# Extracting conclusion sections from PubMed abstracts for rapid key assertion integration in biomedical research

Matthias Samwald (1, 2)

1) Digital Enterprise Research Institute (DERI), NUI Galway, Galway, Ireland
2) Konrad Lorenz Institute for Evolution and Cognition Research, Altenberg, Austria

This work is dedicated to the *HypER (Hypotheses, Evidence & Relationships)* community and the *W3C Semantic Web Health Care and Life Science Interest Group*

## Abstract / Summary

Key assertions are extracted from "conclusions" sections of PubMed abstracts and converted into Semantic Web / Linked Data format. The results are made accessible via files, a SPARQL endpoint, and a  faceted search interface. Conclusion sections are identified as valuable resources for machine-augmented key assertion identification and integration in the biomedical domain.  Results are discussed and opportunities for future work and cooperation are highlighted.

## Introduction

A common challenge faced by biomedical researchers and clinicians is to quickly get an overview of publications for a certain biomedical topic, and to identify relevant, valid facts, research trends and contradictory findings. One search strategy to address this challenge is to do a PubMed search, look at the first few dozen results and quickly skim over the conclusions in the abstracts of the most recent publications. Of course, this only gives a shallow summary of the contents of each publication, and it makes the judgment of the validity of each claim rather dubious. Nonetheless, this search strategy is useful to get an overview of relevant findings, to see how different biological phenomena relate to each other,  and to identify starting points for further investigation. From here on, I will refer to this process as "key assertion identification" and "key assertion integration". The goal of this work is to facilitate rapid key assertion identification / integration over large biomedical literature collections by technical means, enabling researchers and clinicians to make better decisions in a shorter time.

In a sizable fraction of PubMed abstracts, the narrative of the abstract is clearly delineated by explicit section headers (*"INTRODUCTION:", "METHODS:", "RESULTS:", "CONCLUSIONS:"*). The conclusion sections of biomedical abstracts seem like a gold-mine for automated key assertion identification, since the relevant portion of text can be identified easily. A search in PubMed reveals that $\sim$ 1,7 million abstracts  contain the words "conclusion"  or "conclusions" (out of a total of $\sim$ 19 million citations indexed in PubMed). Most of these abstracts really do contain a clearly delineated conclusion section. This means that a huge corpus of biomedical abstracts with explicit conclusion sections exists, covering a broad area of knowledge domains.

The goal of the work described in this document is to test if these explicit conclusion sections can be used as starting points for the creation of structured representation of

biomedical hypotheses.; and to test the coverage and expressiveness of these resources.

## Methods

I wrote a script that does the following:

→ Retrieve PubMed abstracts containing conclusion sections for a certain query. The script could process all ~1,7 million abstracts with explicit conclusion sections, but for this trial, I chose a more restrictive query that retrieves abstracts about emotion and cognition:

*("conclusion"[Title/Abstract] OR "conclusions"[Title/Abstract]) AND (antidepressant OR "Emotions"[Mesh] OR "Behavioral Symptoms"[Mesh] OR "Mood Disorders"[Mesh])*

This yields 58.000 results. Note that removing the constraint for 'conclusion' or 'conclusions' in this query would increase the number of results to 430.000, which means that roughly 1/7th of the abstracts for this topic contain an explicit conclusion sections.

→ Abbreviations that are locally defined in each abstract are expanded to their long forms using the Schwartz & Hearst algorithm (http://biotext.berkeley.edu/software.html). In most abstracts, abbreviations are introduced in the introduction section, e.g.:

*„INTRODUCTION: Seasonal affective disorder (SAD) is common in …"*

while the conclusion sections contain lots of these abbreviated forms that tend to be unintelligible when only the conclusion sections are viewed in isolation, e.g.:

*„CONCLUSIONS: This study shows that SAD is effectively treated with …"*

The script recognizes local abbreviations and expands them, making the conclusion sections better intelligible. E.g., after processing the conclusion now reads

*„CONCLUSIONS: This study shows that Seasonal affective disorder is effectively treated with …"*

→ The conclusion sections are then extracted and are turned into aTags (a simple convention for representing statements and their annotations with Semantic Web standards such as RDF and SIOC, further described in http://hcls.deri.org/atag/ ). For this trial, each aTag was annotated with the MeSH terms associated with the article. In future work, this could be replaced/enhanced with annotations created by automated entity recogntion (each a BioPortal webservice or EBI Whatizit) or manual curation.

## Results

The aTags that were generated by this process are available in Turtle RDF format http://hcls.deri.org/datafeeds/atag/emotion_query_1.ttl (114 MB)

A (shortened) example of one aTag looks like this:

```
<http://purl.org/concept/assertion/pmid/19671904>
 sioc:content "In this study, both higher Mediterranean-type diet adherence and
higher physical activity were independently associated with reduced risk for
Alzheimer disease." ;
sioc:topic <http://purl.org/commons/record/mesh/D038441> ;
sioc:topic <http://purl.org/commons/record/mesh/D000544> ;
sioc:topic <http://purl.org/commons/record/mesh/D012307>   .
<http://purl.org/commons/record/mesh/D038441> skos:prefLabel "Diet,
Mediterranean" .
<http://purl.org/commons/record/mesh/D000544> skos:prefLabel "Alzheimer Disease" .
<http://purl.org/commons/record/mesh/D012307> skos:prefLabel "Risk Factors" .
```

The content of this file is also available in the HCLS Knowledge Base
(http://hcls.deri.org/sparql) and can be queried like this:

*SELECT \* FROM <http://purl.org/concept/assertion/>*
*WHERE {?s ?p ?o} LIMIT 10*

The MeSH URIs (such as http://purl.org/commons/record/mesh/D012307) are already
used by other datasets in the HCLS Knowledge Base, so each aTag is interlinked with
other datasets in the knowledge base. For example, this can be used to query for related
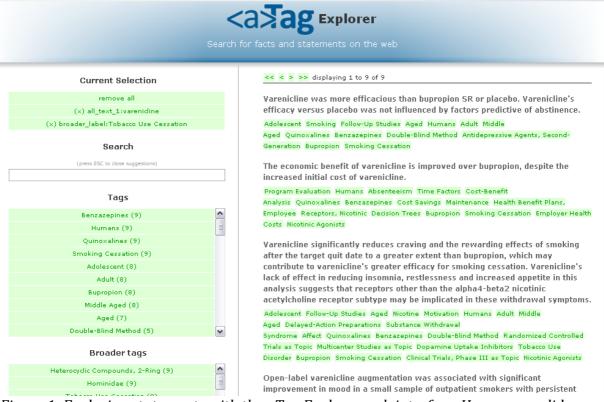PubMed articles or DBpedia entries.



*Figure 1: Exploring statements with the aTag Explorer web interface. Here, a user did a
text search for the drug 'varenicline', then restricted results to those statements that deal
with 'Tobacco Use Cessation' by selecting a facet value. The tags / facet values for each
statement are terms from Semantic Web / Linked data resources such as MeSH and
DBpedia. The 'Broader tags' for each statement are inferred by the system from these
terminologies / ontologies. This makes it possible to identify links between statements that
are not explicitly contained in the source literature.*

Furthermore, a human-friendly interface for convenient faceted browsing of the aTags is the *aTag Explorer*, accessible at http://hcls.deri.org/atag/explorer (Fig. 1, note that this interface currently works with all browsers except Internet Explorer).

The aTag Explorer also contains other statements and definitions from other datasets, such as the SIDER drug side effect database (http://sideeffects.embl.de/), DBpedia (http://dbpedia.org) as well as user-generated content that can be created by any person on the web with the aTag Generator bookmarklet (http://hcls.deri.org/atag/generator/).

First qualitative evaluations of using the statements generated by this work to answer realistic biomedical questions were conducted, using the aTag explorer as a search interface. **Preliminary results are very encouraging, giving results of very good accuracy and satisfying the information needs for each research question, even though the underlying corpus is very limited.**

A subjective comparison of query results produced by the system with other systems that provide sentence-based querying over entire PubMed abstracts was conducted. Examples for such sentence-based, whole-abstract search systems include I-HOP (http://www.ihop-net.org/), Wikigenes (www.wikigenes.org) and MedEvi (http://www.ebi.ac.uk/Rebholz-srv/MedEvi/). While these other systems provide far better coverage, the search results contain a lot of unwanted noise produced by statements derived from introduction, methods and results sections of abstracts, producing results that are often not very relevant, unintelligible outside of the context of the entire text, or very redundant (e.g., introduction sections of abstracts often re-iterate the same fact again and again). In comparison, the statements derived by extracting conclusion section seem to contain far less noise and might provide much better user satisfaction, even though coverage is drastically lower.

## Conclusions / Outlook

Conclusion sections are valuable resources for machine-augmented key assertion identification and integration in the biomedical domain.

More research will be devoted to evaluating the usefulness of the approach described in this paper for answering realistic biomedical research questions. The claims made in this paper need to be further substantiated by more thorough, quantitative empirical analysis.

The results of this simplistic approach to key assertion identification should be combined with more sophisticated methods that make use of subtle linguistic cues in abstracts and full texts, in order to increase the coverage of existing literature, including publications without explicit conclusion sections in the abstract.

These preliminary results will serve as the basis for more extensive work that will be done in cooperation with other members of the HypER (Hypotheses, Evidence & Relationships) community and W3C Health Care and Life Science Interest Group (http://www.w3.org/2001/sw/hcls/).

## Acknowledgements