Information Retrieval in Biomedicine:

Natural Language Processing for Knowledge Integration

Violaine Prince
University Montpellier 2, France & LIRMM-CNRS, France

Mathieu Roche
University Montpellier 2, France & LIRMM-CNRS, France

Medical Information Science
REFERENCE

MEDICAL INFORMATION SCIENCE REFERENCE
Hershey · New York

Director of Editorial Content:	Kristin Klinger
Senior Managing Editor:	Jamie Snavely
Managing Editor:	Jeff Ash
Assistant Managing Editor:	Carole Coulson
Typesetter:	Amanda Appicello
Cover Design:	Lisa Tosheff
Printed at:	Yurchak Printing Inc.

Published in the United States of America by
Information Science Reference (an imprint of IGI Global)
701 E. Chocolate Avenue
Hershey PA 17033
Tel: 717-533-8845
Fax: 717-533-8661
E-mail: cust@igi-global.com
Web site: http://www.igi-global.com/reference

and in the United Kingdom by
Information Science Reference (an imprint of IGI Global)
3 Henrietta Street
Covent Garden
London WC2E 8LU
Tel: 44 20 7240 0856
Fax: 44 20 7379 0609
Web site: http://www.eurospanbookstore.com

Library of Congress Cataloging-in-Publication Data

Information retrieval in biomedicine : natural language processing for
knowledge integration / Violaine Prince and Mathieu Roche, editors.
p. cm.
Includes bibliographical references and index.
Summary: "This book provides relevant theoretical frameworks and the latest
empirical research findings in biomedicine information retrieval as it
pertains to linguistic granularity"--Provided by publisher.
ISBN 978-1-60566-274-9 (hardcover) -- ISBN 978-1-60566-275-6 (ebook) 1.
Medical informatics. 2. Biotechnology--Data processing. 3. Information
storage and retrieval systems--Medical care. I. Prince, Violaine, 1958- II.
Roche, Mathieu, 1975-
R858.I536 2009
651.5'04261--dc22

2008043763

British Cataloguing in Publication Data
A Cataloguing in Publication record for this book is available from the British Library.

All work contributed to this book is new, previously-unpublished material. The views expressed in this book are those of the authors, but not necessarily of the publisher.

Chapter XI
Analyzing the Text of Clinical Literature for Question Answering

Yun Niu
Ontario Cancer Institute, Canada

Graeme Hirst
University of Toronto, Canada

ABSTRACT

The task of question answering (QA) is to find an accurate and precise answer to a natural language question in some predefined text. Most existing QA systems handle fact-based questions that usually take named entities as the answers. In this chapter, the authors take clinical QA as an example to deal with more complex information needs. They propose an approach using Semantic class analysis as the organizing principle to answer clinical questions. They investigate three Semantic classes that correspond to roles in the commonly accepted PICO format of describing clinical scenarios. The three Semantic classes are: the description of the patient (or the problem), the intervention used to treat the problem, and the clinical outcome. The authors focus on automatic analysis of two important properties of the Semantic classes.

INTRODUCTION

The vast increase in online information brings new challenges to the area of information retrieval (IR) in both query processing and answer processing. To free the user from constructing a complicated boolean keywords query, a system should be able to process queries represented in natural language. Instead of responding with some documents relevant to the query, the system should actually answer the questions accurately and concisely. Systems with such characteristics are *question-answering* (QA) systems, which take advantage of high-quality natural language processing and mature technologies in IR. The task of a QA system is to find the answer to a particular natural language question in some predefined text. In this paper, we propose an approach that aims to automatically find answers to clinical questions.

Clinicians often need to consult literature on the latest information in patient care, such as side effects of a medication, symptoms of a disease, or time constraints in the use of a medication. The published medical literature is an important source to help clinicians make decisions in patient treatment (Sackett & Straus, 1998; Straus & Sackett, 1999). For example:

- **Q:** In a patient with a suspected MI does thrombolysis decrease the risk of death if it is administered 10 hours after the onset of chest pain?

An answer to the question can be found in *Clinical Evidence* (CE) (Barton, 2002), a regularly updated publication that reviews and consolidates experimental results for clinical problems:

- **A:** Systematic reviews of RCTs have found that prompt thrombolytic treatment (within 6 hours and perhaps up to 12 hours and longer after the onset of symptoms) reduces mortality in people with AMI and ST elevation or bundle branch block on their presenting ECG.

Studies have shown that searching the literature can help clinicians answer questions regarding patient treatment (Cimino, 1996; Gorman, Ash, & Wykoff, 1994; Mendonça, Cimino, Johnson, & Seol, 2001). It has also been found that if high-quality evidence is available in this way at the point of care—e.g., the patient's bedside —clinicians will use it in their decision making, and it frequently results in additional or changed decisions (Sackett & Straus, 1998; Straus & Sackett, 1999). The practice of using the current best evidence to help clinicians in making decisions on the treatment of individual patients is called *evidence-based medicine* (EBM).

Clinical questions usually represent complex information needs and cannot be answered using a single word or phrase. For a clinical question, it is often the case that more than one clinical trial with different experimental settings will have been performed. Results of each trial provide some evidence on the problem. To answer such a question, all this evidence needs to be taken into account, as there may be duplicate evidence,

Figure 1. Example of a clinical question, with corresponding evidence from Clinical Evidence

Clinical question: Are calcium channel blockers effective in reducing mortality in acute myocardial infarction patients?

Evidence 1: … calcium channel blockers do not reduce mortality, may increase mortality.

Evidence 2: … verapamil versus placebo … had no significant effect on mortality.

Evidence 3: … diltiazem significantly increased death or reinfarction.

Evidence 4: … investigating the use of calcium channel blockers found a non-significant increase in mortality of about 4% and 6%.

partially agreed-on evidence, or even contradictions. A complete answer can be obtained only by synthesizing these multiple pieces of evidence, as shown in Figure 1. In our work, we take EBM as an example to investigate clinical QA. Our targets are questions posed by physicians in patient treatment.

Our task is to find answers to clinical questions automatically. Our work is part of the EPoCare project ("Evidence at Point of Care") at the University of Toronto. The goal of EPoCare is to develop methods for answering clinical questions automatically with CE as the source text. (We do not look at primary medical research text.)

BACKGROUND

Many advances have been made to answer *factoid* questions that have named entities such as a person or location as answers; this is factoid question answering (FQA). For example, the answer to the question *Who was the U.S. president in 1999?* is *Bill Clinton*. However, much less has been understood in finding answers to complex questions that demand synthesis of information, such as clinical QA, which is *non-factoid* QA (NFQA).

We observe two distinct characteristics that differentiate *factoid* QA and *non-factoid* QA.

- Non-factoid questions usually cannot be answered using a word or phrase, such as named entities. Instead, answers to these questions are much more complex, and often consist of multiple pieces of information from multiple sources.
- Compared to *factoid* QA, in which an answer can be judged as *true* or *false*, *non-factoid* QA needs to determine what information is *relevant* in answer construction.

Non-factoid QA is attracting more and more research interest (Diekema, Yilmazel, Chen, Harwell, He, & Liddy, 2004; Niu, Hirst, McArthur, & Rodriguez-Gianolli, 2003; Stoyanov, Cardie, & Wiebe, 2005; DUC, 2005). Unlike FQA, in which the main research focuses on *wh-* questions (e.g. *when*, *where*, *who*) in a rather general domain, most work in NFQA starts with a specific domain, such as terrorism, or a specific type of question, such as opinion-related questions. The complexity of NFQA tasks may account for this difference. In this section, current work in NFQA is reviewed according to different research problems in the QA task that it addresses.

Because the information needs are more complex, some work puts more effort into understanding questions. Hickl et al. (2004), Small et al. (2004) and Diekema et al. (2004) suggest answering questions in an interactive way to clarify questions step by step. In addition, Hickl et al. argue that decomposition of complex scenarios into simple questions is necessary in an interactive system.

Following that work, Harabagiu et al. (2004) derived the intentional structure and the implicatures enabled by it for decomposing complex questions, such as *What kind of assistance has North Korea received from the USSR/Russia for its missile program?* The authors claim that intentions that the user associates with the question may express a set of *intended questions*; and each intended question may be expressed as *implied questions*. The intended questions of this example include *What is the USSR/Russia? What is assistance? What are the missiles in the North Korean inventory?* Then, these intended questions further have implied questions, such as *Is this the Soviet/Russian government? Does it include private firms, state-owned firms, educational institutions, and individuals? Is it the training of personnel? What was the development timeline of the missiles?*

The system HITIQA (High-Quality Interactive Question Answering) (Small, Strzalkowski, Liu, Ryan, Salkin, Shimizu, et al., 2004) also emphasizes interaction with the user to understand their information needs, although it does not attempt to decompose questions. During the interaction, the system asks questions to confirm the user's needs. After receiving *yes* or *no* from the user, the goal of searching is clearer. The interaction is data-driven in that questions asked by the system are motivated by the previous results of information searching (which form the answer space).

Diekema et al. (2004) also suggest having a question-negotiation process for complex. Their QA system deals with real-time questions related to reusable launch vehicles. For example, broad-coverage questions like *How does the shuttle fly?*, and questions about comparison of two elements such as *What advantages/disadvantages does an aluminum alloy have over Ti alloy as the core for a honeycomb design?* are typical in the domain. A question-answering system architecture with a module for question negotiation between the system and the questioner is proposed in the paper.

Berger et al. (2000) describe several interesting models to find the connection between question terms and answer terms. Soricut and Brill (2006) extend Berger's work to answer FAQ-like questions. In their work, although FAQ question and answer pairs are used as training data, the goal is to extract answers from documents on the Web, instead of pairing up existing questions and answers in FAQ corpora. Taking questions and answers as two different languages, a machine translation model is applied in the answer extraction module to extract three sentences that maximize the probability $p(q|a)$ (where q is the question and a is the answer) from the retrieved documents as the answer.

In the system HITIQA, frame structure is used to represent the text, where each frame has some attributes. For example, a general frame has *frame type*, *topic*, and *organization*. During the processing, frames will be instantiated by corresponding named entities in the text. In answer generation, text in the answer space is scored by comparing their frame structures with the corresponding goal structures generated by the system according to the question. Answers consist of text passages from which the zero conflict frames are derived. The correctness of the answers was not evaluated directly. Instead, the system was evaluated by how effective it is in helping users to achieve their information goal. The results of a three-day evaluation workshop validated the overall approach.

Cardie et al. (2004) aim to answer questions about opinions (multi-perspective QA), such as: *Was the most recent presidential election in Zimbabwe regarded as a fair election?* and *What was the world-wide reaction to the 2001 annual U.S. report on human rights?*. They developed an annotation scheme for low-level representation of opinions, and then proposed using opinion-oriented scenario templates to act as a summary representation of the opinions. Possible ways of using the representations in multi-perspective QA are discussed. In related work, Stoyanov, Cardie, and Wiebe (2005) analyzed characteristics of opinion questions and answers and showed that traditional FBQA techniques are not sufficient for multi-perspective QA. Results of some initial experiments showed that using filters that identify subjective sentences is helpful in multi-perspective QA.

The typical work discussed here shows the state-of-the-art in NFQA. Most systems are investigating complex questions in specific domains or of particular types. Although interesting views and approaches have been proposed, most work is at the initial stage, describing the general framework or potential useful approaches to address characteristics of non-factoid QA.

Our work on NFQA is in the medical domain. Clinical QA as an NFQA task presents challenges similar to those of the tasks described in this section. Our work is to investigate these challenges by addressing a key issue: *what information is relevant?* We do not attempt to elicit such information by deriving additional questions, such as performing question decomposition (Hickl et al., 2004) or through interactive QA (Small et al, 2004). Instead, we aim to identify the best information available in a designated source to construct the answer to a given question. To achieve these goals, we propose to use Semantic class analysis in non-factoid QA and use frame structure to represent Semantic classes.

OUR APPROACH FOR CLINICAL QA: SEMANTIC CLASS ANALYSIS

As discussed in the introduction, answers to clinical questions are not named entities and often consist of multiple pieces of information. In response to these major characteristics, we propose frame-based Semantic class analysis as the organizing principle to answer these questions.

Representing Scenarios Using Frames

Clinical questions often describe scenarios. For example, they may describe relationships between clinical problems, treatments, and corresponding clinical outcomes, or they may be about symptoms, hypothesized diseases, and diagnosis processes. To answer these questions, essentially, we need an effective schema to understand scenario descriptions.

Semantic Roles

The Semantics of a scenario or an event are expressed by the Semantic relationships between its participants, and such Semantic relationships are defined by the role that each participant plays in the scenario. These relationships are referred to as *Semantic roles* (Gildea & Jurafsky, 2002), or *conceptual roles* (Riloff, 1999). This viewpoint dates back to frame Semantics, posed by Fillmore (1976) as part of the nature of language. Frame Semantics provides a schematic representation of events or scenarios that have various participants as roles. In our work, we use frames as our representation schema for the Semantic roles involved in questions and answer sources.

Research on Semantic roles has proposed different sets of roles ranging from the very general to the very specific. The most general role set consists of only two roles: PROTO-AGENT and PROTO-PATIENT (Dowty, 1991; Valin & Robert, 1993). Roles can be more domain-specific, such as perpetrators, victims, and physical targets in a terrorism domain. In question-answering tasks, specific Semantic roles can be more instructive in searching for relevant information, and thus more precise in pinpointing correct answers. Therefore, we take domain-specific roles as our targets.

The Treatment Frame

Patient-specific questions in EBM usually can be described by the so-called **PICO** format (Sackett, Straus, Richardson, Rosenberg, & Haynes, 2000). In a *treatment scenario*, **P** refers to the *status of the patient (or the problem)*, **I** means an *intervention*, **C** is a *comparison intervention* (if relevant), and **O** describes the *clinical outcome*. For example, in the following question:

- **Q:** In a patient with a suspected myocardial infarction does thrombolysis decrease the risk of death?

the description of the patient is *patient with a suspected myocardial infarction*, the intervention is *thrombolysis*, there is no comparison intervention, and the clinical outcome is *decrease the risk of death*. Originally, PICO format was developed for therapy questions describing treatment scenarios, and was later extended to other types of clinical questions such as diagnosis, prognosis, and etiology. Representing clinical questions with PICO format is widely believed to be the key to efficiently finding high-quality evidence (Ebell, 1999; Richardson, Wilson, Nishikawa, & Hayward, 1995). Empirical studies have shown that identifying PICO elements in clinical scenarios

Table 1. The treatment frame

P:	a description of the patient (or the problem)
I:	an intervention
O:	the clinical outcome

improves the conceptual clarity of clinical problems (Cheng, 2004).

We found that PICO format highlights several important Semantic roles in clinical scenarios, and can be easily represented using the frame structure. Therefore, we constructed a frame based on it. Since **C** mainly indicates a comparison relation to **I**, we combined the comparisons as one filler of the same slot *intervention* in the frame, connected by a specific relation. We focused on therapy-related questions and built a *treatment frame* that contains three slots, as shown in Table 1.

A slot in a frame designates a *Semantic class* (corresponding to a *Semantic role* or a *conceptual role*), and relations between Semantic classes in a scenario are implied by the design of the frame structure. The treatment frame expresses a cause-effect relation: the *intervention* for the *problem* results in the *clinical outcome*.

When applying this frame to a sentence, we extract constituents in the sentence to fill in the slots in the frame. These constituents are *instances of Semantic classes*. In this paper, the terms *instances of Semantic classes* and *slot fillers* are used interchangeably. Some examples of the instantiated treatment frame are as follows.

- **Sentence**: One RCT [randomized clinical trial] found no evidence that low molecular weight heparin is superior to aspirin alone for the treatment of acute ischaemic stroke in people with atrial fibrillation.
  P: acute ischaemic stroke in people with atrial fibrillation
  I: low molecular weight heparin vs. aspirin
  O: no evidence that low molecular weight heparin is superior to aspirin
- **Sentence**: Subgroup analysis in people with congestive heart failure found that diltiazem significantly increased death or reinfarction.
  P: people with congestive heart failure
  I: diltiazem
  O: significantly increased death or reinfarction
- **Sentence**: Thrombolysis reduces the risk of dependency, but increases the risk of death.
  P: —
  I: thrombolysis
  O: reduces the risk of dependency, but increases the risk of death

The first example states the result of a clinical trial, while the second and third depict clinical outcomes. We do not distinguish the two cases in this study, and treat them in the same manner.

Relationship to Information Extraction

Our approach of Semantic class analysis has a close relation to *information extraction* (IE), in which domain-specific Semantic roles are often explored to identify predefined types of information from text (Riloff, 1999). Our approach shares the view with IE that Semantic classes/roles are the keys to understanding scenario descriptions. Frames are also used in IE as the representation scheme. Nevertheless, in our work, as shown by the above examples of treatment frames, the syntactic constituents of an instance of a Semantic class can be much more complex than those of traditional IE tasks, in which slot fillers are usually named entities (Riloff, 1999; TREC, 2001). Therefore, approaches based on such Semantic classes go beyond named-entity identification, and thus will better adapt to clinical QA. In addition, extracting instances of Semantic classes from text is not the ultimate goal of QA. Frame representation of Semantic classes provides a platform for matching questions to answers in our QA system.

Main Components of a QA System Guided by Semantic Class Analysis

With Semantic class analysis as the organizing principle, we identify four main components of our QA system:

- Detecting Semantic classes in questions and in answer sources
- Identifying properties of Semantic classes
- Question-answer matching: exploring properties of Semantic classes to find relevant pieces of information
- Constructing answers by merging or synthesizing relevant information using relations between Semantic classes

To search for the answer to a question, the question and the text in which the answer may occur will be processed to detect the Semantic classes. A Semantic class can have various properties. These properties can be extremely valuable in finding answers, which we will discuss in detail in the following sections. In the matching process, the question scenario will be compared to an answer candidate, and pieces of relevant information should be identified by exploring properties of the Semantic classes. To construct the answer, relevant information that has been found in the matching process will be merged or synthesized to generate an accurate and concise answer. The process of synthesizing scenarios relies on comparing instances of Semantic classes in these scenarios; for example, two instances might be exactly the same or one might be the hypernym of the other.

In the following sections we will discuss our approaches to automatically detecting two properties of the Semantic classes in the treatment scenario: the cores of the classes and the polarities of clinical outcomes.

CORES OF SEMANTIC CLASSES

In a frame structure, the slots in question and answer frames can be filled with either *complete* or *partial* information. Consider the following example, where parentheses delimit each instance of a Semantic class (a slot filler) and the labels **P** (problem description), **I** (an intervention), **O** (the clinical outcome) indicate the type of the instance:

- **Sentence**: Two systematic reviews in (people with AMI)$_{\mathbf{P}}$ investigating the use of (calcium channel blockers)$_{\mathbf{I}}$ found a (non-significant increase in mortality of about 4% and 6%)$_{\mathbf{O}}$.

  Complete slot fillers:
  P: people with AMI
  I: calcium channel blockers
  O: a non-significant increase in mortality of about 4% and 6%
  Partial slot fillers:
  P: AMI
  I: calcium channel blockers
  O: mortality

The partial slot fillers in this example contain the smallest fragments of the corresponding complete slot fillers that exhibit information rich enough for deriving a reasonably precise answer. We use the term *core* to refer to such a fraction of a slot filler (instance of a Semantic class).

Importance of Cores

As mentioned in the introduction, before the matching process, keyword-based document retrieval is usually performed to find relevant documents that may contain the answer to a given question. Keywords in the retrieval are derived from the question. Cores of Semantic classes can be extremely valuable in searching for such documents for complex question scenarios, as shown in the following example. (The scenario is an example used in usability testing in the EPoCare project at the University of Toronto.)

- **Question scenario**: A physician sees a 7-year-old child with asthma in her office. She is on Flovent and Ventolin currently and was recently discharged from hospital following her fourth admission for asthma exacerbation. During the most recent admission, the dose of Flovent was increased. Her mother is concerned about the impact of the additional dose of steroids on her daughter's growth. This is the question to which the physician wants to find the answer.

For a complex scenario description like this, the answer could be missed or drowned in irrelevant documents found by inappropriate keywords derived from the question. However, the search can be much more effective if we have the information of cores of Semantic classes, for example, *P: asthma, I: steroids, O: growth*.

Similarly, Semantics presented in cores can help filter out irrelevant information that cannot be identified by searching methods based on simple string overlaps.

1. In patients with **myocardial infarction**, do **β blockers** reduce all cause **mortality** and **recurrent myocardial infarction** without adverse effects?
2. In someone with **hypertension** and **high cholesterol**, what management options will decrease his risk of **stroke** and **cardiac events**?

In question 1, the first occurrence of *myocardial infarction* is a disease but the second occurrence is part of the clinical outcome. In question 2, *stroke* is part of the clinical outcome rather than a disease to be treated, as it usually is. Obviously, string matching cannot distinguish between the two cases. By identifying and classifying cores of Semantic classes, the relations between these important Semantic units in the scenarios are very clear. Therefore, documents or passages that do not contain *myocardial infarction* or *stroke* as clinical outcomes can be discarded.

In addition, identifying cores of Semantic classes in documents can facilitate the question-answer matching process. Some evidence relevant to the above question scenario on *asthma* is listed below, where boldface indicates a core:

- **Evidence 1**: A more recent systematic review (search date 1999) found three RCTs comparing the effects of **becolmetasone** and **non-steroidal medication** on linear **growth** in children with **asthma** (200μg twice daily, duration up to maximum 54 weeks) suggesting a short-term decrease in linear **growth** of –1.54 cm a year.
- **Evidence 2**: Two systematic reviews of studies with long-term follow up and a subsequent long-term RCT have found no evidence of **growth retardation** in **asthmatic children** treated with inhaled **steroids**.

The evidence sentences here are from CE (Barton, 2002). The clinical outcomes mentioned in the evidence have very different phrasings — yet both pieces of the evidence are relevant to the question. The pieces of evidence describe two distinct outcomes — that short-term decrease in growth is found and that there is no effect on growth in some long-term studies. Missing either of the outcomes will lead to an incomplete answer for the physician. Here, cores of the Semantic classes provide the only clue that both pieces of evidence are relevant to this question and should be included in the answer. Hence, a complete description of Semantic classes does not have to be found. In fact, such a description with more information could make the matching harder to find because of the different expressions of the outcomes.

Finally, cores of Semantic classes in a scenario are connected to each other by the relations embedded in the frame structure. The frame of the treatment scenario contains a cause–effect relation: an intervention used to treat a problem results in a clinical outcome.

In this section, we propose a method to automatically identify and classify the cores of Semantic classes according to their context in a sentence. We take the treatment frame as an example, in which the goal is to identify cores of *interventions*, *problems*, and *clinical outcomes*. For ease of description, we will use the terms *intervention-core*, *disease-core*, and *outcome-core* to refer to the corresponding cores. We work at the sentence level, i.e., we identify cores in a sentence rather than a clause or paragraph. Two principles are followed in developing the method. First, complete slot fillers do not have to be extracted before core identification. Second, we aim to reduce the need for expensive manual annotation of training data by using a semi-supervised learning approach.

Architecture of the Method

In our approach, we first collect candidates of the target cores from sentences under consideration. For each candidate, we classify it as one of the four classes: *intervention-core*, *disease-core*, *outcome-core*, or *other*. In the classification, a candidate will get a class label according to its context, its UMLS Semantic types, and the syntactic relations in which it participates. Figure 2 shows the architecture of the approach.

Preprocessing

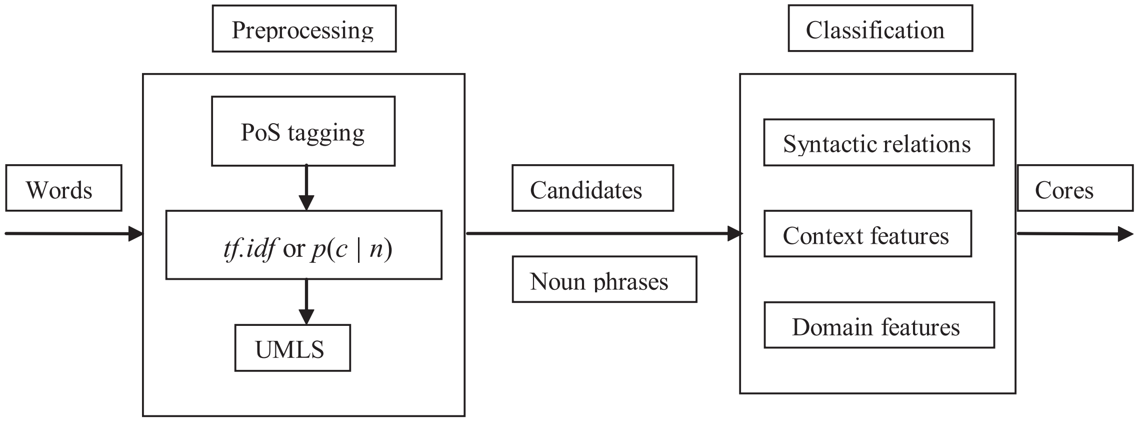
In the preprocessing, all words in the data set are examined. The first two steps are to reduce noise, in which some of the words that are unlikely to be part of real cores are filtered out. Then, the rest are mapped to their corresponding concepts, and these concepts are candidate target cores.

PoS tagging Our observation is that cores of the three types of slot fillers are usually nouns or noun phrases. Therefore, words that are not nouns are first removed from the candidate set. PoS tags are obtained by using Brill's tagger (Brill, 1993).

Filtering out some *bad* nouns This step is the second attempt to remove noise from the candidate set. Nouns that are unlikely to be part of real cores are considered as *bad* candidates. Two research options of measures are used to evaluate how *good* a noun is.

- Extended *tf.idf*. After the *tf.idf* value is calculated for a noun in each document, the highest value of all the documents is taken as the final score of the noun. Nouns with scores lower than a threshold are removed from the candidate set. The threshold was set manually after observing the scores of some nouns that frequently occur in the text. CE text is used to get the score of a noun. For this, 47 sections in CE are segmented to 143 files of about the same size. Each file is treated as a document. This measure is referred to as *tf.idf* in later description.
- Domain specificity. We calculate the conditional probability $p(\mathrm{c}|n) = p(c,n) / p(n)$, where c is the medical class, and n is a noun. It is the probability that a document is in the medical domain c given it contains the noun n. Intuitively, *intervention-cores*, *disease-cores*, and *outcome-cores* are domain-specific, i.e., a document that contains them is very likely to be in the medical domain. For example, *morbidity, mortality,*

Figure 2. Architecture of the approach of core identification



> *aspirin*, and *myocardial infarction* are very likely to occur in a medicine-related context. This measure intends to keep highly medical domain-specific nouns in the candidate set. A noun is a better candidate if the corresponding probability is high. Text from two domains is needed in this measure: medical text, and non-medical text. In our experiment, we used the same 47 sections in *CE* as the medical class text (separated into 143 files of about the same size). For the non-medical class, we used 1000 randomly selected documents from the Reuters-21578 newswire text collection, because newswire stories are mainly in the general domain. Nouns whose probability values are below a threshold (determined in the same manner as in the *tf.idf* measure) are filtered out.

Mapping to concepts To this point, the candidate set consists of nouns. In many cases, nouns are part of noun phrases (concepts) that are better candidates of cores. For example, the phrase *myocardial infarction* is a better candidate of a disease-core than the noun *infarction*. Therefore, the software MetaMap (Aronson, 2001) is used to map a noun to its corresponding concept in the Metathesaurus. All the concepts form the candidates of cores to be classified.

Representing Candidates Using Features

We expect that candidates in the same Semantic class will have similar behavior. Therefore, the idea of the classification is to group together similar candidates. The similarity is characterized by syntactic relations, context information, and Semantic types in UMLS. All features are binary features, i.e., a feature takes value 1 if it is present; otherwise, it takes value 0.

Syntactic Relations

Previous researchers have explored syntactic relations to group similar words (Lin, 1998) and words of the same sense in word sense disambiguation (Kohomban & Lee, 2005). Lin (1998) inferred that *tesguino* is similar to *beer*, *wine*, etc., i.e., it is a kind of drink, by comparing syntactic relations in which each word participates. Kohomban and Lee (2005) determined the sense of a word in a context by observing a subset of all syntactic relations in the corpus that the word participates in. The hypothesis is that different instances of the same sense will have similar relations.

In our work, we need to group cores of the same Semantic class. Such cores may participate in similar syntactic relations while those of different classes will have different relations. For example, intervention-cores often are subjects of sentences, while outcome-cores are often objects.

Candidates in our task are phrases, instead of words as for Lin and Kohomban and Lee. Thus, we extend their approaches of analyzing relations between two words to extract relations between a word and a phrase. This is done by considering all relations between a candidate noun phrase and other words in the sentence. To do that, we ignore relations between any two words in the phrase when extracting syntactic relations. Any relation between a word not in the phrase and a word in the phrase is extracted. We use the Minipar parser (Lin, 1994) to get the syntactic relations between words. After a sentence is parsed, we extract relevant syntactic relations from the output of the parser. A relation is represented using a triple that contains two words (one of them is in the noun phrase and the other is not) and the grammatical relation between them. Figure 3 shows relevant triples extracted from a sentence. Because long-distance relations are considered, the relation between *thrombolysis* and *increases* is captured.

In the feature construction, a triple is taken as a feature. The set of all distinct triples is the syntactic relation feature set in the classification.

Local Context

Context of candidates is also important in distinguishing different classes. For example, a disease-core may often have *people with* in its left context. However, it is very unlikely that the phrase *people with mortality* will occur in the text.

We considered the two words on each side of a candidate (stop-words were excluded). When extracting context features, all punctuation marks were removed except the sentence boundary. The window did not cross boundaries of sentences. We evaluated two representations of context: with and without order. In the ordered case, local context to the left of the phrase is marked by *-L*, and *R-* marks that to the right. The symbols *-L* and *R-* are used only to indicate the order of text. For the candidate *dependency* in Figure 3, the context features with order are: *reduces-L*, *risk-L*, *R-increases*, and *R-chance*. The context features without order are: *reduces*, *risk*, *increases*, and *chance*.

Figure 3. Example of dependency triples extracted from output of Minipar parser

Sentence:

Thrombolysis reduces the risk of dependency, but increases the chance of death.

Candidates:

thrombolysis, dependency, death

Relations:

(thrombolysis subj-of increase), (thrombolysis subj-of reduce)

(dependency pcomp-n-of of)

(death pcomp-n-of of)

This example shows a case where ordered context helps distinguish an intervention-core from an outcome-core. If order is not considered, candidates *thrombolysis* and *dependency* have overlapped context: *reduces* and *risk*. When taking order into account, they have no overlapped features at all — *thrombolysis* has features *R-reduces* and *R-risk*, while *dependency* has features *reduces-L* and *risk-L*.

Domain Features

As described in the *mapping to concepts* step in the preprocessing, at the same time of mapping text to concepts in UMLS, MetaMap also finds their Semantic types. Each candidate has a Semantic type defined in the Semantic Network of UMLS. For example, the Semantic type of *death* is **organism function**, that of *disability* is **pathologic function**, and that of *dependency* is **physical disability**. These Semantic types are used as features in the classification.

Data Set

Two sections of *CE* were used in the experiments. A clinician labeled the text for intervention-cores and disease-cores. Complete clinical outcomes are also identified. Using the annotation as a basis, outcome-cores were labeled by the author. The number of instances of each class is shown in Table 2.

The Model of Classification

Because our classification strategy is to group together similar cores and the cluster structure of the data is observed, we chose a semi-supervised learning model developed by Zhu, Ghahramani, and Lafferty (2003) that explores the cluster structure of data in classification. The general hypothesis of this approach is that similar data points will have similar labels.

A graph is constructed in this model. In the graph, nodes correspond to both labeled and unlabeled data points (candidates of cores), and an edge between two nodes is weighted according to the similarity of the nodes. More formally, let (x_1,y_1), …, (x_l,y_l) be labeled data, where $Y_L = \{y_1, \ldots, y_l\}$ are corresponding class labels. Similarly, let (x_{l+1},y_{l+1}), …, (x_{l+u},y_{l+u}) be unlabeled data, where $Y_U = \{y_{l+1}, \ldots, y_{l+u}\}$ are labels to be predicted. A connected graph $G = (V,E)$ can be constructed, where the set of nodes V correspond to both labeled and unlabeled data points and E is the set of edges. The edge between two nodes i, j is weighted. Weights w_{ij} are assigned to agree with the hypothesis; for example, using a radial basis function (RBF) kernel, we can assign larger edge weights to closer points in Euclidean space.

Zhu, Ghahramani, and Lafferty formulated the intuitive label propagation approach as a problem of energy minimization in the framework of Gaussian random fields, where the Gaussian field is over a continuous state space instead of over a discrete label set. The idea is to compute a *real-valued* function $f\colon V \rightarrow R$ on graph G that minimizes the energy function $E(f) = ½\sum_{ij} w_{ij} (f(i) - f(j))^2$, where i and j correspond to data points in the problem. The function $f = \operatorname{argmin}_f E(f)$ determines the labels of unlabeled data points. This solution can be efficiently computed by direct matrix calculation even for multi-label classification, in which solutions are generally computationally expensive in other frameworks.

Table 2. Number of instances of cores in the whole data set

Intervention-core	Disease-core	Outcome-core	Total
501	153	384	1038

This approach propagates labels from labeled data points to unlabeled data points according to the similarity on the edges, thus it follows closely the cluster structure of the data in prediction. We expect it to perform reasonably well on our data set. It is referred to as *SEMI* in the following description. We use SemiL (Huang, Kecman, & Kopriva, 2006), an implementation of the algorithm using Gaussian random fields in the experiment (default values are used for the parameters unless otherwise mentioned).

Results and Analysis

We first evaluate the performance of the semi-supervised model on different feature sets. Then, we compare the two candidate sets obtained by using *tf.idf* and domain specificity, respectively. Finally, we compare the semi-supervised model to a supervised approach to justify the usage of a semi-supervised approach in the problem.

In all experiments, the data set contains all candidates of cores. Unless otherwise mentioned, the result reported is achieved by using the candidate set derived by $p(c|n)$, the feature set of the combination of syntactic relations, ordered context, and Semantic types, and the distance measure of cosine distance. The result of an experiment is the average of 20 runs. In each run, labeled data is randomly selected from the candidate set, and the rest is unlabeled data whose labels need to be predicted. We make sure all classes are present in labeled data. If any class is absent, we redo the sampling. The evaluation of the Semantic classes is very strict: a candidate is given credit if it gets the same label as given by the annotator and the tokens it contains are exactly the same as marked by the annotator. Candidates that contain only some of the tokens matching the labels given by the annotators are treated as the *other* class in the evaluation.

Experiment 1: Evaluation of Feature Sets

This experiment evaluates different feature sets in the classification. As described above, two options are used in the second step of preprocessing to pick up *good* candidates. Here, as our focus is on the feature set, we report only results on candidates selected by $p(c|n)$. The number of instances of each of the four target classes in the candidate set is shown in Table 3 (the performance of candidate selection will be discussed below).

Figure 4 shows the accuracy of classification using different combinations of four feature sets: syntactic relations, ordered context, un-ordered context, and Semantic types.

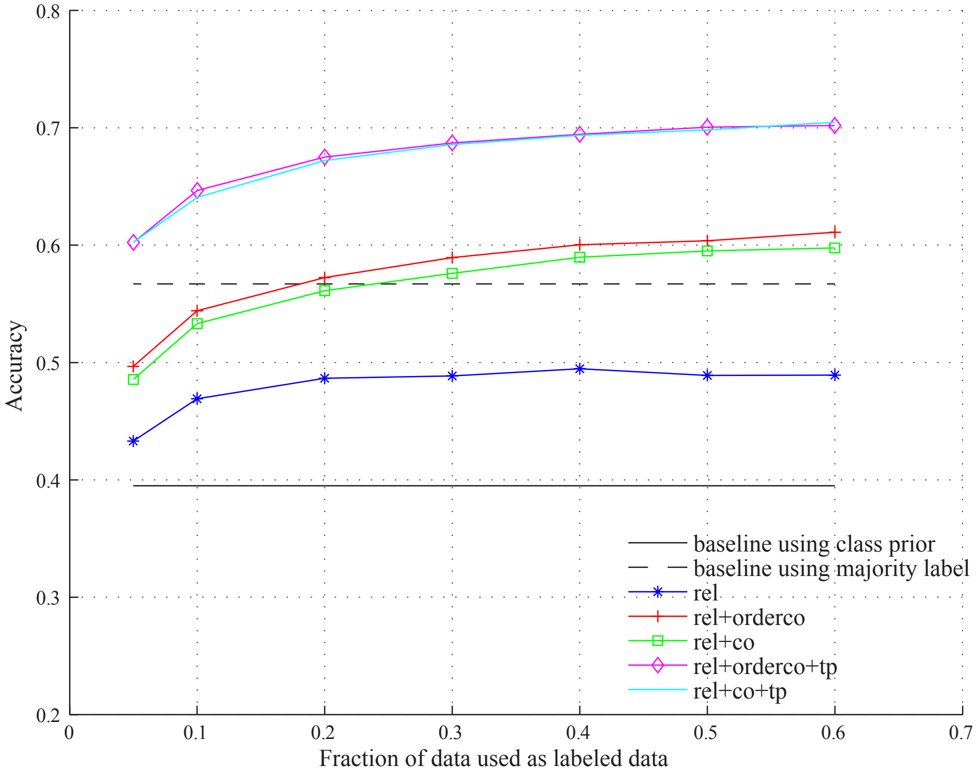
We set a baseline by assigning labels to data points according to the prior knowledge of the distribution of the four classes, which has accuracy of 0.395. Another choice of baseline is to assign the label of the majority class, *others* in this case, to each data point, which produces an accuracy of 0.567. However, all the three classes of interest have accuracy 0 according to this baseline. Thus, this baseline is not very informative in this experiment.

It is clear in the figure that incorporating new kinds of features into the classification resulted in a large improvement in accuracy. Using only syntactic relations (*rel* in the figure) as features, the best accuracy is a little lower than 0.5, which is much higher than the baseline of 0.395. The addition of ordered context (*orderco*) or no-order context features (*co*) improved the accuracy by about 0.1. Adding Semantic type features (*tp*) further improved 0.1 in accuracy. Combining all three kinds of features achieved the best performance. With only 5% of data as labeled data, the whole feature set achieved an accuracy of 0.6, which is much higher than the baseline of 0.395. Semantic type seems to be a very powerful feature set, as it substantially improves the performance on top of the combination of the

Table 3. Number of instances of target classes in the candidate set

Intervention-core	Disease-core	Outcome-core	Others	Total
298	106	209	801	1414

Figure 4. Classification results of candidates



other two kinds of features. Therefore, we took a closer look at the Semantic type feature set by conducting the classification using only Semantic types, and found that the result is even worse than using only syntactic relations. This observation reveals interesting relations among the feature sets. In the space defined by only one kind of features, data points may be close to each other, hence hard to distinguish. Adding another kind sets apart data points in different classes toward a more separated position in the new space. It shows that every kind of feature is informative to the task. The feature sets characterize the candidates from different angles that are complementary in the task. We also see that there is almost no difference between ordered and unordered context in distinguishing the target classes, although ordered context seems to be slightly better when Semantic types are not considered.

Experiment 2: Evaluation of Candidate Sets

In the second step of preprocessing, one of two options can be used to filter out some *bad* nouns

— using the *tf.idf* measure or the domain specificity measure $p(c \mid n)$. This experiment compares the two measures in the core identification task. A third option using neither of the two measures (i.e., skip the second step of preprocessing) is evaluated as the baseline. The first three rows in Table 4 are numbers of instances remaining in the candidate set after preprocessing. The last row shows the numbers of manually annotated true cores, which has been listed in Table 2 and is repeated here for comparison.

tf.idf and Domain Specificity vs. Baseline

As shown in Table 4, there are many fewer instances in the *others* class in the sets derived by *tf.idf* and the probability measure as compared to those derived by the baseline, which shows that the two measures effectively removed some of the *bad* candidates of intervention-core, disease-core, and outcome-core. At the same time, a small number of real cores were removed. Compared to the baseline method, the probability measure kept almost the same number of intervention-cores and disease-cores in the candidate set, while omitting some outcome-cores. It indicates that outcome-cores are less domain-specific than intervention-cores and disease-cores. Compared to the *tf.idf* measure, more intervention-cores and outcome-cores were kept by the conditional probability measure, showing that the probability measuring the domain-specificity of a noun better characterizes the cores of the three Semantic classes. The probability measure is also more robust than the *tf.idf* measure, as *tf.idf* relies more on the content of the text from which it is calculated. For example, if an intervention is mentioned in many documents of the document set, its *tf.idf* value can be very low although it is a good candidate for being an intervention-core.

The precision, recall, and *F*-score of the classification shown in Table 5 confirm the above analysis. The domain specificity measure gets substantially higher *F*-scores than the baseline for all the three classes that we are interested in, using different amounts of labeled data. Compared to *tf.idf*, the performance of the domain specificity measure is much better on identifying intervention-cores (note that $p(c|n)$ picked up more real intervention-cores than *tf.idf*), and slightly better on identifying outcome-cores, while the two are similar on identifying disease-cores.

Baseline vs. the Set of Manually Annotated Cores

As mentioned at the beginning of this subsection, the baseline candidate set was derived by the first (PoS tagging) and third step (mapping from nouns to concepts) in the preprocessing. As shown by Table 4, 62.3% of manually annotated cores are kept in the baseline. We roughly checked about one-third of the total true cores (manually annotated cores) in the data set and found that 80% of lost cores are because MetaMap either extracted more or fewer tokens than marked by the annotator, or it failed to find the concepts. 10% of missing cores are caused by errors of the PoS tagger, and the rest are because some cores are not nouns.

Experiment 3: Comparison of the Semi-Supervised Model and SVMs

In the semi-supervised model, labels propagate along high-density data trails, and settle down at low-density gaps. If the data has this desired structure, unlabeled data can be used to help learning. In contrast, a supervised approach makes use only of labeled data. This experiment compares SEMI to a state-of-the-art supervised approach; the goal is to investigate how well unlabeled data contributes to the classification using the semi-supervised model. We compare the performance of SEMI to support-vector machines (SVMs) when different amounts of data are used as labeled data.

Table 4. Number of candidates in different candidate sets

Measures	Intervention-core	Disease-core	Outcome-core	Others
tf.idf	243	108	194	785
$p(c\|n)$	298	106	209	801
baseline	303	108	236	1330
true cores	501	153	384	—

Table 5. Results of classification on different candidate sets

INT: intervention-core; DIS: disease-core; OUT: outcome-core.

Fraction of		1%			5%			10%			30%			60%		
labeled data		P	R	F	P	R	F	P	R	F	P	R	F	P	R	F
INT	baseline	.44	.69	.53	.51	.83	.63	.53	.87	.66	.58	.90	.70	.59	.92	.72
	tf.idf	.44	.62	.51	.52	.74	.61	.55	.77	.64	.59	.84	.69	.60	.87	.71
	$p(c\|n)$	.51.	.65	**.57**	.60	.83	**.69**	.62	.86	**72**	.65	.90	**.75**	.67	.91	**.77**
DIS	Baseline	.16	.63	.25	.25	.68	.36	.31	.73	.43	.34	.84	.48	.35	.86	.49
	tf.idf	.20	.55	**.29**	.31	.64	**.41**	.34	.70	.46	.39	.82	**.53**	.41	.86	**.55**
	$p(c\|n)$	.18	.56	.27	.30	.66	**.41**	.34	.73	**.47**	.39	.83	**.53**	.41	.87	**.55**
OUT	Baseline	.22	.42	.28	.33	.53	.41	.39	.61	.48	.44	.66	.53	.46	.69	.55
	tf.idf	.30	.43	.35	.43	.56	**.49**	.47	.61	.53	.53	.66	.59	.55	.70	.61
	$p(c\|n)$	.31	.46	**.37**	.43	.56	**.49**	.48	.62	**.54**	.54	.69	**.60**	.56	.71	**.63**

Support Vector Machines

In SVMs, the process of classification given a set of training examples is an optimization procedure searching for the optimal rule that predicts the label of unseen examples with minimum errors. The goal of SVMs is to find an optimal hyperplane so that examples on the same side of the hyperplane will have the same label. The classification task is then to determine on which side of the hyperplane a data point lies. The optimal hyperplane that SVMs chose is the one with the largest margin.

In this experiment, we use OSU SVM (Ma, Zhao, Ahalt & Eads, 2003), a toolbox for Matlab built on top of LIBSVM (Chang & Lin, 2001). LIBSVM is an implementation of SVMs. We use RBF as the kernel method, and set the Sigma value heuristically using labeled data. SVM addresses the problem of unbalanced data using a parameter, which assigns weights to each class in the task. A class with larger weight will get a greater penalty when finding the optimum hyperplane. We set the parameter according to the prior knowledge of the class distribution and give larger weight to a class that contains fewer instances. Default values are used for other parameters.

Comparison of SEMI to SVMs

As shown in Table 6, when there is only a small amount of labeled data (less than 5% of the whole data set), which is often the case in real-world applications, SEMI achieves much better performance than SVMs in identifying all the three classes. For intervention-cores and outcome-cores, with 5% data as labeled data, SEMI outperforms SVMs with 10% data as labeled data. With less than

Table 6. F-score of classification using different models

Candidate set: produced by p(c|n) (see Table 4)
INT: intervention-core; DIS: disease-core; OUT: outcome-core.

Fraction of	labeled data	1%	5%	10%	30%	60%
INT	SEMI	.57	.69	.72	.75	.77
	SVM	.33	.60	.68	.74	.77
DIS	SEMI	.27	.41	.47	.53	.55
	SVM	.33	.60	.68	.74	.77
OUT	SEMI	.37	.49	.54	.60	.63
	SVM	.07	.27	.44	.56	.62

60% data as labeled data, the performance of SEMI is either superior to or comparable to SVMs for intervention-cores and outcome-cores. This shows that SEMI effectively exploits unlabeled data by following the manifold structure of the data. The promising results achieved by SEMI show the potential of exploring unlabeled data in classification.

Related Work

The task of named entity (NE) identification, similar to the core-detection task, involves identifying words or word sequences in several classes, such as proper names (locations, persons, and organizations), monetary expressions, dates, and times. NE identification has been an important research topic ever since it was defined in Message Understanding Conference (MUC, 2003). In 2003, it was taken as the shared-task in the Conference on Computational Natural Language Learning (Erik, Sang & Meulder, 2003). Most statistical approaches use supervised methods to address the problem (Chieu & Ng, 2003; Florian, Ittycheriah, Jing & Zhang, 2003; Klein, Smarr, Nguyen, & Manning, 2003). Unsupervised approaches have also been tried in this task. Cucerzan and Yarowsky (1999) use a bootstrapping algorithm to learn contextual and morphological patterns iteratively. Collins and Singer (1999) tested the performance of several unsupervised algorithms on the problem: modified bootstrapping (DL-Co-Train) motivated by co-training (Blum & Mitchell, 1998), an extended boosting algorithm (CoBoost), and the Expectation Maximization (EM) algorithm. The results showed that DL-CoTrain and CoBoost perform about the same, and both are superior to EM.

Much effort in entity extraction in the biomedical domain has gene names as the target. Various supervised models including naive Bayes, support vector machines, and hidden Markov models have been applied (Ananiadou & Tsujii, 2003). The work most related to our core-identification in the biomedical domain is that of Rosario and Hearst (2004), which extracts *treatment* and *disease* from MEDLINE and examines seven relation types between them using generative models and a neural network. They claim that these models may be useful when only partially labeled data is available, although only supervised learning is conducted in the paper. The best *F*-score of identifying *treatment* and *disease* obtained by using the supervised method is .71. Another piece of work extracting similar Semantic classes is by Ray and Craven (2001), who report an *F*-score of about .32 for extracting *proteins* and *locations* and about .50 for *gene* and *disorder*.

A difficulty of using this approach, however, is in detecting boundaries of the targets. A segmentation step that pre-processes the text is needed. This will be our future work, in which we aim to investigate approaches that perform the segmentation precisely.

As a final point, we want to emphasize the difference between cores and named entities. While the identification of NEs in a text is an important component of many tasks including question answering and information extraction, its benefits are constrained by its coverage. Typically, it is limited to a relatively small set of classes, such as *person*, *time*, and *location*. However, in sophisticated applications, such as the non-factoid medical question answering that we consider, NEs are only a small fraction of the important Semantic units discussed in documents or asked about by users. As shown by the examples in this section, cores of clinical outcomes are often not NEs. In fact, many Semantic roles in scenarios and events that occur in questions and documents do not contain NEs at all. For example, the *test method* in *diagnosis* scenarios, the *means* in a *shipping* event, and the *manner* in a *criticize* scenario may all have non-NE cores. Therefore, it is imperative to identify other kinds of Semantic units besides NEs. Cores of Semantic classes are one such extension that consist of a more diverse set of Semantic units that goes beyond simple NEs.

POLARITY OF CLINICAL OUTCOMES

One of the major concerns in patient treatment is the clinical outcomes of interventions in treating diseases: are they positive, negative or neutral? This polarity information is an inherent property of clinical outcomes. An example of each type of polarity taken from CE is shown below.

- **Positive**: Thrombolysis reduced the risk of death or dependency at the end of the studies.
- **Negative**: In the systematic review, thrombolysis increased fatal intracranial haemorrhage compared with placebo.
- **Neutral**: The first RCT found that diclofenac plus misoprostol versus placebo for 25 weeks produced no significant difference in cognitive function or global status.

Sentences that do not have information on clinical outcomes form another group: no outcome.

- **No outcome**: We found no RCTs comparing combined pharmacotherapy and psychotherapy with either treatment alone.

Polarity information is crucial to answering questions related to clinical outcomes. We have to know the polarity to answer questions about benefits and harms of an intervention. In addition, knowing whether a sentence contains a clinical outcome can help filter out irrelevant information in answer construction. Furthermore, information on negative outcomes can be crucial in clinical decision making. In this section, we discuss the problem of automatically identifying outcome polarity in medical text (Niu, et al., 2005). More specifically, we focus on detecting the presence of a clinical outcome in medical text, and, when an outcome is found, determining whether it is positive, negative, or neutral. We observe that a single sentence in medical text usually describes a complete clinical outcome. As a result, we perform sentence-level analysis in our work.

Related Work

The problem of polarity analysis is also considered as a task of sentiment classification (Pang, Lee & Vaithyanathan, 2002; Pang & Lee, 2003) or Semantic orientation analysis (Turney, 2002): determining whether an evaluative text, such as a movie review, expresses a "favorable" or "unfavorable" opinion. All these tasks are to obtain the orientation of the observed text on a discussion topic. They fall into three categories: detection of the polarity of words, of sentences, and of documents. Among them, as Yu and Hatzivassiloglou (2003) pointed out, the problem at the sentence level is the hardest one.

Turney (2002) has employed an unsupervised learning method to provide suggestions on documents as *thumbs up* or *thumbs down*. The polarity detection is done by averaging the Semantic orientation (SO) of extracted phrases (phrases containing adjectives or adverbs) from a text. The document is tagged as *thumbs up* if the average of SO is positive, and otherwise is tagged as *thumbs down*. In more recent work, Whitelaw, Garg, and Argamon (2005) explore *appraisal groups* to classify positive and negative documents. Similar to phrases used in Turney's work, *appraisal groups* consist of coherent words that together express the polarity of opinions, such as "extremely boring", or "not really very good". Instead of calculating the mutual information, a lexicon of *adjectival appraisal groups* (groups headed by an appraising adjective) is constructed semi-automatically. These groups are used as features in a supervised approach using SVMs to detect the sentiment of a document. Pang et al. (2002) also deal with the task at document level. The sentiment classification problem was treated as a text classification issue and a variety of machine-learning techniques were explored to classify movie reviews into positive and negative. A series of lexical features were employed on these classification strategies in order to find effective features. Pang et al. found that support vector machines perform the best among three classification strategies. The main part of Yu and Hatzivassiloglou's work (2003) is at the sentence level, and hence is closest to our work. They first separate facts from opinions using a Bayesian classifier, then use an unsupervised method to classify opinions as positive, negative, and neutral by evaluating the strength of the orientation of words contained in a sentence.

The polarity information we are observing relates to clinical outcomes instead of the personal opinions studied by the work mentioned above. Therefore, we expect differences in the expressions and the structures of sentences in these two areas. For the task in the medical domain, it will be interesting to see whether domain knowledge will help. These differences lead to new features in our approach.

A Supervised Approach for Clinical Outcome Detection and Polarity Classification

Since SVMs have been shown also very effective in many other classification tasks, we investigate SVMs in sentence-level analysis to detect the presence of a clinical outcome and determine its polarity.

In our approach, each sentence as a data point to be classified is represented by a vector of features. In the feature set, we use words themselves as they are very informative in related tasks such as sentiment classification and topic categorization. In addition, we use contextual information to capture changes described in clinical outcomes, and use generalized features that represent groups of concepts to build more regular patterns for classification.

We use binary features in most of the experiments except for the *frequency* feature in one of our experiments. When a feature is present in a sentence, it has a value of 1; otherwise, it has a value of 0. Among the features in our feature set, UNIGRAMS and BIGRAMS have been used in previous sentiment classification tasks, and the rest are new features that we developed.

Unigrams

A sentence is composed of words. Distinct words (unigrams) can be used as the features of a sentence. In previous work on sentiment classification (Pang et al., 2002; Yu & Hatzivassiloglou, 2003) unigrams are very effective. Following this work, we also take unigrams as features. We use unigrams occurring more than 3 times in the data set in the feature set, and they are called UNIGRAMS in the following description.

Context Features

Our observation is that outcomes often express a change in a clinical value (Niu and Hirst, 2004). In the following example, *mortality* was *reduced*.

- In these three postinfarction trials ACE inhibitor versus placebo significantly *reduced mortality, readmission for heart failure, and reinfarction.*

The polarity of an outcome is often determined by how a change happens: if a **bad** thing (e.g., mortality) was **reduced**, then it is a positive outcome; if a **bad** thing was **increased**, then the outcome is negative; if there is no change, then we get a neutral outcome. We tried to capture this observation by adding context features – BIGRAMS, two types of CHANGE PHRASES (MORE/LESS features and POLARITY-CHANGE features), and NEGATIONS.

Bigrams

Bigrams (two adjacent words) are also used in sentiment classification. In that task, they are not so effective as UNIGRAMS. When combined with UNIGRAMS, they do not improve the classification accuracy (Pang et al., 2002; Yu & Hatzivassiloglou, 2003). However, in our task, the context of a word in a sentence that describes the change in a clinical value is important in determining the polarity of a clinical outcome. Bigrams express the patterns of pairs, and we expect that they will capture some of the changes. Therefore, they are used in our feature set. As with UNIGRAMS, bigrams with frequency greater than 3 are extracted and referred to as BIGRAMS.

Change Phrases

We developed two types of new features to capture the trend of changes in clinical values. The collective name CHANGE PHRASES is used to refer to these features. To construct these features, we manually collected four groups of words by observing several sections in CE: those indicating **more** (*enhanced, higher, exceed*, ...), those indicating **less** (*reduce, decline, fall,* ...), those indicating **good** (*benefit, improvement, advantage*, ...), and those indicating **bad** (*suffer, adverse, hazards*, ...).

- MORE/LESS **features**. This type of feature emphasizes the effect of words expressing "changes". The way the features are generated is similar to the way that Pang et al. (2002) add negation features. We attached the tag _MORE to all words between the **more**-words and the following punctuation mark, or between the **more**-words and another **more** (or **less**) word, depending on which one comes first. The tag _LESS was added similarly. This way, the effect of the "change" words is propagated.
  - The first systematic review found that ß blockers significantly reduced_LESS the_LESS risk_LESS of_LESS death_LESS and_LESS hospital_LESS admissions_LESS.
  - Another large RCT (random clinical trial) found milrinone versus placebo increased_MORE mortality_MORE over_MORE 6_MORE months_MORE.
- POLARITY-CHANGE **features.** This type of feature addresses the co-occurrence of **more/less** words and **good/bad** words, i.e., it detects whether a sentence expresses the idea of "change of polarity". We used four features for this purpose: MORE GOOD, MORE BAD, LESS GOOD, and LESS BAD. As this type of feature aims for the "changes" instead of "propagating the change effect", we used a

smaller window size to build these features. To extract the first feature, a window of four words on each side of a **more**-word in a sentence was observed. If a **good**-word occurs in this window, then the feature MORE GOOD was set to 1. The other three features were set in a similar way.

Negations

Most frequently, negation expressions contain the word *no* or *not*. We observed several sections of CE and found that *not* is usually used in a way that does not affect the polarity of a sentence, as shown in the following examples, so it is not included in the feature set:

- However, disagreement for uncommon but serious adverse safety outcomes has **not** been examined.
- The first RCT found fewer episodes of infection while taking antibiotics than while **not** taking antibiotics.
- The rates of adverse effects seemed higher with rivastigmine than with other anticholinesterase drugs, but direct comparisons have **not** been performed.

The case for *no* is different: it often suggests a neutral polarity or no clinical outcome at all:

- There are **no** short or long term clinical benefits from the administration of nebulised corticosteroids …
- One systematic review in people with Alzheimer's disease found **no** significant benefit with lecithin versus placebo.
- We found **no** systematic review or RCTs of rivastigmine in people with vascular dementia.

We develop the NEGATION features to take into account the evidence of the word *no*. To extract the features, all the sentences in the data set are first parsed by the Apple Pie parser (Sekine, 1997) to get phrase information. Then, in a sentence containing the word *no*, the noun phrase containing *no* is extracted. Every word in this noun phrase except *no* itself is attached by a _*NO* tag.

Semantic Types

Using category information to represent groups of medical concepts may relieve the data sparseness problem in the learning process. For example, we found that diseases are often mentioned in clinical outcomes as **bad** things:

- A combined end point of death or disabling stroke was significantly lower in the accelerated-t-PA group…

Thus, all names of specific diseases in the text are replaced with the tag DISEASE.

Intuitively, the occurrences of Semantic types, such as **pathologic function** and **organism function**, may be different in different polarity of outcomes, especially in the *no outcome* class as compared to the other three classes. To verify this intuition, we collect all the Semantic types in the data set and use each of them as a feature. They are referred to as SEMANTIC TYPES. Thus, in addition to the words contained in a sentence, all the medical categories mentioned in a sentence are also considered. The Unified Medical Language System (UMLS) is used as the domain knowledge base for extracting Semantic types of concepts. The software MetaMap (Aronson, 2001) is incorporated for mapping concepts to their corresponding Semantic types in the UMLS Metathesaurus.

Experiments

We carried out several experiments on two text sources: CE and Medline abstracts. Compared to CE text, Medline has a more diverse writing style as different abstracts have different authors. The performance of the supervised classification approach on the two sources was compared to find out if there is any difference. We believe that these experiments will lead to better understanding of the polarity detection task.

Outcome Detection and Polarity Classification in CE Text

Experimental Setup

The data set of sentences in all the four classes was built by collecting sentences from different sections in CE (sentences were selected so that the data set is relatively balanced). The number of instances in each class is shown in Table 7. The data set was labeled manually by three graduate students, and each sentence was labeled by one of them. We used the OSU SVM package (Ma et al., 2003) with an RBF kernel for this experiment. The σ value was set heuristically using training data. Default values were used for other parameters in the package.

Results and Analysis

Table 8 shows the results of the five feature sets used for classification. The accuracy is the average of 50 runs of the experiment. In each run, 20% of the data is selected randomly as the test set, and the rest is used as the training set. With just UNIGRAMS as features, we get 76.9% accuracy, which is taken as the baseline. The addition of BIGRAMS in the feature set results in an increase of about 2.5% in accuracy, which corresponds to 10.8% of relative error reduction. CHANGE PHRASES lead to a very small improvements and NEGATIONS do not improve the performance on top of BIGRAMS. Note that CHANGE PHRASES tend to capture the impact of context, and bigrams also contain context information. It could be that some effect of CHANGE PHRASES has already been captured by bigrams. Also, since the target classes are different in the two tasks, CHANGE PHRASES may be more important in distinguishing positive from negative outcomes. The SEMANTIC TYPES features further improve the performance on top of the combination of other features, which shows that generalization is helpful.

Which class is the most difficult to detect, and why? To answer these questions, we further examine the errors in every class. The precision, recall and *F*-score of each class are shown in Table 9 (it is the result of one run of the experiment). It is clear in the table that the negative class has the lowest precision and recall. A lot of errors occur in distinguishing negative from no-outcome classes. We studied the incorrectly classified sentences and found some interesting cases. Some of the errors occur because descriptions of diseases in the no-outcome class are often identified as negative. These sentences are difficult in that they contain negative expressions (e.g., *increased risk*), yet do not belong to the negative class:

- Lewy body dementia is an insidious impairment of executive functions with Parkinsonism, visual hallucinations, and fluctuating cognitive abilities and increased risk of falls or autonomic failure.

Negative samples are sometimes assigned a positive label when a sentence has phrasings that seem to contrast, as shown in the following example:

Table 7. Number of instances in each class (CE)

Positive	Negative	Neutral	No-outcome	Total
472	338	250	449	1509

- The mean increase in height in the budesonide group was 1.1 cm less than in the placebo group (22.7 vs 23.8 cm, P= 0005); …

In this sentence, the clinical outcome of impaired growth is expressed by comparing height increase in two groups, which is less explicit and hard to capture.

Outcome Detection and Polarity Classification in Medline

With Medline abstracts, we evaluate two tasks: the first one is two-way classification that aims to detect the presence of clinical outcomes. In this task, a sentence is classified into two classes: containing a clinical outcome or not. The second task is the four-way classification, i.e., identifying whether an outcome is positive, negative, neutral, or the sentence does not contain an outcome.

Experimental Setup

We collected 197 abstracts from Medline that were cited in CE. The number of sentences in each class is listed in Table 10. The data set was annotated with the four classes of polarity information by two graduate students. Each single sentence was annotated by one of them.

In this experiment, again, 20% of the data was randomly selected as test set and the rest was used as the training data. The averaged accuracy was obtained from 50 runs. We again used the OSU SVM package for this experiment; parameters were set in the same manner.

Results and Analysis

Results of the two tasks are shown in Table 11. Not surprisingly, the performance on the two-way classification is better than on the four-way task. For both tasks, we see a similar trend in accuracy as for CE text (see Table 8). The accuracy goes up as more features are added, and the complete feature set has the best performance. Compared to UNIGRAMS, the combination of all features significantly improves the performance in both tasks (paired *t*-test, *p* values <0.0001). With just UNIGRAMS as features, we get 80.1% accuracy for the two-way task. The addition of BIGRAMS in the feature set results in an increase of 1.6 percentage points in accuracy, which corresponds to 8.0% of relative error reduction as compared to UNIGRAMS. Similar improvements are observed in the four-way task. The SEMANTIC TYPES features also slightly reduce the error rate.

Compared to the results on CE text in Table 8, the four-way classification task tends to be more difficult on Medline text. This can be observed by comparing the improvement of adding all other features to UNIGRAMS. As we mentioned in section 5.3, Medline abstracts have a more diverse writing style because they are written by different authors. This could be a factor that makes the classification task more difficult. However, the general performance of features on Medline abstracts and

Table 8. Results of the four-way classification with different feature sets in CE

Features	Accuracy (%)	Relative Error Reduction (%) (to UNIGRAMS)
(1) UNIGRAMS	76.9	—
(1)+(2) BIGRAMS	79.4	10.8
(1)+(2)+(3) CHANGE PHRASES	79.6	11.7
(1)+(2)+(3)+(4) NEGATIONS	79.6	11.7
(1)+(2)+(3)+(4)+(5) SEMANTIC TYPES	80.6	16.0

Table 9. Classification results of each class on CE data

	Positive	Negative	Neutral	No Outcome
Precision (%)	86.8	73.1	79.2	76.8
Recall (%)	83.2	73.1	76.0	82.0
F-score (%)	85.0	73.1	77.6	79.3

Table 10. Number of instances in each class (Medline)

Positive	Negative	Neutral	No Outcome	Total
469	122	194	1513	2298

Table 11. Results of two-way and four-way classification with different feature sets (Medline)

RER=Relative Error Reduction (compared to unigrams)

	two-way		four-way	
Features	Accuracy (%)	RER (%)	Accuracy (%)	RER (%)
(1) UNIGRAMS	80.1	—	75.5	—
(1)+(2) BIGRAMS	81.7	8.0	77.4	7.8
(1)+(2)+(3) CHANGE PHRASES	82.0	9.5	77.6	8.6
(1)+(2)+(3)+(4) NEGATIONS	81.9	9.0	77.6	8.6
(1)+(2)+(3)+(4)+(5) SEMANTIC TYPES	82.5	12.1	78.3	11.4

CE text is similar, which shows that the feature set is relatively robust. In our outcome detection and polarity classification task, UNIGRAMS are very effective features, as has been previously shown in the context of sentiment classification problems. This shows that information in words is very important for the polarity detection task. Context information represented by BIGRAMS and CHANGE PHRASES is also valuable in our task (see Table 8 and Table 11). The effectiveness of BIGRAMS is different from the results obtained by Pang et al. (2002) and Yu & Hatzivassiloglou (2003). In their work, adding bigrams does not make any difference in the accuracy, or even is slightly harmful in some cases. This indicates the difference in the expression of polarity in clinical outcomes and the polarity in opinions. Generalization features (SEMANTIC TYPES in Table 8 and Table 11) are also helpful in our task.

Discussion

The Performance Bottleneck in Polarity Classification

As described in section 5.1, supervised approaches have been used in sentiment classification. Features used in these approaches usually include: *n*-grams, PoS tags, and features based on words with Semantic orientations (e.g., adjectives such as *good*, *bad*). In all such studies, a common observation is that unigrams are very effective, while adding more features does not gain much.

- In the task of detecting polarity of documents (Pang et al., 2002), the best performance is obtained using unigrams.
- In the sentence-level opinion/fact classification task (Yu & Hatzivassiloglou, 2003), various features based on Semantic orientation of words are tried, including counts of Semantically oriented words, the polarity of the head verbs and the average Semantic orientation score of the words in the sentence. A gold standard set is built which includes 400 sentences labeled by one judge. In the opinion class, the only result better than the performance of unigrams is obtained by combining all features, which results in only 0.01 improvement in precision. Similarly, not much is achieved by adding all other features in detecting facts.
- In the work of Whitelaw, Garg, and Argamon (2005), the best performance of the approach is achieved by the combination of unigrams with the appraisal groups, which is 3% higher in accuracy than using unigrams alone.

From all this work, we observe a *performance bottleneck* problem in the polarity classification task: various features have been developed; however, adding more features does not gain much in classification accuracy, and it may even hurt the performance. In our task, although the context and generalization features significantly improve the performance compared to unigrams, we observe a similar *performance bottleneck* problem.

Analysis of the Problem

The bottleneck problem shows that additional features have much overlap with unigram features, and they may add noise to the classification. We further analyzed the data, and found that most words in a sentence do not contribute to the classification task. Instead, they can be noise that cannot be removed by adding more features. This could be a crucial reason of the bottleneck discussed above.

To verify this hypothesis, we conducted some additional experiments on the Medline data set of 2298 sentences. From each sentence in the data set, we manually extracted some words that fully determine the polarity of the sentence. We refer to these words by *extractions* in the following description. For those sentences that do not contain outcomes, nothing is extracted. The following examples are some sentences with different polarity and the extractions from them. These extractions form another data set, which we call the *extraction set*.

- **Sentence**: Treatment with reperfusion therapies and achievement of TIMI 3 flow are associated with increased short- and medium-term survival after infarction.
  Extraction: Increased short- and medium-term survival
- **Sentence**: In all three studies, a significant decrease in linear growth occurred in children treated with beclomethasone compared to those receiving placebo or non-steroidal asthma therapy.
  Extraction: Decrease in linear growth occurred
- **Sentence**: The doxazosin arm, compared with the chlorthalidone arm, had a higher risk of stroke.
  Extraction: A higher risk of stroke
- **Sentence**: Prednisolone treatment had no effect on any of the outcome measures.
  Extraction: No effect
- **Sentence**: There was no significant mortality difference during days 0–35, either among all randomised patients or among the pre-specified subset presenting within 0–6 h of pain onset and with ST elevation on the electrocardiogram in whom fibrinolytic treatment may have most to offer.
  Extraction: No significant mortality difference

We performed the four-way classification task on this extraction set. We constructed UNIGRAM features based on the extraction set and used them in the classification. Using 80% of the data as the training data and the rest as the test data, we achieved an accuracy of 93.3%, which is much higher than the accuracy of the four-way classification task on the original sentence set (75.5%).

The fact that we do not extract any words from no-outcome sentences may make the task easier. Therefore, we removed from the extraction set all sentences that do not contain an outcome, and reran the experiment. This task has three target classes: positive, negative or neutral. We obtained an accuracy of 82.2%. However, performing the three-way classification on the original sentence set only achieves 70.7% accuracy.

The results clearly show that irrelevant words actually introduce a lot of noise in the polarity detection task. Therefore, a new direction of research on the task is to conduct feature selection to remove words that do not contribute to the classification.

We took a closer look at the extraction set and found that the extractions usually form a sequence or several sequences in a sentence. Because hidden Markov models and conditional random fields are effective models for sequence detection, they will be explored in the future work of this research.

Summary

In this section, we discussed a supervised approach to identifying an inherent property of clinical outcomes — their polarity. Polarity information is important to answer questions related to clinical outcomes. We analyzed this problem from various aspects:

- We developed features to represent context information and explored domain knowledge to get generalized features. The results show that adding these features significantly improves the classification accuracy.
- We showed that the feature set has consistent performance on two different text sources, CE and Medline abstracts.
- We compared outcome polarity detection to sentiment classification according to different performance of context features on the two tasks. We found that although bigram features have almost no effect on the sentiment classification task, they improve the classification accuracy of identifying presence and polarity of clinical outcomes.
- We identified a *performance bottleneck* problem in the polarity classification task using a supervised approach. In both the sentiment classification and the outcome polarity detection, we observed that adding more features on top of the unigram features does not lead to major improvement in accuracy. We found a crucial reason for this — the noise in the feature set is not removed by adding more features. We proposed to use hidden Markov models or conditional random fields to conduct feature selection and thus to remove noise from the feature set.

CONCLUSION

Clinical question-answering is a complex task in which multiple pieces of information are often needed to construct a complete answer. We have proposed a novel approach guided by Semantic class analysis to deal with the complicated information needs. This approach consists of four major components:

- Detecting Semantic classes in questions and answer sources
- Identifying properties of Semantic classes
- Question-answer matching: exploring prop-

erties of Semantic classes to find relevant pieces of information
- Constructing answers by merging or synthesizing relevant information using relations between Semantic classes

We focused on three Semantic classes that correspond to roles in the commonly accepted PICO format of describing clinical scenarios. The three classes are: the problem of the patient, the intervention used to treat the problem, and the clinical outcome. In this paper, we have described our approach to automatically identifying two important properties of the three Semantic classes.

ACKNOWLEDGMENT

This project is supported by grants from Bell University Laboratories at the University of Toronto and by a grant from the Natural Sciences and Engineering Research Council of Canada.

REFERENCES

Ananiadou, S., & Tsujii, J. (Eds.) (2003). *Proceedings of the ACL 2003 Workshop on Natural Language Processing in Biomedicine*. Stroudsburg, PA, USA: Association for Computational Linguistics.

Aronson, A. R. (2001). Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. In S. Bakken (Ed.), *American Medical Informatics Association Symposium* (pp. 17–21). Bethesda, MD, USA: American Medical Informatics Association.

Barton, S. (Ed.). (2002). *Clinical Evidence.* London, England: BMJ Publishing Group.

Berger, A., Caruana, R., Cohn, D., Freitag, D., & Mittal, V. (2000). Bridging the lexical chasm: statistical approaches to answer-finding. In N. J. Belkin, P. Ingwersen, & M. Leong (Eds.), *23rd International Conference on Research and Development in Information Retrieval* (pp. 192–199). New York, NY, USA: Association for Computing Machinery Press.

Blum, A., & Mitchell, T. (1998). Combining labeled and unlabeled data with co-training. In *Proceedings of the 11th Annual Conference on Computational Learning Theory* (pp. 92–100). New York, NY, USA: Association for Computing Machinery Press.

Brill, E. (1993). A c*orpus-based approach to language learning.* Unpublished doctoral dissertation, University of Pennsylvania, Philadelphia, PA.

Cardie, C., Wiebe, J., Wilson, T., & Litman, D. (2004). Combining low-level and summary representations of opinions for multi-perspective question answering. In L. Greenwald, Z. Dodds, A. Howard, S. Tejada, & J. Weinberg (Eds.), *AAAI Spring Symposium: New Directions in Question Answering* (pp. 20–27). Menlo Park, CA, USA: AAAI Press.

Chang, C. C., & Lin, C. J. (2001). *LIBSVM — A library for support vector machines*. http://www.csie.ntu.edu.tw/~cjlin/libsvm/.

Cheng, G. Y. (2004). A study of clinical questions posed by hospital clinicians. *Journal of the Medical Library Association*, *93*(4), 445–458.

Chieu, H. L., & Ng, H. T. (2003). Named entity recognition with a maximum entropy approach. In W. Daelemans & M. Osborne (Eds.), 7*th Conference on Computational Natural Language Learning* (pp. 160–163). Stroudsburg, PA, USA: Association for Computational Linguistics.

Cimino, J. J. (1996). Linking patient information systems to bibliographic resources. *Methods of Information in Medicine*, *35*(2), 122–126.

Collins, M., & Singer, M. (1999). Unsupervised models for named entity classification. In *Proceedings of the 1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora* (pp. 189–196). Stroudsburg, PA, USA: Association for Computational Linguistics.

Cucerzan, S., & Yarowsky, D. (1999). Language independent named entity recognition combining morphological and contextual evidence. In *Proceedings of the 1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora* (pp. 90–99). Stroudsburg, PA, USA: Association for Computational Linguistics.

Diekema, A., Yilmazel, O., Chen, J., Harwell, S., He, L., & Liddy, E. D. (2004). What do you mean? Finding answers to complex questions. In L. Greenwald, Z. Dodds, A. Howard, S. Tejada, and J. Weinberg (Eds.), *AAAI Spring Symposium: New Directions in Question Answering* (pp. 87–93). Menlo Park, CA, USA: AAAI Press.

Dowty, D. R. (1991). Proto-roles and argument selection. *Language*, *67*(3), 547–619.

DUC. (2005). *Document Understanding Conference.* http://duc.nist.gov/duc2005.

Ebell, M. H. (1999). Information at the point of care: answering clinical questions. *Journal of the American Board of Family Practice*, *12*(3), 225–235.

Fillmore, C. J. (1976). Frame Semantics and the nature of language. *Annals of the New York Academy of Sciences: Conference on the Origin and Development of Language and Speech*, *280*, 20–32.

Florian, R., Ittycheriah, A., Jing, H., & Zhang, T. (2003). Named entity recognition through classifier combination. In W. Daelemans & M. Osborne (Eds.), *7th Conference on Computational Natural Language Learning* (pp. 168–171). Stroudsburg, PA, USA: Association for Computational Linguistics.

Gildea, D., & Jurafsky, D. (2002). Automatic labeling of Semantic roles. *Computational Linguistics*, *28*(3), 245–288.

Gorman, P., Ash, J., & Wykoff, L. (1994). Can primary care physicians' questions be answered using the medical journal literature? *Bulletin of the Medical Library Association*, *82*(2), 140–146.

Harabagiu, S., Maiorano, S., Moschitti, A., & Bejan, C. (2004). Intentions, implicatures and processing of complex questions. In S. Harabagiu and F. Lacatusu (Eds.), *Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics, Workshop on Pragmatics of Question Answering* (pp. 31–42). Stroudsburg, PA, USA: Association for Computational Linguistics.

Hickl, A., Lehmann, J., Williams, J., & Harabagiu, S. (2004). Experiments with interactive question answering in complex scenarios. In S. Harabagiu & F. Lacatusu (Eds.), *Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics, Workshop on Pragmatics of Question Answering* (pp. 60–69). Stroudsburg, PA, USA: Association for Computational Linguistics.

Huang, T. M., Kecman, V., & Kopriva, I. (2006). *Kernel based algorithms for mining huge data sets*. Berlin, Germany: Springer.

Klein, D., Smarr, J., Nguyen, H., & Manning, C. D. (2003). Named entity recognition with character-level models. In W. Daelemans & M. Osborne (Eds.), *7th Conference on Computational Natural Language Learning* (pp. 180–183). Stroudsburg, PA, USA: Association for Computational Linguistics.

Kohomban, U. S., & Lee, W. S. (2005). Learning Semantic classes for word sense disambiguation.

In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics* (pp. 34–41). Stroudsburg, PA, USA: Association for Computational Linguistics.

Lin, D. (1994). Principar — an efficient, broad-coverage, principle-based parser. In *Proceedings of the 15th International Conference on Computational Linguistics* (pp. 482–488). Stroudsburg, PA, USA: Association for Computational Linguistics.

Lin, D. (1998). Automatic retrieval and clustering of similar words. In *Proceedings of the 17th International Conference on Computational Linguistics* (pp. 768–774). Stroudsburg, PA, USA: Association for Computational Linguistics.

Ma., J., Zhao, Y., Ahalt, S., & Eads, D. (2003). *OSU SVM classifier Matlab toolbox*. http://svm.sourceforge.net/docs/3.00/api/.

Mendonça, E. A., Cimino, J. J., Johnson, S. B., & Seol, Y. H. (2001). Accessing heterogeneous sources of evidence to answer clinical questions. *Journal of Biomedical Informatics*, *34*, 85–98.

MUC. (2003). Message Understanding Conference. http://www.cs.nyu.edu/cs/faculty/grishman/muc6.html

Niu, Y. (2007). *Analysis of Semantic classes: toward non-factoid question answering.* Unpublished doctoral dissertation, University of Toronto, Toronto, Canada.

Niu, Y., & Hirst, G. (2004). Analysis of Semantic classes in medical text for question answering. In D. Mollá & J. L. Vicedo, *42nd Annual Meeting of the Association for Computational Linguistics, Workshop on Question Answering in Restricted Domains* (pp. 54–61). Stroudsburg, PA, USA: Association for Computational Linguistics.

Niu, Y., & Hirst, G. (2007). Identifying cores of Semantic classes in unstructured text with a semi-supervised learning approach. In *Proceedings of Recent Advances in Natural Language Processing 2007* (pp. 418–424).

Niu, Y., Zhu, X.D., & Hirst, G. (2006). Using outcome polarity in sentence extraction for medical question-answering. In Daniel Masys (Ed.), *American Medical Informatics Association 2006 Annual Symposium* (pp. 599–603). Bethesda, MD, USA: American Medical Informatics Association.

Niu, Y., Zhu, X., Li, J., & Hirst, G. (2005). Analysis of polarity information in medical text. In C. P. Friedman (Ed.), *American Medical Informatics Association 2005 Annual Symposium* (pp. 570–574). Bethesda, MD, USA: American Medical Informatics Association.

Niu, Y., Hirst, G., McArthur, M., and Rodriguez-Gianolli, P. (2003). Answering clinical questions with role identification. In Ananiadou, S., & Tsujii, J. (Eds.), *41st Annual Meeting of the Association for Computational Linguistics, Workshop on Natural Language Processing in Biomedicine* (pp. 73–80). Stroudsburg, PA, USA: Association for Computational Linguistics.

Pang, B., & Lee, L. (2003). A sentimental education: sentiment analysis using subjectivity smmarizaiton based on minimum cuts. In *Proceedings of the 42th Annual Meeting of the Association for Computational Linguistics* (pp. 271–278). Stroudsburg, PA, USA: Association for Computational Linguistics.

Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up? sentiment classification using machine learning techniques. In *Proceedings of 2002 Conference on Empirical Methods in Natural Language Processing* (pp. 79–86). PA, USA: Association for Computational Linguistics.

Ray, S., & Craven, M. (2001). Representing sentence structure in hidden Markov models for information extraction. In B. Nebel (Ed.), *17th International Joint Conferences on Artificial*

Intelligence (pp. 1273–1279). San Fransisco, CA, USA: Morgan Kaufmann Publishers Inc.

Richardson, W. S., Wilson, M. C., Nishikawa, J., & Hayward, R. S. (1995). The well-built clinical question: a key to evidence-based decisions. *ACP Journal Club*, *123*(3), 12–13.

Riloff, E. (1999). Information extraction as a stepping stone toward story understanding. In A. Ram & K. Moorman (Eds.), *Computational Models of Reading and Understanding*. Cambridge, MA, USA: The MIT Press.

Rosario, B., & Hearst, M. A. (2004). Classifying Semantic relations in bioscience texts. In *Proceedings of 42nd Annual Meeting of the Association for Computational Linguistics* (pp. 431–438). Stroudsburg, PA, USA: Association for Computational Linguistics.

Sackett, D. L., & Straus, S. E. (1998). Finding and applying evidence during clinical rounds: the "evidence cart". *Journal of the American Medical Association*, *280*(15), 1336–1338.

Sackett, D. L., Straus, S. E., Richardson, W. S., Rosenberg, W., & Haynes, R. B. (2000). *Evidence-based medicine: How to practice and teach EBM.* Edinburgh: Harcourt Publishers Limited.

Sekine, S. (1997). *Apple Pie Parser.* http://nlp.cs.nyu.edu/app/.

Small, S., Strzalkowski, T., Liu, T., Ryan, S., Salkin, R., Shimizu, N., Kantor, P., Kelly, D., Rittman, R., Wacholder, N., & Yamrom, B. (2004). HITIQA: Scenario based question answering. In S. Harabagiu and F. Lacatusu (Eds.), *Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics, Workshop on Pragmatics of Question Answering* (pp. 52–59). Stroudsburg, PA, USA: Association for Computational Linguistics.

Soricut, R., & Brill, E. (2006). Question answering using the Web: Beyond the factoid. *Information Retrieval — Special Issue on Web Information Retrieval, 9*, 191–206.

Stoyanov, V., Cardie, C., & Wiebe, J. (2005). Multi-perspective question answering using the OpQA Corpus. In *Proceedings of Human Language Technology Conference/Conference on Empirical Methods in Natural Language Processing* (pp. 923–930). Stroudsburg, PA, USA: Association for Computational Linguistics.

Straus, S. E., & Sackett, D. L. (1999). Bring evidence to the point of care. *Journal of the American Medical Association, 281,* 1171–1172.

Tjong Kim Sang, E.F. & De Meulder, F. (2003). Introduction to the CoNLL-2003 shared task: language-independent named entity recognition. *Proceedings of Conference on Computational Natural Language Learning* (pp. 142–147). Stroudsburg, PA, USA: Association for Computational Linguistics.

TREC. (2001). *Text REtrieval Conference*. http://trec.nist.gov/.

Turney, P. (2002). Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics* (pp. 417–424). Stroudsburg, PA, USA: Association for Computational Linguistics.

Valin, V., & Robert, D. (1993). A synosis of role and reference grammar. In Robert, D., & Valin, V. (Ed.), *Advances in Role and Reference Grammar* (pp. 1–166). Amsterdam: John Benjamins Publishing Company.

Whitelaw, C., Garg, N., & Argamon, S. (2005). Using appraisal groups for sentiment analysis. In *Proceedings of the 14th ACM International Conference on Information and Knowledge management* (pp. 625–631). New York, NY, USA: Association for Computing Machinery Press.

Yu, H., & Hatzivassiloglou, V. (2003). Towards answering opinion questions: separating facts from opinions and identifying the polarity of opinion sentences. In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing* (pp. 129–136). Stroudsburg, PA, USA: Association for Computational Linguistics.

Zhu, X., Ghahramani, Z., & Lafferty, J. (2003). Semi-supervised learning using Gaussian fields and harmonic functions. In T. Fawcett & N. Mishra, (Eds.) *The 20th International Conference on Machine Learning* (pp. 912–919). Menlo Park, CA, USA: AAAI Press.

ENDNOTE

[1] This chapter summarizes work that was published earlier in Niu, 2007; Niu & Hirst, 2004; Niu & Hirst, 2007; Niu, Hirst, McArthur & Rodriguez-Gianolli, 2003; Niu, Zhu, Li & Hirst, 2005; and Niu, Zhu & Hirst, 2006. Some con-tent in this chapter is reprinted from Niu & Hirst, 2004 and Niu, Hirst, McArthur & Rodriguez-Gianolli, 2003, with permission from the American Medical Informatics Association. Some content is reprinted from Niu, Zhu, Li & Hirst, 2005 and Niu, Zhu & Hirst, 2006 with permission from the Association for Computational Linguistics.