Biomedical Natural Language Processing
Natural Language Processing (NLP)

The scope of NLP ranges from theoretical Computational Linguistics topics to highly practical Language Technology topics. The focus of the series is on new results in NLP and modern alternative theories and methodologies. For an overview of all books published in this series, please see http://benjamins.com/catalog/nlp

Editor
Ruslan Mitkov
University of Wolverhampton

Advisory Board
Gloria Corpas
University of Malaga

Robert Dale
Macquarie University, Sydney

Eduard Hovy
University of Southern California

Alma Kharrat
Microsoft Research

Richard Kittredge
GoGenTex Inc, Ithaca

Lori Lamel
LIMSI, CNRS

Inderjeet Mani
Yahoo! Labs, Sunnyvale, USA

Carlos Martín-Vide
Rovira i Virgili Un., Tarragona

Rada Mihalcea
University of Michigan

Andrei Mikheev
Daxtra Technologies

Roberto Navigli
Universita di Sapienza, Roma

John Nerbonne
University of Groningen

Nicolas Nicolov
Microsoft Research

Kemal Oflazer
Carnegie Mellon University, Qatar

Constantin Orăsan
University of Wolverhampton

Manuel Palomar
University of Alicante

Khalil Simaan
University of Amsterdam

Richard Sproat
Google Research

Key-Yih Su
Behaviour Design Corp.

Benjamin Tsou
The Hong Kong Institute of Education

Yorick Wilks
Florida Institute of Human and Machine Cognition

Editorial Assistant
Miranda Chong
University of Wolverhampton

Volume 11
Biomedical Natural Language Processing
by Kevin Bretonnel Cohen and Dina Demner-Fushman
Biomedical Natural Language Processing

Kevin Bretonnel Cohen
University of Colorado, School of Medicine

Dina Demner-Fushman
National Library of Medicine

John Benjamins Publishing Company
Amsterdam / Philadelphia
Biomedical Natural Language Processing / Kevin Bretonnel Cohen and Dina Demner-Fushman.

Includes bibliographical references and index.


© 2014 – Kevin Bretonnel Cohen & Dina Demner-Fushman

No part of this book may be reproduced in any form, by print, photoprint, microfilm, or any other means, without written permission from the publisher.

John Benjamins Publishing Co. · P.O. Box 36224 · 1020 ME Amsterdam · The Netherlands
John Benjamins North America · P.O. Box 27519 · Philadelphia PA 19118-0519 · USA
Acknowledgments

Our thanks go to our colleagues whose innovative biomedical NLP research provided the foundations for this book and who reviewed some or all of it. Here is an incomplete list of those whose work made this book possible: Sophia Ananiadou, Alan Aronson, Christian Blaschke, Wendy Chapman, Mark Craven, Laurie Damianos, Marcelo Fiszman, Noemie Elhadad, John Ely, Carol Friedman, Ken Fukuda, Bob Futrelle, Aaron Gabow, Rob Gaizauskas, Graciela Gonzalez, Jorg Hakenberg, Marti Hearst, Larry Hunter, Helen Johnson, Cem Kaner, Judith Klavans, Zhiyong Lu, Kathy McKeown, Preslav Nakov, John Pestian, Thomas Rindflesch, Larry Smith, Padmini Srinivasan, Junichi Tsujii, Alfonso Valencia, Karin Verspoor, Ellen Voorhees, John Wilbur, Pete White, Hong Yu, and Pierre Zweigenbaum.

We are very grateful to Matthew Simpson and Ivo Georgiev for their thorough and thoughtful reviews of the book.
# Table of contents

## List of figures

| List of figures | XI |

## CHAPTER 1

### Introduction to natural language processing

<table>
<thead>
<tr>
<th>1.1 Some definitions</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1.1 Computational linguistics</td>
<td>1</td>
</tr>
<tr>
<td>1.1.2 Natural language processing</td>
<td>1</td>
</tr>
<tr>
<td>1.1.3 Text mining</td>
<td>2</td>
</tr>
<tr>
<td>1.1.4 Usage of these definitions in practice</td>
<td>2</td>
</tr>
<tr>
<td>1.2 Levels of document and linguistic structure and their relationship to natural language processing</td>
<td>2</td>
</tr>
<tr>
<td>1.2.1 Document structure</td>
<td>2</td>
</tr>
<tr>
<td>1.2.2 Sentences</td>
<td>4</td>
</tr>
<tr>
<td>1.2.3 Tokens</td>
<td>4</td>
</tr>
<tr>
<td>1.2.4 Stems and lemmata</td>
<td>5</td>
</tr>
<tr>
<td>1.2.5 Part of speech</td>
<td>5</td>
</tr>
<tr>
<td>1.2.6 Syntactic structure</td>
<td>6</td>
</tr>
<tr>
<td>1.2.7 Semantics</td>
<td>6</td>
</tr>
</tbody>
</table>

## CHAPTER 2

### Historical background

| 2.1 Early work in the medical domain | 7 |
| 2.2 The emergence of the biological domain | 9 |
| 2.3 Clinical text mining | 10 |
| 2.4 Types of users of biomedical NLP systems | 11 |
| 2.5 Resources and tools | 12 |
| 2.6 Legal and ethical issues | 17 |
| 2.7 Is biomedical natural language processing effective? | 18 |

## CHAPTER 3

### Named entity recognition

| 3.1 Overview | 21 |
| 3.2 The crucial role of named entity recognition in BioNLP tasks | 22 |
| 3.3 Why gene names are the way they are | 22 |
3.4 An example of a rule-based gene NER system: KeX/PROPER 25
3.5 An example of a statistical disease NER system 28
3.6 Evaluation 29

CHAPTER 4
Relation extraction 31
4.1 Introduction 31
4.1.1 Protein–protein interactions as an information extraction target 31
4.2 Binarity of most biomedical information extraction systems 33
4.3 Beyond simple binary relations 33
4.4 Rule-based systems 37
4.4.1 Co-occurrence 38
4.4.2 Example rule-based systems 39
4.4.3 Machine learning systems 41
4.5 Relations in clinical narrative 43
4.5.1 MedLEE 44
4.6 SemRep 45
4.6.1 NegEX 48
4.7 Evaluation 49

CHAPTER 5
Information retrieval/document classification 51
5.1 Background 51
5.1.1 Growth in the biomedical literature 51
5.1.2 PubMed/MEDLINE 52
5.2 Issues 54
5.3 A knowledge-based system that disambiguates gene names 55
5.4 A phrase-based search engine, with term and concept expansion and probabilistic relevance ranking 58
5.5 Full text 59
5.6 Image and figure search 60
5.7 Captions 61
5.7.1 Evaluation 61

CHAPTER 6
Concept normalization 63
6.1 Gene normalization 63
6.1.1 The BioCreative definition of the gene normalization task 64
6.2 Building a successful gene normalization system 65
6.2.1 Coordination and ranges 66
6.2.2 An example system 67
6.3 Normalization and extraction of clinically pertinent terms 71
6.3.1 MetaMap UMLS mapping tools 71
# Table of contents

**CHAPTER 7**  
**Ontologies and computational lexical semantics**  
7.1 Unified Medical Language System (UMLS)  77  
7.1.1 The Gene Ontology  80  
7.2 Recognizing ontology terms in text  80  
7.3 NLP for ontology quality assurance  81  
7.4 Mapping, alignment, and linking of ontologies  82  

**CHAPTER 8**  
**Summarization**  
8.1 Medical summarization systems  87  
8.1.1 Overview of medical summarization systems  87  
8.1.2 A representative medical summarization system: Centrifuser  88  
8.2 Genomics summarization systems  90  
8.2.1 Sentence selection for protein–protein interactions  93  
8.2.2 EntrezGene SUMMARY field generation  94  

**CHAPTER 9**  
**Question-answering**  
9.1 Principles  95  
9.1.1 Question analysis and formal representation  95  
9.1.1.1 Clinical questions  95  
9.1.2 Formal representation of questions  96  
9.1.3 Domain model-based question representation  97  
9.1.3.1 Genomics and translational research questions  99  
9.1.4 Answer retrieval  100  
9.1.5 Answer extraction and generation  101  
9.1.5.1 Reference answer formats for clinical questions  101  
9.1.5.2 Entity-extraction approaches to answer generation  102  
9.2 Applications  103  
9.2.1 Question analysis and query formulation  104  
9.2.2 Knowledge Extraction  105  
9.2.2.1 Population Extractor  105  
9.2.2.2 Problem Extractor  106  
9.2.2.3 Intervention Extractor  106  
9.2.2.4 Outcome Extractor  107  
9.2.2.5 Clinical Task classification  108  
9.2.2.6 Strength of Evidence classification  111  
9.2.2.7 Document scoring and ranking  112  
9.2.3 Question–Document frame matching (PICO score)  113  
9.2.3.1 Answer generation  115  
9.2.4 Semantic clustering  115
List of figures

Figure 2.1 MEDLINE citation in MEDLINE format 15
Figure 3.1 Balanced F-scores of the submissions to the BioCreAtIvE Task 1A (gene mention finding evaluation) 27
Figure 4.1 Complex of Lys48-linked di-Ubiquitin with the UBA2 domain from hHR23A. (PDB code 1zo6) 32
Figure 4.2 Histogram Comparison of Co-Occurrence Measures. Histogram of the number of proteins assigned a given confidence value by the co-occurrence measures 39
Figure 4.3 Relations between proteins in the Drosophila Pelle system 40
Figure 4.4 Precision versus recall of the subcellular localization extraction in Craven & Kumlien (1999) 42
Figure 4.5 Sample MedLEE XML output 46
Figure 5.1 The number of indexed citations added to MEDLINE during each fiscal year since 1995 52
Figure 6.1 Data processing in GNAT 67
Figure 7.1 The concept “Lou Gehrig’s Disease” and some of its hypernyms 79
Figure 9.1 Overall architecture of the CQA-1.0 system 103
Figure 10.1 A partial list of the command-line options for SVMLight 122
CHAPTER 1

Introduction to natural language processing

As the intended audience for this book is natural language processing specialists who want to move into the biomedical domain, this chapter provides only an overview of the field. Computational biologists who would like a more solid background in the fundamentals of natural language processing and computational linguistics than is provided in this chapter should consult Jurafsky & Martin (2008). However, the chapter provides background on some of the basic issues, and contains information on natural language processing fundamentals that are specific to the biomedical domain that should prove helpful to experienced natural language processing specialists, as well.

1.1 Some definitions

Natural language processing, text mining, and computational linguistics are frequently used as synonyms. However, from a research perspective, it is useful to be aware of differences in the kinds of questions that are asked and the kinds of answers that are produced by each field.

1.1.1 Computational linguistics

Computational linguistics per se deals with developing computationally testable models of the human language faculty. Questions in computational linguistics research typically deal with some aspect of that faculty, such as the minimum necessary grammatical power for describing natural language syntax. Their answer is typically formal.

1.1.2 Natural language processing

Natural language processing is a subfield of linguistics and computer science that deals with computer applications whose input is natural language. Natural language processing research typically deals with building applications that process language at some given level of linguistic structure. For example, a natural language
processing application might be built to tag the parts of speech of words. The answer to a “question” in natural language processing research is typically a value for some performance metric.

1.1.3 Text mining

Text mining is the production of applications that perform specific tasks to meet information needs or provide services. A text mining application might use a natural language processing application for labelling parts of speech to extract statements about protein–protein interactions from scientific journal articles.

1.1.4 Usage of these definitions in practice

In practice, these terms are often used interchangeably, and there is significant cross-talk between the three research communities. For example, the annual meeting of the Association for Computational Linguistics is full of papers on natural language processing, and in the biomedical field, the terms natural language processing and text mining are typically used synonymously.

1.2 Levels of document and linguistic structure and their relationship to natural language processing

1.2.1 Document structure

Most research in natural language processing deals with newswire text, which has a relatively simple document structure. Document structure is a more complicated issue when dealing with biomedical text. The two main genres of interest in biomedical natural language processing are scientific journal articles and clinical documents. Each presents its own set of challenges.

Scientific journal articles are typically separated into an abstract and an article body. To date, the vast majority of research on journal articles has dealt with abstracts. However, in recent years, there has been a move towards more work on full-text articles, i.e. including the article bodies. This has introduced a new set of challenges. The structure and content of abstracts and journal articles are demonstrably different (Cohen et al. 2010a). Structurally, article bodies contain longer sentences than abstracts, and make much heavier use of parenthesized material. This parenthesized material presents a number of problems for natural language
processing applications (Cohen, Christiansen, & Hunter 2011). On a linguistic level, there are statistically significant differences in the incidence of passivization, negation, and pronominal anaphora (words like *it*) between abstracts and article bodies, with article bodies showing greater use of passivization and negation and abstracts showing higher use of pronominal anaphora. The density of semantic classes varies between the two, as well; for example, mutations are mentioned far more frequently in article bodies than in abstracts, while drugs and diseases are mentioned somewhat more frequently in abstracts than in article bodies. These differences between abstracts and article bodies correspond to differences in the performance of extant natural language processing tools. For example, part of speech taggers perform statistically significantly better on abstracts than on article bodies, and gene mention recognizers perform worse on article bodies – sometimes drastically worse.

Article bodies are typically separated into sections, with a common set of sections being an introduction, a materials and methods section, a results section, and a discussion section. Materials and methods sections are notorious sources of false positives for many types of applications, and they are often omitted from processing in system evaluations, but the ability to handle them is crucial for extracting the information on methods and on biological context that biologists find crucial in interpreting experiments and the output of text mining applications.

Although article bodies present a number of additional challenges as compared to abstracts, there is abundant evidence that the ability to process them will be crucial to reaching the full potential of biomedical text mining, and they are increasingly the subject of biomedical natural language processing research (Garten & Altman 2009; Lin 2009; Agarwal & Yu 2009; Czarnecki et al. 2012).

Clinical documents are typically structured, but there is an enormous amount of variability in this structure between different types of documents (e.g. discharge summaries versus nursing notes), of which there is a myriad. Document structures for the same type of document, e.g. discharge summaries, may vary from hospital to hospital, from department to department within the same hospital, and even from physician to physician within the same department. The ability to segment clinical documents into sections and to label the sections appropriately is crucial. For example, in determining a patient's current problems, it is crucial to differentiate between the history section and the diagnosis section. A small number of tools for segmenting clinical documents exists, but much more research is required in this area, and clinical natural language processing practitioners will often find the construction of an ad hoc document segmenter to be an unavoidable first step in building an application. Demner-Fushman *et al.* (2011) describes an iterative procedure for building such a segmenter for a range of clinical document types.
1.2.2 Sentences

After segmenting the document, sentence segmentation, i.e. finding the boundaries between sentences, is frequently the next step. Doing this for newswire text is approximately as difficult as a homework problem. However, biomedical text presents additional difficulties. For example, sentences in biomedical journal articles can begin with a lower-case letter if they begin with the name of a gene and a mutant form of the gene is being discussed, for example

\[ \text{eya-clones de-repress hth}. \]

In clinical documents, the situation is even more difficult – in clinical text, it may not be apparent what constitutes a sentence at all. For example, x-ray reports, one of the earliest subjects of medical natural language processing research, are replete with text like

\[ \text{Clear lungs. No evidence of focal pneumonia.} \]

This text clearly contains two separate textual segments. Each makes a separate assertion. However, neither is typical of an English sentence, both lacking verbs, for instance. Clinical text often also features lists consisting of items that are semantically and syntactically discrete but lack any punctuation, e.g. “problem lists” and lists of current medications:

- Active Problem List
- Erythema multiforma SJ syndrome
- Acute renal failure
- Transaminitis
- Chronic pain secondary to oral lesions
- Right arm DVT

Chronic Problem List
hx strep infection

1.2.3 Tokens

Tokens are individual elements in the text. These include discrete words, but also punctuation marks, which in most cases must be separated from the words to which they are attached. Tokenization may also include removal of affixes like \( n't \), genitive markers, and the like, depending on the design of the application. In biomedical text, as in other textual genres, tokenization can be problematic. For example, in journal articles, it is generally the case that hyphenated words are kept as a single unit. However, there are cases where the correct semantic interpretation
requires splitting them. For example, in a journal article, \textit{freac1-freac7} has the interpretation \textit{freac1, freac2, freac3, etc.}, through \textit{freac7}, and the sequence \textit{freac1-freac7} should probably be split into the three tokens \textit{freac1 – freac7}. The sequence \textit{IL3-5} has the interpretation \textit{IL3 and IL5}, and again, should probably be split into three tokens. Clinical text presents its own set of tokenization challenges and associated difficulties of semantic interpretation. For example, –\textit{fever} has the interpretation \textit{no fever present}, and the hyphen and word should probably be separated into two tokens.

\subsection*{1.2.4 Stems and lemmata}

For some applications it is useful to reduce words to their “stems.” \textit{Stem} is a term with no linguistic definition; it is purely an artifact of natural language processing. It is a normalized form of a word without inflectional and sometimes without derivational morphemes that is produced by some natural language processing application, which might not necessarily correspond to any form in the actual language. For example, for the inputs \textit{phosphorylate, phosphorylates, phosphorylated, phosphorylating, and phosphorylation}, a commonly used stemmer produces the form \textit{phosphoryl}. Note that this is not an actual English word and does not even accord with our intuitions about what the base of the words might be, i.e. \textit{phosphorylat}. However, stemmers can be useful as features for machine learning applications and as components of rules for rule-based systems.

In contrast to a stem, a lemma has a clear linguistic definition – actually, two. According to one definition, the lemma of a word is the set of inflectional forms of that word. For example, the lemma of \textit{translocate} would be \textit{translocate, translocates, translocated, translocating}. On the other definition, a lemma is the base form of a set of inflected words, such that the lemma of \textit{translocate, translocates, translocated, translocating} would be \textit{translocate}. Lemmatization is a more difficult task than stemming, and is much less often attempted. A tool specialized for biomedical journal articles is the BioLemmatizer (Liu \textit{et al.} 2013).

\subsection*{1.2.5 Part of speech}

\textit{Lexical categories}, more commonly known as \textit{parts of speech}, are categories that determine the inflectional morphemes (e.g. plurals for nouns, past tense for verbs, etc.) that a word can take and the slots that it can fill in grammatical structures (e.g., an adjective can appear between \textit{the} and a noun). Natural language processing systems typically assume about eighty parts of speech. We get from the eight parts of speech of traditional English grammar to the eighty parts of speech of natural
language processing by subdividing the traditional ones. For example, singular nouns and plural nouns are considered to be two separate parts of speech. Part of speech taggers require retraining to work well on biomedical text, whether it is journal articles or clinical text.

1.2.6 Syntactic structure

Syntactic parsers generally require retraining to work well on biomedical journal articles. Nonetheless, the syntactic structures of biomedical journal articles are probably similar to the structures of scientific publications in general, which have been studied by Biber et al. (1999). A number of approaches to simplifying the parsing problem in biomedical text have been taken. A common one is to “detokenize” some semantic or syntactic class of entity, e.g. by finding gene names like breast cancer associated 1 and turning them into a single token breast_cancer-associated_1. Another common approach is to delete parenthesized material, although Cohen, Christiansen, & Hunter (2011) demonstrated that this is sometimes informationally lossy.

1.2.7 Semantics

Whereas lexical semantics is still understudied in biomedical text, semantics is widely used as synonym of domain knowledge. The biomedical domain is one of the few domains with very rich machine readable sources of domain knowledge captured in databases and ontologies, discussed in detail in Chapter 7.
CHAPTER 2

Historical background

This chapter will provide a brief history of the field, from early medical work in the 1960s through genomic work in the late 2000s. A brief review of available resources (lexical resources, tools, etc.) will cover those that are so ubiquitous (e.g. PubMed/MEDLINE, Entrez Gene, MeSH, the Gene Ontology, and GENIA) that it is difficult to have an extended discussion of BioNLP without being familiar with them. Ethical and legal issues that are specific to the biomedical domain will be covered, including HIPAA regulations (which govern privacy and access to medical records, among other things).

2.1 Early work in the medical domain

The earliest work on biomedical natural language processing was done with inputs from the medical domain. The early work on extraction and synthesis of useful information from text in clinical records and biomedical publications was motivated by increasing concerns over the quality and rising costs of medical care, and the need to keep up with biomedical research. The need for keeping up with a professional field is traditionally satisfied through library services. It is not surprising, therefore, that with the increasing volume and availability of information in the 1960s came an understanding that “the full text of the entire library ought to be read and searched for each request” and that this task could be accomplished only by a computerized system (Swanson 1960). Similarly, computer-based systems were needed to facilitate improvements in medical care services, epidemiological and clinical research, and planning of medical care resources. Parties involved in these services need information extracted from medical records (for the most part kept as free text) and relevant valid findings of medical research extracted from the literature (US Congress 1977).

Subsequently, the observation that some logical relations between entities could be implied through piecing together facts scattered in different sets of publications (Swanson 1986b) brought about text mining in its strict sense of discovery of new facts and hypothesis generation. Text-based knowledge discovery relies on text mining in its broader sense and uses extraction of overtly stated pertinent information to infer relationships between biomedical entities and processes.