Brat2BioC: conversion tool between brat and BioC

Antonio Jimeno Yepes[1,2], Mariana Neves[3,4], Karin Verspoor[1,2]

[1]NICTA Victoria Research Lab, Melbourne VIC 3010, Australia

[2]Department of Computing and Information Systems, University of Melbourne, Melbourne VIC 3010, Australia

[3]Humboldt-Universität zu Berlin, WBI, Berlin, Germany

[4]Berlin Brandenburg Center for Regenerative Therapies, Charité, Berlin, Germany

Introduction

Interoperability between text mining solutions requires sharing information, specifically resources such as annotated corpora, in a common format. Several formats are available that have been used in the biomedical natural language processing (BioNLP) community, though no single standard has emerged. The BioC formalism [1] is intended to fill this gap, by providing tools to work with BioC, in addition to the proposed format itself. Translation of annotations of commonly used formats into BioC allows reusing existing annotated corpora with BioC solutions. The standoff *brat* (brat rapid annotation tool) format[1] is one of the more commonly used formats. For instance it has been used in the BioNLP shared task series [2]. Several corpora have been made available in the brat format, including the Human Variome Project corpus[2] and the CellFinder corpus[3] [3]. We have prepared a software solution, named Brat2BioC, that translates annotations originally in brat format into BioC and vice versa. The Brat2BioC tool is available in bitbucket at https://bitbucket.org/nicta_biomed/brat2bioc.

Methods

The Brat2BioC tool was developed in the Java programming language, using provided BioC code[4] to model the data using BioC objects and to serialize and deserialize BioC files.

Several differences exist between the two formats. These include the physical division of data and annotations among various files, and the representational choices for entity and relation annotations. These differences need to be resolved in order to perform the mapping between the two formats.

[1] Brat standoff annotation: http://brat.nlplab.org/standoff.html

[2] Human Variome Project corpus: http://www.opennicta.com/home/health/variome

[3] CellFinder corpus: http://www.informatik.hu-berlin.de/forschung/gebiete/wbi/resources/cellfinder

[4] BioC java: http://www.ncbi.nlm.nih.gov/CBBresearch/Dogan/BioC/BioC_Java_1.0.tar.gz

File representation

The brat format for annotated documents assumes that the raw text of documents appears in one file and annotations associated with that raw text appear in a separate file or files. Typically, one file is provided for each document and several files are provided for the annotations. On the other hand, BioC can handle annotations of several documents and document passages within the same file. In our implementation, the set of document files from the source brat files are converted to a single BioC file.

For an annotated corpus in BioC, all documents and annotations can be integrated into the same file, while the brat format requires a file for each document text ("*.txt") and one file for the annotations on a given document ("*.ann"; for the BioNLP Shared Tasks there are typically two annotation files, "*.a1" and "*.a2"). An additional difference is that brat has no explicit mechanism for representing internal document structure. In most existing uses of brat, a single source text is divided into several smaller files, each corresponding to a section of the source document. The name of those files typically is used to convey the meta-data about the source document and the section of that document that the file corresponds to. For instance the file name "2265717-01-Abstract-p01.txt" indicates that the file contains the first paragraph in the abstract of the document with PubMed identifier (PMID) 2265717. In some cases, as in the BioNLP shared task 2009, the name of the file is just the PMID, indicating that the file contains the title and abstract text associated with that PMID.

As mentioned above, file extensions in brat indicate the type of data in a file. In our BioC conversion, we capture this information through an *infon* object that specifies the extension of the source file in which the annotation was found (*a1*, *a2* or *ann*). The implementation offered by the BioC C++ code approaches this by generating several files, but we have preferred a more compact approach to the problem, thus requiring just one file to be generated.

When converting a BioC file into brat files, the extension of the annotation file(s) should be provided. If the extension information is not provided in the BioC file, by default the annotations are added to a file that is given a name corresponding to the value of the *id* tag of the document, and with default extension ann.

In our mapping from brat to BioC, the BioC document *id* tag is set to the name of the brat source file without the extension. This is a convention commonly used in several shared tasks and annotation efforts using the brat format. The document text, in txt brat files, is entered as a single passage tag in the BioC format. No assumptions are made about the intrinsic structure of the text documents since this structure is not defined in the brat format. An example of the high-level structure of a brat document mapped to BioC format is presented below in Figure 1:

```
<document><id>2265717-01-Abstract-p01</id><passage><offset>0</offset>  <text>**
IGNORE LINE **...</text>
```

Figure 1. Document text example

Conversion of different types of annotations from brat to BioC

Information provided by the brat format can be mapped into the BioC representation due to is flexibility but this flexibility implies that there are some BioC features that are not available in the brat format. In this section, we explain the conversion decisions for annotation types are explored and some examples are provided. Brat has several annotation types that have been modelled as BioCAnnotation and BioCRelation objects. In brat, the type of annotation is denoted by the first letter of the first token denoting as well the identifier. The same notation is used to denote the different type of annotations as in BioC. The identifier from the brat file is considered as the identifier of the BioC object.

We have compared our initial conversion proposal with the one proposed by Yifan Peng, Vijay Shanker and Cathy Wu [4], used in their iSimp tool[5]. We found several differences. The first one is that they separate a brat document text into different passages according to newlines, while we just enter the text in a single *passage* tag. The second is that we initially used an *infon* tag to store an event trigger instead of storing it in a *node* tag. We adjusted our proposal to store the event trigger in a *node* tag, since the event trigger is already declared as an entity. Furthermore, they use an *infon* tag to specify the type of BioCRelation being modelled, to explicitly distinguish event, relation, equivalence, and event modification. We have also adopted this representational choice. Finally, we have included an *infon* tag to specify the file extension of the annotation file (e.g. a1, a2 or ann). In the Peng et al proposal, the annotation type is instead used to identify the annotation file extension required to convert the BioC annotation back into the brat format. This dependency might be problematic if several file extensions are used in the future to define different sets of annotations for a given document.

Some questions remain about the best way to model document content in BioC arising from the difference identified in the application of the *passage* tag. The choice of the granularity of a "passage" in a document would seem to vary depending on what kind of text is annotated. While having a passage for each newline in the input may be appropriate for a short document such as an abstract and where newlines are consistently used to separate paragraphs, for some documents a different level of granularity could be more appropriate. For instance, a *passage* could more appropriately be an entire section/subsection within a document, or a set of paragraphs defined some other way. If the input text does not use newlines consistently to separate paragraphs (such as in the case for a LaTeX document, which uses two newlines rather than one to separate paragraphs), a *passage* might appropriately correspond to multiple input lines. A possible

[5]iSimp tool: http://research.bioinformatics.udel.edu/isimp/

solution for this would be to allow some specification of the appropriate definition of *passage* for a given conversion via a configuration parameter. This is left for future work.

Entity annotation

Entity annotation in brat is mapped to the BioCAnnotation entity in BioC as shown in Figure 2. The type of the entity is provided with an *infon* tag with key value type and value the type of the entity. Start and end of the entity is mapped to offset and length in the BioC format. Support is provided for split entities by using several location entries in BioC.

```
brat
T1      disease 54 68   Lynch syndrome
BioC
<annotation id="T1">
<infon key="type">disease</infon>
<infon key="file">ann</infon>
<location offset="54" length="14"></location>
<text>Lynch syndrome</text>
</annotation>
```

Figure 2. Example of entity annotation in brat and BioC

Relation annotation

A brat relation is encoded as a BioCRelation object in BioC as shown in Figure 3, and the brat id is used to identify the relation. Brat relations are binary, so only two nodes are created. The relation type is encoded as an *infon* object with *key* value *relation type* and the tag value contains the type of relation denoted in the brat format. Each related entity is encoded using the *node* tag, indicating the identifier of the entity in the *refid* attribute and the type of entity in the *role* attribute.

```
brat
R1_1    relatedTo body-part:T14 disease:15
BioC
<relation id="R1_1">
<infon key="type">relation</infon>
<infon key="relation type">relatedTo</infon>
<infon key="file">ann</infon>
<node refid="T14" role="body-part"></node>
<node refid="T15" role="disease"></node>
</relation>
```

Figure 3. Example of relation annotation in brat and BioC

Event annotation

Events contain a relation between a trigger entity and one or more entities. This annotation type has been encoded using the BioCRelation as shown in Figure 4. The trigger and its type are encoded in a *node* tag while the related entities have been modelled as nodes as well. The relation id denotes the event identifier in the original file.

```
brat
E21     Negative_regulation:T48 Theme:E23
BioC
<relation id="E21">
<infon key="type">event</infon>
<infon key="file">a2</infon>
<infon key="event type">Negative_regulation</infon>
<node role="trigger" refid="T48"/>
<node role="Theme" refid="E23"/>
</relation>
```

Figure 4. Example of event annotation in brat and BioC

Equivalence annotation

The equivalence entity relates to several entities, expressing that they are semantically equivalent. A BioCRelation is used to model the equivalence, mapping the related entities to node tags, without specific role. The *id* is set to *Equiv*. This is shown below in Figure 5.

```
brat
*       Equiv T6 T7
BioC
<relation id="Equiv">
<infon key="file">a2</infon>
<infon key="type">equiv</infon>
<node role="" refid="T6"/>
<node role="" refid="T7"/>
</relation>
```

Figure 5. Example of entity annotation in brat and BioC

Attribute and modification annotation

This annotation type defines an attribute of another brat annotation. The same specification can work on several annotations. We have defined it as a *BioCRelation* and specified the type of the annotation using the attribute *type* in an *infon* tag. An example is shown below in Figure 6.

```
brat
M2      Negation E14
<relation id="M2">
<infon key="file">a2</infon>
<infon key="type">Negation</infon>
<node role="" refid="E14"/>
</relation>
```

Figure 6. Example of attribute and modification annotation in brat and BioC

Normalization annotations

In addition to the boundaries of the entities, brat allows linking an identifier from a given resource to the annotated entities. The information provided as the annotation id, type of the annotation, the reference to the resource (in the example, Wikipedia) and a string linked to it are modelled using tags and attributes from the BioCRelation object. An example is shown below in Figure 7.

```
brat
N1      Reference T1 Wikipedia:534366       Barack Obama
BioC
<relation id="N1">
<infon key="file">a2</infon>
<infon key="string">Barack Obama</infon>
<infon key="type">Reference</infon>
<node role="Wikipedia:534366" refid="T1"/>
</relation>
```

Figure 7. Example of entity annotation in brat and BioC

Note annotations

Brat allows adding annotations on the entities. The type and string are encoded as *infon* tags. The annotation on which the note is added is specified in a *node* tag. An example is shown below in Figure 8.

Results

We have applied the conversion tool to existing corpora available in the brat format. The Brat2BioC tool is available from https://bitbucket.org/nicta_biomed/brat2bioc. The processed corpora include the HVP corpus [5], the BioNLP Shared Task 2009, 2011 and 2013, available from https://bitbucket.org/nicta_biomed/brat2bioc/downloads. The developed solution has been compared to the code available from the BioC website performing the transformation of the 2009

shared task data. Our software covers a larger set of brat annotations, thus it can deal with a large set of corpora.

```
brat
#1	AnnotatorNotes T1	this annotation is suspect
BioC
<relation id="#1">
<infon key="file">a2</infon>
<infon key="string">this annotation is suspect</infon>
<infon key="type">AnnotatorNotes</infon>
<node role="" refid="T1"/>
</relation>
```

Figure 8. Example of note annotation in brat and BioC

In addition, Brat2BioC has been used to convert a large set of corpora which are available for visualization on the WBI repository[6]. This repository allows on-line visualization of more than 20 popular corpora on the biomedical natural language processing domain and annotations range from named-entities (e.g., genes and drugs) and binary relationships (e.g., protein-protein interactions) to biomedical events (e.g., phosphorylation). Most of these were converted to the BioC format and made available for download from repository's page, including the AIMed, BioInfer, BioText, CellFinder, Drug-Drug Interaction Extraction 2011, Drug-Drug Interaction Extraction 2013, GeneReg, Genia, GETM, GREC, HPDR50, IEPA, LLL, OSIRIS and SNP Corpus corpora. We have not converted those corpora whose license does not allow their redistribution and or those which are only available for download after license agreement (e.g., the SCAI chemical compound corpus).

Conclusions

We have developed a tool to perform the conversion of the brat format into BioC. This conversion required analysing the way the information can be modelled in each system and explored the limitations of each of the annotation formalisms. Some possible configuration parameters, such as the extension for the generated annotation file, and a specification of the appropriate definition of a passage for the corpus, have been identified.

Acknowledgements

This work was supported by Australian Federal and Victoria State Governments and the Australian Research Council through the ICT Centre of Excellence program, National ICT Australia (NICTA).

[6]WBI repository: http://corpora.informatik.hu-berlin.de

References

1. Comeau, D et al (2013 to appear) “BioC: A Minimalist Approach to Interoperability for Biomedical Text Processing.” Database: The Journal of Biological Databases and Curation.
2. Kim, Jin-Dong, et al. (2009) "Overview of BioNLP'09 shared task on event extraction." In Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing: Shared Task. Association for Computational Linguistics.
3. Neves, M., Damaschun, A., Kurtz, A., & Leser, U. (2012). “Annotating and evaluating text for stem cell research.” In *Third Workshop on Building and Evaluation Resources for Biomedical Text Mining*.
4. Peng Y, Tudor C, Torii M , Wu CH, Vijay-Shanker K. (2013) “Enhance Interoperability of iSimp by Using the BioC Format.” Submitted to the BioCreative IV workshop.
5. Verspoor K, et al. (2013) "Annotating the biomedical literature for the human variome." Database: the journal of biological databases and curation, bat019, doi:10.1093/database/bat019.