Improving Interoperability of Text Mining Tools with BioC

Ritu Khare, Chih-Hsuan Wei, Yuqing Mao, Robert Leaman, Zhiyong Lu*

National Center for Biotechnology Information, 8600 Rockville Pike, Bethesda, Maryland, USA
*Corresponding Author: Tel: 301-594-7089, E-mail: zhiyong.lu@nih.gov

Abstract

The lack of interoperability among text mining tools is a major bottleneck in creating more complex applications. Despite the availability of numerous methods and techniques for various text mining tasks, combining different tools requires substantial efforts and time. In response, BioC offers a minimalistic approach to tool interoperability by stipulating minimal changes to existing tools and applications. In this study, we introduce several state-of-the-art text mining tools (for recognizing and annotating genes, diseases, mutations, species, and chemicals in biomedical text) developed at the National Center for Biotechnology Information (NCBI), and modify these tools to make them BioC compatible. We find that only minimal changes were required in order to build the BioC versions of our tools via using the BioC family of format and functions. Through this work, we improved the interoperability of our tools, and anticipate serving a wider community for building more sophisticated applications. Our toolkit was created through participating in the BioCreative IV Interoperability Track and is publicly available at http://www.ncbi.nlm.nih.gov/CBBresearch/Lu/Demo/tmTools.

Introduction

There is an increasing demand of text mining tools in the biomedical and life sciences domain. Many recent BioNLP challenge tasks (1-5) are focused on extracting structured information from scientific articles and clinical notes. Research groups around the world are developing a variety of standalone text mining tools. Typically, a tool is developed using a certain preferred data representation, programming conventions as determined by the individual research group. In order to build complex text mining applications or pipelines, it is often required to combine multiple tools, possibly designed by different groups. The current practice of independent tool development poses a hindrance to tool interoperability and integration. In order to use a new tool or a new dataset, text mining researchers spend a substantial amount of time developing algorithms for processing the new data format. This heterogeneity in data representations slows down the development of powerful applications and thereby leads to inefficiencies in research and innovation.

There have been quite a few efforts to promote interoperability among text analytics tools. Unstructured information management architecture (UIMA)(6-8) and General Architecture for

Text Engineering (GATE) (9) are two notable proposals that prescribe using a predefined framework to develop text mining applications to achieve interoperability among independently developed tools. Development of UIMA- or GATE-compliant applications requires the entire tool to be (re-)written into framework specific constructs. The steep learning curve associated with these frameworks keeps them from being broadly accepted as a development and data sharing standard (10). Motivated by this, a recent effort in this direction, BioC (11), is based on a minimalist approach in that it offers interoperability by stipulating minimal changes in existing applications or datasets. The goals of BioC are simplicity, reusability, interoperability and wide use. In a nutshell, BioC is a family of XML formats that define how to present text documents and annotations. BioC also provides tools to read and write documents in the BioC format in two widely used programming languages.

In this paper, we present our efforts on using BioC to re-package the suite of text mining software and web-based tools (12-18) developed at the biomedical text mining group at the National Center for Biotechnology Information (NCBI). Specifically, we wrap five stand-alone biomedical named entity recognition (NER) tools, one web-based annotation tool, and one annotated text corpus, into BioC. All tools are aimed toward accelerating the biomedical discovery and manual curation of biological databases, and by making them BioC compatible, we expect them to serve a wider community.

Method

In this section, we first introduce the NCBI suite of tools that comprises six tools for concept recognition and annotation, and an annotated text corpus for Gene Ontology concept recognition. Then, we describe the key steps and challenges in creating a BioC compatible version of the tools and the text corpus.

Our Concept Recognition and Annotation Toolkit

At NCBI, we have developed several NER tools for automatically recognizing key biomedical concepts such as chemicals, diseases, genes, mutations, and species, from scientific publications. Each tool accepts a PubMed or PMC full-text article as an input and identifies the biomedical entities at either mention-level, or at both mention and concept level. Figure 1 provides a summary of our concept recognition and annotation toolkit.

- *DNorm*(1,15) is an open-source software tool to identify and normalize disease mentions from biomedical texts. *DNorm* is based on pair-wise learning to rank and is the first technique to use machine learning for disease normalization. This tool was developed in Java.

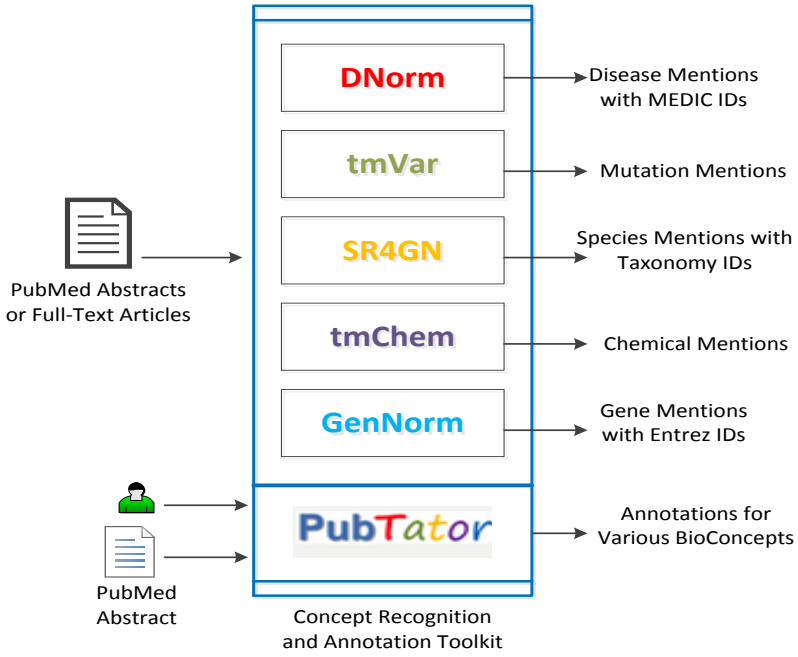

Figure 1. Visual Summary of NCBI Concept Recognition and Annotation Toolkit

- *tmVar*(14) is a machine learning system for mutation recognition to assist biomedical curation. It is based on conditional random fields and identifies many types of mutations and sequence variants in protein, gene, DNA, and RNA levels for biomedical curation. This tool was developed in Perl and uses the CRF++ module developed in C++.
- *SR4GN*(12) is a species recognition tool optimized for the gene normalization task. It is a rule-based system that identifies species from full-texts and pairs them with corresponding gene or protein mentions. This tool was developed in Perl.
- *tmChem*(17) is a machine learning based NER system for chemicals. The system is designed to identify a wide variety of chemical mentions in literature, including identifiers, brand and trade names and also systematic formats. The system uses conditional random fields with a rich feature set and rule-based post processing modules for resolving local abbreviations and improving consistency. This tool was developed in Java.
- *GenNorm*(13) is a rule-based tool to for gene recognition and performs gene name recognition, species assignation and species-specific gene normalization. *GenNorm* addresses the challenging issues of orthologous gene ambiguity and intra-species gene ambiguity. This tool was developed in Perl.

Based on the above NER tools, we also developed a web-based annotation tool called *PubTator* (16,19,20) for assisting manual curation. *PubTator* is in sync with PubMed and supports annotation of biomedical entities and their relationships in PubMed articles.

The BC4GO corpus

More recently, we developed the *BC4GO* corpus (18) (not shown in the figure), a corpus of 200 full-text articles along with their gene ontology (GO) annotations describing genes and gene product attributes across species and databases. As annotations, the corpus presents the evidence sentences along with the gene/protein entities, GO terms, and GO evidence codes. The corpus was developed with eight expert biocurators using a web-based annotation tool. This is the official corpus for the BioCreative IV Track-4 GO Task (21), which tackles the challenge of automatic GO annotation through literature analysis.

Building BioC Compatible Tools

The BioC family of XML formats and functions comprises the following three items:

(i) The XML Document Type Definition (DTD) that defines how to present text document and annotations in higher-level semantics to share common information. It allows many different annotations to be represented, including sentences, tokens and named entities. The general BioC format recommends keeping the text of the article and the corresponding annotations in separate files, namely BioC article file and BioC annotation file.

(ii) A key file to describe the lower level semantics of the elements in the BioC annotation file. The key file describes how data should be interpreted in the BioC annotation file, and needs to be created for a specific type of data

(iii) C++ and Java libraries that include functions and classes to read and write documents in BioC format and to hold the documents in memory.

To comply our tools with BioC, we modified the input and output formats of the tools, i.e., by adding BioC as a new option, and translated the articles and the annotations into BioC article files and BioC annotation files, respectively.

Concept Recognition Tools

The main challenge faced when converting these various concept recognition tools to BioC was to define an appropriate key file. Since the semantics of all these tools are similar to *PubTator* in terms of the type of data, we used the same key file, **PubTator.key**. The same key file is used for interpreting the input full-text articles/abstracts, and the output articles/abstracts with annotations.

The mutation recognition tool, *tmVar*, originally accepts the PubTator format, free text, and the PMC XML format. The output format is the PubTator format. For *GenNorm* and *SR4GN*, the input formats are free text, PMC XML format, and the GenNorm format[1], and the output format is the GenNorm format. To make these tools compatible with BioC, we added the BioC format

[1] http://www.ncbi.nlm.nih.gov/CBBresearch/Lu/Demo/Summary/Format.html#GenNorm

as a new option for input/ output. The main difference between the custom-defined format and the BioC format is the offset calculation. The custom-defined format calculates the offsets for separate sentences, and to translate to the BioC format, we had to re-calculate the global offset for the each mention. Accordingly, we added the new elements, mention, and paragraph, in the key file.

The previous output format for *tmChem* was the BioCreative IV CHEMDNER format, which is essentially a delimited format for representing one NER mention on each line. *DNorm* is a relatively new tool and did not previously have a default output format. Since both tools are built on top of BANNER (22), input compatibility with BioC only required writing a single new dataset loading class in BANNER to read BioC. Modifying the output required modifying the class containing the main method to output the BioC format.

PubTator

The original input/output format for *PubTator* is a pre-defined format that we refer to as the PubTator format[2]. To make *PubTator* BioC compatible, we added a new format option giving users the option to input and output in the BioC format. For our purposes, we also slightly modified the original BioC format. The original BioC recommends keeping only the annotations, and not the passages, in the BioC annotation file. However, this would require users to upload two files when importing annotations to *PubTator*. Hence, in our version of BioC compatible *PubTator*, the annotations are appended after the article passages in the BioC annotation file. The other concepts recognition tools and corpus still follow the original BioC format. We defined the **PubTator.key** file that describes specific attributes such as bioconcept, identifier, offset, and mentions.

BC4GO Corpus

First, the 200 full-text articles of the *BC4GO* corpus, originally in the PMC XML data model format, were converted to the BioC format. Then, we extracted annotated sentences from downloaded HTML files from the tool and identified their offsets. Finally, for each article we created a corresponding BioC annotation file for the associated GO annotations. For the gene entity, we provide both the gene mention as appeared in text and its corresponding NCBI Gene identifier. In the BioC released corpus, each article is named by its PubMed identifier, e.g. "20130316.xml." The annotation file associated with the article file shares the same PMID in the file name, e.g. "annotation_20130316.xml." The annotation file includes all annotations of the article; each annotation has a unique ID and is defined by four distinct elements: gene, go-term, go-evidence, and type. We define separate key files to describe the full-text articles and the annotation files with GO annotations, namely **pmc_go.key** and **go_annotation.key**, respectively. There were certain challenges in creating the BioC compatible version of the *BC4GO* corpus.

[2] http://www.ncbi.nlm.nih.gov/CBBresearch/Lu/Demo/PubTator/import.example.html

```
<collection>
    <source>Example</source>
    <date>1999-Jan-1</date>
    <key>PubTator.key</key>
    <document>
        <id>20085714</id>
        <passage>
            <infon key="type">title</infon>
            <offset>0</offset>
            <text>Autosomal-dominant striatal degeneration is caused by a mutation in the
                phosphodiesterase 8B gene.</text>
        </passage>
        <passage>
            <infon key="type">abstract</infon>
            <offset>98</offset>
            <text>Autosomal-dominant striatal degeneration (ADSD) is an autosomal-dominant movement
                disorder affecting the striatal part of the basal ganglia. ADSD is characterized by
                bradykinesia, dysarthria, and muscle rigidity. These symptoms resemble idiopathic
                Parkinson disease, but tremor is not present. Using genetic linkage analysis, we
                have mapped the causative genetic defect to a 3.25 megabase candidate region on
                chromosome 5q13.3-q14.1. A maximum LOD score of 4.1 (Theta = 0) was obtained at
                marker D5S1962. Here we show that ADSD is caused by a complex frameshift mutation
                (c.94G>C+c.95delT) in the phosphodiesterase 8B (PDE8B) gene, which results in a loss
                of enzymatic phosphodiesterase activity. We found that PDE8B is highly expressed in
                the brain, especially in the putamen, which is affected by ADSD. PDE8B degrades
                cyclic AMP, a second messenger implied in dopamine signaling. Dopamine is one of the
                main neurotransmitters involved in movement control and is deficient in Parkinson
                disease. We believe that the functional analysis of PDE8B will help to further
                elucidate the pathomechanism of ADSD as well as contribute to a better understanding
                of movement disorders.</text>
        </passage>
    </document>
</collection>
```

Figure 2. A snippet of the BioC article file for PMID 20085714

The first challenge was in creating the BioC annotation file using the user annotations downloaded from the web-based annotation tool. We observed encoding discrepancies in the article file and the downloaded file. The original file in PMC XML format is encoded in ASCII, which is also the encoding convention for the BioC format. However, the annotation results downloaded from the Web were encoded using Unicode. For example, the term “neurexin-1α” (see PMID:22262843 in corpus) would read “neurexin-1alpha” in ASCII but “neurexin-1I+-“ in Unicode. In order to maintain consistency between the BioC article and annotation files, we translated the Unicode characters back to ASCII using a neighbor matching method as described in (18).

Another challenge was presenting those evidence sentences that contain multiple discontinuous sentences, possibly from different passages in the article (see the evidence sentence for GO:1990124 in PMID 18695045 in corpus). We addressed this challenge by linking these disjoint evidence sentences using the same annotation ID for recognition, i.e., they are treated as one whole evidence sentence for a GO term.

```
<annotation>
    <infon key="type">Mutation</infon>
    <offset>679</offset>
    <length>8</length>
    <text>c.95delT</text>
    <id>c|DEL|95|T</id>
</annotation>
<annotation>
    <infon key="type">Gene</infon>
    <offset>696</offset>
    <length>20</length>
    <text>phosphodiesterase 8B</text>
    <id>8622</id>
</annotation>
<annotation>
    <infon key="type">Gene</infon>
    <offset>718</offset>
    <length>5</length>
    <text>PDE8B</text>
    <id>8622</id>
</annotation>
<annotation>
    <infon key="type">Gene</infon>
    <offset>810</offset>
    <length>5</length>
    <text>PDE8B</text>
    <id>8622</id>
</annotation>
<annotation>
    <infon key="type">Disease</infon>
    <offset>898</offset>
    <length>4</length>
    <text>ADSD</text>
    <id>609161</id>
</annotation>
<annotation>
    <infon key="type">Gene</infon>
    <offset>904</offset>
    <length>5</length>
    <text>PDE8B</text>
    <id>8622</id>
</annotation>
<annotation>
    <infon key="type">Chemical</infon>
    <offset>919</offset>
    <length>10</length>
    <text>cyclic AMP</text>
</annotation>
```

Figure 3. A snippet from the BioC annotation file for PMID 20085714 (integrated result of applying our five concept recognition tools on the abstract). The offset element is the global offset.

One limitations of the corpus released in BioC is that the BioC annotation file of an article would not contain an evidence sentence that is located in the "Acknowledgement" section of the article (see PMID 18695045) because this section is not provided in the original PMC XML file for the article. Also, in some cases, such as footnotes, incomplete sentences were created due to the additional space characters in the original PMC XML files. Such cases were manually processed to create a consistent BioC annotation files.

Results

The new BioC versions of all tools and the common **PubTatory.key** file are made publicly available. The key file is also shown in Figure 5 in the Appendix section.

To describe the outputs of our concept recognition tools, we use a PubMed abstract (PMID 20085714) that contains mentions of multiple biomedical entities, including genes, mutations, chemicals, and diseases, as a running example. A snippet of the BioC article file for this example is shown in Figure 2, and the integrated results from all the tools are displayed in Figure 3 showing a snippet of the BioC annotation file.

The BioC version of the *BC4GO* corpus, with 200 BioC article files and 200 BioC annotation files, can be downloaded at the BioCreative IV Track 4 task's official webpage, http://www.biocreative.org/tasks/biocreative-iv/track-4-GO/. The key files, **pmc_go.key** and **go_annotation.key** are submitted as part of the BioCreative IV Track 1 submission. These key files are also shown as Figures 6 and 7 in the Appendix section. A snippet of the BioC annotation file corresponding to the PubMed article with PMID 23840682 is shown in Figure 4.

```
<collection>
    <source>GO_Annotation</source>
    <date>20130316</date>
    <key>go_annotation.key</key>
    <document>
        <id>23840682</id>
        <passage>
            <infon key="type">abstract</infon>
            <offset>89</offset>
            <annotation id="23840682_1">
                <infon key="gene">emb16(100170235)</infon>
                <infon key="go-term">embryo development|GO:0009790</infon>
                <infon key="goevidence">IMP</infon>
                <infon key="type">GOA</infon>
                <location offset="415" length="114"/>
                <text>The emb16 mutation arrests embryogenesis at transition stage and allows the
                    endosperm to develop largely normally.</text>
            </annotation>
```

Figure 4. A Snippet from the file **annotation_23840682.xml** from the *BC4GO* corpus

PubTator.key: A BioC format for PubTator and all equipped tools (i.e., tmChem, DNorm, tmVar, SR4GN or GenNorm).

The goal of this collection is to provide easy access to the text and its bio-concept annotation of PMC articles. All of the text in an article is easily accessible. Some of the other information in an article is also available.

collection: a group of PubMed documents split into title, abstract and other passages

source: PubMed or PubMed Central

date: Date document downloaded from PubTator

document: Title, abstract and other passages from a PubMed or PMC reference

id: PubMed id

passage: Title, abstract and other passages

infon["type"]: "title", "abstract" and other passages

offset: Title has an offset of zero, while the other passages (e.g., abstract) are assumed to begin after the previous passages and one space

annotation: One bio-concept of the passage as determined by the tmChem, DNorm, tmVar, SR4GN or GenNorm

infon["type"]: "Gene", "Species", "Disease", "Chemical" or "Mutation"

id: The bio-concept identifiers which are detected by DNorm,tmVar, SR4GN and GenNorm

offset: A document offset to where the bio-concept begins in the passage. The global offset within the document

length: The length of the bio-concept in the passage

text: Mention of the bio-concept

Figure 5. The **PubTator.key** file

Discussion and Conclusions

The goal of this study was to improve the interoperability of our NER tools using the recently developed BioC Family of XML formats and classes. The NCBI suite of tools consists of several competition winning, high-performing tools for concept recognition and annotation. For example, *GenNorm* obtained the highest performance in the BioCreative III Gene Normalization task (23), and *DNorm* achieved the best results the 2013 ShARe/CLEF shared task for

normalizing disease names in clinical notes (1). Also, the *tmVar* tool for mutation recognition delivers over 90% F-measure on multiple benchmarking test sets; and the PubMed-like, color-coded interface of *PubTator* makes it a highly usable annotation tool for human biocurators. In addition to accelerating knowledge discovery and assisting manual curation, the NCBI text mining toolkit is capable of solving other important and challenging problems in the biomedical domain. For instance, text mining mutation information is very critical for the analysis and interpretation of sequence variations in complex diseases in the post-genomic era. Disease recognition is important for many lines of inquiry, including etiology (e.g. gene-disease relationships) and clinical aspects (e.g. diagnosis, prevention, and treatment). Gene and specifies recognition could be useful for protein-protein interaction extraction.

```
pmc_go.key: A BioC format for PubMed Central (PMC) articles.

The goal of this collection is to provide easy access to the full-text of
PMC articles.

   collection:  PMC articles articles selected for the GO annotation
         track of BioCreative IV

            source:     PMC

            date:       yyyymmdd. Date articles downloaded from PMC.

            document:    PMC article

            id:         PubMed id

            passage:     Title, abstract and other passages

                        infon["type"]: "title", "abstract" and other passages

                        offset:   Title has an offset of zero, while the other
                        passages (e.g., abstract) are assumed to begin after the previous
                        passages and one space

                        text:     The ASCII text of the passage.
```

Figure 6. The **pmc_go.key** file

```
go_annotation.key: A BioC format for PubMed Central (PMC) article annotations.

The goal of this collection is to provide easy access to the text of
PMC article annotationss. All of the text in an article is easily accessible.
Some of the other information in an article is also available.

 collection:  Annotations of PMC articles articles selected for the GO annotation track of BioCreative IV

        source:     PMC and GO annotations made by professional GO curators

        date:       yyyymmdd. Date articles downloaded from PMC.

        document:   PMC article

                id:         PubMed id

                passage:    Title, abstract and other passages

                        infon["type"]:  "title", "abstract" and other passages

                        offset:     Title has an offset of zero, while the other passages (e.g., abstract)
                                    are assumed to begin after the previous passages and one space

                        annotation:  The evidence sentence of the passage as determined by the
                                    curator

                                    infon["type"]:  "gene","go-term","goevidence" and "type" of the
                                            annotation (typically "GOA").

                                    offset:     A document offset to where the evidence sentence begins
                                            in the passage. The global offset within the document

                                    length:     Length of the evidence sentence

                                    text:       ASCII text of the evidence setence
```

Figure 7. The **go_annotation.key** file

Our experience shows that only minimal changes were required to re-package the NCBI suite of text mining tools with BioC. Also, reading and writing to BioC format was fairly straightforward as the functions and classes are already provided by the BioC authors in two widely used programming languages. For each tool, the primary developers modified their respective tools, and confirmed the simplicity and learnability of the BioC format. The primary challenge was to create the key files for the tools. However, it was a one-time effort since all the six concept

recognition and annotation tools can use a common key file for defining their BioC annotation files. The released **PubTator.key** file could also evolve as a standard key file for concept recognition and annotation tasks as recommended in (24). All our tools are freely available and ready to be re-used by a wider community of researchers in text mining, bioinformatics, and biocuration communities.

Through this study, we promote the interoperability of our tools, not only with each other, but also with the tools and datasets developed by several other groups worldwide. The tools, although developed in different programming languages such as Java, Perl, and C++, are now capable of sharing their inputs/outputs with each other, without any additional programming efforts. Our tools in BioC can interact with other state-of-the-art tools to build much more powerful applications. For example, a modular text mining pipeline of various BioC compatible tools for NER, normalization, annotation, and relationship extraction, could be developed to build sophisticated systems, e.g., an integrative disease-centered system connecting the biological and clinical aspects, providing information from causes (gene-mutation-disease relationship) to treatment (drug-disease relationships) of diseases by mining/annotating unstructured (biomedical literature, clinical notes, etc.) and structured resources (datasets released by organizations and research groups). In the future, we anticipate much broader usage of these tools as further efforts are invested in publicizing BioC.

Acknowledgements

We would like to thank Don Comeau, Rezarta Dogan and John Wilbur for their discussion and help with the BioC tools and in particular, their help on preparing the PMC articles in BioC XML format for the BioCreative IV GO task. This research was supported by the Intramural Research Program of the NIH, National Library of Medicine.

References

1. Leaman, R., Khare, R., Lu, Z. (2013) NCBI at 2013 ShARe/CLEF eHealth Shared Task: Disorder Normalization in Clinical Notes with DNorm. *Conference and Labs of the Evaluation Forum 2013 Working Notes*.

2. Lu, Z., Kao, H.Y., Wei, C.H.*, et al.* (2011) The gene normalization task in BioCreative III. *BMC bioinformatics*, **12 Suppl 8**, S2.

3. Morgan, A.A., Lu, Z., Wang, X.*, et al.* (2008) Overview of BioCreative II gene normalization. *Genome biology*, **9 Suppl 2**, S3.

4. Krallinger, M., Vazquez, M., Leitner, F.*, et al.* (2011) The Protein-Protein Interaction tasks of BioCreative III: classification/ranking of articles and linking bio-ontology concepts to full text. *BMC bioinformatics*, **12 Suppl 8**, S3.

5. Mork, J.G., Bodenreider, O., Demner-Fushman, D.*, et al.* (2010) Extracting Rx information from clinical narrative. *Journal of the American Medical Informatics Association : JAMIA*, **17**, 536-539.

6. Ferrucci, D., Lally, A. (2004) UIMA: An Architectural Approach to Unstructured Information Processing in the Corporate Research Environment. *Natural Language Engineering*, **10**, 327-348.
7. Ferrucci, D., Lally, A., Gruhl, D.*, et al.* (2006) Towards an Interoperability Standard for Text and Multi-Modal Analytics. IBM Research.
8. Kano, Y., Baumgartner, W.A., Jr., McCrohon, L.*, et al.* (2009) U-Compare: share and compare text mining tools with UIMA. *Bioinformatics*, **25**, 1997-1998.
9. GATE : General Architecture for Text Engineering. The University of Sheffield.
10. Stubbs, A. (2011) MAE and MAI: lightweight annotation and adjudication tools. *Proceedings of the 5th Linguistic Annotation Workshop*, pp. 129-133.
11. Comeau, D.C., Doğan, R.I., Ciccarese, P.*, et al.* (2013) BioC: A Minimalist Approach to Interoperability for Biomedical Text Processing. *Database: The Journal of Biological Databases and Curation*.
12. Wei, C.H., Kao, H.Y., Lu, Z. (2012) SR4GN: a species recognition software tool for gene normalization. *PloS one*, **7**, e38460.
13. Wei, C.H., Kao, H.Y. (2011) Cross-species gene normalization by species inference. *BMC bioinformatics*, **12 Suppl 8**, S5.
14. Wei, C.H., Harris, B.R., Kao, H.Y.*, et al.* (2013) tmVar: a text mining approach for extracting sequence variants in biomedical literature. *Bioinformatics*, **29**, 1433-1439.
15. Leaman, R., Islamaj Dogan, R., Lu, Z. (2013) DNorm: Disease Name Normalization with Pairwise Learning to Rank. *Bioinformatics*.
16. Wei, C.H., Kao, H.Y., Lu, Z. (2013) PubTator: a web-based text mining tool for assisting biocuration. *Nucleic acids research*, **41**, W518-522.
17. Leaman, R., Wei, C.-H., Lu, Z. (2013) NCBI at the BioCreative IV CHEMDNER Task: Recognizing chemical names in PubMed articles using tmChem. *Proceedings of BioCreative IV*.
18. Auken, K.V., Schaeffer, M.L., McQuilton, P.*, et al.* (2013) Corpus Construction for the BioCreative IV GO Task. *Proceedings of BioCreative IV*.
19. Wei, C.H., Harris, B.R., Li, D.*, et al.* (2012) Accelerating literature curation with text-mining tools: a case study of using PubTator to curate genes in PubMed abstracts. *Database : the journal of biological databases and curation*, **2012**, bas041.
20. Wei, C.-H., Kao, H.-Y., Lu, Z. (2012) PubTator: A PubMed-like interactive curation system for document triage and literature curation. *Proceedings of BioCreative 2012 workshop*, Washington DC, USA, pp. 145-150.
21. Mao, Y., Auken, K.V., Li, D.*, et al.* (2013) The Gene Ontology Task at BioCreative IV. *Proceedings of the BioCreative IV Workshop*, Bethesda, MD.
22. Leaman, R., Gonzalez, G. (2008) BANNER: an executable survey of advances in biomedical named entity recognition. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, 652-663.
23. Wei, C.-H., Kao, H.-Y. (2010) Inference network method on cross species gene normalization in full-text articles. *Procceding of BioCreative III Workshop*, Bethesda, Maryland, pp. 73-81.
24. Arighi, C.N., Carterette, B., Cohen, K.B.*, et al.* (2013) An overview of the BioCreative 2012 Workshop Track III: interactive text mining task. *Database : the journal of biological databases and curation*, **2013**, bas056.