An End-to-End PDF Toolchain for Marking Up Scientific Documents

Sanna Hulkkonen, Oliver Ray
University of Bristol,UK
National Institute of Informatics, Japan
sh16145.2016@my.bristol.ac.uk, csxor@bristol.ac.uk

Abstract

This paper proposes a system for making sentence-level semantic enrichment of scientific publications more user-friendly by developing an end-to-end toolchain for augmenting PDFs with automatically determined textual annotations and visual highlights. The aim is to categorise each sentence according to a given classification scheme and display the labels in a visually appealing way that preserves document structure and formatting while allowing users to work with standard PDF tools they are already accustomed to. This is in contrast to existing approaches which provide an XML representation of document content obtained by abstracting away formatting and structural details in order to focus on the raw text. In particular, we present a toolchain that automatically marks up each sentence in the body of a PDF with a Core Scientific Concept category using a classifier trained with a corpus of papers on social insect biology that we manually labelled. Preliminary testing with domain experts provides anecdotal evidence that end-users do find such automatically derived sentence classifications useful and that they prefer to work directly with marked up PDFs.

Keywords: Portable Document Format, Mark Up, Core Scientific Concepts

1. Introduction

This work is concerned with the semantic enrichment of scientific publications using sentence-level classifications like Argumentative Zoning (AZ) (Teufel, 1999) or Core Scientific Concepts (CoreSC) (Liakata et al., 2012). Its specific focus is on finding sentence labels with a combination of Natural Language Processing (NLP) and Machine Learning (ML) and projecting the labels back onto the original document using PDF manipulation tools.

Our goal is to display sentence labels in a visually intuitive way that preserves document formatting while allowing users to work with standard PDF software they are already accustomed to. This is in contrast to existing approaches which provide an XML representation of a document's content by abstracting away formatting details so the plain text of each sentence can be enclosed within semantic tags. While XML is not a suitable filetype for most end-users to work with, the sentence labels can be conveniently visualised using a text-based annotation interface like Brat (Stenetorp et al., 2012), as shown in Figure 1:

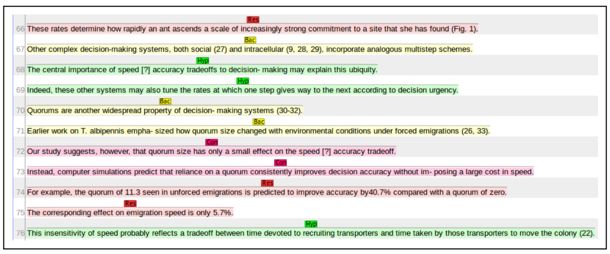


Figure 1: Typical view of sentence mark up.

Our main contribution is developing an end-to-end toolchain that inserts CoreSC annotations directly into the original PDF using classifiers trained on a corpus of papers from social insect biology that we labelled ourselves. Preliminary tests with domain experts suggest they do find such annotations useful and that they prefer to work with marked up PDFs of the sort illustrated in Figure 2:

Figure 2: Proposed view of sentence mark up.

2. System

Our initial work for generating marked up PDFs of the form shown in Figure 2 is detailed in (Hulkkonen, 2017). The motivation for that work comes from a requirement to present social insect biologists with intuitive sentence-level annotations highlighting some key aspects of a scientific investigation as formalised by the CoreSC ontology (Liakata et al., 2012) .

To do this, we used the Sapienta system (Liakata et al., 2012) to convert PDFs into XML and classify sentences among 11 CoreSC categories which we used 4 colours to highlight: Motivation, Goal, Object, Hypothesis, Background, Method, Model, Experiment and Observation, Result, Conclusion. This resulted in an output of the form shown in Figure 1.

After that we used pdftotext to extract bounding boxes of the words in the classified sentences in order to project the annotations back onto the PDF as shown in Figure 2.

But, while this is mostly satisfactory, we discovered some incompatibilities between Sapienta's XML conversions (that use the PDFX utility for textual content extraction) and our back projections (that use pdftotext for bounding box extraction) which mean that some highlighting errors are unavoidable using this method.

To get around this, we re-implemented our approach using a prototype PDF toolkit called PDFNLT that is currently being developed by the Aizawa lab at the Japanese National Institute of Informatics (NII). The advantage of this system is that, given a PDF as input, it produces as output an XHTML version of the document that includes bounding box information along with a CSV file that indicates which words belong to which sentences.

To give users the ability to manually edit sentence labels, we helped the PDFNLT team develop the web interface shown in Figure 3. In essence the left hand pane corresponds to textual view of Figure 1 (and is well suited to relabelling a selected sentence by overwriting the relevant text box) while the right hand pane corresponds to graphical view of Figure 2 (and is well suited to visualising the context of a selected sentence through highlighting).

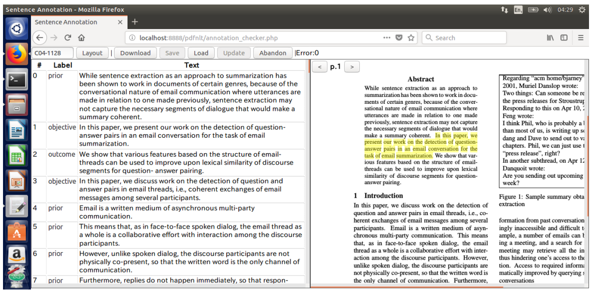
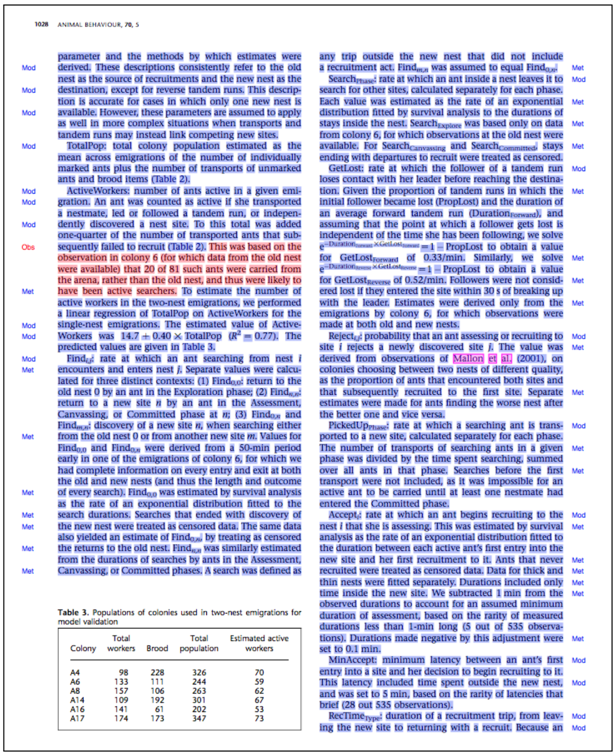


Figure 3: Developmental web interface that combines the content-based and presentation-based approaches.

3. Results

Although our visualisation tool can be used with any pre-trained sentence classifier, we also wanted to test the effect of retraining Sapienta's CRF model. So we manually annotated 5,300 sentences from 27 papers from our ant corpus. We asked three domain experts to verify the annotations of one document; and, in this small sample, we found an inter-annotator agreement above 90%. We then used 19 papers for training and 8 for testing. These tests showed that retraining significantly improved accuracy and further gains were also achieved by debugging some of Sapienta's XML conversion and sentence splitting code.

The domain experts stated they preferred to work with highlighted PDFs rather than an XML format as they liked to see


1028 ANIMAL BEHAVIOUR, 70, 5

parameter and the methods by which estimates were derived. These descriptions consistently refer to the old nest as the source of recruitments and the new nest as the destination, except for reverse tandem runs. This description is accurate for cases in which only one new nest is available. However, these parameters are assumed to apply as well in more complex situations when transports and tandem runs may instead link competing new sites.

TotalPop: total colony population estimated as the mean across emigrations of the number of individually marked ants plus the number of transports of unmarked ants and brood items (Table 2).

ActiveWorkers: number of ants active in a given emigration. An ant was counted as active if she transported a nestmate, led or followed a tandem run, or independently discovered a nest site. To this total was added one-quarter of the number of transported ants that subsequently failed to recruit (Table 2). This was based on the observation in colony 6 (for which data from the old nest were available) that 20 of 81 such ants were carried from the arena, rather than the old nest, and thus were likely to have been active searchers. To estimate the number of active workers in the two-nest emigrations, we performed a linear regression of TotalPop on ActiveWorkers for the single-nest emigrations. The estimated value of ActiveWorkers was $14.7 + 0.40 \times$ TotalPop ($R^2 = 0.77$). The predicted values are given in Table 3.

$Find_{i,j}$: rate at which an ant searching from nest i encounters and enters nest j. Separate values were calculated for three distinct contexts: (1) $Find_{0,0}$: return to the old nest 0 by an ant in the Exploration phase; (2) $Find_{n,n}$: return to a new site n by an ant in the Assessment, Canvassing, or Committed phase at n; (3) $Find_{0,n}$ and $Find_{m,n}$: discovery of a new site n, when searching either from the old nest 0 or from another new site m. Values for $Find_{0,0}$ and $Find_{0,n}$ were derived from a 50-min period early in one of the emigrations of colony 6, for which we had complete information on every entry and exit at both the old and new nests (and thus the length and outcome of every search). $Find_{0,0}$ was estimated by survival analysis as the rate of an exponential distribution fitted to the search durations. Searches that ended with discovery of the new nest were treated as censored data. The same data also yielded an estimate of $Find_{0,n}$, by treating as censored the returns to the old nest. $Find_{n,n}$ was similarly estimated from the durations of searches by ants in the Assessment, Canvassing, or Committed phases. A search was defined as any trip outside the new nest that did not include a recruitment act. $Find_{m,n}$ was assumed to equal $Find_{0,n}$.

$Search_{Phase}$: rate at which an ant inside a nest leaves it to search for other sites, calculated separately for each phase. Each value was estimated as the rate of an exponential distribution fitted by survival analysis to the durations of stays inside the nest. $Search_{Explore}$ was based only on data from colony 6, for which observations at the old nest were available. For $Search_{Canvassing}$ and $Search_{Committed}$, stays ending with departures to recruit were treated as censored.

GetLost: rate at which the follower of a tandem run loses contact with her leader before reaching the destination. Given the proportion of tandem runs in which the initial follower became lost (PropLost) and the duration of an average forward tandem run ($Duration_{Forward}$), and assuming that the point at which a follower gets lost is independent of the time she has been following, we solve $e^{-Duration_{Forward} \times GetLost_{Forward}} = 1 - PropLost$ to obtain a value for $GetLost_{Forward}$ of 0.33/min. Similarly, we solve $e^{-Duration_{Reverse} \times GetLost_{Reverse}} = 1 - PropLost$ to obtain a value for $GetLost_{Reverse}$ of 0.52/min. Followers were not considered lost if they entered the site within 30 s of breaking up with the leader. Estimates were derived only from the emigrations by colony 6, for which observations were made at both old and new nests.

$Reject_{i,j}$: probability that an ant assessing or recruiting to site i rejects a newly discovered site j. The value was derived from observations of Mallon et al. (2001), on colonies choosing between two nests of different quality, as the proportion of ants that encountered both sites and that subsequently recruited to the first site. Separate estimates were made for ants finding the worse nest after the better one and vice versa.

$PickedUp_{Phase}$: rate at which a searching ant is transported to a new site, calculated separately for each phase. The number of transports of searching ants in a given phase was divided by the time spent searching, summed over all ants in that phase. Searches before the first transport were not included, as it was impossible for an active ant to be carried until at least one nestmate had entered the Committed phase.

$Accept_i$: rate at which an ant begins recruiting to the nest i that she is assessing. This was estimated by survival analysis as the rate of an exponential distribution fitted to the duration between each active ant's first entry into the new site and her first recruitment to it. Ants that never recruited were treated as censored data. Data for thick and thin nests were fitted separately. Durations included only time inside the new site. We subtracted 1 min from the observed durations to account for an assumed minimum duration of assessment, based on the rarity of measured durations less than 1-min long (5 out of 535 observations). Durations made negative by this adjustment were set to 0.1 min.

MinAccept: minimum latency between an ant's first entry into a site and her decision to begin recruiting to it. This latency included time spent outside the new nest, and was set to 5 min, based on the rarity of latencies that brief (28 out 535 observations).

$RecTime_{Type}$: duration of a recruitment trip, from leaving the new site to returning with a recruit. Because an

Table 3. Populations of colonies used in two-nest emigrations for model validation

Colony	Total workers	Brood	Total population	Estimated active workers
A4	98	228	326	70
A6	133	111	244	59
A8	157	106	263	62
A14	109	192	301	67
A16	141	61	202	53
A17	174	173	347	73

Figure 4: Example highlighting an important observation buried within a large block of text about the methodology.

each sentence in the context of the surrounding textual and graphical cues. As shown in Figure 4 above, the highlighting even led to the discovery of an important observation within a whole page of methodology that would not have been so easily found without our tool. The observation described a seemingly commonplace behaviour (social carries initiated outside the nest) of whose existence our biological collaborators were unfamiliar.

4. Acknowledgements

We thank Goran Topic and Akiko Aizawa for the development of PDFNLT; we thank Maria Liakata and James Ravenscroft for help with the deployment of Sapienta; we thank Nigel Franks, Ana Sendova-Franks and Alan Worley for their domain expert feedback; and we thank Simone Teufel and the anonymous reviewers for useful feedback.

5. Bibliographical References

Hulkkonen, S. (2017). Automatically classifying the content of research publications. Master's thesis, Univ. of Bristol, UK.

Liakata, M. et al. (2012). Automatic recognition of conceptualization zones in scientific articles and two life science applications. *Bioinformatics*, 28(7):991–1000.

Stenetorp, P. et al. (2012). Brat: A web-based tool for nlp-assisted text annotation. In *Proc. of the Demonstrations at the 13th Conf. of the Europ. Chap. of the Assoc. for Comp. Linguistics*, EACL12, pages 102–107. Assoc. for Comp. Linguistics.

Teufel, S. (1999). *Argumentative Zoning: Information Extraction from Scientific Text*. Ph.D. thesis, Univ. of Edinburgh.