Extraction of Document Structure for Genomics Documents

David Eichmann,[1,2]

[1]School of Library and Information Science
[2]Institute for Clinical and Translational Science
The University of Iowa
david-eichmann@uiowa.edu

The University of Iowa participated only in the genomics track in TREC-2006. Our work concentrated almost entirely on exploring how accurately we could regenerate the logical structure of each of the documents in the corpus from their HTML instantiations. This year's work is hence primarily infrastructure building, with little in the way of support for the track's specific tasks in place.

1 – Document Structure Extraction

We are taking as our foundational assumption that effective information retrieval tasks in the broad domain of biomedical literature must address the singular nature of scholarly communication and the effect this has upon a document corpus. The corpus for a typical TREC task is comprised of a temporal sequence of news documents exhibiting little, if any, internal structure. Reduction of a document to a single term vector is viable because of that document's inherent *aboutness* regarding the story being reported. A scholarly paper, on the other hand, has distinct components, each fulfilling a specific function and exhibiting a correspondingly specific syntactic structure. A query to retrieve documents reporting new results regarding a particular biological organism should not retrieve candidates that mention that organism in the references or background sections, but rather ones that mention the organism in the results and discussion sections.

This structure-centric assumption leads to the following goals:

- identification and extraction of all prefatory material (i.e., title, authors, affiliations, etc.);
- identification of section and subsection headers;
- identification and extraction of figures, tables and their respective legends;
- identification and extraction of bibliographic citations; and
- transformation of all HTML character entity occurrences and their inline image surrogates into the corresponding Unicode character.

Our intent here is to create as clean a presentation of the narrative of the paper as possible, while retaining the ability to use figure legends, bibliographic citations, etc. as discrete and value-adding document elements.

For an arbitrary HTML document retrieved from the Web, extraction of its internal structure (if any) is a daunting task. The documents in the genomics corpus are fortunately not a random collection of HTML files, but are rather members of a limited set of document format classes roughly corresponding to the journal in which they were published. In sampling each journal source from the collection, we established that only two distinct, non-overlapping format classes contained 135,207 of the 162,259 documents in the collection (83%). Further examination of the remaining documents yielded markup hints that decomposed the entire corpus into six format classes of content-carrying documents and one junk class of documents carrying no useful content. Table 1 lists the document pools by inferred generator.

Table 1: Format Classes in Genomics Corpus

Format Classes	Corpus Occurrences	Exclusions	Contributing Documents	Cumulative Coverage
'BIBL' anchor for references	77827	none	77827	48.0%
Electronic Press Engine	57380	none	57380	83.3%
'cbot' links	29797	BIBL, EPE	6568	87.4%
'breadcrumb' links	3214	BIBL, EPE	2138	88.7%
'mosaic' comments	58566	BIBL, EPE, cbot, breadcrumb	4084	91.2%
'other article' links	131676	BIBL, EPE, cbot, breadcrumb, mosaic	14053	99.9%
small files (ignored)	3067	none	0	n/a

We then constructed format class-specific structure extractors for each of the format classes, based upon our previous work on extracting document structure from PDF versions of scholarly papers [1]. Each extractor generates a document object comprised of a sequence of sections, which are in turn comprised of subsections (if present), and each (sub)section is comprised of a sequence of paragraphs. Figures, tables and references are attached to the document as separate elements from the text. All navigation bars and other HTML artifacts are removed. Each paragraph instantiated hence reflects a reader's intuition of a paragraph - a contiguous sequence of sentences with no intervening document structures. All paragraphs generated in this manner are no larger than those defined by the track's 'legal span' definition, do not cross paragraph boundary tags (hence making them legal spans for submission), and in many cases are much smaller in size than the containing official legal span.

2 – Topic Processing

We adapted our work for question answering and novelty detection [2, 3] to the classification of topics. Each topic is parsed and the parse tree matched against a set of templates constructed for this domain, as shown in Table 2. The extraction patterns identify salient elements of the topic definition as well as identifying the topic class.

Table 2: Topic Classes and Extraction Patterns

Topic Class	Extraction Pattern
simple role	VP <1 AUX <2 [NP <1 [NP <1 DT <2 /role/] <2 [PP <1 IN <2 (NP)] <3 [PP <1 IN <2 (NP)]]
definition role	VP <1 AUX <2 [NP <1 [NP <1 DT <2 /role/] <2 [PP <1 IN <2 (NP)] <3 [PRN <2 (NP)] <4 [PP <1 IN <2 (NP)]]
	VP <1 AUX <2 [NP <1 [NP <1 DT <2 /role/] <2 [PP <1 IN <2 (NP)]] <3 [PRN <2 (NP)] <4 [PP <1 IN <2 (NP)]
located mutation action	SQ <2 [NP <1 [NP <1 /mutations/] <2 [PP <1 IN <2 (NP)]] <3 [VP <1 (VB) <2 (NP) <3 [PP <1 IN <2 (NP)]]
mutation action	SQ <2 [NP <1 [NP <1 /mutations/] <2 [PP <1 IN <2 (NP)]] <3 [VP <1 (VB) <2 (NP)]
interaction function action	SQ <2 (NP <-1 /interaction(s)?/) <3 [VP <1 (VB) <2 (NP)]
	SQ <2 [NP <1 [NP <1 /interaction(s)?/] <2 [PP <1 IN <2 [NP <1 (/.*/) <2 CC <3 (/.*/)]]] <3 [VP <1 (VB) <2 (NP)]
function action	SQ <2 (NP) <3 [VP <1 (VB) <2 (NP)]
interaction contribution	SQ <2 [NP <1 (NP) <2 CC <3 (NP <-1 /interaction(s)?/)] <3 [VP <1 / contribute/ <2 [PP <1 TO <2 (NP)]]
	SQ <2 (NP) <3 [VP <1 /interact/ <2 [PP <1 IN <2 (NP)] <3 [S << (NP)]]
located contribution	SQ <2 (NP) <3 [VP <1 /contribute/ <2 [PP <1 TO <2 (NP)] <3 [PP <1 IN <2 (NP)]]
contribution	SQ <2 (NP) <3 [VP <1 /contribute/ <2 [PP <1 TO <2 (NP)]]

3 – Official Runs

Extracted terms are expanded with synonyms identified using UMLS. Paragraphs as defined in section 1, or some contraction of paragraphs, formed the baseline unit of retrieval for our submissions. We detect sentence boundaries on each paragraph. Each sentence is then matched against the vector of expanded terms for each active vector (as specified by the topic class). Our first submission (UIowa06Gen01) returns any paragraph that matches at least one term from each active vector for the current topic. Our second submission (UIowa06Gen02) then contracts the starting and ending positions of the passage to be returned to eliminate leading and trailing sentences that do no contain at least one term from each active vector for the current topic. Our final submission (UIowa06Gen03) constrains our first submission to only those paragraphs that are at least 300 characters in length.

Table 3 presents our official results. For the passage level measure, the contracted passage run outperforms the other two, indicating that contraction can be an effective means of improving precision. For the aspect level measure, the length limited run outperforms the other two, an interesting result, given that a topic-compliant passage must address the same range of MeSH concepts as the topic specification. For the document level measure, the full paragraph run outperforms the other two, implying that our structure extraction scheme is generating paragraphs that of relatively high value, even when those paragraphs are relatively long.

Table 3: Official Results (all runs are automatic)

ID	Description	Passage MAP	Aspect MAP	Document MAP
UIowa06Gen01	NLP processing of question, entire paragraph returned as result	0.0039	0.0199	0.0234
UIowa06Gen02	NLP processing of question, paragraphs contracted to only those sentences mentioning query terms	0.0044	0.0187	0.0200
UIowa06Gen03	NLP processing of question, entire paragraphs returned as long as they are at least 300 characters in length	0.0039	0.0219	0.0198

4 – Conclusions

When combined with our previous work on PDF versions of papers [1], we now have a flexible means of extracting useful document structure from a broad range of scholarly sources. The scores for our submitted runs, while relative modest compared to other submissions, appear to indicate that our approach yields relative ordering of the three measures that holds potential for substantial improvement. The framework is easily extensible with new format classes. Our work on the genomics track for next year will now concentrate on biomedical entity and relationship recognition.

References

[1] Bradshaw, S, Light, M. and D. Eichmann, "(Bee)Dancing on the Boundaries Between PIM and GIM," *SIGIR Workshop on Personal Information Management*, Seattle, WA August 10-11, 2006.

[2] Eichmann, D., Y. Zhang, S. Bradshaw, X. Y. Qiu, L. Zhou, P. Srinivasan, A. Sehgal and H. Wang, "Novelty, Question Answering and Genomics: The University of Iowa Response," *Thirteenth Conference on Text Retrieval*, NIST, Washington, D.C., November 17-19, 2004.

[3] Eichmann, D., and P. Srinivasan, "Experiments in Questions and Relationships at The University of Iowa," *Fourteenth Conference on Text Retrieval*, NIST, Washington, D.C. November 16-18, 2005.