Normalisation with the BRAT rapid annotation tool

Pontus Stenetorp[1] Sampo Pyysalo[2] Goran Topić[3]
Sophia Ananiadou[2] and Akiko Aizawa[1,3]
[1]Department of Computer Science, University of Tokyo, Tokyo, Japan
[2]{National Centre for Text Mining, University of Manchester}, Manchester, UK
[3]National Institute of Informatics, Tokyo, Japan
{pontus,smp}@is.s.u-tokyo.ac.jp
sophia.ananiadou@manchester.ac.uk
{goran_topic,aizawa}@nii.ac.jp

Abstract

We introduce new functionality for the BRAT rapid annotation tool, focusing on support for the manual annotation of text with *normalisation* annotations that identify entries in external resources such as ontologies and entity databases. The tool is available under an open-source license at *http://brat.nlplab.org*

1 Introduction

The identification of the real-world entities that are referred to in text is an important part of analysing the meaning of text. This challenge is addressed in various ways in natural language processing and manual text annotation efforts, with specific task formulations termed variously as e.g. normalisation, entity linking, grounding, and wikification (Morgan et al., 2008; McNamee and Dang, 2009; Mihalcea and Csomai, 2007). Broadly, these tasks involve assigning unique identifiers corresponding to entries in some ontology or database resource to mentions of relevant entities in text. Examples include associating the appropriate Wikipedia entries with expressions referring to specific people in news articles (e.g. "*Barack Obama*", "*Obama*", "*the president*") and the assignment of Entrez Gene[1] or UniProt[2] identifiers to mentions of gene and protein names in scientific publications. The importance of normalisation is well recognised also in biomedical text mining, where gene name normalisation has been pursued in several shared tasks (Smith et al., 2008; Arighi et al., 2011) and tools for the normalisation of e.g. chemicals (Batchelor and Corbett, 2007) and organisms (Gerner et al., 2010; Wang et al., 2010; Naderi et al., 2011) are available.

These tasks involve numerous challenges not only for automatic analysis but also for manual annotation, pursued e.g. to create gold standard annotations for the training and evaluation of automatic methods. Various tools have been introduced to help human annotators deal with the often overwhelming size of the resources involved in normalisation and to assist in maintaining the quality and consistency of created annotations (Rinaldi et al., 2010; Arighi et al., 2011). However, such tools are frequently oriented toward specific tasks and resources, and in many cases only limited consideration has been given to generality or usability.

We introduce normalisation annotation functionality for the brat rapid annotation tool (BRAT), a general web-based tool for manual text annotation. We extend the capabilities of the tool to support a new annotation primitive, *normalisation*, and introduce multiple new functions to the BRAT server and client software to allow the tool to be applied to a broad selection of tasks involving annotations that associate spans of text with external resources such as ontologies and entity databases.

2 Features

2.1 BRAT base features

BRAT is a recently introduced open-source tool for manual text annotation (Stenetorp et al., 2012). The tool seeks to be general-purpose, and can be configured to perform e.g. entity mention annotation, binary or n-ary relation annotation, and dependency syntactic annotation, among other tasks. BRAT has

[1]http://www.ncbi.nlm.nih.gov/gene/
[2]http://www.uniprot.org/

RID:EID	Resource	Entry name/term
FB:/en/barack_obama	Freebase	Barack Obama
UniProt:Q8NEY8	UniProt	Periphilin-1
GO:0016310	GO	phosphorylation
FMA:61830	FMA	Cerebral cortex

Table 1: Example references to external resources.

been applied in various annotation efforts, including several targeting biomedical text (Ohta et al., 2012; Neves et al., 2012). The system is implemented using a client-server architecture, with the Python server and the JavaScript client communicating using Asynchronous JavaScript and XML (AJAX).

2.2 Normalisation annotation primitive

The original version of BRAT supported five annotation primitives: text spans, binary relations, n-ary associations, attributes, and free-form text comments. We introduce an additional annotation primitive for normalisation. Like other primitives, each normalisation has an identifier, unique within the document. Each normalisation is associated with exactly one annotation, i.e. the one for which it assigns an external resource identifier; any number of normalisation annotations can be associated with an annotation, allowing for normalisation towards multiple external resources. The primary information carried by each normalisation annotation consists of two parts, a resource identifier (RID) and an entry identifier (EID). By convention, we write these as `RID:EID` for short, following usage in e.g. OBO (Smith et al., 2007). See Table 1 for examples.

The RID is not on its own sufficient to uniquely identify a resource: for example, `GO` could alternatively identify a Government Organisation resource. Thus, we do not rely on the RIDs to identify resources, but rather require the system to be configured to associate each RID with a uniform resource identifier (URI) that identifies the resource,[3] an approach similar to that of e.g. Courtot et al. (2011).

2.3 Visualisation

We extend the existing BRAT annotation visualisation functionality to display additional information

[3]It is thus immaterial which specific strings are used as RIDs: one could equally well use e.g. `GO` for Wikipedia and `FMA` for UniProt. We use conventional labels here for clarity.

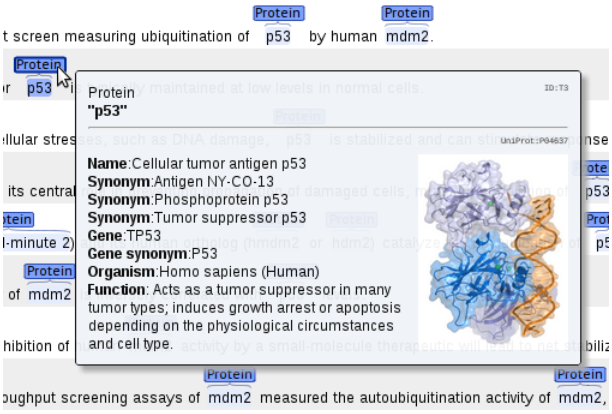

Figure 1: "Pop-up" with information from an external resource entry identified through normalisation.

on each normalised annotation based on the contents of the external resource entry referred to. As this information can potentially be very rich, we chose an implementation where information available via normalisation is displayed in a "pop-up" only when the user places the mouse over a normalised annotation (Figure 1). To avoid unnecessary computational and network overhead, this information is fetched from the BRAT server only when needed for display. To support visualisation of normalised data involving very large numbers of images, we further decouple the part of the server providing basic normalisation information from that serving the images, thus avoiding the need to store images separately on each BRAT server.

2.4 Ontology-based annotation

The most direct way to create annotations that are associated with specific entries in external resources in BRAT is to configure the annotation type system to directly use terms mapping to such entries. When set up this way, the standard dialog for selecting an annotation type serves also to associate the created annotation with the relevant external resource entry (Figure 2).

This approach is most appropriate for small resources or medium-sized resources with clear structure, and is intended to be used in particular for annotation with reference to ontologies organised primarily in *is-a* hierarchies. For larger ontologies and for resources without structure, navigating a dialog of this type becomes inefficient.

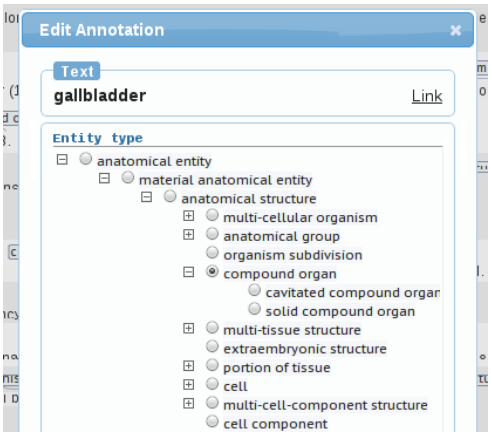

Figure 2: Entity annotation dialog with a configuration generated from CARO, a small upper-level ontology of anatomy with 48 terms.

2.5 Normalisation using large resources

To allow BRAT to be used for normalisation annotations using large or unstructured resources, we allow these annotations to be created either by directly entering resource and entry IDs – the latter presumably first identified separately e.g. using some resource-specific search functionality – or by searching by entry name or synonym using newly introduced database search functionality within BRAT (Figure 3). While resource-specific search tools can be very well tailored to the task, search functionality within the annotation tool can provide better integration and frees users from dealing with (frequently opaque) identifiers.

3 Implementation

3.1 Search

In the design of the new functionality, we aimed to create a system capable of supporting rapid lookup and flexible search of moderately large databases – millions to tens of millions of strings – on standard desktop systems. To allow the system to be used to perform approximate-matching search on such resources, we implemented the search functionality using a recently introduced fast approximate string matching algorithm, SimString (Okazaki and Tsujii, 2010), in addition to a standard SQL database.

Search is implemented in two steps: first, strings input by the user are queried in a SimString database

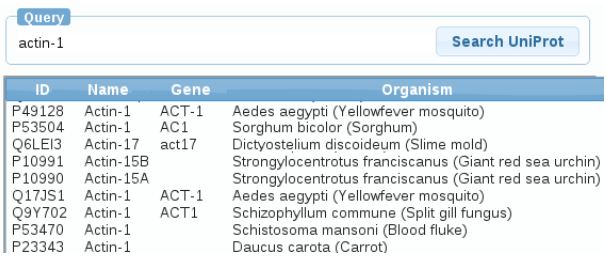

Figure 3: New BRAT resource search dialog with query results against the UniProt protein database.

to fetch a set of strings that approximately[4] match the input. These strings are then filtered to remove weak matches using a slower but more sensitive matching algorithm based on edit distance with a custom cost matrix, and for each of the filtered strings, the set of entries involving the string are then queried from a standard SQL database using exact string matching to retrieve the full data associated with each entry. The data is then presented to the user for the selection of the intended entry.

3.2 Ontology and database integration

The set of ontologies and databases against which normalisation could potentially be performed in annotation tasks is open-ended, and it is not possible to anticipate and support all reference resource formats. To reduce the demands on users developing conversions between reference resources and the BRAT normalisation system, we follow two complementary approaches, first, introducing an intermediate representation and tools for input into the system databases, and, second, conversions from a number of prominent standard resource formats into the intermediate representation.

Well-established formats such as the Open Biomedical Ontologies (OBO) Foundry[5] (Smith et al., 2007) OBO format and UniProt XML are supported "out of the box" by providing conversion and database creation scripts. Support for other formats such as the Freebase[6] DB (Bollacker et al., 2008) format is planned, and will be made available as part of our additions to the annotation tool.

[4] By default, we use the `overlap` match option with a 0.7 similarity threshold.

[5] http://www.obofoundry.org/

[6] http://www.freebase.com/

Task	Time
SimString DB initialisation	82 sec
SQL DB initialisation	45 min
SimString DB lookup	34 sec (294 lookups/sec)
SQL DB lookup	31 sec (323 lookups/sec)

Table 2: Resource requirements for creating a database of 1,7M strings and performing lookup of 10,000 strings.

4 Evaluation

We next briefly present basic performance measurements of the implementation of normalisation functionality in BRAT. Practical performance is strongly dependent on a number of factors such as database size and the machine on which the system is run, and these measurements should only be taken as broadly indicative of the general level of performance.

Tests were run on a sub-set of 730,000 entries from the total of 1,7 million strings contained in the UniProt database. A relatively low-powered laptop with a dual-core 1.33GHz processor and 2GB of memory was used as the reference system during testing. Table 2 shows the wall-clock time costs for creating the databases and querying 10,000 strings. We note that the only non-trivial cost is the creation of the SQL database, a step which only needs be performed once during the system setup.

5 Conclusions

We have introduced normalisation functionality added to the BRAT rapid annotation tool. We discussed theoretical and technical motivations for our choice of implementation, and demonstrated how our proposed additions can be used to support annotation projects aiming to normalise against smaller ontologies as well as large-scale resources. A small-scale evaluation indicated that the implementation scales to resources of over a million strings with only modest resource requirements.

The new version of BRAT including all functionality presented in this work is available under the MIT open-source license at *http://brat.nlplab.org*

Acknowledgements

This work was supported by the Royal Swedish Academy of Sciences and the UK Biotechnology and Biological Sciences Research Council (BBSRC BB/G013160/1).

References

C. Arighi, P. Roberts, S. Agarwal, S. Bhattacharya, G. Cesareni, et al. 2011. BioCreative III interactive task: an overview. *BMC Bioinformatics*.

C.R. Batchelor and P.T. Corbett. 2007. Semantic enrichment of journal articles using chemical named entity recognition. In *Proceedings of the Demonstrations at ACL 2007*.

K. Bollacker, C. Evans, P. Paritosh, T. Sturge, and J. Taylor. 2008. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of SIGMOD 2008*.

M. Courtot, F. Gibson, A.L. Lister, J. Malone, D. Schober, R.R. Brinkman, and A. Ruttenberg. 2011. MIREOT: the minimum information to reference an external ontology term. *Applied Ontology*.

M. Gerner, G. Nenadic, and C. Bergman. 2010. LINNAEUS: A species name identification system for biomedical literature. *BMC Bioinformatics*.

P. McNamee and H.T. Dang. 2009. Overview of the tac 2009 knowledge base population track. In *Proceedings of TAC*.

R. Mihalcea and A. Csomai. 2007. Wikify!: linking documents to encyclopedic knowledge. In *Proceedings of CIKM 2007*.

A. Morgan, Z. Lu, X. Wang, A. Cohen, J. Fluck, et al. 2008. Overview of BioCreative II gene normalization. *Genome Biology*.

N. Naderi, T. Kappler, C.J.O. Baker, and R. Witte. 2011. OrganismTagger: detection, normalization and grounding of organism entities in biomedical documents. *Bioinformatics*.

M. Neves, A. Damaschun, A. Kurtz, and U. Leser. 2012. Annotating and evaluating text for stem cell research. In *Proceedings of BioTxtM*.

T. Ohta, S. Pyysalo, J. Tsujii, and S. Ananiadou. 2012. Open-domain anatomical entity mention detection. In *Proceedings of DSSD*.

N. Okazaki and J. Tsujii. 2010. Simple and Efficient Algorithm for Approximate Dictionary Matching. In *Proceedings of Coling 2010*.

F. Rinaldi, S. Clematide, G. Schneider, M. Romacker, and T. Vachon. 2010. ODIN: An Advanced Interface for the Curation of Biomedical Literature. *Nature Precedings*.

B. Smith, M. Ashburner, C. Rosse, J. Bard, W. Bug, et al. 2007. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nature Biotechnology*.

L. Smith, L.K. Tanabe, R.J. Ando, C.J. Kuo, I.F. Chung, et al. 2008. Overview of BioCreative II gene mention recognition. *Genome Biology*.

P. Stenetorp, S. Pyysalo, G. Topić, T. Ohta, S. Ananiadou, and J. Tsujii. 2012. BRAT: a Web-based Tool for NLP-Assisted Text Annotation. In *Proceedings of the Demonstrations at EACL 2012*.

X. Wang, J. Tsujii, and S. Ananiadou. 2010. Disambiguating the species of biomedical named entities using natural language parsers. *Bioinformatics*, 26(5):661–667.