

POSTER PRESENTATION

Open Access

# Medie and Info-pubmed: 2010 update

Tomoko Ohta<sup>1\*</sup>, Takuya Matsuzaki<sup>1</sup>, Naoaki Okazaki<sup>1</sup>, Makoto Miwa<sup>1</sup>, Rune Sætre<sup>1</sup>, Sampo Pyysalo<sup>1</sup>, Jun'ichi Tsujii<sup>1,2,3</sup>

From Workshop on Advances in Bio Text Mining  
Ghent, Belgium. 10-11 May 2010

## Introduction

In the recent decades, high-throughput screening methods were established, bringing forth major breakthroughs in the fields of molecular biology and biomedicine. Since researchers in these fields need to interpret an enormous quantity of data and the publication rates of scientific articles are exploding, demands on text mining technology are growing with each passing year.

Medie (<http://www-tsujii.is.s.u-tokyo.ac.jp/medie/>) and Info-pubmed (<http://www-tsujii.is.s.u-tokyo.ac.jp/info-pubmed/>) were developed as a response to these information needs. Medie is a general-purpose integrated Pubmed search engine and Info-pubmed is a targeted system for finding information about the interactions of key biomedical entities.

In this work, the first update of these systems since their introduction, we present multiple extensions of the systems based on recent advances in biomedical text mining.

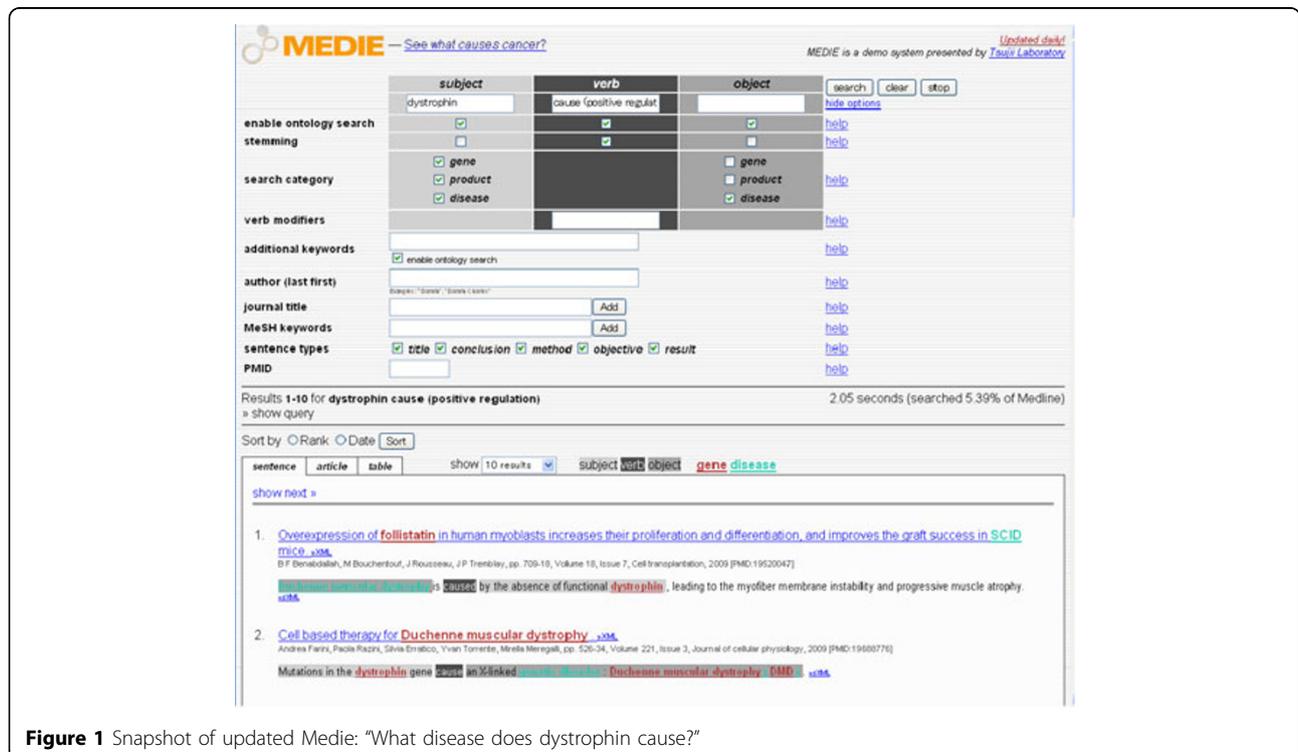


Figure 1 Snapshot of updated Medie: "What disease does dystrophin cause?"

\* Correspondence: okap@is.s.u-tokyo.ac.jp

<sup>1</sup>Department of Computer Science, University of Tokyo, Tokyo, Japan  
Full list of author information is available at the end of the article

## Extensions of Medie and Info-pubmed

Medie and Info-pubmed are based on deep syntactic analysis of sentence structure. To allow users to take advantage of the latest parsing technology, the current release integrates an improved parser [1].

In an extension of semantic search capabilities, the updated Medie system incorporates extended ontology-based search that allows the query verb to be replaced by any GENIA event ontology (<http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/>) term. Such searches are expanded to the set of verbs annotated as expressing the given event in GENIA corpus [2]: for example, a search for *Positive regulation* will now match *activate*, *induce*, etc.

To allow more focused searches, we incorporated the section labeling method of Hirohata et al. [3], creating search options limiting queries to specific types of sentences such as *methods*, *results* and *conclusions*. The indexing system and search options were further augmented with Pubmed annotation metadata, allowing searches to be limited by MeSH terms, author, or journal.

The initial release of Info-pubmed implemented search for automatically detected protein-protein interactions. We have extended this search capability to include gene-disease associations [4], allowing the system to be used also to study the epidemiological connections of biomolecules.

Finally, we have extended the coverage of both systems to the entire PubMed and added scheduled update modules that perform daily updates of the system database, fully automating data access, analysis and indexing.

Figure 1 shows an example search result on Medie illustrating a number of the newly introduced functions.

## Conclusions

We have introduced extended and updated functionality for Medie and Info-pubmed, search systems integrating state-of-the-art text mining technology. The updates allow advanced semantic searches of the latest published information in all of Pubmed.

### Author details

<sup>1</sup>Department of Computer Science, University of Tokyo, Tokyo, Japan.

<sup>2</sup>School of Computer Science, University of Manchester, Manchester, UK.

<sup>3</sup>National Centre for Text Mining, University of Manchester, Manchester, UK.

Published: 6 October 2010

### References

1. Ninomiya T, Matsuzaki T, Miyao Y, Tsujii J: **A log-linear model with an n-gram reference distribution for accurate HPSG parsing.** *Proceedings of IWPT 2007* 2007, Prague, Czech Republic.
2. Kim JD, Ohta T, Tsujii J: **Corpus annotation for mining biomedical events from literature.** *BMC Bioinformatics* 2008, **9**:10, [ISSN 1471-2105].
3. Hirohata K, Okazaki N, Ananiadou S, Ishizuka M: **Identifying Sections in Scientific Abstracts using Conditional Random Fields.** *Proceedings of the*

*Third International Joint Conference on Natural Language Processing (IJCNLP 2008), Hyderabad, India 2008*, 381-388.

4. Chun Hw, Tsuruoka Y, Kim JD, Shiba R, Nagata N, Hishiki T, Tsujii J: **Extraction of Gene-Disease Relations from MedLine using Domain Dictionaries and Machine Learning.** *Proceedings of The Pacific Symposium on Biocomputing (PSB), Maui, Hawaii, USA 2006*, 4-15.

doi:10.1186/1471-2105-11-S5-P7

**Cite this article as:** Ohta et al.: Medie and Info-pubmed: 2010 update. *BMC Bioinformatics* 2010 **11**(Suppl 5):P7.

**Submit your next manuscript to BioMed Central  
and take full advantage of:**

- Convenient online submission
- Thorough peer review
- No space constraints or color figure charges
- Immediate publication on acceptance
- Inclusion in PubMed, CAS, Scopus and Google Scholar
- Research which is freely available for redistribution

Submit your manuscript at  
[www.biomedcentral.com/submit](http://www.biomedcentral.com/submit)

