What can NLP tell us about BioNLP?

Attapol Thamrongrattanarit, Michael Shafir, Michael Crivaro, Bensiin Borukhov, Marie Meteer
Department of Computer Science
Brandeis University
Waltham, MA 02453, USA
{tet, mshafir, mcrivaro, bborukhov, mmeteer}@brandeis.edu

Abstract

The goal of this work is to apply NLP techniques to the field of BioNLP in order to gain a better insight into the field and show connections and trends that might not otherwise be apparent. The data we analyzed was the proceedings from last decade of BioNLP workshops. Our findings reveal the prominent research problems and techniques in the field, their progression over time, the approaches that researchers are using to solve those problems, insightful ways to categorize works in the field, and the prominent researchers and groups whose works are influencing the field.

1 Introduction

Thanks to improving technology and the discovery of stronger statistical methods, natural language processing techniques have more power than ever to give us insights into real datasets too large for humans to efficiently process. In the field of BioNLP, we see that natural language processing has a wide range of applications within the medical domain from analysis of clinical data to literature. With the increasing amount of publications in this growing field, building a classification structure is helpful both for categorizing papers in a sensible way and for recognizing the trends that brought the field to where it is today. Understanding the current nature of the field can show us where the most effort is needed, while taking a look at where the field has been can highlight successes and even unanswered questions.

As the use of NLP in the medical domain has expanded in recent years so has the amount of freely-available online research. With this wealth of information comes a problem, however, as it is not truly feasible for humans to read through all the research out there and classify it in a way that will capture the less-obvious trends and the finer relationships between seemingly-disconnected works. Instead, we propose that statistical methods can help us discover both the most reasonable way to partition the field and also see how the research has changed over the past decade. The longer term goal for the work is to contribute to a "map" of the field that can be a community resource, such as www.medlingmap.org, described in Meteer, et al. (2012).

Schuemie et al. (2009) used clustering techniques to analyze the domain of Medical Informatics. They processed a large number of Medline abstracts to find a subset of the journals classified as "Medical Informatics" whose content was sufficiently related to constitute a basis for the field. Using hierarchical clustering, they determined that such a group of journals exists and, as we might expect, the rest of the journals were largely disconnected. They also used this cluster of journals as the basis for a topic modeling task. Analyzing the articles from their new basis of journals, they found three very strong, topic-based clusters, each comprised of three sub-clusters. Overall, Schuemie et al. (2009) demonstrated how it is possible to gain a great deal of insight into the nature of a field by using statistical methods over that field's literature. More recently, Gupta and Manning (2011) used automatic methods to tag documents for "focus," "technique," and "domain" by examining

over 15,000 ACL abstracts. This level of categorization is useful because it expands beyond the simple notion of the ”topic” to implicitly show if a work, for example, is about an application of named-entity recognition or if it simply uses NER to achieve a greater task. The techniques demonstrated by Gupta and Manning could be very enlightening if applied to the BioNLP proceedings, though in this paper we refrain from drawing conclusions about individual papers. Instead, we will relate them through the topics extracted from the full-text proceedings.

For our task, we look to the ACL and NAACL-associated workshops on NLP applications in the medical domain. Entering its 11th year, the BioNLP workshop (under a variety of names) has given us ten rich and varied proceedings in addition to a pair of more focused shared tasks. All in all, the workshops have produced over 270 unique papers. Our data of 270 documents was small relative to (Schuemie et al., 2009) 6.3 million documents; therefore, we chose to expand our analysis to the full text of the documents instead of just the abstracts. Additionally, using the full papers allowed us to capture information about document content that abstracts alone could not provide.

2 Methods and Results

2.1 Pipeline Architecture

We implemented a document processing pipeline that would allow our approaches to be generalizable, easily reproducible, and extendable. Each of our analytic processes was integrated into this pipeline and parameterized to allow us proper flexibility for empirical experimentation. The pipeline works by managing the interaction between a configurable set of data layers and a configurable set of processing stages over those layers. It supports saving and loading its internal state between stages. In addition, layers and stages follow specific templates that reduce the amount of code to write and maintain. The ordering and activation of each stage is also parameterized. This pipeline allowed us to quickly and efficiently experiment with various approaches and combine them. The sample implementation of this pipeline is available publicly at github.com/attapol/mapping_bionlp.

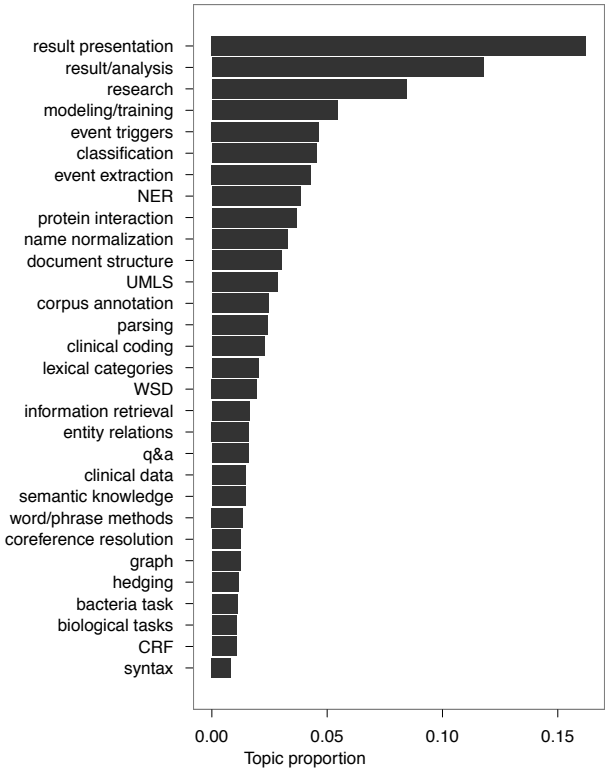

Figure 1: Average topic proportion across all the documents output by the LDA model

2.2 Preprocessing

The papers from the BioNLP workshop are all available freely from the ACL Anthology Archive [1]. We first extracted the text from the PDF files using pdf2text unix tool and then tagged them all for title, authors, places of origin, abstract, content, and references. In all cases, the abstract, content, and references were separated automatically using a script, and the places had to be hand-annotated. Papers from 2004 onward (starting with the first BioLINK workshop) have complete BibTeX entries that allowed us to automatically extract the titles and authors, but for 2002 and 2003 this work had to be done manually. Since we wanted to perform our analysis solely on the prose of the papers, and not on any of the numerical data, we filtered out portions of the text containing elements such as tables, graphs, footnotes, and URLs. We also filtered out stopwords (as defined by the NLTK package (Bird and Loper, 2004) for Python).

[1] aclweb.org

2.3 Topic Modeling

Using the Mallet toolkit (McCallum, 2002), we were able to generate topics from our cleaned data using the Latent Dirichlet Allocation (LDA) model. This approach allows us to represent each document as a vector of topic proportions instead of a bag of words, which prevents the problem of sparsity. When we set the number of topics to 30, the system output a set of distinct topics that seem to describe a range of tasks and methods within the domain of BioNLP. The topics generated by the LDA model reflect areas of study that are being pursued, techniques that are being applied, and resources that are being consulted in the field. A list of the generated topics along with the associated keywords is shown in Table 1 and the distributions of the topics across the entire document set is displayed in Figure 1.

Additionally, we found that the topics generated by LDA were more informative about the full content of a work than those generated by TF-IDF as TF-IDF would often give too much weight to specific examples over general concepts. For example, TF-IDF tended to select specific names of resources and ontologies rather than general terms. For example, it selected "Frame-net" instead "ontology" and "RadLex" instead of "lexicon". We concluded that, while interesting, TF-IDF results were not strongly suited for capturing an overall glimpse of the field. However, we think that TF-IDF can be much more useful in its more traditional capacity of finding document-specific keywords; we aim to use these indices to partially automate keyword generation for MedlingMap (Meteer et al., 2012), which is our accompanying project.

2.4 Topic Correlation

While looking at the topic proportions for each of our LDA topics overall can help us paint a broad picture of the field, it can also help to look at the relationship between these topics as they occur in the documents. Some topics appear highly ranked in nearly all papers, such as the topic that is characterized by terms such as "system" and "results", and the topic that includes "precision" and "recall" because they reflect the performance evaluation convention in the field. However, most topics are only dominant in a small subset of the papers. Some topics refer to tasks (e.g. named-entity recognition, hedging) and others refer to techniques (e.g. CRFs, parsing). We can look at how often pairs of task-related topic and method-related topic co-occur to see if researchers in the community are using certain techniques in conjunction with solving certain problems. We first turned a topic proportion vector into a binary vector where each element indicates which topic is discussed more extensively than average. Then, we counted the co-occurrences of tasks and methods of interest. To demonstrate this, we computed the number of papers that substantially discuss event extraction in conjunction with parsing, graph, lexical categories, or semantic knowledge (Figure 2). This topic comparison method provides a means of visualizing how researchers in the field are approaching BioNLP problems. It reveals that parsing and graph-based methods are commonly used in biological event extraction, while lexical categories and semantic knowledge are not as central to many of the approaches to this task. Moving forward, tracking how these correlations change over time will provide an insightful reflection of the field's progress on the task in a more meaningful way than evaluation scores alone. While a deeper analysis of all of such trends is beyond the scope of this paper, it certainly warrants further investigation.

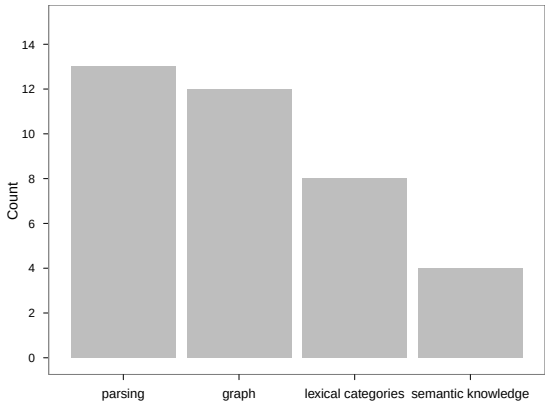

Figure 2: The bar plot shows the frequency of the co-occurrences between the event extraction topic and some of the method-related topics.

Table 1: The resulting topics and their associated keywords generated by LDA model with 30 topics

Topic Name	Keywords
Event Extraction	event, task, extraction, types, data, annotation
Coreference Resolution	anaphora, resolution, referring, links, antecedent
Graph	graph, relationships, nodes, edges, path, constraint, semanics
Clinical Coding	medical, data, codes, patients, notes, reports
Hedging	negation, scope, cues, speculative, hedge, lexical
Clinical Data	condition, historical, clinical, temporal, reports, context
Bacteria Task	bacteria, names, location, organisms, taxonomic, host, roles, type
Entity Relations	relations, entities, feature, static, renaming, annotated, pairs
Document Structure Analysis	rst, classification, abstracts, identification, data, terms
Q&A	question, answer, structure, passage, evidence, purpose
Event Triggers	triggers, dependency, binding, type, training, token, detection
Semantic Knowledge	semantic, frame, structures, argument, patterns, domain, types
Protein Interaction	protein, patterns, interaction, extraction, biological
Parsing	dependency, parser, tree, syntactic, structures, grammar, link
Name Normalization	gene, names, dictionary, normalization, protein, database, synonyms
Named Entity Recognition	entity, named, word, recognition, features, class, protein
Information Retrieval	search, queries, interface, text, retrieval, document
Corpus Annotation	corpus, annotation, guidelines, agreement, papers
Lexical Categories	semantic, categories, resources, simstring, lexical, gazetteer, features
Research	text, figure, knowledge, domain, research, complex, processing
CRF	crf, skip, chain, linear, dependency, words, edges, sentence
Result Discussion	system, based, results, set, table, test, shown, approach
Biological Tasks	species, disease, mutation, mentions, features, entities, acronym
UMLS	terms, semantic, phrases, umls, concepts, ontology, corpus
Word/Phrase Methods	words, measures, morphological, tag, token, chunking, form
WSD	disambiguation, sense, word, semantic, wsd, ambiguous
Result Analysis	found, number, precision, recall, cases, high, related, results
Classification	features, training, data, classification, set, learning, svm
Modeling/Training	training, data, model, tagger, performance, corpus, annotated
Syntax	attachment, pps, np, fragments, pp, noun, vp, nos, pattern

2.5 Trends within the subdisciplines in Biomedical NLP Literature

Our analysis of temporal trends builds on the idea proposed by (Hall et al., 2008) in their analysis of the changing trends in the field of computational linguistics over time. In their approach, they attempted, among other things, to analyze which topics were up and coming in the field and which were becoming less popular. Given their sound results, we decided to perform the same kind of trend analysis over the BioNLP topics. For many of our 30 topics, there was little change in the topic frequency over time. Considering the relative youth of the BioNLP field, this result is not entirely surprising. We did, however, find a few topics that have undergone notable changes in these past ten years, as observable in Figure 3. In particular, we found that two topics have seen surges of activity in recent years, whereas there were three topics that started out strong in the early years but that have since petered off. The two topics that have gained popularity in the past few years both involve biomedical events. Specifically, one such topic is primarily about event extraction tasks, and the other is about event triggers and the more fine-grained roles one needs to tag to categorize such events. The popularity of these two tasks is hardly surprising, given that they were the focus of the 2009 and 2011 shared tasks which were about working with events in both general and detailed ways. We do notice, however, that the growing trends continue in 2010 as well, when there was no shared task, and so we can see that events are of great interest in the field at present even without the added incentive of the shared tasks. It is reasonable to suggest that the 2009 BioNLP Shared Task in event extraction generated interest in the topic that continued through 2010 and 2011. Two more topics originally saw their popularity rise in the early years, but have

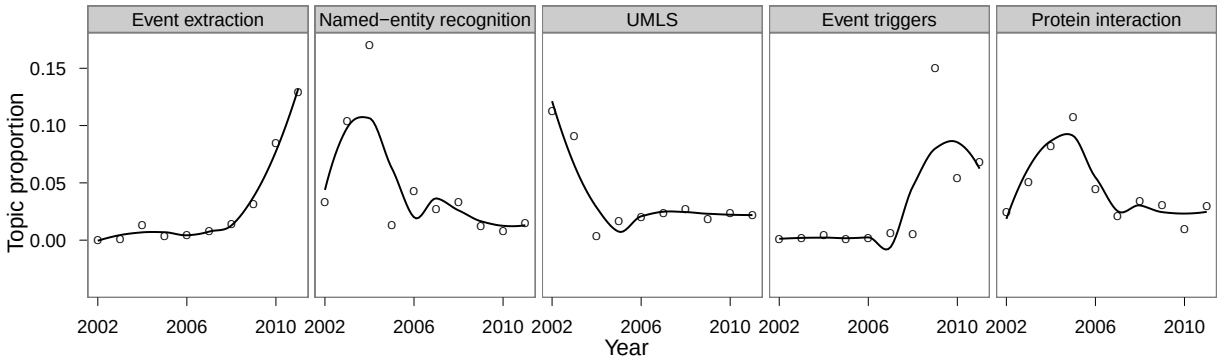

Figure 3: Topic proportions for some topics have gone through dramatic changes, which reflect how research interest and methodology evolve over time.

since seen it fade. Each of these is a specific task: named-entity recognition, which dropped off after 2004, and protein interaction, which saw a sharp decline after 2005. Although a detailed causal analysis is beyond the scope of this paper, we might wonder what accounts for these drops in topic proportion. The explanation that seems most likely is that great strides were made in these areas early on, but we have since reached a plateau in advancements. As such, the research has moved elsewhere. The only topic to see a steady decrease from the start was the topic associated with the Unified Medical Language System. In general, we can view a trend associated with a resource differently from one associated with a task. Above, when discussing tasks, we saw where the research currently has been heading and where it has been. With a resource, we could consider an upward trend to represent either an increased number of applications to a task or perhaps an expansion of the resource itself. In the case of UMLS, the downward trend likely suggests that the field has moved away from this particular resource, either because it does not apply as well to newer tasks or because it has been replaced with something more powerful.

2.6 Cluster Analysis

Our next step with the LDA-generated topics was to run a *k*-means clustering algorithm. We used the same topic proportion vector and a Euclidean metric to create the feature space for clustering. We used the standard *k*-means function in the statistical language R (R Development Core Team, 2010). The assumption of the LDA model biases each topic proportion vector to be sparse (Blei et al., 2003), and this turns out to be true in our data set. Therefore, we chose the number of clusters to match the number of topics so that the document space can be partitioned proportionally to its dimensionality. This clustering provides us with a useful schema for document classification within the domain of BioNLP. We can use the clusters as a guide for how to organize the current papers, and we can also view the clusters as a guide for how to select relevant research to build future work on. Clusters bring together related papers from different research groups and multiple workshops, such as those shown in Table 2. In all of these examples, the selection of these sets of papers simply based on keyword search would be very difficult, since many of the key terms are going to be present in a much larger set of documents.

2.7 Author Relation Analysis

As an additional task, we investigated the connections between authors in the BioNLP proceedings. Eggers et al. (2005) used a graph to visualize who was being cited by whom in ISI publications. There, the hope was to identify which authors worked within the same subdisciplines by examining clusters within the citation graph. By examining who cited whom in the BioNLP publications, we hoped instead to uncover the authors of the most influential papers, both within our own clusters and outside the scope of the BioNLP workshops. In our model, which can be viewed in Figure 4, we constructed a

List of papers assigned to the cluster where the most discussed topic is parsing (44.74% on average)
A Comparative Study of Syntactic Parsers for Event Extraction
Analysis of Link Grammar on Biomedical Dependency Corpus Targeted at Protein-Protein Interactions
On the unification of syntactic annotations under the Stanford dependency scheme
A Transformational-based Learner for Dependency Grammars in Discharge Summaries
A Study on Dependency Tree Kernels for Automatic Extraction of Protein-Protein Interaction

List of papers assigned to the cluster where the most discussed topic is clinical data (48.74% on average)
Applying the TARSQI Toolkit to Augment Text Mining of EHRs
Temporal Annotation of Clinical Text
Extracting Distinctive Features of Swine (H1N1) Flu through Data Mining Clinical Documents
ConText: An Algorithm for Identifying Contextual Features from Clinical Text
Distinguishing Historical from Current Problems in Clinical Reports – Which Textual Features Help?

Table 2: Two sample clusters from running *k*-means clustering algorithm on the corpus

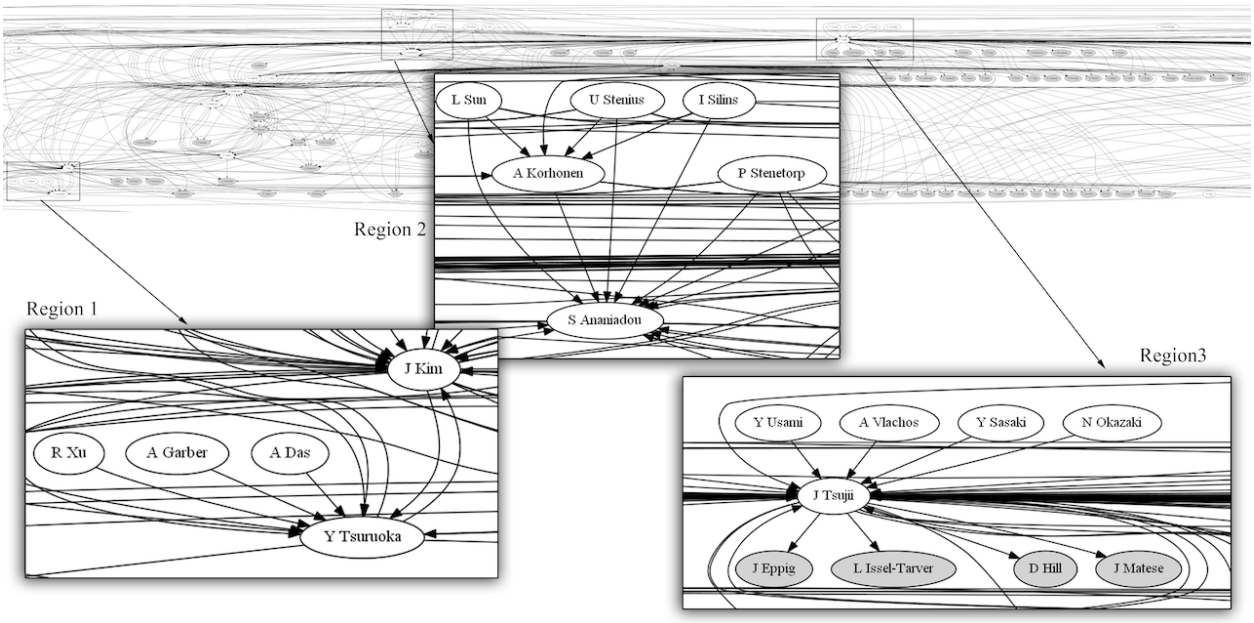

Figure 4: Citation relation graph. Each node represents an author whose papers are either published in the BioNLP proceedings or are cited by one of the papers in the proceedings. Each edge represents a citation activity.

directed graph of author citations from the BioNLP workshops and shared tasks. We disregarded the author ordering within each paper and gave the same weights for all authors whose names appear on the paper. In this graph, a node points to another node if that author cited the other author at least three times. Additionally, a white node signifies an author who published in the BioNLP workshop between 2008 and 2011, whereas a grey node is someone who did not, but was cited in papers during that time span. As can be seen in Figure 4 above, which is itself only a piece of the complete graph, this graph is rather large and complex, showing us a large degree of interconnectedness and interdependence in the field. Simply from the density of the lines, we can find some of the most influential figures, such as Jun'ichi Tsujii, shown in Region 3 and Yoshimasa Tsuruoka, shown in Region 2. Unsurprisingly, Tsujii's node is bustling with activity, as a very large number of authors cite works with Tsujii as an author, and his own prolific authorship (or co-authorship) naturally has him citing a variety of authors. The white nodes near his own show the authors who published BioNLP papers and primarily referenced his works, whereas the grey nodes near his show people who didn't publish, but who Tsujii cited in the proceedings multiple

times. Thus, proximity can also be very telling in a graph like this. Since nodes with a heavier reliance on one another tend to end up closer to one another, we can also observe something of a "citation hierarchy" in sections of the graph. Region 2 is a prime example of this notion. We observe Ananiadou at the bottom with a large number of incoming edges. Above her node, we see Korhonen, who cites Ananiadou but is also cited by a number of other authors herself. Finally, above Korhonen there are a series of single nodes who cite her (and Ananiadou) but are without incoming edges of their own. We can think of this as something of a "local hierarchy", consisting of authors who are closely connected, with the more heavily-cited (and heavily-citing) easy to pick out.

3 Next Steps

The work described here provides a snapshot into the field. Underlying the work is a toolset able to reproduce the results on new sets of data to continue tracking the trends, topics, and collaborations. However, to be really useful to the research community, the results need to be captured in a way that can facilitate searches in this domain and support ongoing research. In order to do this, we are in the process of incorporating the results presented here in a content management system, MedLingMap (Meteer et al., 2012), which supports faceted indexing. Research in search interface design has shown that techniques which can create hierarchical faceted metadata stuctures of a domain significantly increase the ability of users to efficiently access documents in the collection (Stoica et al., 2005). The techniques described here can be fed into MedLingMap to create much of the metadata required to efficiently navigate the space.

4 Conclusion

In this report, we have outlined a variety of methods that can be used to gain a better understanding of BioNLP as a field. Our use of topic modeling demonstrates that the field already has several well-defined tasks, techniques, and resources, and we showed that we can use these topics to gain insight into the major research areas in the field and how those efforts areas are progressing. We put forth that this analysis could be powerful in recognizing when a problem has been effectively "solved", when a technique falls out of favor, and when a resource grows outdated. At the same time, we can see rising trends, such as how the 2009 shared task spurred an obvious 2010 interest in event extraction, and the correlations in the field between certain approaches and certain tasks. Through clustering, we were able to show that these topics also can help us separate the documents from the field into distinctive groups with a common theme, which can aid in building a database for current documents and classifying future ones. Finally, we ended with an analysis of author relations based on citation frequency and demonstrated how such a structure can be useful in identifying influential figures through their works.

As a further benefit of this work, we propose to use it to create a more lasting resource for the community that makes these results available to support search and and navigation in the bio-medical NLP field.

References

Andrew McCallum. 2002. MALLET: A Machine Learning for Language Toolkit. http://mallet.cs.umass.edu.

David Blei, Andrew Ng, and Michael Jordan. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993-1022

David Hall, Dan Jurafsky, and Christopher D. Manning. 2008. Studying the history of ideas using topic models. *In EMNLP.*

MJ Schuemie, JL Talmon, PW Moorman, and JA Kors 2009. Mapping the domain of medical informatics. *Methods Inf Med* 48:76-83.

Marie Meteer, Bensiin Borukhov, Michael Crivaro, Michael Shafir, and Attapol Thamrongrattanarit 2012. MedLingMap: Growing a resource for the Bio-Medical NLP field.

R Development Core Team 2010. R: A language and environment for statistical computing. http://www.R-project.org.

S. Eggers, Z. Huang, H. Chen, L. Yan, C. Larson, A. Rashid, M. Chau, and C. Lin. 2005 Mapping Medical Informatics Research Medical Informatics: Knowledge Management and Data Mining in BioMedicine. Springer Science+Business Media, Inc.

S Gupta, and C. Manning 2011. Analyzing the dynamics of research by extracting key aspects of scientific papers. Proceedings of IJCNLP.

Steven Bird, Edward Loper, and Ewan Klein. 2009. Natural Language Processing with Python. OReilly Media Inc.

Emilia Stoica, Marti A. Hearst, and Megan Richardson. 2007. Automating creation of hierarchical faceted metadata structure. Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT 2007).