Extraction and generalisation of variables from scientific publications

Erwin Marsi, Pinar Öztürk
Department of Computer and Information Science
Norwegian University of Science and Technology (NTNU)
{emarsi,pinar}@idi.ntnu.no

Abstract

Scientific theories and models in Earth science typically involve changing *variables* and their complex interactions, including correlations, causal relations and chains of positive/negative feedback loops. Variables tend to be complex rather than atomic entities and expressed as noun phrases containing multiple modifiers, e.g. *oxygen depletion in the upper 500 m of the ocean or timing* and *magnitude of surface temperature evolution in the Southern Hemisphere in deglacial proxy records*. Text mining from Earth science literature is therefore significantly different from biomedical text mining and requires different approaches and methods. Our approach aims at automatically locating and extracting variables and their direction of variation: *increasing*, *decreasing* or just *changing*. Variables are initially extracted by matching tree patterns onto the syntax trees of the source texts. Next, variables are generalised in order to enhance their similarity, facilitating hierarchical search and inference. This generalisation is accomplished by progressive pruning of syntax trees using a set of tree transformation operations. Text mining results are presented as a browsable variable hierarchy which allows users to inspect all mentions of a particular variable type in the text as well as any generalisations or specialisations. The approach is demonstrated on a corpus of 10k abstracts of Nature publications in the field of Marine science. We discuss experiences with this early prototype and outline a number of possible improvements and directions for future research.

1 Introduction

Text mining of scientific literature originates from efforts to cope with the ever growing flood of publications in biomedicine (Swanson, 1986; Swanson, 1988; Swanson and Smalheiser, 1997; Hearst, 1999; Ananiadou et al., 2006; Zweigenbaum et al., 2007; Cohen and Hersh, 2005; Krallinger et al., 2008; Rodriguez-Esteban, 2009; Zweigenbaum and Demner-Fushman, 2009; Ananiadou et al., 2010; Simpson and Demner-Fushman, 2012; Ananiadou et al., 2014). Consequently the resulting approaches, methods, tools and applications – as well as data, corpora and evaluation tasks – are rooted in the paradigm of biomedical research and its conceptual framework. Typical source text consists of abstracts from PubMed or full-text articles from PubMed Central. Standard tasks include recognition, normalisation and mapping of biological entities (e.g., genes, proteins, drugs, symptoms and diseases), extraction of biological relations (e.g., protein-protein interaction, disease-gene associations or drug-drug interaction) or bio-event extraction (e.g., regulation or inhibition events and their participants). There are extensive ontologies like the Gene Ontology (Consortium, 2001), annotated corpora like the GENIA (Kim et al., 2003) and BioInfer (Pyysalo et al., 2007) corpora and dedicated shared tasks including BioCreative (Hirschman et al., 2005) and BioNLP (Pyysalo et al., 2012). In short, there is a whole infrastructure supporting biomedical text mining (Cohen and Hunter, 2008).

Text mining is now spreading out to other scientific disciplines, notably in the humanities and social sciences (O'Connor et al., 2011), holding the promise for knowledge discovery from large text collections. Our own research targets text mining in the field of Earth science, more specifically in Oceanography or Marine science, with a focus on climate change. As text mining efforts in this

Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 505–511,
Lisbon, Portugal, 17-21 September 2015.

area are extremely rare (Ekstrom and Lau, 2008; Vossen et al., 2010; Zhang et al., 2013; Marsi et al., 2014; Aamot, 2014), it is not surprising that a corresponding infrastructure is mostly lacking. In addition, however, we found that due to significant differences between the conceptual frameworks of biomedicine and marine science, simply "porting" the biomedical text mining infrastructure to another domain will not suffice.

One major difference is that the biomedical entities of interest are relatively well defined – genes, proteins, organisms, species, drugs, diseases, etc. – and typically expressed as proper nouns. In contrast, defining the entities of interest in marine science turns out to be much harder. Not only does it seem to be more open-ended in nature, the entities themselves tend to be complex and expressed as noun phrases containing multiple modifiers, giving rise to examples like *oxygen depletion in the upper 500 m of the ocean* or *timing and magnitude of surface temperature evolution in the Southern Hemisphere in deglacial proxy records*.

Given the difficulties with entities, we propose to concentrate first on text mining of events, leaving entities underspecified for the time being. Theories and models in marine science are characterised by changing variables and their complex interactions, including correlations, causal relations and chains of positive/negative feedback loops. Many marine scientists are interested in finding evidence – or counter-evidence – in the literature for events of change and their relations. Here we present ongoing work to automatically locate and extract *variables* and their direction of variation: *increasing*, *decreasing* or just *changing*. Examples are given in Table 1.

Since many of these changing variables are long and complex expressions, their frequency of occurrence tends to be low, making the discovery of relations among different variables harder. As a partial solution to this problem, we propose progressive pruning of syntax trees using a set of tree transformation operations. For example, generalising *oxygen depletion in the upper 500 m of the ocean* to *oxygen depletion in the ocean* and subsequently to the much more frequent *oxygen depletion*. Text mining results are then presented as a browsable variable hierarchy which allows users to inspect all mentions of a particular variable type in the text as well as any generalisations or specialisations.

2 Variable extraction

Our text material consists of 10k abstracts from journals published by Nature Publishing Group. Search terms obtained from domain experts were used to query Nature's OpenSearch API[1] for publications in a limited range of relevant journals, after 1997, retrieving records including title and abstract. The top-10k abstracts matching most search terms were selected for further processing with CoreNLP (Manning et al., 2014), including tokenisation, sentence splitting, POS tagging, lemmatisation and parsing. Lemmatised parse trees were obtained by substituting terminals with their lemmas. The resulting new corpus contains 9,586 article abstracts, 59,787 sentences and approximately 4M tokens.

Methods for information extraction broadly rely on either knowledge-based pattern matching or supervised machine learning (Sarawagi, 2008). Although ML approaches are currently dominant in IE research, rule-based systems have several advantages, including: (a) the rules are interpretable and thus suitable for rapid development and domain transfer; and (b) humans and machines can contribute to the same model (Valenzuela-Escárcega et al., 2015). In our case, patterns offered more flexibility in exploring the domain, whereas the manual annotation required for ML demands more commitment to a precise definition of entities, relations and events, which we found hard to achieve at this stage. Tree pattern matching is applied to lemmatised syntax trees using the Tregex engine (Levy and Andrew, 2006), which supports a compact language for writing regular expressions over trees; see Table 1 for examples of patterns and matching phrases. For instance, the pattern for a decreasing variable is defined as a noun phrase (NP) that is immediately dominated (>) by a verb phrase (VP), which in turn is headed by (<<#) the lemma *reduce*. Similarly, the pattern for increase describes an NP dominated by a prepositional phrase (PP) that is headed by the preposition *in* or *of*; in addition, this PP must be preceded by an NP sister node (`$,`) headed by the lemma *increase*.

Patterns were generated by instantiating a small set of hand-written pattern templates, drawing from manually created lists of verbs and nouns ex-

[1] `http://www.nature.com/developers/documentation/api-references/opensearch-api`

Table 1: Examples of tree patterns and matching variables

Direction:	**Tree pattern:**	**Matched variable in sentence:**
Change	`NP <- (/NN/=d1 < variability $ /NN/) !$. PP`	Thus **the annual, Milankovitch and continuum temperature variability** together represent the response to deterministic insolation forcing.
Increase	`NP > (PP <<# (in|of) $, (NP <<# increase))`	The record reveals a linear increase in **annual temperature between 1958 and 2010** by 2.4 +/-1.2 degreesC ...
Decrease	`NP > (VP <<# reduce )`	Some researchers have observed that abundant natural gas substituting for coal could reduce **carbon dioxide (CO2) emissions**.

pressing change, increase or decrease. The patterns cover expression as a main verb (*X increases, something increases X)*, attributive use of verbs (*increasing temperature, temperature is increasing*), head of NP (*a temperature increase*) or NP with PP modifier (*increase in temperature*). The total number of patterns is 320: 90 for change, 122 for increase, 108 for decrease (see supplements for a full list). The total number of matched variables in the corpus is 21,817: 9,352 for change, 7,400 for increase and 5,065 for decrease.

Some variables do not exactly correspond to a node, i.e., not every variable is a valid syntactic phrase. For instance, the pattern for Change in Table 1 matches the NP *the annual, Milankovitch and continuum temperature variability*, whereas the actual variable is *the annual, Milankovitch and continuum temperature*. This is corrected in a post-processing step that deletes the *variability* node from the extracted subtree and substring. For this purpose, the pattern contains an assignment of the name d1 to the node directly dominating the lemma *variability* (`/NN/=d1 < variability`)), allowing a corresponding tree operation to delete this node, which is implemented using the Tsurgeon counterpart of Tregex.

3 Variable generalisation

Since many of the extracted variables are long and complex expressions, their frequency is low. The most frequent variables are generic terms (*climate* 1207, *temperature* 156, *global climate* 73), but over 66% is unique. This evidently impedes the discovery of relations among variables. As a partial solution to this problem, variables are generalised by progressive pruning of syntax trees using a set of tree transformation operations.

Figure 1 shows an example of generalisation by iterative tree pruning. The first transformation STRIP INIT DT strips the initial determiner from the NP. Next, COORD 3.1 deletes everything but the first conjunct from a coordinated structure of three NPs, resulting in *annual temperature* , which is finally reduced to just *temperature* by stripping the premodifier (STRIP PREMOD 1). An analogous procedure is applied to the other two conjuncts of the coordinated structure.

Tree transformations are implemented using Tsurgeon (Levy and Andrew, 2006): Tregex patterns match the syntactic structures of interest, whereas an associated Tsurgeon operation deletes selected nodes (see supplements for details). The transformations are ordered in four groups. The first group handles coordination of two to four conjuncts (cf. Figure 1) – at the phrase level or the lexical level – as well as cases of ellipsis (e.g. *hailstorm frequency and intensity* into *hailstorm frequency* and *hailstorm intensity*). The second group strips bracketed material in parenthetical and list structures. The third group deletes non-restrictive relative clauses and other non-restrictive modifiers preceded by a comma. The final group progressively strips premodifiers (mainly adjectives) from left to right and postmodifiers (PPs, relative clauses) from right to left . Since different transformation may arrive at the same generalisation (e.g. *temperature* in Figure 1), duplicates are filtered out. After filtering, 150,716 variables remained, which is 4.86 times the number of originally extracted variables.

As mentioned, the point of generalisation is to find relations among variables. In Table 1, for example, both *the annual, Milankovitch and continuum temperature variability* and *annual temperature between 1958 and 2010* are generalised to *annual temperature*. However, many generalised variables are unique and thus serve no purpose in relating variables. Retaining only original variables and generalised variables with at least two mentions yields a total of 17,613 variable types.

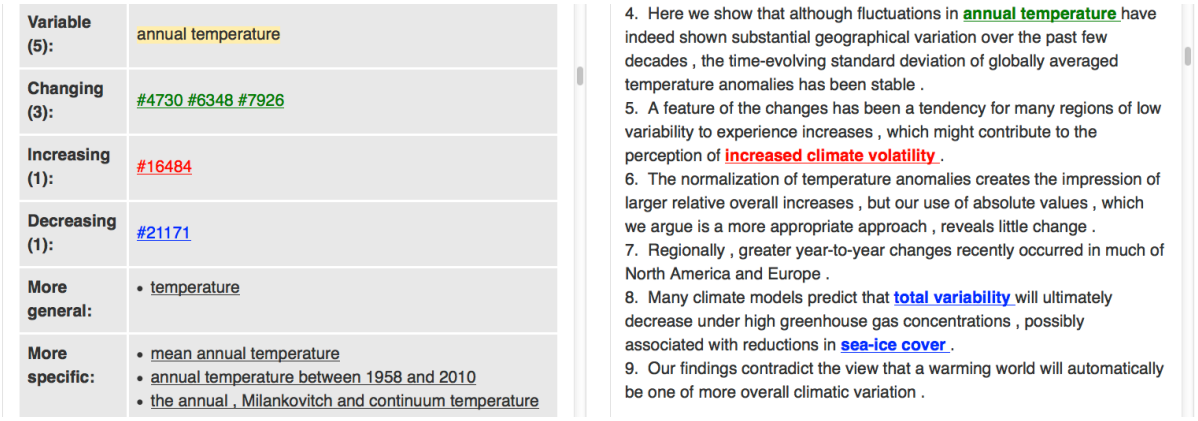

Figure 2: Partial screenshot of user interface showing variable type hierarchy (left) and linked variable mentions in text (right) where colour encodes change (green), increase (red) or decrease (blue)

the annual , Milankovitch and continuum temperature
  STRIP INIT DT → annual , Milankovitch and continuum temperature
    COORD 3.1 → annual temperature
      STRIP PREMOD 1 → temperature
    COORDI 3.2 → Milankovitch temperature
      STRIP PREMOD 1 → temperature
    COORD 3.3 → continuum temperature
      STRIP PREMOD 1 → temperature

Figure 1: Example of generalisation by iterative tree pruning

4 User interface

The output of the text mining step can be regarded as a directed graph where the nodes are variable *types* and the edges point from a more specific variable to a more general variable (as a result of a particular tree transformation). Each variable type is also linked to a set of tokens, i.e. variable mentions in the text which are either changing, increasing or decreasing. Figure 2 shows how this information is presented to the user in a browser (see supplements for full version). The left panel lists the variable types, ordered from most general to most specific and, secondary, on decreasing token frequency. Links point to more specific/general variables types, as well as to changing/increasing/decreasing variable mentions in the text. The right panel shows the source text, where colour encodes changing (green), increasing (red) or decreasing (blue) variable mentions, which are linked to their most specific variable type. This setup allows users to quickly explore variables, for example, finding abstracts containing a variable of interest and from there to related variables.

5 Discussion

We have argued that the paradigm established in biomedical text mining does not transfer directly to other scientific domains like Earth science. A new approach was proposed for extracting variables and their direction of variation (increasing, decreasing or just changing), focusing on events rather than entities. A generic system based on syntactic pattern matching and tree transformations was described for extraction and subsequent generalisation of variable events. Text mining results are presented in an innovative way as a browsable hierarchy ranging from most general to most specific variables, with links to their textual instances. In addition, a first text corpus in marine science was produced, including automatically annotated change events. Our corpus as well as the extracted variables are publicly available[2]. We think our approach to extraction is generalisable to other domains where the entities of interest are common nouns or complex noun phrases rather the proper nouns, e.g. in nanotechnology & nanoscience (Kostoff et al., 2007).

To the best of our knowledge, there are currently no other systems for text mining in Earth science which we can compare our results with, nor are there any benchmark data sets for our task. Most related is (Marsi et al., 2014), but their definition of variables is more restricted and their pilot corpus is too small for evaluation purposes. Reporting on our ongoing work now, future work will include an evaluation by asking domain exports to judge the correctness of extracted variables as well

[2] https://dl.dropboxusercontent.com/u/2370516/emnlp15_corpus.zip

as their generalisations in the given context.

Preliminary observations indicate that most problems originate from syntactic parsing errors, in particular well-known ambiguities in coordination and PP-attachment. As a result, patterns may either fail to match or match unintentionally, yielding incomplete or incoherent variables. Since many sentences are long, complex and domain-specific, it comes as no surprise that the parser often fails to correctly resolve well-known ambiguities in coordination and PP-attachment. However, with pattern matching on strings and/or POS tags instead of syntax trees, determining boundaries of variables would be problematic. False positives also occur because of different semantics of the same pattern, e.g. *change in western Europe* is unlikely to mean literally that the European continent is changing, neither does *changes in less than a few thousand years* imply that past years are changing.

At the same time, certain false negatives are beyond the power of pattern matching. For instance, variation may be entailed rather than explicitly stated: *ocean acidification* entails increasing acidity of ocean water and *Arctic warming* entails increasing temperature in the Arctic region. This is closely related to textual entailment (Androutsopoulos and Malakasiotis, 2010; Dagan et al., 2006), requiring inference in combination with domain knowledge. A related matter is negation (*no increase in global temperature*), which can even be expressed in non-trivial ways (*temperature remained constant*) (Morante and Daelemans, 2009). Variables were also found to be recursive or embedded, expressing "a change of a change". For example, *reduce subseasonal temperature variance* implies both a change in temperature as well as a decrease of this temperature change. The current visualisation falls short in these cases, as HTML browsers cannot render a link in a link.

Generalisation by tree pruning appears to work quite well as long as the parse is correct. However, pruning by itself is insufficient and should be supplemented with other methods. For instance, linking named entities like species, chemicals or locations to unique concepts in appropriate ontologies/taxonomies would support generalisations such as *iron* is a *metal* or a *diatom* is a *plankton*. Generalisation also bears a strong resemblance to other text-to-text generation tasks such as paraphrasing (Androutsopoulos and Malakasiotis, 2010), sentence compression (Jing, 2000) and sentence simplification (Shardlow, 2014). Given suitable training data, ML approaches may therefore be applied, e.g. (Knight and Marcu, 2002; Cohn and Lapata, 2009).

The most general variables are probably too generic to be of much help to a user, e.g. *concentration*, *rate*, *level*, etc. Likewise, *climate* is by far the most frequent changing variable due to the frequently occurring collocation *climate change*. In addition, variables often contain references to previously mentioned entities – anaphoric *it* being the ultimate example of this – suggesting a need for co-reference resolution (Miwa et al., 2012).

Yet another future direction is to structurally model variables as opposed to a possibly over-simplified generalisation. Similar to nominal SRL, one can define relevant arguments including frequency (e.g. annual), temporal scope (between 1958 and 2010), location, etc. The most generic variables mentioned earlier in fact provide a good basis for such modelling.

Extraction and generalisation of variables provides a basis for building systems supporting knowledge discovery. One approach is mining associations between variables frequently co-occurring in the same sentence or abstract (Jenssen et al., 2001; Hashimoto et al., 2012)) More precise results can be expected by extracting causal relations between change events (Chang and Choi, 2005; Blanco et al., 2008; Raja et al., 2013). Pairs of change events – causally or otherwise associated – obtained from different publications can be chained together, possibly in combination with domain knowledge, in order to generate new hypotheses, as pioneered in the work on literature-based knowledge discovery (Swanson, 1986; Swanson, 1988; Swanson and Smalheiser, 1997). Automatic extraction and generalisation of variables from scientific publications thus paves the way for future research on text mining in Earth science.

Acknowledgments

Financial aid from the European Commission (OCEAN-CERTAIN, FP7-ENV-2013-6.1-1; no: 603773) is gratefully acknowledged. We thank Murat Van Ardelan for sharing his knowledge of Marine science and the anonymous reviewers for their valuable comments.

References

Elias Aamot. 2014. Literature-based discovery for oceanographic climate science. In *Proceedings of the Student Research Workshop at the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1–10, Gothenburg, Sweden, April. Association for Computational Linguistics.

Sophia Ananiadou, Douglas B. Kell, and Jun I. Tsujii. 2006. Text mining and its potential applications in systems biology. *Trends Biotechnol*, 24(12):571–579, December.

Sophia Ananiadou, Sampo Pyysalo, Jun'ichi Tsujii, and Douglas B. Kell. 2010. Event extraction for systems biology by text mining the literature. *Trends in biotechnology*, 28(7):381–390, July.

Sophia Ananiadou, Paul Thompson, Raheel Nawaz, John McNaught, and Douglas B Kell. 2014. Event-based text mining for biology and functional genomics. *Briefings in functional genomics*, page elu015.

Ion Androutsopoulos and Prodromos Malakasiotis. 2010. A survey of paraphrasing and textual entailment methods. *Journal of Artificial Intelligence Research*, 38:135–187, May.

Eduardo Blanco, Nuria Castell, and Dan I Moldovan. 2008. Causal relation extraction. In *LREC*, pages 310–313.

Du-Seong Chang and Key-Sun Choi. 2005. Causal relation extraction using cue phrase and lexical pair probabilities. In *Natural Language Processing–IJCNLP 2004*, pages 61–70. Springer.

Aaron M. Cohen and William R. Hersh. 2005. A survey of current work in biomedical text mining. *Briefings in Bioinformatics*, 6(1):57–71, March.

Kevin Bretonnel Cohen and Lawrence Hunter. 2008. Getting Started in Text Mining. *PLoS Comput Biol*, 4(1):e20+, January.

Trevor Cohn and Mirella Lapata. 2009. Sentence compression as tree transduction. *J. Artif. Int. Res.*, 34(1):637–674.

The Gene Ontology Consortium. 2001. Creating the gene ontology resource: Design and implementation. *Genome Research*, 11(8):1425–1433, August.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2006. The PASCAL Recognising Textual Entailment challenge. *Machine Learning Challenges*, pages 177–190.

Julia A Ekstrom and Gloria T Lau. 2008. Exploratory text mining of ocean law to measure overlapping agency and jurisdictional authority. In *Proceedings of the 2008 international conference on Digital government research*, pages 53–62. Digital Government Society of North America.

Chikara Hashimoto, Kentaro Torisawa, Stijn De Saeger, Jong H. Oh, and Jun'ichi Kazama. 2012. Excitatory or inhibitory: A new semantic orientation extracts contradiction and causality from the web. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, EMNLP-CoNLL '12, pages 619–630, Stroudsburg, PA, USA. Association for Computational Linguistics.

Marti A. Hearst. 1999. Untangling text data mining. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics*, ACL '99, pages 3–10, Stroudsburg, PA, USA. Association for Computational Linguistics.

Lynette Hirschman, Alexander Yeh, Christian Blaschke, and Alfonso Valencia. 2005. Overview of BioCreAtIvE: critical assessment of information extraction for biology. *BMC Bioinformatics*, 6(Suppl 1):S1+.

T. K. Jenssen, A. Laegreid, J. Komorowski, and E. Hovig. 2001. A literature network of human genes for high-throughput analysis of gene expression. *Nat Genet*, 28:21–28.

Hongyan Jing. 2000. Sentence reduction for automatic text summarization. In *Proceedings of the sixth conference on Applied natural language processing*, pages 310–315. Association for Computational Linguistics.

J. D. Kim, T. Ohta, Y. Tateisi, and J. Tsujii. 2003. GENIA corpus—a semantically annotated corpus for bio-textmining. *Bioinformatics*, 19(suppl 1):i180–i182, July.

Kevin Knight and Daniel Marcu. 2002. Summarization beyond sentence extraction: A probabilistic approach to sentence compression. *Artificial Intelligence*, 139(1):91–107.

Ronald N. Kostoff, Raymond G. Koytcheff, and Clifford G.Y. Lau. 2007. Global nanotechnology research literature overview. *Technological Forecasting and Social Change*, 74(9):1733 – 1747. Three Special Sections: Assessment of China's and India's Science and Technology Literature Nanotechnology Policy Minding the Gap: Previewing the Potential of Breakthrough Technologies.

Martin Krallinger, Alfonso Valencia, and Lynette Hirschman. 2008. Linking genes to literature: text mining, information extraction, and retrieval applications for biology. *Genome Biology*, 9(Suppl 2):S8+, September.

Roger Levy and Galen Andrew. 2006. Tregex and Tsurgeon: tools for querying and manipulating tree data structures. In *Proceedings of the fifth international conference on Language Resources and Evaluation*, pages 2231–2234.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60.

Erwin Marsi, Pinar Oztürk, Elias Aamot, Gleb Sizov, and Murat V Ardelan. 2014. Towards text mining in climate science: Extraction of quantitative variables and their relations. In *Proceedings of the Fourth Workshop on Building and Evaluating Resources for Health and Biomedical Text Processing*, Reykjavik, Iceland.

Makoto Miwa, Paul Thompson, and Sophia Ananiadou. 2012. Boosting automatic event extraction from the literature using domain adaptation and coreference resolution. *Bioinformatics*, 28(13):1759–1765.

Roser Morante and Walter Daelemans. 2009. Learning the scope of hedge cues in biomedical texts. In *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing*, pages 28–36. Association for Computational Linguistics.

Brendan O'Connor, David Bamman, and Noah Smith. 2011. Computational text analysis for social science: Model assumptions and complexity. In *Proceedings of the Second Workshop on Computational Social Science and the Wisdom of the Crowds (NIPS 2011)*.

Sampo Pyysalo, Filip Ginter, Juho Heimonen, Jari Bjorne, Jorma Boberg, Jouni Jarvinen, and Tapio Salakoski. 2007. BioInfer: a corpus for information extraction in the biomedical domain. *BMC Bioinformatics*, 8:50.

Sampo Pyysalo, Tomoko Ohta, Rafal Rak, Dan Sullivan, Chunhong Mao, Chunxia Wang, Bruno Sobral, Jun'ichi Tsujii, and Sophia Ananiadou. 2012. Overview of the ID, EPI and REL tasks of BioNLP shared task 2011. *BMC Bioinformatics*, 13(Suppl 11):S2+, June.

Kalpana Raja, Suresh Subramani, and Jeyakumar Natarajan. 2013. Ppinterfinder–a mining tool for extracting causal relations on human proteins from literature. *Database (Oxford)*, 2013:bas052.

Raul Rodriguez-Esteban. 2009. Biomedical Text Mining and Its Applications. *PLoS Comput Biol*, 5(12):e1000597+, December.

Sunita Sarawagi. 2008. Information extraction. *Found. Trends databases*, 1(3):261–377, March.

Matthew Shardlow. 2014. A survey of automated text simplification. *International Journal of Advanced Computer Science and Applications*, 4(1).

Matthew S. Simpson and Dina Demner-Fushman. 2012. Biomedical Text Mining: A Survey of Recent Progress. In Charu C. Aggarwal and ChengXiang Zhai, editors, *Mining Text Data*, pages 465–517. Springer US.

Don R. Swanson and Neil R. Smalheiser. 1997. An interactive system for finding complementary literatures: a stimulus to scientific discovery. *Artificial Intelligence*, 91(2):183–203, April.

Don R. Swanson. 1986. Fish oil, raynaud's syndrome, and undiscovered public knowledge. *Perspectives in biology and medicine*, 30(1):7–18.

Don R. Swanson. 1988. Migraine and magnesium: eleven neglected connections. *Perspectives in Biology and Medicine*, 31(4):526–557.

Marco A. Valenzuela-Escárcega, Gustave Hahn-Powell, Thomas Hicks, and Mihai Surdeanu. 2015. A domain-independent rule-based framework for event extraction. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Assian Federation of Natural Language Processing: Software Demonstrations (ACL-IJCNLP)*.

Piek Vossen, German Rigau, Eneko Agirre, Aitor Soroa, Monica Monachini, and Roberto Bartolini. 2010. KYOTO: an open platform for mining facts. In *Proceedings of the 6th Workshop on Ontologies and Lexical Resources* , pages 1–10, Beijing, China, August. Coling 2010 Organizing Committee.

Ce Zhang, Vidhya Govindaraju, Jackson Borchardt, Tim Foltz, Christopher Ré, and Shanan Peters. 2013. GeoDeepDive: Statistical inference using familiar data-processing languages. In *Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data*, SIGMOD '13, pages 993–996, New York, NY, USA. ACM.

Pierre Zweigenbaum and Dina Demner-Fushman. 2009. Advanced Literature-Mining Tools. In David Edwards, Jason Stajich, and David Hansen, editors, *Bioinformatics*, pages 347–380. Springer New York.

Pierre Zweigenbaum, Dina Demner-Fushman, Hong Yu, and Kevin B. Cohen. 2007. Frontiers of biomedical text mining: current progress. *Briefings in Bioinformatics*, 8(5):358–375, September.