SEMEDICO: A Comprehensive Semantic Search Engine for the Life Sciences

Erik Faessler
Jena University Language & Information Engineering (JULIE) Lab
Friedrich-Schiller-Universität Jena
Jena, Germany
erik.faessler@uni-jena.de

Udo Hahn
Jena University Language & Information Engineering (JULIE) Lab
Friedrich-Schiller-Universität Jena
Jena, Germany
udo.hahn@uni-jena.de

Abstract

SEMEDICO is a semantic search engine designed to support literature search in the life sciences by integrating the semantics of the domain at all stages of the search process—from query formulation via query processing up to the presentation of results. SEMEDICO excels with an ad-hoc search approach which directly reflects relevance in terms of information density of entities and relations among them (events) and, a truly unique feature, ranks interaction events by certainty information reflecting the degree of factuality of the encountered event.

1 Introduction

The exponential growth of scientific publications in the life science domain (Lu, 2011) has inspired a wide range of information retrieval services over the last decade (for a brief survey, see Section 2). Simple term-based retrieval techniques, including frequency-based approaches based on TF-IDF scores, rapidly hit their limits given the enormous complexity of the sublanguage in the life sciences, not only due to the sheer vocabulary size (amounting to millions of specialized terms) but also due to factors such as excessive ambiguity, non-canonicity of complex phrases, extensive terminological paraphrasing, etc.

Overly long hit lists returned for standard queries in PUBMED (Lu, 2011), the most prominent literature hub for life scientists, make focused search strategies a major desideratum. Current search mechanisms are unable to distinguish between semantically tightly bound informational units, like semantic relations (events) between entities (e.g., protein-protein interactions), and much looser relations between entities, like co-occurrence of search terms within the same paragraph or entire document.

Hence, in order to improve literature search, a retrieval system should take into account the domain knowledge of the domain under scrutiny, connect it with the contents of the publications in the document collection in a meaningful way, decide which information pieces to present with high priority and display them to the user in an easily digestable way. However, existing search engines only partially match these requirements.

As an alternative, we here present the semantic search engine SEMEDICO. It features a front-end with interactive disambiguation for query concepts that share a common name with other concepts, including abbreviations which have been automatically extracted from documents. Due to the incorporation of several life science ontologies (see Section 3) all subordinates of search terms are included in a search. This semantic enrichment not only plays a major role in retrieving relevant documents (implicitly, all subordinates are OR-ed) but also supports searchers in the formulation of adequate queries since it makes conceptual neighborhoods lucid, thus easing query formulation.

At the back end side, gene interactions are scored relative to the degree of factuality explicitly expressed in the document (*"we have evidence for the interaction of X and Y"* is a stronger claim than *"X might potentially interact with Y"* and will thus be ranked higher than the second statement; see Section 5). For ranking, we also take into account the proximity of occurrences of search terms within well-defined document portions. We deem shorter text passages populated by several query terms to be more informative to the researcher than wider dispersed term occurrences. The most informative units, from this perspective, are tightly connected semantic relations where constituent entities are syntactically

Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics-System Demonstrations, pages 91–96
Vancouver, Canada, July 30 - August 4, 2017.
https://doi.org/10.18653/v1/P17-4016

related as well. This means that SEMEDICO prefers shorter passage matches over larger ones and scores document hits accordingly. In the final hit list, matching entities and relations are highlighted in order to orient the reader immediately to the relevant text parts (as defined by the query).

2 Related Semantic Search Engines

Several search engines for the life sciences have been developed to address the needs of researchers (for a survey, cf. Lu (2011)). A common characteristic of these systems is the incorporation of the semantics of the underlying domain, by design, in terms of domain-specific terminologies, thesauri and ontologies. GOPUBMED (Doms and Schroeder, 2005) integrates the Medical Subject Headings (MESH),[1] GENE ONTOLOGY (GO)[2] and UNIPROT.[3] It allows to browse PUBMED citations taxonomically structured by the MESH and GO. Search results also include hits for taxonomic descendants of search concepts, as does SEMEDICO. However, GOPUBMED does not integrate any relational information (such as protein-protein interactions) or factuality detection and operates on PUBMED abstracts only.

FACTA+ (Tsuruoka et al., 2011) recognizes a range of biomedical entity types (genes/proteins, diseases, symptoms, drugs, enzymes and compounds) in MEDLINE abstracts and analyzes documents for biomedical event triggers (Kim et al., 2008). FACTA+ offers multiple search modes. The *Find Associated Concepts* mode finds indirect associations between entities in the spirit of Swanson's notion of undiscovered public knowledge (Swanson, 1986) and thus is not the focus of this comparison. The *View Documents* mode is the information retrieval part of the system and lists highlighted MEDLINE titles and abstracts. This mode retrieves keyword-based results without making use of conceptual knowledge. FACTA+ detects event triggers and also gene mentions, but it does not include the gene arguments in its event model. Thus, one cannot search specifically, for example, a regulation of the gene *BRCA1*. SEMEDICO, on the other hand, exploits its ontological resources for concept synonyms, recognizes event trigger-argument structures and stores them as searchable items in the index.

[1] https://www.nlm.nih.gov/mesh/
[2] http://www.geneontology.org/
[3] http://www.uniprot.org/

QUETZAL (Coppernoll-Blach, 2011) stores hundreds of millions (250 million as of 2011) subject-verb-object relations that are matched against query terms to produce focused sentence-level retrieval results. In this regard, QUETZAL shares the basic idea of SEMEDICO that semantic relations between query terms are more relevant than longer text passages mentioning the query terms only in a loosely connected way. QUETZAL includes arbitrary relations of all kinds rather than domain-specific types of relations like SEMEDICO. The advantage of this approach is a higher domain coverage. On the downside, QUETZAL's restriction fails to account for a large number of interactions which are expressed using nouns, e.g., *"the regulation of mTOR"*. QUETZAL does not incorporate factuality information to the best of our knowledge.

FERRET's (Srinivasan et al., 2015) focus lies on the exploration of sentence-level gene-centric relationships in MEDLINE citations. The system performs gene name disambiguation and allows for query expansion via gene homologues. Retrieved sentences contain findings for gene-gene or gene-keyword pairs. SEMEDICO, in contradistinction, flexibly searches genes in a larger variety of text segment block sizes, including sentences.

POLYSEARCH2 (Liu et al., 2015) finds associations between an extensive range of entity types. Given a query with a specified entity class, the user may ask for relationships to another entity class. POLYSEARCH2 searches associations in a wide range of resources, including PUBMED, PUBMED CENTRAL, Wikipedia and life-science related databases. It does not support ad hoc free-text queries and does not employ dedicated recognition tools for entities or relations and always operates on the sentence level, much in contrast to SEMEDICO.

GENEVIEW (Thomas et al., 2012) employs an large variety of named entity recognition tools to automatically annotate different entity classes in MEDLINE and PUBMED CENTRAL, including SNPs, species, chemicals, histone modifications, genes, protein-protein interactions (PPIs) and more. Document scoring includes field length normalization, such that term matches in titles achieve higher scores than comparable matches across a whole section. In this way, GENEVIEW implements the idea that shorter text portions with entity matches are more relevant than longer stretches

of texts in a similar way as SEMEDICO does, but is restricted to formal title, abstract and full text sections. Unlike SEMEDICO, which automatically searches gene name query terms within molecular events, GENEVIEW requires the exact database identifier (e.g., NCBI GENE ID to search for a gene or an input of the form PPI:GENEID to search for PPIs including the given gene ID). There is no possibility to rank PPIs according to the degree of factuality.

HYPOTHESISFINDER (Malhotra et al., 2013) is one of the few life science search engines besides SEMEDICO that employs factuality statements. Accordingly, it provides the user with speculative sentences from MEDLINE matching a keyword query. Its goal is to explicitly provide speculative statements in order to find scientific hypotheses, yet there is no ranking for factuality in the sense of SEMEDICO, nor makes it use of sophisticated entity or event extraction methods.

3 Resources Used in SEMEDICO

Literature input for SEMEDICO comes from two sources, *viz.* more than 27 million life science abstracts from MEDLINE/PUBMED[4][5] and approximately 1,5 million life science full texts from the open access subset of PUBMED CENTRAL. They are stored in a POSTGRESQL database.[6]

Domain knowledge for the life sciences is gathered from several terminological and ontological resources. Each document from MEDLINE is indexed with entries from the Medical Subject Headings (MESH), a hierarchically organized thesaurus with rather general entries at the top (e.g., "Anatomy") and quite specific entries at the hierarchy's leaves (e.g., "Ankle"). SEMEDICO makes use of the MESH headings as encoded in the original XML files, while it also recognizes mentions of MESH entry terms within the document text by its named entity recognizers.

Another extensively used resource is the NCBI GENE database.[7] Our gene recognition and normalization engine (see Section 4) maps gene mentions in document text to unique NCBI GENE database entries to handle gene name synonymy and ambiguity. Additionally, SEMEDICO integrates the GENE ONTOLOGY and the GENE REGULATION ONTOLOGY (GRO)[8] for the semantic description of different types of gene events.

All resources are stored in a NEO4J[9] graph database for direct access to their hierarchical structure. All terminologies, ontologies and databases are converted into a common JSON format. This format is then imported into NEO4J using a custom NEO4J server plugin.

4 Text Analytics

The complete document set of all MEDLINE/PUBMED abstracts and PMC full texts (roughly, 28,5m documents) is represented in SEMEDICO's index. Before indexing, each document undergoes an extensive text analytics as depicted in Figure 1. The goal is to identify textual units referring to gene/protein mentions, MESH headings, ontology concepts, gene interaction events and associated factuality markers.

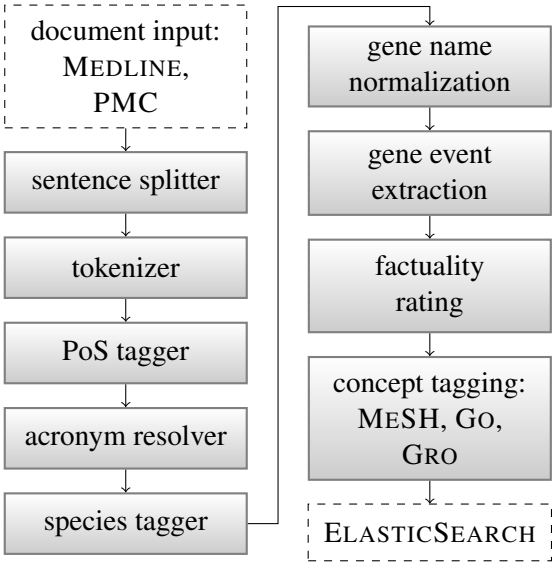

Figure 1: SEMEDICO's text analytics pipeline.

The morpho-syntactic analysis includes the resolution of acronyms (Schwartz and Hearst, 2003). This step is crucial for the interactive disambiguation feature of SEMEDICO. For most of these tasks, we employ JCORE (Hahn et al., 2016), our UIMA (Unstructured Information Management Architecture)[10] component repository.

Semantic analysis includes species tagging by the LINNAEUS tagger (Gerner et al., 2010), gene mention tagging and normalization using GENO

[4] https://www.ncbi.nlm.nih.gov/pubmed
[5] https://www.nlm.nih.gov/databases/download/pubmed_medline.html
[6] https://www.postgresql.org/
[7] https://www.ncbi.nlm.nih.gov/gene
[8] https://bioportal.bioontology.org/ontologies/GRO
[9] https://neo4j.com/
[10] https://uima.apache.org/

(Wermter et al., 2009), gene / protein event recognition with BIOSEM (Bui et al., 2013) and identification of event confidence ratings using the factuality rating determined by Hahn and Engelmann (2014). For BIOSEM, we use a model trained on the BIONLP SHARED TASK 2011 (Kim et al., 2011) training data that includes abstracts as well as full texts. MESH, GO and GRO concepts are tagged by a dictionary component. We then store the annotation results together with the original, raw documents in the document database.

In a last step, the analysis results are sent to an ELASTICSEARCH cluster for indexing. We use a custom ELASTICSEARCH plugin to have ELASTICSEARCH accept a term format that allows to exactly specify index terms within the ELASTICSEARCH index. This way, the exact linguistic analysis results are channeled into the index.

5 Document Indexing and Scoring

All concepts, i.e., entities like species, MESH headings, GO or GRO concepts, are indexed including their taxonomical ascendants such that a search for *Dementia* also includes text mentions of *Alzheimer's Disease* or *Huntington's Disease*.

As the basic document scoring algorithm, ELASTICSEARCH's *TF-IDF* scoring function is used. Additionally to this concept-centric scoring strategy, SEMEDICO splits MEDLINE citations and PMC full texts into their titles, sentences, abstract sections, paragraphs, full text sections, table and figure captions and the complete document text, if applicable. In a technically similar manner, relations between genes/proteins are extracted directly from the documents and stored as searchable items in the ELASTICSEARCH index as nested documents, still being connected to the original document. Relations are stored with information about the event types playing a role in the gene/protein interaction (e.g., Binding, Phosphorylation, Positive/Negative Regulation, etc.) and the actual gene/protein arguments involved.

Additionally, each relation item in the index is assigned an ordinal value representing the factuality status of the relation as expressed by the authors through explicit linguistic signals using epistemic modalities (such as *'could'*, *'probably'*, *'we believe'*, etc.). Based on experiments described in Hahn and Engelmann (2014), each lexical indicator for the expression of factuality is assigned an empirically determined "likelihood" value which is subsequently transferred to each relation that carries such an epistemic labeling. The lowest likelihood value is issued when a negation is encountered because the authors express the firm belief that such a statement is false. If no epistemic modalities are detected in a sentence, we assign the highest likelihood.

SEMEDICO uses such factuality information to rank gene interaction relations according to their certainty, by default prioritizing statements with a higher factuality rating over lower ones. The final document score for the result list ranking is derived from the individual text portion and relation scores the document has, weighted by ELASTICSEARCH field length normalization on the basis of the spatial proximity of the text portions in which search terms co-occur. This way, SEMEDICO prefers query matches on shorter text passages over those in larger ones.

6 Web Application

SEMEDICO is realized as an APACHE TAPESTRY 5 web application.[11] Its start page presents itself with an input field for query input. It expects the user to enter query terms and prompts suggestions derived from the items in the NEO4J concept database (see Section 3) as soon as the user types into the input field (see Figure 2).

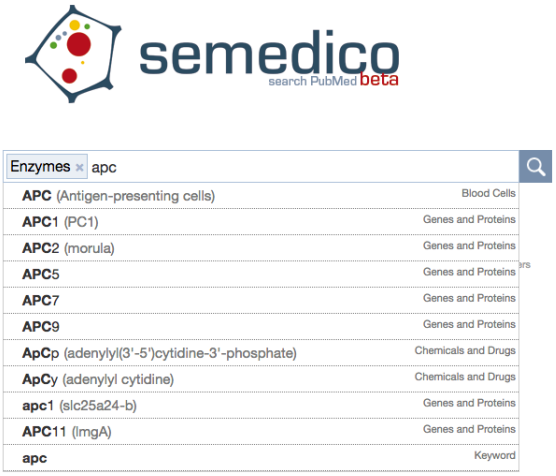

Figure 2: SEMEDICO finds suggestions in the concept database.

We use an adapted version of the JQUERY TOKEN PLUGIN[12] to segment the query into "tokens" to clarify what is searched for. A token

[11] http://tapestry.apache.org/
[12] http://loopj.com/jquery-tokeninput/

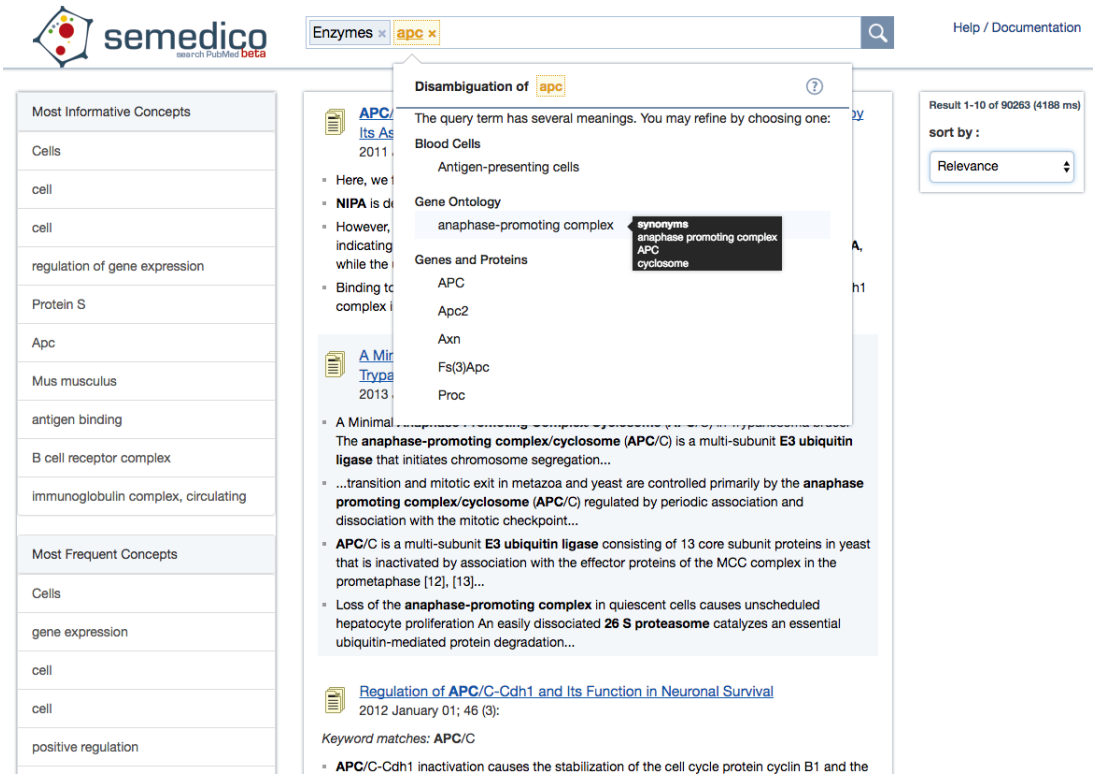

Figure 3: SEMEDICO highlights query concept matches in document snippets and allows explicit disambiguation of concept names.

may consist of multiple words and either represents a database concept or a keyword that cannot be (or, as decided by the user, should not be) resolved to a concept name.

If the user does not select any of these suggestions, SEMEDICO automatically recognizes concepts in the query. For query portions which can be mapped to multiple concepts, SEMEDICO assigns a specific graphical styling to the query tokens in question after the search process and displays disambiguation options when the cursor is hovered over the tokens (see Figure 3). All disambiguation options are concepts from the database and contain their synonyms as tooltips to help the user in the disambiguation process.

SEMEDICO makes extensive use of highlighting to clarify at first glance why a document match was deemed relevant. Since SEMEDICO does not only search for the exact query concepts but also for their taxonomic subordinates, subordinate matches are also highlighted. For example, Figure 3 shows that a search for *Enzymes* also leads to matches like *E3 ubiquitin ligase*, which is a taxonomic descendant of the *Enzymes* heading in the MESH. Again, matches in shorter text snippets are expected to be more valuable to the user than those in larger text portions, and are thus displayed to the user with higher preference. On the left side, SEMEDICO shows concepts occurring in the document result list, sorted either by frequency or by using the ELASTICSEARCH "Significant Terms Aggregation".[13] These concepts may be added to the current query for refinement.

Clicking on an article title opens a new page showing the abstract with highlighted search concept matches and, for PMC hits, a list of highlighted full text matches, showing the highest ranking query matches without the need for further search within a – possibly very long – document. Links to PUBMED, PMC and publisher full text sources allow easy access to the original publication.

7 Conclusion

We presented SEMEDICO, a semantic search engine for PUBMED and PUBMED CENTRAL that assists users with query formulation by con-

[13] https://www.elastic.co/guide/en/elasticsearch/guide/2.x/significant-terms.html

cept suggestion, recognition and interactive disambiguation. SEMEDICO covers multiple levels of semantics, from simple abbreviation resolution over entity recognition to relation extraction for gene interaction events. Sentences are tagged for varying degrees of factuality and relations are ranked by scoring these degrees. The semantic units are further scored by varying levels of textual proximity—first, looking for explicitly expressed gene relations, co-occurrences of query concepts within sentences, paragraphs or even larger text blocks. All sources of evidence are translated into a measure of semantic tightness between query concepts. Furthermore, the ranking reflects a preference for grouping query terms together in a closer textual context, while textually more dispersed co-occurrences are sorted on lower ranks.

Acknowledgments.

This work is part of the Collaborative Research Center AQUADIVA (SFB 1076: AQUADIVA) established at Friedrich-Schiller-Universität Jena and funded by *Deutsche Forschungsgemeinschaft* (DFG). Initial funding of SEMEDICOwas due to the *German Ministry of Education and Research* (BMBF) for the *Jena Centre of Systems Biology of Ageing* (JENAGE) (grant no. 0315581D).

References

Quoc-Chinh Bui, David Campos, Erik M. van Mulligen, and Jan A. Kors. 2013. A fast rule-based approach for biomedical event extraction. In *BioNLP 2013 — Proceedings of the BioNLP Shared Task 2013 Workshop @ ACL 2013. Sofia, Bulgaria, August 9, 2013*. pages 104–108.

Penny Coppernoll-Blach. 2011. Quertle: The conceptual relationships alternative search engine for PubMed. *Journal of the Medical Library Association* 99(2):176–177.

Andreas Doms and Michael Schroeder. 2005. GoPubMed: exploring PubMed with the Gene Ontology. *Nucleic Acids Research* 33(Suppl 2):W783–W786.

Martin Gerner, Goran Nenadic, and Casey M. Bergman. 2010. Linnaeus: a species name identification system for biomedical literature. *BMC Bioinformatics* 11:#85.

Udo Hahn and Christine Engelmann. 2014. Grounding epistemic modality in speakers' judgments. In Duc-Nghia Pham and Seong-Bae Park, editors, *Trends in Artificial Intelligence. PRICAI 2014 — Proceedings of the 13th Pacific Rim International Conference on Artificial Intelligence. Gold Coast, Australia, 1-5 Dec, 2014*. Springer, number 8862 in Lecture Notes in Artificial Intelligence, pages 654–667.

Udo Hahn, Franz Matthies, Erik Faessler, and Johannes Hellrich. 2016. UIMA-based JCoRe 2.0 goes GitHub and Maven Central: state-of-the-art software resource engineering and distribution of NLP pipelines. In *LREC 2016 — Proceedings of the 10th International Conference on Language Resources and Evaluation. Portorož, Slovenia, 23-28 May 2016*. pages 2502–2509.

Jin-Dong Kim, Tomoko Ohta, and Jun'ichi Tsujii. 2008. Corpus annotation for mining biomedical events from literature. *BMC Bioinformatics* 9:10.

Jin-Dong Kim, Sampo Pyysalo, Tomoko Ohta, Robert Bossy, Ngan Nguyen, and Jun'ichi Tsujii. 2011. Overview of BioNLP Shared Task 2011. In *BioNLP 2011 — Proceedings of the BioNLP Shared Task 2011 Workshop @ ACL-HLT 2011. Portland, Oregon, USA, 24 June 2011*. pages 1–6.

Yifeng Liu, Yongjie Liang, and David S. Wishart. 2015. PolySearch2: a significantly improved text-mining system for discovering associations between human diseases, genes, drugs, metabolites, toxins and more. *Nucleic Acids Research* 43(W1):W535–W542.

Zhiyong Lu. 2011. Pubmed and beyond: a survey of web tools for searching biomedical literature. *Database: The Journal of Biological Databases and Curation* page #baq036.

Ashutosh Malhotra, Erfan Younesi, Harsha Gurulingappa, and Martin Hofmann-Apitius. 2013. 'HypothesisFinder:' a strategy for the detection of speculative statements in scientific text. *PLoS Computational Biology* 9(7):e1003117.

Ariel S. Schwartz and Marti A. Hearst. 2003. A simple algorithm for identifying abbreviation definitions in biomedical text. In *PSB 2003 – Proceedings of the Pacific Symposium on Biocomputing 2003. Kauai, Hawaii, USA, January 3-7, 2003*. pages 451–462.

Padmini Srinivasan, Xiao-Ning Zhang, Roxane Bouten, and Caren Chang. 2015. Ferret: a sentence-based literature scanning system. *BMC Bioinformatics* 16(1):#198.

Don R. Swanson. 1986. Fish oil, Raynaud's Syndrome, and undiscovered public knowledge. *Perspectives in Biology and Medicine* 30(1):7–18.

Philippe E. Thomas, Johannes Starlinger, Alexander Vowinkel, Sebastian Arzt, and Ulf Leser. 2012. GeneView: a comprehensive semantic search engine for PubMed. *Nucleic Acids Research* 40(W1):W585–W591.

Yoshimasa Tsuruoka, Makoto Miwa, Kaisei Hamamoto, Jun'ichi Tsujii, and Sophia Ananiadou. 2011. Discovering and visualizing indirect associations between biomedical concepts. *Bioinformatics* 27(13):i111–i119.

Joachim Wermter, Katrin Tomanek, and Udo Hahn. 2009. High-performance gene name normalization with GeNo. *Bioinformatics* 25(6):815–821.