TR-NLP-UT-2006-3*

GENIA Corpus Curation Framework

Jin-Dong Kim, Jun'ichi Tsujii

University of Tokyo,
University of Manchester,
National Centre for Text Mining

Date of Release: 15 November 2006

* All technical reports from Tsujii Laboratory can be found at http://www-tsujii.is.s.u-tokyo.ac.jp/TR/.

1. Introduction

To make *text mining* (*TM*) more successful, text data from various sources need to be exploited. For example, *Medline* includes vast amount of abstracts of journal articles. *PubMed Central (PMC)* includes full journal articles although they are still in limited volume. *Entrez Gene* includes description of genes written in English.

Although it may be true that much useful information is contained in such text data, since it is written in natural language, the information is not directly accessible by computers. The text data are thus often processed and the results are stored together with the text data to raise the accessibility to the information contained in it. We call such process *annotation* and the processor *annotator*.

There are many software tools which perform text analysis for *natural language processing* (*NLP*), *information extraction* (*IE*), etc., which are all potential annotators. Human experts also can perform annotation with a help of a bunch of supporting tools.

A problem arises when we try to deal with text data from multiple sources: they may be encoded in different formats, e.g., XML-compliant formats, PDF, etc. Since we cannot make every annotator capable of processing all such formats, we usually employ a conversion process by which to take texts from source documents and encode them in an intermediate format which is assumed to be understandable by annotators. Such a collection of texts we usually call a corpus.

A typical way of constructing a corpus is to collect a large amount of texts, to encode them in a convenient format and to forget about the source documents.

***

We are in need of a framework to maintain the link between the texts curated in a corpus and their source documents, which is why the GENIA Corpus Curation Framework comes up.

2. GENIA Corpus Curation Framework

The *GENIA Corpus Curation Framework* (*GCCF*) defines an XML-based encoding format for text data to be used for NLP, IE or TM, and it also defines devices required to maintain the link between curated text data, the corresponding source documents and the annotations which are produced by annotators to the text data.

The fundamental objects of GCCF are *excerpt*s. An excerpt includes a text block from a *source document* while preserving most of additional information which is assumed to be useful for NLP, IE or TM, e.g., emphasized texts, embedded figures or tables, cross-references, etc.

Each excerpt is enclosed in a wrapping element which contains all the information to maintain the link to the source text block so that when required the annotation to the text can be re-attributed to the source text block.

The link between an excerpt and the corresponding source text block is facilitated in two folds: a pointer from excerpt to source and a sync data. The pointer of an excerpt to its source is the URL of the source text block written using XPointer, which means the source text has to be stored on a persistent database which is accessible through the net.

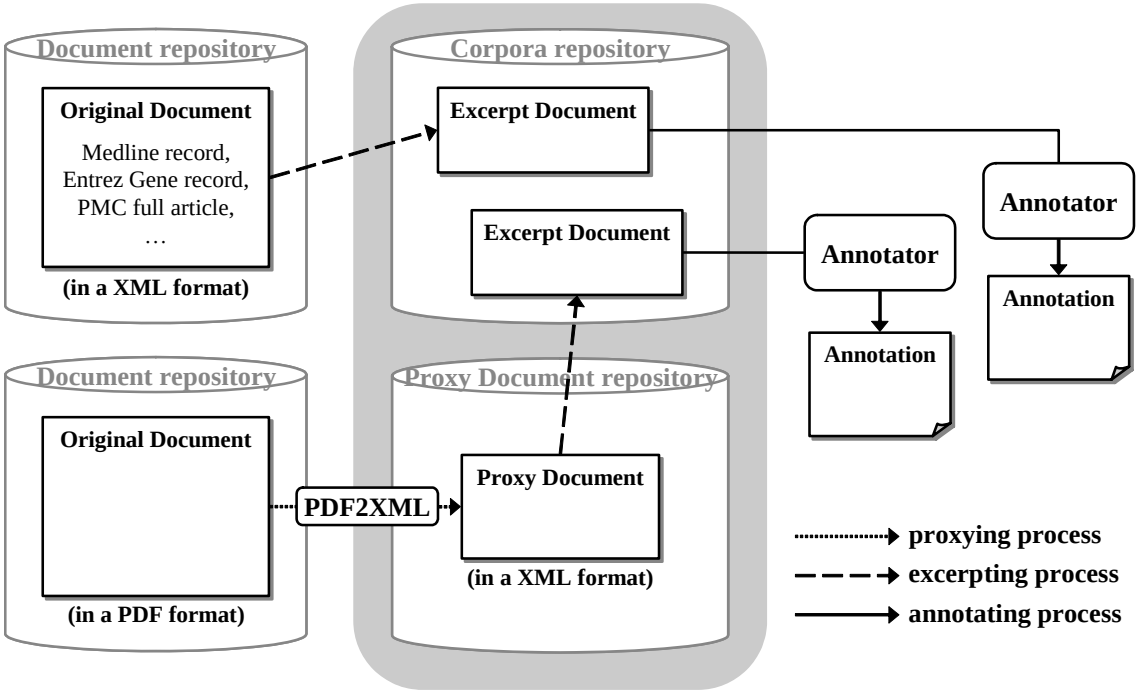
The sync between an excerpt and its source text block is established by maintaining a hash code which is a digest of the content.

The use of URL to point the source text block requires the source document to accessible from the net, and the use of XPointer to point the specific block requires the source document to be encoded in an XML-compliant format, both of which are not always the case.

When an *original document* is encoded in a none-XML-compliant format, e.g., PDF, we need to convert it to XML to make it applicable for GCCF. In that case, we call the XML version of original document a *proxy document* to differentiate it from the original document. The different of the original and proxy document is not only in the format but also the in the maintaining authority. Original documents are maintained by its copyright holder while the proxy document is maintained by the creator of the proxy document.

To sum up, the source document of an excerpt is either the original document or a proxy document of the original document

An *original document* is a document maintained by its copyright holder or delegate. Usually, publishers, aggregators or libraries take the role. In GCCF, each original document needs to be stored in a persistent database which is accessible through the net.



3. Consideration of Existing Formats

3.1. Medline record

The Medline database indexes bibliographic information of journal articles of life science. NCBI provides efetch utility with which each Medline record can be retrieved

by its PubMed ID (PMID). Accordingly, each record has its URL which looks like as follows:
http://www.ncbi.nlm.nih.gov/entrez/eutils/efetch.fcgi?db=pubmed&retmode=xml&rettype=medline&id=7641692.

The URI calls the efetch utility to retrieve from PubMed the Medline record of PMID 7641692 in XML encoding. Figure 3-1 shows the retrieved XML document. Note that it only shows selected parts of the documents with the other parts collapsed into ellipsis symbols. Since we are interested in constructing a corpus of natural language text, the ArticleTitle and AbstractText elements will be extracted from the records to become the target of annotation.

```
<?xml version="1.0"?>
<!DOCTYPE PubmedArticleSet PUBLIC "-//NLM//DTD PubMedArticle, 1st
November 2004//EN"
"http://www.ncbi.nlm.nih.gov/entrez/query/DTD/pubmed_041101.dtd">
<PubmedArticleSet>
<PubmedArticle>
   <MedlineCitation Owner="NLM" Status="MEDLINE">
      <PMID>7641692</PMID>
      …
      <Article PubModel="Print">
         …
         <ArticleTitle>TEXT</ArticleTitle>
         …
         <Abstract>
            <AbstractText>TEXT</AbstractText>
         </Abstract>
         …
      </Article>
      …
   </MedlineCitation>
    …
</PubmedArticle>
</PubmedArticleSet>
```

Figure 3-1. A Medline record in its native XML encoding

3.2. Entrez Gene Record

Entrez Gene is a database of gene-specific information. An Entrez Gene record contains rich information about a gene covering its locus, properties, homology, etc. From a perspective of NLP or IE, we are interested in the text description contained in the **Entrezgene_summary** element. The **Gene-ref_locus** and **Gene-ref_desc** elements contain the official symbol and name of the gene respectively.

<?xml-stylesheet type="text/css" href="xml-default.css"?>
<Entrezgene>
<Entrezgene_track-info>…</Entrezgene_track-info>
<Entrezgene_type value="protein-coding">6</Entrezgene_type>
<Entrezgene_source>…</Entrezgene_source>
<Entrezgene_gene>
…
<Gene-ref_locus>*NFKB1***</Gene-ref_locus>**
<Gene-ref_desc>*nuclear factor of kappa light polypeptide gene enhancer in B-cells 1 (p105)***</Gene-ref_desc>**
…
</Entrezgene_gene>
<Entrezgene_prot>…</Entrezgene_prot>
<Entrezgene_summary>*This gene encodes a 105 kD protein which can undergo cotranslational processing by the 26S proteasome to produce a 50 kD protein. The 105 kD protein is a Rel protein-specific transcription inhibitor and the 50 kD protein is a DNA binding subunit of the NF-kappa-B (NFKB) protein complex. NFKB is a transcription regulator that is activated by various intra- and extra-cellular stimuli such as cytokines, oxidant-free radicals, ultraviolet irradiation, and bacterial or viral products. Activated NFKB translocates into the nucleus and stimulates the expression of genes involved in a wide variety of biological functions. Inappropriate activation of NFKB has been associated with a number of inflammatory diseases while persistent inhibition of NFKB leads to inappropriate immune cell development or delayed cell growth.***</Entrezgene_summary>**
<Entrezgene_location>…</Entrezgene_location>
<Entrezgene_gene-source>…</Entrezgene_gene-source>
<Entrezgene_locus>…</Entrezgene_locus>
<Entrezgene_properties>…</Entrezgene_properties>
<Entrezgene_homology>…</Entrezgene_homology>
<Entrezgene_comments>…</Entrezgene_comments>
</Entrezgene>

Figure 3-2. An Entrez Gene record in its native XMLencoding

3.3. PMC record

PubMed Central (PMC) is an archive of life science journal literature services by NCBI. It functions as a digital library providing free or open access to the full-text of journal articles. All or part of journal articles from many publishers (including BioMed Central, PLoS, Elsevier) are being archived in PMC.

```
<?xml version="1.0"?>
<!DOCTYPE pmc-articleset PUBLIC "-//PMC//DTD ARTICLE SET 2.0//EN">
<pmc-articleset>
<article xmlns:xlink="http://www.w3.org/1999/xlink" article-type="research-article">
<front>article-title, abstract, …</front>
<body>sec, fig, table-wrap, supplementary-material</body>
<back>fig, table-wrap, fn, …</back>
</article>
</pmc-articleset>
```

Figure 3-3. The Structure of the PubMed Central record

Figure 3-3 shows the top-level structure of the PMC record retrieved by the efetch utility. In XML encoding of PMC archiving, one article consists of its front matter (**front**), body (**body**) and back matter (**back**). Of course, our primal interest is in the **body** element which contains the full-text of the article. The title (**article-title**) and abstract (**abstract**) from the front matter also contain important information about the content of the article. Other sources of natural language text include footnotes (**fn**) and captions (**caption**) of figures (**fig**), tables (**table-wrap**) and supplementary-materials (**supplementary-material**).

The **abstract** and **body** elements consists of one or more **sec** (section) elements which usually starts with one **title** element followed by one or more **p** (paragraph) elements. The **sec** element is recursive and may contain other **sec** elements to present its sub-sections.

The **caption** element is similar to the **sec**. It may start with a **title** followed by one or more **p** elements, but it is not recursive. The **fn** element contains one or more **p** elements. The **article-title** and **title** elements contain text which is not divided into paragraphs.

The text in the context of PMC article is rather a rich text including the following tags:

- Emphasizing texts
  - bold, italic, underline, overline, strike, monospace
- Subscripts and superscripts
  - sub, sup
- Non-text objects
  - inline-graphic, inline-formula, display-formula
- Cross-references
  - xref, ext-link

```
article_title := rich-text
abstract := sec+
body := sec+
caption := title?, p+
fn := p+
sec := title, (p|fig|table-wrap|supplementary-material|sec)+
title := (text|bold|italic|underline|overline|strike|monospace|sub|sup|xref|ext-link)+
p := (text|bold|italic|underline|overline|strike|monospace|sub|sup|xref|ext-link
      |inline-graphic|inline-formula|display-formula)+
```

Figure 3-4. Document Model of a PMC article

As a summary, Figure 3-4 shows the document model of a PMC article described so far[†]. We are going to take excerpts of the first five elements: **article**, **abstract**, **body**, **caption** and **fn**. The **abstract** and **body** elements consist of a number of **sec** element. The **sec** and **caption** elements consist of a **title** and a number of **p** elements. Additionally, the **sec** element may include **table-wrap** and **supplementary-material** elements, the content of which will be ignored except for the caption. The **fn** element consists of a number of **p** elements without **title**. The text in **p** or **title** elements may be emphasized by **bold**, **italic**, **underline**, **overline**, **strike**, **monospace**, **sub** or **sup** tags. The **p** element may also include **inline-graphic**, **inline-formula** and **display-formula** elements, the content of which will be ignored. A region of text may be associated with other elements, e.g., fig, table-wrap, fn, etc. by **xref** tag or outside resources by **ext-link** tag.

Figure 3-5 shows a paragraph encoded in NLM XML encoding format.

```
<p>Having observed the increased IL-17 production in RA PBMC, it was important to
know which signal transduction pathways were involved. As illustrated in Fig. <xref
ref-type="fig" rid="F3">3</xref>, an significant decrease in anti-CD3-induced IL-17
production was observed when co-incubated with NF-κB inhibitor, PDTC and
dexamethasone in comparison with anti-CD3 alone (38 ± 5 and 54 ± 11 versus 98 ± 19
pg/ml, respectively; <italic>P</italic> &lt; 0.05).</p>
```

Figure 3-5. A paragraph taken from a PMC article

4. Excerpts

The excerpt is the most fundamental unit of GCCF containing a text block. The excerpt document is a physical unit corresponding to a file on a file system. One excerpt document may contain one or more excerpts. For practical reason, an excerpt document is expected to contain excerpts from one source document. The root element of an excerpt document is **excerptset**. An **excerpt** has its own ID (**id**) and information of the

[†] The document structure defined in NLM is much more complex. We have simplified the document model after examining 44 articles from 17 journals. The articles were retrieved from PMC by using the MeSH term “blood cells” and “transcription factors”.

source document (**srcDB** and **srcrecID**) and excerpted text block (**srctxtURL**). It also maintain the sequence of excerpts (**seqprev** and **seqnext**). Section 4.1, 4.2 and 4.3 show examples of excerpts taken from Medline, Entrez Gene and PubMed Central respectively.

4.1. Excerpts from Medline

From a Medline record, two excerpts corresponding to the **ArticleTitle** and **AbstractText** elements are taken. The **texttype** are specified as *article-title* and *abstract* respectively.

<?xml version="1.0" encoding="UTF-8"?>
<excerptset>
<excerpt id="25" srcDB="Medline" srcrecID="1299224" date="08SEP2006" texttype="article-title" seqnext="26" srctxtURL="http://www.ncbi.nlm.nih.gov/entrez/eutils/efetch.fcgi?db=pubmed&retmode=xml&rettype=medline&id=1299224#xpointer(/PubmedArticleSet[1]/PubmedArticle[1]/MedlineCitation[1]/Article[1]/ArticleTitle[1])"
>Ablation of transplanted HTLV-I Tax-transformed tumors in mice by antisense inhibition of NF-kappa B [published erratum appears in Science 1993 Mar 12;259(5101):1523]**</excerpt>**
< excerpt id="26" srcDB="Medline" srcrecID="1299224" date="08SEP2006" texttype="abstract" seqprev="25" srctxtURL="http://www.ncbi.nlm.nih.gov/entrez/eutils/efetch.fcgi?db=pubmed&retmode=xml&rettype=medline&id=1299224#xpointer(/PubmedArticleSet[1]/PubmedArticle[1]/MedlineCitation[1]/Article[1]/Abstract[1]/AbstractText[1])" >Mice transgenic for the human T cell leukemia virus (HTLV-I) Tax gene develop fibroblastic tumors that express NF-kappa B-inducible early genes. In vitro inhibition of NF-kappa B expression by antisense oligodeoxynucleotides (ODNs) inhibited growth of these culture-adapted Tax-transformed fibroblasts as well as an HTLV-I-transformed human lymphocyte line. In contrast, antisense inhibition of Tax itself had no apparent effect on cell growth. Mice treated with antisense to NF-kappa B ODNs showed rapid regression of transplanted fibrosarcomas. This suggests that NF-kappa B expression may be necessary for the maintenance of the malignant phenotype and provides a therapeutic approach for HTLV-I-associated disease.**</excerpt>**
</excerptset>

Figure 4-1. Excerpts taken from a Medline record

4.2. Excerpts from Entrez Gene

From an Entrez Gene record, the **Gene-ref_locus** element which contains the official symbol of the gene name and the **Entrezgene_summary** element which contains a description of the gene are take to make two excerpts. The **texttype** is specified as *symbol* and *description* respectively

<?xml version="1.0" encoding="UTF-8"?>
<excerptset>
<excerpt id="12563" srcDB="EntrezGene" srcrecID="4790" date="08SEP2006" texttype="symbol" seqnext="12564" srctxtURL="http://www.ncbi.nlm.nih.gov/entrez/eutils/efetch.fcgi?db=gene&retmode=xml&id=4790#xpointer(/Entrezgene[1]/Entrezgene_gene[1]/Gene-ref[1]/Gene-ref_locus[1])">*NFKB1***</excerpt>**
<excerpt id="12564" srcDB="EntrezGene" srcrecID="4790" date="08SEP2006" texttype="description" seqprev="12563" srctxtURL="http://www.ncbi.nlm.nih.gov/entrez/eutils/efetch.fcgi?db=gene&retmode=xml&id=4790#xpointer(/Entrezgene[1]/Entrezgene_summary[1])">*Mice transgenic for the human T cell leukemia virus (HTLV-I) Tax gene develop fibroblastic tumors that express NF-kappa B-inducible early genes. In vitro inhibition of NF-kappa B expression by antisense oligodeoxynucleotides (ODNs) inhibited growth of these culture-adapted Tax-transformed fibroblasts as well as an HTLV-I-transformed human lymphocyte line. In contrast, antisense inhibition of Tax itself had no apparent effect on cell growth. Mice treated with antisense to NF-kappa B ODNs showed rapid regression of transplanted fibrosarcomas. This suggests that NF-kappa B expression may be necessary for the maintenance of the malignant phenotype and provides a therapeutic approach for HTLV-I-associated disease.*****</excerpt>**
</excerptset>

Figure 4-2. Excerpts taken from a Entrez Gene record

4.3. Excerpts from PMC

From a PMC article, **article-title**, **abstract**, **body**, **caption** and **fn** elements are taken as excerpts. The **texttype** is specified as *article-title*, *abstract*, *article-body*, *fig-caption* and *footnote* respectively. Figure 4-3 shows the skeleton of an excerpt document for a PMC article.

```
<?xml version="1.0" encoding="UTF-8"?>
<excerptset>
<excerpt id="58391" srcDB="PMC" srcrecID="1134656" date="08SEP2006"
texttype="article-title"  seqnext="58392"
srctxtURL="http://www.ncbi.nlm.nih.gov/entrez/eutils/efetch.fcgi?db=pmc&retm
ode=xml&id=1134656#xpointer(/pmc-articleset[1]/article[1]/front[1]/article-
meta[1]/title-group[1]/article-title[1])" >…</excerpt>
<excerpt id="58392" srcDB="PMC" srcrecID="1134656" date="08SEP2006"
texttype="abstract"  seqprev="58391" seqnext="58393"
srctxtURL="http://www.ncbi.nlm.nih.gov/entrez/eutils/efetch.fcgi?db=pmc&retm
ode=xml&id=1134656#xpointer(/pmc-articleset[1]/article[1]/front[1]/article-
meta[1]/abstract[1])" >…</excerpt>
<excerpt id="58393" srcDB="PMC" srcrecID="1134656" date="08SEP2006"
texttype="article-body"  seqprev="58392" seqnext="58394"
srctxtURL="http://www.ncbi.nlm.nih.gov/entrez/eutils/efetch.fcgi?db=pmc&retm
ode=xml&id=1134656#xpointer(/pmc-articleset[1]/body[1])" >…</excerpt>
<excerpt id="58394" srcDB="PMC" srcrecID="1134656" date="08SEP2006"
texttype="fig-caption"  seqprev="58393" seqnext="58395"
srctxtURL="http://www.ncbi.nlm.nih.gov/entrez/eutils/efetch.fcgi?db=pmc&retm
ode=xml&id=1134656#xpointer(/pmc-
articleset[1]/back[1]/sec[1]/fig[1]/caption[1])">…</excerpt>
…
</excerptset>
```

Figure 4-3. Excerpts taken from a PMC article

Note that the excerpts from PMC article do not contain just simple texts. They contain kind of rich texts including tags for structuring, emphasizing, cross-referencing, etc according to the document model shown in Figure 3-4. Thus, annotators to process those excerpts need to know the document model.

Sometimes a part of text is associated with other objects using xref. In such a case, the **associate** tag is used to explicitly state the association. Figure 4-4 shows an example of association.

We clustered the mouse/human ortholog pairs based on their expression profiles across the 140 mouse and human tissues (Methods). We computed clusters at cut-off levels ranging from Pearson's correlation coefficient (hereafter termed PCC) 0.61 to 0.99. At the lowest applied cut-off, 57% of all ortholog pairs in the data were members of a cluster. The cluster sizes were distributed in a skewed manner, with a predominant formation of small clusters (Figure **<associate ref="58394">**<xref ref-type="fig" rid="F1">1A</xref>**</associate>**). All analyses hereafter were performed at a PCC = 0.75 cut-off, which generated 160 clusters with 2407 ortholog pairs. This relatively stringent cut-off was chosen to reduce the number of non-relevant genes in the clusters. A higher cut-off did not seem reasonable given the noise-level of the microarray experiments (as judged by the Pearson correlation between replicated samples and between probes that are annotated for the same gene, data not shown).

Figure 4-4. Example of association with another excerpt.