presents the evaluation of the quality of annotations done by the human annotator, whether the annotation propagation tool was used or not. Black font pinpoints the best results. This evaluation allows us to determine the impact of annotation propagation on an annotation task.

Category	No propagation			Propagation		
	P	R	F	P	R	F
Anatomy	0.979	0.979	0.979	**1.000**	**1.000**	**1.000**
Chemical	0.987	0.961	0.973	**1.000**	**1.000**	**1.000**
Concentration	1.000	1.000	1.000	1.000	1.000	1.000
Disorders	0.607	0.895	0.723	**1.000**	**0.947**	**0.973**
Dosage	0.941	**1.000**	**0.970**	**0.969**	0.969	0.969
Duration	0.750	**1.000**	0.857	**0.895**	0.944	**0.919**
Frequency	0.867	0.813	0.839	**1.000**	0.813	**0.897**
Function	0.872	0.774	0.820	**0.946**	**0.981**	**0.963**
Job	0.778	0.875	0.824	**1.000**	**1.000**	**1.000**
Mode	0.783	**0.947**	0.857	**1.000**	0.895	**0.944**
Procedure	0.917	0.595	0.721	**1.000**	**0.676**	**0.807**
Signsymptom	0.905	0.791	0.845	**0.934**	**0.949**	**0.941**
Time	0.769	0.769	0.769	**1.000**	**1.000**	**1.000**
Weight	1.000	1.000	1.000	1.000	1.000	1.000
Overall	0.895	0.853	0.873	**0.964**	**0.948**	**0.956**

Table 4 – Evaluation of annotations quality whether the annotation propagation tool was used or not (P=Precision, R=Recall, F=F-measure). Black font pinpoints the best results

4. Discussion

Evolution of number of annotations As presented in Figure 2, the human annotation without using the propagation tool follows a diagonal (red line). This observation shows a regular number of annotations along the annotation process. As expected, the use of the propagation tool allows to rapidly increase the number of annotations (green line) from the first files.

We also produced more annotations when using the propagation annotation tool, for a total number of 640 annotations, than not using it (620 annotations).

Annotation time As shown in Table 3, our propagation annotation tool allowed us to annotate a corpus by reducing annotation time of about 31.7% (i.e., a gain of 13 minutes) in comparison with the same annotation task without using the propagation tool.

Moreover, according to Figure 3, the propagation annotation tool allows the user to keep a consistent annotation speed along the annotation process (green line) while not using such a tool, the human annotator spends more time and looses time to annotate a few files (either because of high number of annotations to be done on those files, or because of fatigue and weariness while annotating the corpus).

Annotation quality According to Table 4, we achieved a better annotation quality using our propagation annotation tool: precision increases by 6.9 points, recall by 9.5 points, and F-measure by 8.3 points. All categories benefit from this propagation annotation processing. Nevertheless, we observed that three categories obtained lower recall values when using the propagation annotation tool: *Dosage* (-3.1 pts), *Duration* (-5.6 pts) and *Mode* (-5.2 pts). Those decreasing values are due to missing annotations (false negatives), which also implies a lower number of true positives.

This observation highlights the fact that the human annotator was too much confident with the propagation annotation tool and did not pay attention to new annotations that have not been observed in previous files, making it impossible to propagate this annotation: *depuis presque un an* ("for almost a year"). Missing annotations also concern parts where propagations were not made due to the configuration of the tool (only annotations composed of at least three characters are propagated, see section 2.2.2.): *un seul comprimé* ("a single tablet"), the dosage *un* has not been propagated and the human annotator thus missed the mode of administration *comprimé* ("tablet") since there was no existing annotation in its context. Nevertheless, since those two missing annotations occur on the same file, one can not rule out a loss of attention of the human annotator when processing this file.

Comparison In comparison with existing propagation annotation tools, our system does not rely on external resources (e.g., ontologies, lexicon, etc.) as done by Swift et al. (2004), Zonta Pastorello Jr et al. (2010) or Lansdall-Welfare et al. (2012). Our tool only focuses on existing annotations done on previous files. This ensures both annotation consistency and annotation quality since no out-of-domain annotations can be made.

Moreover, our tool automatically propagates annotations without any interactive system as done by Voutilainen (2012). This allows a faster propagation annotation process. Nevertheless, this type of propagation is not suitable for ambiguous annotations such as part-of-speech annotations, where the context must be taken into account in order to choose the right category. In addition, this kind of automatic annotation propagation method, based on an identical token pairing without any interaction from the user, is not appropriate to process overlap annotations as a generic entity and its more specific version (e.g., "arm" vs. "left arm"). In such a case, both generic and specific versions will be propagated, leading the user to remove the specific version in each generic version found in the corpus.

Errors propagation At last, an automatic annotation propagation process can propagate errors (e.g., correct part associated with a wrong category, or incorrect annotation done while not needed). To process this issue, we designed a second script to propagate removal of annotations. This allows the user to rapidly correct errors made when propagating existing annotations. If ambiguous annotations must be processed, we consider a more sophisticated annotation propagation tool must be used.

5. Conclusions

In this paper, we presented the tool we designed to propagate existing annotations produced through the BRAT rapid annotation tool. Our experiments revealed the human annotator spent 31.7% less time when using the annotation propagation tool. Nevertheless, quality of annotations decreased, either due to a too much confidence in this tool or in a loss of concentration of the human annotator.

This tool can be used, either to propagate annotations on remaining files to be annotated, or in addition to pre-annotation systems, in order to manage annotations hard to process with rules or statistical approaches (namely, longer annotations such as address or hospital name). Such categories can be hard to process using rules or statistical models due to the size of the annotation, the difficulty to identify correct frontiers of annotation, or because elements from this category vary too much along the corpus, making it difficult to capture a robust representation of those elements.

As a future work, we plan to make this propagation annotation process more dynamic through a better integration of our tool in the BRAT annotation tool, which would make this annotation propagation process closer to active learning approaches.

6. Acknowledgments

This work was supported by the ANSM (French National Agency for Medicines and Health Products Safety) through the Vigi4MED project[2] (grant ANSM-2013-S-060).

7. References

Budnik, M., Poignant, J., Besacier, L., and Quénot, G. (2014). Automatic propagation of manual annotations for multimodal person identification in TV shows. In *Proc of Content-Based Multimedia Indexing*, pages 1–4, Klagenfurt, Austria.

Chevallet, J.-P., Maillot, N., and Lim, J.-H. (2006). Concept propagation based on visual similarity application to medical image annotation. *LNCS*, 4182:514–521.

Grouin, C. and Névéol, A. (2014). De-identification of clinical notes in French: towards a protocol for reference corpus development. *J Biomed Inform*, 50:151–61.

Grouin, C., (2015). *Guide d'annotation des effets secondaires rapportés par les patients sur les réseaux sociaux*.

Kantner, C., Kutter, A., Hildebrandt, A., and Püttcher, M. (2011). How to get rid of the noise in the corpus: Cleaning large samples of digital newspaper texts. *International Relations Online Working Paper Series*, 2011/2.

Lafferty, J. D., McCallum, A., and Pereira, F. C. N. (2001). Conditional Random Fields: Probabilistic models for segmenting and labeling sequence data. In *Proc of ICML*, pages 282–9, Williamstown, MA.

Lansdall-Welfare, T., Flaounas, I., and Christianini, N. (2012). Scalable corpus annotation by graph construction and label propagation. In *Proc of International Conference on Pattern Recognition Applications and Methods*, pages 25–34.

Lindberg, D. A., Humphreys, B. L., and McRay, A. T. (1993). The Unified Medical Language System. *Methods Inf Med*, 32(4):281–91.

Morlane-Hondère, F., Grouin, C., and Zweigenbaum, P. (2016). Identification of adverse drug reactions from social media. In *Proc of LREC*, Portorož, Slovenia.

Patton, R. M. and Potok, T. E. (2006). Characterizing large text corpora using a maximum variation sampling genetic algorithm. In *Proc of GECCO*, Seattle, WA.

Stenetorp, P., Pyysalo, S., Topić, G., Ohta, T., Ananiadou, S., and Tsujii, J. (2012). BRAT: a web-based tool for NLP-assisted text annotation. In *Proc of EACL Demonstrations*, pages 102–7, Avignon, France. ACL.

Swift, M. D., Dzikovska, M. O., Tetreault, J. R., and Allen, J. F. (2004). Semi-automatic syntactic and semantic corpus annotation with a deep parser. In *Proc of LREC*, Lisbon, Portugal.

Voutilainen, A. (2012). Improving corpus annotation productivity: a method and experiment with interactive tagging. In *Proc of LREC*, pages 2097–2102, Istanbul, Turkey.

Zonta Pastorello Jr, G., Daltio, J., and Bauzer Medeiros, C. (2010). A mechanism for propagation of semantic annotations of multimedia content. *Journal of Multimedia*, 5(4):332–342.

[2]Vigi4MED: *Vigilance dans les forums sur les médicaments*.