

CORAAL – Towards Deep Exploitation of Textual Resources in Life Sciences

Vít Nováček, Tudor Groza, Siegfried Handschuh

Digital Enterprise Research Institute (DERI)
National University of Ireland, Galway
IDA Business Park, Lower Dangan, Galway, Ireland
E-mail: `FirstName.LastName@deri.org`

Abstract. Prominent biomedical literature search tools like ScienceDirect, PubMed Central or MEDLINE allow for efficient retrieval of resources based on key words. Due to vast amounts of data available in life sciences, key word search is not always sufficient, though. One would often welcome more intelligent search for knowledge, i.e., for concepts and their mutual relations. This is, however, still a major challenge, since getting the necessary machine-readable knowledge manually is virtually impossible in large scale, while its automatic extraction is not particularly reliable. We have researched a novel framework actually enabling practical exploitation of automatically extracted knowledge, though. On the top of the framework, we implemented CORAAL, a prototype for knowledge-based biomedical literature search. This paper describes its essential principles, innovative capabilities and current results.

1 Introduction

Digital content processing has no doubt introduced a whole lot of new possibilities of dealing with scientific publications. It makes knowledge much more open and exploitable than in the old “paper times”. However, one still needs to go manually through a lot of possibly irrelevant content very often before actually finding the right answers. If we are to make the next step, it is necessary to process knowledge (i.e., concepts and their mutual relations), and not just data or shallow meta-data (i.e., chunks of free text, titles or author names). Substantial automation of such meaning-intensive information processing is hardly possible with the current industry-strength technologies (e.g., full-text search), since they lack proper support for extraction, representation and processing of knowledge implicitly present in texts. As an illustration, imagine for instance finding a support of the claim that *acute granulocytic leukemia* is different from *T-cell leukemia*. With the current solutions, it is easy to find articles that contain both or either of the terms, however, the number of results may be quite high (e.g., 593 on PubMed). It is tedious or even impossible to go through all of them in order to find out which of them actually mention the two leukemias being different.

Methods for automated knowledge extraction than can dig more than mere key words from text exist, however, their results are deemed to be too noisy and sparse to be exploited by the current state of the art without significant manual post-processing [1]. We have recently researched a novel framework for effortless exploitation of automatically extracted knowledge that makes use of similarity-based knowledge representation and respective light-weight inference services [2]. We combined the framework with our repository for semantically inter-linked publications [3], delivering a prototype knowledge-based publication search engine – CORAAL (*COntent extended by emeRgent and Asserted Annotations of Linked publication data*). The tool essentially *extracts* asserted publication meta-data together with the knowledge implicitly present in the respective text, *integrates* the emergent content with existing domain knowledge and *exposes* it via a multiple-perspective search&browse interface. This way we allow for fine-grained publication search combined with convenient and effortless large scale exploitation of the knowledge associated with and hidden in the publication texts.

The rest of the paper is organised as follows. Section 2 describes the data used in the current CORAAL deployment, as well as the tool’s essential technological principles and capabilities. Section 3 reports on experiments assessing the applicability of CORAAL and quality of the knowledge served to users. Related work is analysed in Section 4. We discuss the potential of the delivered work, conclude the paper and outline future directions in Section 5.

2 Method

Here we describe the data processed by CORAAL, and summarise the essential technical principles of the prototype.

2.1 Inputs and Outputs

Input As of March 2009, we have processed 11,761 Elsevier journal articles from the provided XML repositories that were related to cancer research and treatment. The access to the articles was provided within the Elsevier Grand Challenge competition (cf. <http://www.elseviergrandchallenge.com>). The domain was selected so due to the expertise of our sample users and testers from Masaryk Oncology Institute in Brno, Czech Republic. We processed articles evenly distributed across the journals in the following list: 1. *FEBS Letters*; 2. *Biochemical Pharmacology*; 3. *Cancer Genetics and Cytogenetics*; 4. *Cell*; 5. *Trends in Cell Biology*; 6. *Experimental Cell Research*; 7. *Controlled Clinical Trials*; 8. *Molecular Aspects of Medicine*; 9. *Advanced Drug Delivery Reviews*; 10. *Gene*; 11. *Trends in Genetics*; 12. *Genomics*; 13. *Leukemia Research*; 14. *Journal of Microbiological Methods*; 15. *Trends in Microbiology*; 16. *Journal of Molecular Biology*; 17. *Oral Oncology*; 18. *European Journal of Pharmacology*. From the article repository, we extracted the knowledge and publication metadata for

further processing by CORAAL. Besides the publications themselves, we employed legacy machine-readable vocabularies for the refinement and extension of the extracted knowledge (currently, we use the NCI and EMTREE thesauri – see <http://www.cancer.gov/cancertopics/terminologyresources> and <http://www.embase.com/emtree/>, respectively).

Output CORAAL exposes two data-sets as an output of the publication processing: (1) We used a **triple store** containing publication meta-data (citations, their contexts, structural annotations, titles, authors and affiliations) associated with respective full-text indices. The resulting store contained 7,608,532 of RDF subject-predicate-object statements [4] describing the input articles. This included 247,392 publication titles and 374,553 authors (both from full-texts and references processed). (2) We employed a custom EUREEKA **knowledge base** [2] with facts of various certainty extracted and inferred from the article texts and the seed life science thesauri. Directly from the articles, 215,645 concepts were extracted (and analogically extended). Together with the data from the initial thesauri, the domain lexicon contained 622,611 terms, referring to 347,613 unique concepts. The size of the emergent knowledge base was 4,715,992 weighed statements (ca. 99 and 334 extracted and inferred statements per publication in average, respectively). The contextual meta-knowledge related to the statements, namely provenance information, amounted to more than 10,000,000 additional statements (should it be expressed in RDF triples). Query evaluation on the produced content takes usually fractions and at most units of seconds.

2.2 Core Technologies and Capabilities

The publications, their meta-data and full-text were stored and indexed within our KONNEX framework for linked publication data processing [3]. After parsing the input XML article representations, the XML meta-data and structural annotations were quite straightforwardly integrated in the KONNEX RDF repository. Full-text information regarding the articles' content, titles, authors and references were managed using multiple Lucene IR indices (cf. <http://lucene.apache.org/java/docs/>).

Exploitation of the publication knowledge was tackled by our novel EUREEKA framework for emergent (e.g., automatically extracted) knowledge processing [2]. The framework de facto builds on a simple triple model [4], a widely-used part of the Semantic Web [5] standards. However, we extended the subject-predicate-object triples by positive or negative heuristic certainty measures and organised them in so called conceptual matrices, concisely representing every positive and negative relation of an entity to other entities. Metrics can be easily defined on the conceptual matrices. The metrics then serve as a natural basis for gradual concept similarities that define basic light-weight empirical semantics in EUREEKA [2]. On the top of the similarity-based semantics, we implemented simple, yet quite practical inference services of two basic types:

1. *retrieval* of knowledge similar to an input concept, and/or its *extension* by means of similar stored content; 2. fixed-point rule-based *materialisation* of implicit relations, and/or complex *querying* (similarity as a basis for soft variable unification and for approximate fixed-point computation). The inference algorithms have anytime behaviour and it is possible to programmatically adjust their completeness/efficiency trade-off. Technical details of the solution are out of scope regarding this paper, but one can find them in [2].

We applied our prototype to: (i) automated extraction of machine-readable knowledge bases from particular life science article texts; (ii) integration, refinement and extension of the extracted knowledge within one large emergent knowledge base; (iii) exposure of the processed knowledge via a query-answering and faceted browsing interface, tracking the article provenance of particular statements.

For the initial knowledge extraction, we used a NLP-based heuristics stemming from [6,7] in order to process chunk-parsed texts into subject-predicate-object-score quads. The scores were derived from aggregated absolute and document frequencies of subject/object and predicate terms. The extracted quads encoded three major types of ontological relations between concepts: (1) taxonomical—*type*—relationships; (2) concept difference (i.e., negative *type* relationships); (3) “facet” relations derived from verb frames in the input texts (e.g., *has part*, *involves* or *occurs in*). About 27,000 facet relations were extracted. A taxonomy was imposed on them, considering the head verb of the respective phrase as a more generic relation (e.g., *involves expression of* was assumed to be a type of *involves*). Also, several artificial concepts were introduced to restrict the semantics of some most frequent relations. Namely, (positive) *type* was considered transitive and anti-symmetric, and *same as* was set transitive and symmetric. Also, *part of* was assumed transitive and inverse of *has part* for the current deployment. Note that the *has part* relation has rather general semantics within the extracted knowledge, i.e., its meaning is not strictly physically mereological, it can refer also to, e.g., conceptual parts or possession of entities.

The emergent quads were processed as follows:

(I) *addition* – The extracted quads were incrementally added into an emergent knowledge base K , using a fuzzy aggregation of the respective conceptual matrices. As a seed defining the basic domain semantics (i.e., synonymy and core taxonomy of K), we used the EMTREE and NCI thesauri.

(II) *closure* – After the addition of new facts into K , we computed its materialisation according to RDFS entailment rules [8] ported to the format specified in [2].

(III) *extension* – All the extracted concepts were analogically extended by means of similar stored knowledge.

We exposed the content of the eventual knowledge base via a query-answering module. It was returning answer statements sorted according to their relevance scores [2] and similarity to the query. Answers were provided by intersection of publication provenance sets corresponding to the respective statements’ subject and object terms. The module supported queries in the following form:

$t \mid s : (NOT)?p : o(AND s : (NOT)?p : o)^*$, where *NOT* and *AND* stands for negation and conjunction, respectively. s, o, p may be either variable—anything starting with the ? character or even the ? character alone—or a lexical expression. t may be lexical expressions only. The ? and * wildcards mean zero or one and zero or more occurrences of the preceding symbols, respectively, | stands for or. Only one variable name is currently allowed to appear within a single query statement and across a statement conjunction.

Example queries and respective selected answers are as follows:

```

QUERY: ? : type : breast cancer ~> ANSWER: <cystosarcoma phylloides
: TYPE : breast cancer>1 ...
QUERY: rapid antigen testing : part of : ? AND ? : type : clinical
study ~> ANSWER: <dicom study : USE : protein info>0.8 AND <initial
study : INVOLVED : patients>0.9 ...
QUERY: acute granulocytic leukemia : NOT type : T-cell leukemia ~>
ANSWER: <acute granulocytic leukemia : TYPE : T-cell leukemia>-0.7
...

```

The sample answers above are presented in the statement syntax specified in [2] (with rounded degrees). In CORAAL itself, the statements are presented in more human readable way, very similarly to the query syntax. They are also provided by the following types of meta-information: (1) *source* provenance – articles relevant to the statement; (2) *context* provenance – sub-domain of life sciences the statement relates to (determined according to the main topic of the journal that contained the articles the statement was extracted from); (3) *certainty* – a real number meaning how certain the system is that the statement holds and is relevant to the query (values between 0 and 1; derived from the absolute value of the respective statement degree and from the actual similarity of the statement to the query); (4) *inferred* – a boolean value determining whether the statement was inferred or not (i.e., directly extracted).

More can be checked out at <http://coraal.deri.ie:8080/coraal> (points to an online interface of CORAAL deployed on the sample cancer research publication data).

3 Experiments and Evaluation

This section reports on a user-based applicability test of CORAAL and an experiment aimed at assessment of the exposed knowledge quality.

3.1 Applicability Tests with Experts

We prepared five tasks¹ to be worked out with both CORAAL and a base-line application (ScienceDirect or PubMed) by four sample users. Our hypothesis

¹ E.g., find all authors who support the fact that the *acute granulocytic leukemia* and *T-cell leukemia* are different.

was that the users should perform better with CORAAL than with the baseline, since the tasks were focused rather on structured knowledge than than on a plain text-based search.

Using a questionnaire and additional structured interview, we evaluated three major features: (i) the degree to which the queries were considered realistic by the users; (ii) the number of successfully accomplished parts of particular tasks; (iii) the usability. The tasks were deemed rather realistic – the average value was above 4 on the scale from 1 to 6 (worst to best). The success rate of the task accomplishment was 60.7% and 10.7% when using CORAAL and the base-line application, respectively. This clearly confirms our hypothesis regarding improvement over the state-of-the-art. Still, users experienced a lot of frustration related to tasks they were not able to solve with CORAAL. Most sources of the frustration were eliminated by development of a new, better integrated and more intuitive user interface. Further improvements in the user performance were achieved after brief interactive educational sessions. In the beginning, users were just let to play, relying only on an online tutorial. For users given a short interactive lecture about the general features of the CORAAL user interface and query language, the performance was about 75% better and the frustration diminished accordingly.

3.2 Quality of the Exposed Knowledge

We evaluated quality of representative sample answers provided by CORAAL on the cancer research publication data-set. To do so, we picked 100 random concepts and generated 100 random statement queries based on the actually extracted content. We let a committee of domain experts vote on the relevance of respective concept and statement queries to their day-to-day work and used the following most relevant ones to evaluate the CORAAL answers:

Concept queries: myelodysplastic syndrome; p53; BAC clones; primary cilia; colorectal cancer

Statement queries: ? : type : breast cancer; ? : part of : immunization; ? : NOT type : chronic neutrophilic leukemia; rapid antigen testing : part of : ? AND ? : type : clinical study; ? : as : complementary method AND ? : NOT type : polymerase chain reaction

We used the traditional notions of precision and recall for the answer quality evaluation, with average results summed up in Table 1. For a base-line com-

Q. type/measure	P_s	R_s	F_s	P_d	R_d	F_d
concepts	0.474	0.143	0.183	0.496	0.154	0.234
concepts (base)	0.591	0.031	0.056	0.405	0.061	0.102
statements	0.719	0.583	0.586	0.704	0.489	0.541
statements (base)	0.169	0.053	0.067	0.216	0.145	0.171

Table 1. Precision/recall results summary

parison, we employed state-of-the-art Semantic Web technologies – crisp RDFS inference [8] and SPARQL querying² on the same data as processed by CORAAL

² Cf. <http://www.w3.org/TR/rdf-sparql-query/>.

(setting degrees to 1.0 and omitting negative statements, though, since neither RDFS nor SPARQL support uncertainty and negation).

P , R , F in Table 1 columns stands for precision, recall and F-measure (computed as $\frac{2(P \cdot R)}{P+R}$), respectively. The s and d subscripts indicate retrieved *statement* and corresponding *provenance document* precision (or recall), respectively. Base-line results for concept and statement queries are given in the respective *base lines*. Particular precision/recall values were computed as follows. Let C be the corpus of the publications processed by CORAAL. $P_s = \frac{CSR}{ASR}$, $R_s = \frac{CSR}{CSA}$, where CSR , ASR is a number of correct and all answer statements returned by CORAAL, respectively. CSA is the number of all correct statements relevant to the query, as entailed by C data. $P_d = \frac{RDR}{ADR}$, $R_d = \frac{RDR}{RDA}$, where RDR , ADR is a number of relevant and all correct statement provenance publications returned, respectively. RDA is the number of all publications in C relevant to the query and its correct answers.

The degrees in the answer statements were taken into account in this way: if their absolute value was lower than 0.5, i.e., indicating substantial lack of heuristic confidence, the respective statement was deemed neither correct, nor incorrect, and was not considered in the precision/recall computation. Statements originating solely from the initial thesauri were discarded, too. First 400 results were only examined when more eligible answers were available. The results' relevance and numbers of the gold-standard statements and/or publications were determined by domain experts. They did so in a detailed analysis of the C article corpus via a full-text search. They examined both explicit and implicit knowledge in the paragraph contexts of the query and answer terms, as well as in the related NCI and EMTREE thesauri entries. Unequivocal agreement of evaluators was required at all times.

In terms of F-measure, CORAAL clearly outperformed the base-line. The difference was more than two and three-fold regarding F_s for concept and statement queries, respectively. Similarly, F_d was more than eight and three times higher. The base-line precision was higher for P_s and concept queries, though. This was caused by the absence of (partially incorrect) negative statements in the base-line results. On the other hand, recall of CORAAL was much higher due to approximate answer retrieval, and also due to the presence of negative and analogically inferred relations. CORAAL's precision for statement queries was higher due to the support for soft evaluation of both rules and queries – some incorrect crisp statements computed by the base-line were filtered out in CORAAL due to low certainty either in the intermediate, or in the eventual result. Generally better results for statement queries were caused by the fact that only statements directly related to the variable instances conforming to the query structure were taken into account. For concept-only queries, all resulting statements were considered.

The CORAAL results may still be considered rather poor when compared to the gold standard (i.e., F-measure for concept queries around 0.2). However, one must realise that the construction of the gold standard took two working days of an expert committee only for the 10 sample queries. The CORAAL knowledge

base was produced in about the same time for much larger amount of data. Using the faceted browsing provided by the CORAAL user interface, one can find relevant answers very quickly despite of some remaining noise in the purely automatically acquired knowledge. This is a reasonable and unprecedented trade-off according to our expert evaluators and potential users.

4 Related Work

Approaches tackling problems related to those addressed by the core technologies powering CORAAL are analysed in [2, 3]. Here we offer an overview of systems targeting similar problems to those tackled by our framework. Figure 1 organises relevant applications in a plot with two axes – *effort* and *benefit* (the placement is only orientational, though, as it does not reflect any formal measure related to the particular systems). The *effort* axis indicates how much more or less manual effort must the creators and/or maintainers of a tool spend before it can perform sufficiently, or before it can be ported to a new domain. The *benefit* axis reflects how much benefit users get when searching for the knowledge hidden in publications with a tool.



Fig. 1. Informative comparison of selected systems

The state-of-the-art applications like ScienceDirect or PubMed Central require almost no effort in order to expose arbitrary life science publications for search (therefore we used them as a base-line in the user-centric experiment). However, the benefit they provide is rather limited when compared to cutting-edge approaches aimed at utilising also the publication knowledge within the query construction and/or result visualisation. Such innovative solutions may require much more a priori effort in order to work properly, though.

FindUR [9], Melisa [10] and GoPubMed [11] are ontology-based front-ends to a traditional publication full-text search. They allow either for effective restriction and intelligent visualisation of the query results (GoPubMed), or for focusing the queries onto particular topics based on an ontology (FindUR and Melisa). FindUR and Melisa use a Description Logics [12] ontology built from scratch and a custom ontology based on MeSH (cf. <http://www.nlm.nih.gov/mesh/>), respectively. GoPubMed dynamically extracts parts of the Gene Ontology (cf.

<http://www.geneontology.org/>) relevant to the query, which are then used for restriction and a sophisticated visualisation of the classical PubMed search results. None of the tools, nevertheless, offers querying for or browsing of arbitrary publication knowledge – terms and relations not present in the systems’ rather static ontologies simply cannot be reflected in the search. On the other hand, CORAAL works on any domain and extracts arbitrary knowledge from publications automatically, although the offered benefits may not be that high due to possibly higher level of noisiness.

Textpresso [13] is quite similar to CORAAL concerning searching for relations between concepts in particular chunks of text. However, the underlying ontologies and their instance sets have to be provided manually, whereas CORAAL can operate with or even without any legacy ontology. Moreover, the system’s scale regarding the number of publications’ full-texts and concepts covered is much lower than for CORAAL.

From the overview of the related cutting-edge systems, it is obvious that the biggest challenge is a reliable automation of more expressive content acquisition. Contrary to CORAAL, none of the related systems addresses this problem appropriately, which makes them either poorly scalable, or difficult to port to a new domain. This is why we were not even able to use the related systems for a base-line comparison in our domain-specific application scenario – we simply could not adapt them so that they would be able to perform reasonably, both due to technical difficulties and lack of necessary human/time resources.

5 Discussion

In this paper, we have presented CORAAL – a unique combination of a publication repository enhanced by semantic links [3] and an engine for automated extraction, integration and exploitation of knowledge contained in the publication texts [2]. We have shown that the tool has promising results in real-world tasks related to biomedical literature search. Due to substantial automation, we are able to process large amounts of publications in more scalable and efficient way than possible with the state of the art. The potential of CORAAL has also recently been proven by the fact that it was selected as one of the four Elsevier Grand Challenge finalists (cf. <http://www.elseviergrandchallenge.com>).

Note that besides the presented application to literature search, CORAAL can directly be deployed in any use case involving the need for more efficient search in large amounts of textual data. For instance, one could deploy CORAAL in a hospital and feed it with patient records. Appropriate medical ontologies and/or diagnostic rules can be imported into CORAAL to support additional refinement and inference within the patient data. The knowledge scattered among large amounts of patient records can then be integrated and exposed in the same intelligent way as presented in this paper.

Despite of the promising results, there are still certain reserves. We plan to extend the current knowledge processing framework powering CORAAL to a distributed solution, which will significantly improve scalability (from tens of

thousands to millions of publications and beyond). In order to complement our automated approach by the wisdom of the crowds, we have to propose sound mechanisms for easy user involvement in the emergent knowledge (in)validation, updates, and general maintenance. Last but not least, we intend to continue in our cooperation with various groups of biomedical experts, who will help us to realise the CORAAL's promise in agile R&D settings.

Acknowledgments This work has been supported by the 'Líon', 'Líon II' projects funded by SFI under Grants No. SFI/02/CE1/I131, SFI/08/CE/I1380, respectively. We acknowledge much appreciated help from Ioana Hulpus, who developed the initial user interface for CORAAL. Eventually, we are very grateful to our evaluators: Doug Foxvog, Peter Gréll, MD, Miloš Holánek, MD, Matthias Samwald, Holger Stenzhorn and Jiří Vyskočil, MD.

References

1. Bechhofer, S., et al.: Tackling the ontology acquisition bottleneck: An experiment in ontology re-engineering (2003) At <http://tinyurl.com/96w7ms>, Apr'08.
2. Nováček, V.: Towards an efficient knowledge-based publication data exploitation: An oncological literature search scenario. Technical Report DERI-TR-2009-03-23, DERI, NUIG (2009) Available at <http://tinyurl.com/csh3rf>.
3. Groza, T., Handschuh, S., Moeller, K., Decker, S.: KonneXSALT: First steps towards a semantic claim federation infrastructure. In: The Semantic Web: Research and Applications (Proceedings of ESWC 2008), Springer-Verlag (2008) 80–94
4. Manola, F., Miller, E.: RDF Primer. (2004) Available at (November 2008): <http://www.w3.org/TR/rdf-primer/>.
5. Berners-Lee, T., Hendler, J., Lassila, O.: The semantic web. *Sci. Am.* **5** (2001)
6. Maedche, A., Staab, S.: Discovering conceptual relations from text. In: Proceedings of ECAI 2000, IOS Press (2000)
7. Voelker, J., Vrandečić, D., Sure, Y., Hotho, A.: Learning disjointness. In: Proceedings of ESWC'07, Springer (2007)
8. Brickley, D., Guha, R.V.: RDF Vocabulary Description Language 1.0: RDF Schema. (2004) Available at (Feb 2006): <http://www.w3.org/TR/rdf-schema/>.
9. McGuinness, D.L.: Ontology-enhanced search for primary care medical literature. In: Proceedings of the Medical Concept Representation and Natural Language Processing Conference. (1999) 16–19
10. Abasolo, J.M., Gómez, M.: M.: Melisa: An ontology-based agent for information retrieval in medicine. In: Proceedings of the First International Workshop on the Semantic Web (SemWeb2000). (2000) 73–82
11. Dietze, H., et al.: Gopubmed: Exploring pubmed with ontological background knowledge. In: Ontologies and Text Mining for Life Sciences, IBFI (2008)
12. Baader, F., Calvanese, D., McGuinness, D.L., Nardi, D., Patel-Schneider, P.F.: The Description Logic Handbook: Theory, implementation, and applications. Cambridge University Press, Cambridge, USA (2003)
13. Müller, H.M., Kenny, E.E., Sternberg, P.W.: Textpresso: an ontology-based information retrieval and extraction system for biological literature. *PLoS Biology* **2**(11) (2004)