Information Extraction

Information Extraction

Sunita Sarawagi

Indian Institute of Technology
CSE, Mumbai 400076
India
sunita@iitb.ac.in

the essence of knowledge

Boston – Delft

Foundations and Trends® in Databases

Published, sold and distributed by:
now Publishers Inc.
PO Box 1024
Hanover, MA 02339
USA
Tel. +1-781-985-4510
www.nowpublishers.com
sales@nowpublishers.com

Outside North America:
now Publishers Inc.
PO Box 179
2600 AD Delft
The Netherlands
Tel. +31-6-51115274

The preferred citation for this publication is S. Sarawagi, Information Extraction, Foundation and Trends® in Databases, vol 1, no 3, pp 261–377, 2007

ISBN: 978-1-60198-188-2

Foundations and Trends® in Databases

Volume 1 Issue 3, 2007

Editorial Board

Editorial Scope

Foundations and Trends® in Databases covers a breadth of topics relating to the management of large volumes of data. The journal targets the full scope of issues in data management, from theoretical foundations, to languages and modeling, to algorithms, system architecture, and applications. The list of topics below illustrates some of the intended coverage, though it is by no means exhaustive:

- Data Models and Query Languages
- Query Processing and Optimization
- Storage, Access Methods, and Indexing
- Transaction Management, Concurrency Control and Recovery
- Deductive Databases
- Parallel and Distributed Database Systems
- Database Design and Tuning
- Metadata Management
- Object Management
- Trigger Processing and Active Databases
- Data Mining and OLAP
- Approximate and Interactive Query Processing
- Data Warehousing
- Adaptive Query Processing
- Data Stream Management
- Search and Query Integration
- XML and Semi-Structured Data
- Web Services and Middleware
- Data Integration and Exchange
- Private and Secure Data Management
- Peer-to-Peer, Sensornet and Mobile Data Management
- Scientific and Spatial Data Management
- Data Brokering and Publish/Subscribe
- Data Cleaning and Information Extraction
- Probabilistic Data Management

Information for Librarians
Foundations and Trends® in Databases, 2007, Volume 1, 4 issues. ISSN paper version 1931-7883. ISSN online version 1931-7891. Also available as a combined paper and online subscription.

Foundations and Trends® in
Databases
Vol. 1, No. 3 (2007) 261–377

DOI: 10.1561/1500000003

now
the essence of knowledge

Information Extraction

Sunita Sarawagi

Indian Institute of Technology, CSE, Mumbai 400076, India,
sunita@iitb.ac.in

Abstract

The automatic extraction of information from unstructured sources has opened up new avenues for querying, organizing, and analyzing data by drawing upon the clean semantics of structured databases and the abundance of unstructured data. The field of information extraction has its genesis in the natural language processing community where the primary impetus came from competitions centered around the recognition of named entities like people names and organization from news articles. As society became more data oriented with easy online access to both structured and unstructured data, new applications of structure extraction came around. Now, there is interest in converting our personal desktops to structured databases, the knowledge in scientific publications to structured records, and harnessing the Internet for structured fact finding queries. Consequently, there are many different communities of researchers bringing in techniques from machine learning, databases, information retrieval, and computational linguistics for various aspects of the information extraction problem.

This review is a survey of information extraction research of over two decades from these diverse communities. We create a taxonomy of the field along various dimensions derived from the nature of the

extraction task, the techniques used for extraction, the variety of input resources exploited, and the type of output produced. We elaborate on rule-based and statistical methods for entity and relationship extraction. In each case we highlight the different kinds of models for capturing the diversity of clues driving the recognition process and the algorithms for training and efficiently deploying the models. We survey techniques for optimizing the various steps in an information extraction pipeline, adapting to dynamic data, integrating with existing entities and handling uncertainty in the extraction process.

Contents

1

Introduction

Information Extraction refers to the automatic extraction of structured information such as entities, relationships between entities, and attributes describing entities from unstructured sources. This enables much richer forms of queries on the abundant unstructured sources than possible with keyword searches alone. When structured and unstructured data co-exist, information extraction makes it possible to integrate the two types of sources and pose queries spanning them.

The extraction of structure from noisy, unstructured sources is a challenging task, that has engaged a veritable community of researchers for over two decades now. With roots in the Natural Language Processing (NLP) community, the topic of structure extraction now engages many different communities spanning machine learning, information retrieval, database, web, and document analysis. Early extraction tasks were concentrated around the identification of named entities, like people and company names and relationship among them from natural language text. The scope of this research was strongly influenced by two competitions, the Message Understanding Conference (MUC) [57, 100, 198] and Automatic Content Extraction (ACE) [1, 159] program. The advent of the Internet considerably increased the extent and diversity of applications depending on various forms of information

extraction. Applications such as comparison shopping, and other automatic portal creation applications, lead to a frenzy of research and commercial activity on the topic. As society became more data oriented with easy online access to both structured and unstructured data, new applications of structure extraction came around.

To address the needs of these diverse applications, the techniques of structure extraction have evolved considerably over the last two decades. Early systems were rule-based with manually coded rules [10, 127, 181]. As manual coding of rules became tedious, algorithms for automatically learning rules from examples were developed [7, 43, 60, 195]. As extraction systems were targeted on more noisy unstructured sources, rules were found to be too brittle. Then came the age of statistical learning, where in parallel two kinds of techniques were deployed: generative models based on Hidden Markov Models [3, 20, 25, 189] and conditional models based on maximum entropy [26, 118, 135, 143, 177]. Both were superseded by global conditional models, popularly called Conditional Random Fields [125]. As the scope of extraction systems widened to require a more holistic analysis of a document's structure, techniques from grammar construction [191, 213] were developed. In spite of this journey of varied techniques, there is no clear winner. Rule-based methods [72, 113, 141, 190] and statistical methods [32, 72, 146, 220] continue to be used in parallel depending on the nature of the extraction task. There also exist hybrid models [42, 59, 70, 89, 140, 173] that attempt to reap the benefits of both statistical and rule-based methods.

1.1 Applications

Structure extraction is useful in a diverse set of applications. We list a representative subset of these, categorized along whether the applications are enterprise, personal, scientific, or Web-oriented.

1.1.1 Enterprise Applications

News Tracking: A classical application of information extraction, which has spurred a lot of the early research in the NLP community, is automatically tracking specific event types from news sources.

The popular MUC [57, 100, 198] and ACE [1] competitions are based on the extraction of structured entities like people and company names, and relations such as "is-CEO-of" between them. Other popular tasks are: tracking disease outbreaks [99], and terrorist events from news sources. Consequently there are several research publications [71, 98, 209] and many research prototypes [10, 73, 99, 181] that target extraction of named entities and their relationship from news articles. Two recent applications of information extraction on news articles are: the automatic creation of multimedia news by integrating video and pictures of entities and events annotated in the news articles,[1] and hyperlinking news articles to background information on people, locations, and companies.[2]

Customer Care: Any customer-oriented enterprise collects many forms of unstructured data from customer interaction; for effective management these have to be closely integrated with the enterprise's own structured databases and business ontologies. This has given rise to many interesting extraction problems such as the identification of product names and product attributes from customer emails, linking of customer emails to a specific transaction in a sales database [19, 44], the extraction of merchant name and addresses from sales invoices [226], the extraction of repair records from insurance claim forms [168], the extraction of customer moods from phone conversation transcripts [112], and the extraction of product attribute value pairs from textual product descriptions [97].

Data Cleaning: An essential step in all data warehouse cleaning processes is converting addresses that are stored as flat strings into their structured forms such as road name, city, and state. Large customer-oriented organizations like banks, telephone companies, and universities store millions of addresses. In the original form, these addresses have little explicit structure. Often for the same person, there are different address records stored in different databases. During warehouse construction, it is necessary to put all these addresses in a standard canonical format where all the different fields are identified and duplicates

[1] http://spotlight.reuters.com/.
[2] http://www.linkedfacts.com.

removed. An address record broken into its structured fields not only enables better querying, it also provides a more robust way of doing deduplication and householding — a process that identifies all addresses belonging to the same household [3, 8, 25, 187].

Classified Ads: Classified ads and other listings such as restaurant lists is another domain with implicit structure that when exposed can be invaluable for querying. Many researchers have specifically targeted such record-oriented data in their extraction research [150, 156, 157, 195].

1.1.2 Personal Information Management

Personal information management (PIM) systems seek to organize personal data like documents, emails, projects and people in a structured inter-linked format [41, 46, 74]. The success of such systems will depend on being able to automatically extract structure from existing predominantly file-based unstructured sources. Thus, for example we should be able to automatically extract from a PowerPoint file, the author of a talk and link the person to the presenter of a talk announced in an email. Emails, in particular, have served as testbeds for many extraction tasks such as locating mentions of people names and phone numbers [113, 152], and inferring request types in service centers [63].

1.1.3 Scientific Applications

The recent rise of the field of bio-informatics has broadened the scope of earlier extractions from named entities, to biological objects such as proteins and genes. A central problem is extracting from paper repositories such as Pubmed, protein names, and their interaction [22, 32, 166]. Since the form of entities like Gene and Protein names is very different from classical named entities like people and companies, this task has helped to broaden the techniques used for extraction.

1.1.4 Web Oriented Applications

Citation Databases: Many citation databases on the web have been created through elaborate structure extraction steps from sources

ranging from conference web sites to individual home pages. Popular amongst these are Citeseer [126], Google Scholar[3] and Cora [144]. The creation of such databases requires structure extraction at many different levels starting from navigating web sites for locating pages containing publication records, extracting individual publication records from a HTML page, extracting title, authors, and references from paper PDFs, and segmenting citation strings into individual authors, title, venue, and year fields. The resulting structured database provides significant value added in terms of allowing forward references, and aggregate statistics such as author-level citation counts.

Opinion Databases: There are innumerable web sites storing unmoderated opinions about a range of topics, including products, books, movies, people, and music. Many of the opinions are in free text form hidden behind Blogs, newsgroup posts, review sites, and so on. The value of these reviews can be greatly enhanced if organized along structured fields. For example, for products it might be useful to find out for each feature of the product, the prevalent polarity of opinion [131, 167]. See [160] for a recent survey.

Community Websites: Another example of the creation of structured databases from web documents is community web sites such as DBLife [78] and Rexa[4] that tracks information about researchers, conferences, talks, projects, and events relevant to a specific community. The creation of such structured databases requires many extraction steps: locating talk announcements from department pages, extracting names of speakers and titles from them [189], extracting structured records about a conference from a website [111], and so on.

Comparison Shopping: There is much interest in creating comparison shopping web sites that automatically crawl merchant web sites to find products and their prices which can then be used for comparison shopping [87]. As web technologies evolved, most large merchant web sites started getting hidden behind forms and scripting languages. Consequently, the focus has shifted to crawling and extracting information

[3] http://www.scholar.google.com.
[4] http://rexa.info.

from form-based web sites [104]. The extraction of information from form-based web sites is an active research area not covered in this survey.

Ad Placement on Webpages: Suppose a web site wants to place advertisements of a product next to the text that both mentions the product and expresses a positive opinion about it. Both of these subtasks: extracting mentions of products and the type of opinion expressed on the product are examples of information extraction tasks that can facilitate the burgeoning Internet ad placement industry [29].

Structured Web Searches: Finally, a grand challenge problem for information extraction is allowing structured search queries involving entities and their relationships on the World Wide Web. Keyword searches are adequate for getting information about entities, which are typically nouns or noun phrases. They fail on queries that are looking for relationships between entities [45]. For example, if one wants to retrieve documents containing text of the form "Company X acquired Company Y", then keywords alone are extremely inadequate. The only obvious keyword is "acquired", and one has to work hard to introduce related words like "Corp" etc. to get the required documents. Research prototypes for answering such kinds of queries are only starting to appear [39, 196, 197].

1.2 Organization of the Survey

Given the broad scope of the topic, the diversity of communities involved and the long history, compiling an exhaustive survey on structure extraction is a daunting task. Fortunately, there are many short surveys on information extraction from different communities that can be used to supplement what is missed here [71, 98, 104, 139, 142, 153, 154, 178, 209, 212].

We provide a taxonomy of the field by categorizing along different dimensions and alongside scope out what is covered in this survey. We layout the field of information extraction along the following five dimensions.

(1) The type of structure extracted (entities, relationships, lists, tables, attributes, etc.).
(2) The type of unstructured source (short strings or documents, templatized or open-ended).
(3) The type of input resources available for extraction (structured databases, labeled unstructured data, linguistic tags, etc.).
(4) The method used for extraction (rule-based or statistical, manually coded or trained from examples).
(5) The output of extraction (annotated unstructured text, or a database).

These are discussed in Sections 1.3 through 1.7.

1.3 Types of Structure Extracted

We categorize the type of structure extracted from an unstructured source into four types: entities, relationships between entities, adjectives describing entities, and higher-order structures such as tables and lists.

1.3.1 Entities

Entities are typically noun phrases and comprise of one to a few tokens in the unstructured text. The most popular form of entities is *named entities* like names of persons, locations, and companies as popularized in the MUC [57, 100], ACE [1, 159], and CoNLL [206] competitions. Named entity recognition was first introduced in the sixth MUC [100] and consisted of three subtasks: proper names and acronyms of persons, locations, and organizations (ENAMEX), absolute temporal terms (TIMEX) and monetary and other numeric expressions (NUMEX). Now the term entities is expanded to also include generics like disease names, protein names, paper titles, and journal names. The ACE competition for entity relationship extraction from natural language text lists more than 100 different entity types.

Figures 1.1 and 1.2 present examples of entity extractions: Figure 1.1 shows the classical IE task of extracting person, organization,

According to Robert Callahan, president of Eastern's flight attendants union, the past practice of Eastern's parent, Houston-based Texas Air Corp., has involved ultimatums to unions to accept the carrier's terms
The unstructured source

According to ⟨Per⟩ **Robert Callahan** ⟨**/Per**⟩, president of ⟨Org⟩ **Eastern's** ⟨**/Org**⟩ igh t attendants union, the past practice of ⟨Org⟩ **Eastern's** ⟨**/Org**⟩ parent, ⟨Loc⟩ **Houston** ⟨**/Loc**⟩-based ⟨Org⟩ **Texas Air Corp.** ⟨**/Org**⟩, has involved ultimatums to unions to accept the carrier's terms
Annotated entities

Robert Callahan ⟨Employee_Of⟩ Eastern's Texas Air Corp. ⟨Located_In⟩ Houston
Relationships

Fig. 1.1 Traditionally named entity and relationship extraction from plain text (in this case a news article). The extracted entities are bold-faced with the entity type surrounding it.

House number	Building	Road	City	State	Zip
4089	Whispering Pines	Nobel Drive	San Diego	CA	92122

Fig. 1.2 Text segmentation as an example of entity extraction from address records.

and location entities from news articles; Figure 1.2 shows an example where entity extraction can be treated as a problem of segmenting a text record into structured entities. In this case an address string is segmented so as to identify six structured entities. More examples of segmentation of addresses coming from diverse geographical locations appear in Table 1.1.

We cover techniques for entity extraction in Sections 2 and 3.

1.3.2 Relationships

Relationships are defined over two or more entities related in a predefined way. Examples are "is employee of" relationship between a person and an organization, "is acquired by" relationship between pairs of companies, "location of outbreak" relationship between a disease

Table 1.1 Sample addresses from different countries. The first line shows the unformatted address and the second line shows the address broken into its elements.

#	Address text [Segmented address]
0	M. J. Muller, 71, route de Longwy L-4750 PETANGE [recipient: M. J. Muller] [House#: 71] [Street: route de Longwy] [Zip: L-4750] [city:PETANGE]
1	Viale Europa, 22 00144-ROMA RM [Street: Viale Europa] [House#: 22] [City: ROMA] [Province: RM] [Zip: 00144-]
2	7D-Brijdham Bangur Nagar Goregaon (W) Bombay 400 090 [House#: 7D-] [Building: Brijdham] [Colony: Bangur Nagar] [Area: Goregaon (W)] [City: Bombay] [Zip: 400 090]
3	18100 New Hamshire Ave. Silver Spring, MD 20861 [House#: 18100], [Street: New Hamshire Ave.], [City: Silver Spring,], [State: MD], [Zip: 20861]

and a location, and "is price of" relationship between a product name and a currency amount on a web-page. Figure 1.1 shows instances of the extraction of two relationships from a news article. The extraction of relationships differs from the extraction of entities in one significant way. Whereas entities refer to a sequence of words in the source and can be expressed as annotations on the source, relationships are not annotations on a subset of words. Instead they express the associations between two separate text snippets representing the entities.

The extraction of multi-way relationships is often referred to as record extraction. A popular subtype of record extraction is event extraction. For example, for an event such as a disease outbreak we extract a multi-way relationship involving the "disease name", "location of the outbreak", "number of people affected", "number of people killed", and "date of outbreak." Some record extraction tasks are trivial because the unstructured string implies a fixed set of relationships. For example, for addresses, the relation "is located in" is implied between an extracted street name and city name.

In Section 4, we cover techniques for relationship extraction concentrating mostly on binary relationships.

Another form of multi-way relationship popular in the natural language community is Semantic Role Labeling [124], where given a

predicate in a sentence, the goal is to identify various semantic arguments of the predicate. For example, given a predicate *accept* in the sentence "He accepted the manuscript from his dying father with trembling hands" the extraction task is to find the role-sets of the predicate consisting of the "acceptor", "thing accepted", and "accepted-from". We will not cover semantic role labeling in this survey, and refer the reader to [124] to know more about this topic.

1.3.3 Adjectives Describing Entities

In many applications we need to associate a given entity with the value of an adjective describing the entity. The value of this adjective typically needs to be derived by combining soft clues spread over many different words around the entity. For example, given an entity type, say restaurants, or music bands, we need to extract parts of a Blog or web-page that presents a critique of entities of such type. Then, we would like to infer if the critique is positive or negative. This is also called opinion extraction and is now a topic of active research interest in many different communities. We will not cover this topic in this survey but instead refer the reader to [160] for a current and exhaustive survey.

1.3.4 Structures such as Lists, Tables, and Ontologies

The scope of extraction systems has now expanded to include the extraction of not such atomic entities and flat records but also richer structures such as tables, lists, and trees from various types of documents. For example, [109, 134, 164] addresses the identification of tables from documents, [62, 85, 156] considers the extraction of elements of a list, and [130] considers the extraction of ontologies. We will not be able to cover this topic in the survey to contain its scope and volume. On the topic of table extraction there is an extensive research literature spanning many different communities, including the document analysis [84, 109, 134, 222], information retrieval [164], web [62, 96], database [36, 165], and machine learning [164, 216] communities. A survey can be found in [84].

1.4 Types of Unstructured Sources

We classify the type of unstructured source along two dimensions: the basic unit of granularity on which an extractor is run, and the heterogeneity in style and format across unstructured documents.

1.4.1 Granularity of Extraction

Record or Sentences: The most popular form of extraction is from small text snippets that are either unstructured records like addresses, citations and classified ads [3, 25, 151, 163, 195] or sentences extracted from a natural language paragraph [1, 26, 57, 100, 159, 206]. In the case of unstructured records, the data can be treated as a set of structured fields concatenated together, possibly with a limited reordering of the fields. Thus, each word is a part of such structured field and during extraction we just need to segment the text at the entity boundaries. In contrast, in sentences there are many words that do not form part of any entity of interest.

Paragraphs and Documents: Many other extraction tasks make it necessary to consider the context of multiple sentences or an entire document for meaningful extractions. Popular examples include extractions of events from news articles [57, 100], extraction of part number and problem description from emails in help centers, extraction of a structured resume from a word file, extraction of title, location and timing of a talk from talk announcements [189] and the extraction of paper headers and citations from a scientific publication [163].

The techniques proposed in this survey mostly assume the first kind of source. Typically, for extracting information from longer units the main challenge is designing efficient techniques for filtering only the relevant portion of a long document. Currently, this is handled through hand-coded heuristics, so there is nothing specifically to cover in a survey on the handling of longer units.

1.4.2 Heterogeneity of Unstructured Sources

An important concern that has a huge impact on the complexity and accuracy of an extractor is how much homogeneity is there in

the format and style of the unstructured documents. We categorize them as:

Machine Generated Pages: On the easy end of the spectrum we have highly templatized machine generated pages. A popular source in this space is HTML documents dynamically generated via database backed sites. The extractors for such documents are popularly known as wrappers. These have been extensively studied in many communities [11, 184, 16, 17, 67, 103, 106, 123, 133, 149, 156], where the main challenge is how to automatically figure out the layout of a page with little or no human input by exploiting mostly the regularity of HTML tags present in the page. In this survey we will not be able to do justice to the extensive literature on web wrapper development.

Partially Structured Domain Specific Sources: The most studied setting for information extraction is where the input source is from within a well-defined scope, say news articles [1, 57, 100, 159, 206], or classified ads [151, 195], or citations [25, 163], or resumes. In all these examples, there is an informal style that is roughly followed so that it is possible to develop a decent extraction model given enough labeled data, but there is lot more variety from one input to another than in machine generated pages. Most of the techniques in this survey are for such input sources.

Open Ended Sources: Recently [14, 37, 86, 192], there is interest in extracting instances of relationships and entities from open domains such as the web where there is little that can be expected in terms of homogeneity or consistency. In such situations, one important factor is to exploit the redundancy of the extracted information across many different sources. We discuss extractions from such sources in the context of relationship extraction in Section 4.2.

1.5 Input Resources for Extraction

The basic specification of an extraction task includes just the types of structures to be extracted and the unstructured sources from which

it should be extracted. In practice, there are several additional input resources that are available to aid the extraction.

1.5.1 Structured Databases

Existing structured databases of known entities and relationships are a valuable resource to improve extraction accuracy. Typically, there are several such databases available during extraction. In many applications unstructured data needs to be integrated with structured databases on an ongoing basis so that at the time of extraction a large database is available. Consider the example of portals like DBLife, Citeseer, and Google Scholar. In addition to their own operational database of extracted publications, they can also exploit external databases such as the ACM digital library or DBLP. Other examples include the use of a sales transactions database and product database for extracting fields like customer id and product name in a customer email; the use of a contact database to extract authoring information from files in a personal information management system; the use of a postal database to identify entities in address records.

1.5.2 Labeled Unstructured Text

Many extraction systems are seeded via labeled unstructured text. The collection of labeled unstructured text requires tedious labeling effort. However, this effort is not totally avoidable because even when an extraction system is manually coded, a ground truth is necessary for evaluating its accuracy. A labeled unstructured source is significantly more valuable than a structured database because it provides contextual information about an entity and also because the form in which an entity appears in the unstructured data is often a very noisy form of its occurrence in the database.

We will discuss how labeled data is used for learning entity extraction models in Sections 2.3 and 3.4 and for relationship extraction in Section 4.1. In Section 4.2, we show how to learn a model using only a structured database and a large corpus of unlabeled corpus. We discuss how structured databases are used in conjunction with labeled data in Sections 2 and 3.

1.5.3 Preprocessing Libraries for Unstructured Text

Many extraction systems crucially depend on preprocessing libraries that enrich it with linguistic or layout information that serve as valuable anchors for structure recognition.

Natural Language Text: Natural language documents are often analyzed by a deep pipeline of preprocessing libraries, including,

- *Sentence analyzer and tokenizer* that identifies the boundaries of sentences in a document and decomposes each sentence into tokens. Tokens are obtained by splitting a sentence along a predefined set of delimiters like spaces, commas, and dots. A token is typically a word or a digit, or a punctuation.
- *Part of speech tagger* that assigns to each word a grammatical category coming from a fixed set. The set of tags includes the conventional part of speech such as noun, verb, adjective, adverb, article, conjunct, and pronoun; but is often considerably more detailed to capture many subtypes of the basic types. Examples of well-known tag sets are the Brown tag set which has 179 total tags, and the Penn treebank tag set that has 45 tags [137]. An example of POS tags attached to a sentence appears below:

  The/DT University/NNP of/IN Helsinki/NNP hosts/VBZ ICML/NNP this/DT year/NN

- *Parser* that groups words in a sentence into prominent phrase types such as noun phrases, prepositional phrases, and verb phrases. A context free grammar is typically used to identify the structure of a sentence in terms of its constituent phrase types. The output of parsing is a parse tree that groups words into syntactic phrases. An example of a parse tree appears in Figure 4.1. Parse trees are useful in entity extraction because typically named entities are noun phrases. In relationship extraction they are useful because they provide valuable linkages between verbs and their arguments as we will see in Section 4.1.

- *Dependency analyzer* that identifies the words in a sentence that form arguments of other words in the sentence. For example, in the sentence "Apple is located in Cupertino", the word "Apple" and "Cupertino" are dependent on the word "located". In particular, they respectively form the subject and object argument of the word "located". The output of a dependency analyzer is a graph where the nodes are the words and the directed edges are used to connect a word to words that depend on it. An example of a dependency graph appears in Figure 4.2. The edges could be typed to indicate the type of dependency, but even untyped edges are useful for relationship extraction as we will see in Section 4.

Many of the above preprocessing steps are expensive. The shift is now for selective preprocessing of only parts of the text. Many shallow extractions are possible without subjecting a sentence to the full preprocessing pipeline. Also, some of these preprocessing steps, example parsing, are often erroneous. The extraction system needs to be robust to errors in the preprocessing steps to avoid cascading of errors. This problem is particularly severe on ill-formed sentences of the kind found in emails and speech transcripts.

GATE [72] and UIMA [91] are two examples of frameworks that provide support for such preprocessing pipelines. Many NLP libraries are also freely available for download such as IBM's Languageware,[5] libraries from the Stanford NLP group,[6] and several others listed under the OpenNLP effort.[7]

Formatted Text: For formatted text such as a pdf document and a web-page, there is often a need for understanding the overall structure and layout of the source before entity extraction. Two popular preprocessing steps on formatted documents are, extracting items in a list-like environment and creating hierarchies of rectangular regions comprising logical units of content. Much work exists in this area in the document

[5] http://www.alphaworks.ibm.com/tech/lrw.
[6] http://nlp.stanford.edu/software/.
[7] http://opennlp.sourceforge.net/.

analysis community [139] and elsewhere [40, 85, 157, 191]. We will not discuss these in this survey.

1.6 Methods of Extraction

We categorize the method used for information extraction along two dimensions: hand-coded or learning-based and rule-based or statistical.

1.6.1 Hand-coded or Learning-based

A hand-coded system requires human experts to define rules or regular expressions or program snippets for performing the extraction. That person needs to be a domain expert and a programmer, and possess descent linguistic understanding to be able to develop robust extraction rules. In contrast, learning-based systems require manually labeled unstructured examples to train machine learning models of extraction. Even in the learning-based systems, domain expertise is needed in identifying and labeling examples that will be representative of the actual deployment setting. It is also necessary to possess an understanding of machine learning to be able to choose between various model alternatives and also to define features that will be robust on unseen data. The nature of the extraction task and the amount of noise in the unstructured data should be used to decide between a hand-coded and a learning-based system. An interesting commentary that quantitatively and qualitatively compares the two sides can be found in [127].

1.6.2 Rule-based or Statistical

Rule-based extraction methods are driven by hard predicates, whereas statistical methods make decisions based on a weighted sum of predicate firings. Rule-based methods are easier to interpret and develop, whereas statistical methods are more robust to noise in the unstructured data. Therefore, rule-based systems are more useful in closed domains where human involvement is both essential and available. In open-ended domains like fact extraction from speech transcripts, or opinion extraction from Blogs, the soft logic of statistical methods is more appropriate. We will present both rule-based techniques for entity

extraction in Section 2 and statistical techniques for entity and relationship extraction in Sections 3 and 4, respectively.

1.7 Output of Extraction Systems

There are two primary modes in which an extraction system is deployed. First, where the goal is to identify all mentions of the structured information in the unstructured text. Second, where the goal is to populate a database of structured entities. In this case, the end user does not care about the unstructured text after the structured entities are extracted from it. The core extraction techniques remain the same irrespective of the form of the output. Therefore, in the rest of the survey we will assume the first form of output. Only for a few types of open ended extractions where redundancy is used to improve the reliability of extractions stored in a database is the distinction important. We briefly cover this scenario in Sections 4.2 and 5.4.3.

1.8 Challenges

Large scale deployments of information extraction models raises many challenges of accuracy, performance, maintainability, and usability that we elaborate on next.

1.8.1 Accuracy

The foremost challenge facing the research community, in spite of more than two decades of research in the field, is designing models that achieve high accuracy of extraction. We list some of the factors that contribute to the difficulty of achieving high accuracy in extraction tasks.

Diversity of Clues: The inherent complexity of the recognition task makes it crucial to combine evidence from a diverse set of clues, each of which could individually be very weak. Even the simplest and the most well-explored of tasks, Named Entity recognition, depends on a myriad set of clues including orthographic property of the words, their part of speech, similarity with an existing database of entities, presence of specific signature words and so on. Optimally combining these different

modalities of clues presents a nontrivial modeling challenge. This is evidenced by the huge research literature for this task alone over the past two decades. We will encounter many of these in the next three sections of the survey. However, the problem is far from solved for all the different types of extraction tasks that we mentioned in Section 1.3.

Difficulty of Detecting Missed Extractions: The accuracy of extraction comprises of two components: precision, that measures the percent of extracted entries that are correct, and recall, that measures the percent of actual entities that were extracted correctly. In many cases, precision is high because it is easy to manually detect mistakes in extractions and then tune the models until those mistakes disappear. The bigger challenge is achieving high recall, because without extensive labeled data it is not even possible to detect what was missed in the large mass of unstructured information.

Increased Complexity of the Structures Extracted: New tasks requiring the extraction of increasingly complex kinds of entities keep getting defined. Of the recent additions, it is not entirely clear how to extract longer entities such as the parts within running text of a Blog where a restaurant is mentioned and critiqued. One of the challenges in such tasks is that the boundary of the entity is not clearly defined.

1.8.2 Running Time

Real-life deployment of extraction techniques in the context of an operational system raises many practical performance challenges. These arise at many different levels. First, we need mechanisms to efficiently filter the right subset of documents that are likely to contain the structured information of interest. Second, we need to find means of efficiently zooming into the (typically small) portion of the document that contains the relevant information. Finally, we need to worry about the many expensive processing steps that the selected portion might need to go through. For example, while existing database of structured entries are invaluable for information extraction, they also raise performance challenges. The order in which we search for parts of a compound entity or relationship can have a big influence on running time. These and other performance issues are discussed in Section 5.1.

1.8.3 Other Systems Issues

Dynamically Changing Sources: Extraction models take time and effort to build and tune to specific unstructured sources. When these sources change, a challenge to any system that operates continuously on that source is detecting the change and adapting the model automatically to the change. We elaborate on this topic in Section 5.2.

Data Integration: Although in this survey we will concentrate primarily on information extraction, extraction goes hand in hand with the integration of the extracted information with pre-existing datasets and with information already extracted. Many researchers have also attempted to jointly solve the extraction and integration problem with the hope that it will provide higher accuracy than performing each of these steps directly. We elaborate further in Section 5.3.

Extraction Errors: It is impossible to guarantee perfect extraction accuracy in real-life deployment settings even with the latest extraction tools. The problem is more severe when the sources are extremely heterogeneous, making it impossible to hand tune any extraction tool to perfection. One method of surmounting the problem of extraction errors is to require that each extracted entity be attached with confidence scores that correlate with the probability that the extracted entities are correct. Normally, even this is a hard goal to achieve. Another challenging issue is how to represent such results in a database that captures the imprecision of extraction, while being easy to store and query. In Section 5.4, we review techniques for managing errors that arise in the extraction process.

Section Layout

The rest of the survey is organized as follows. In Section 2, we cover rule-based techniques for entity extraction. In Section 3, we present an overview of statistical methods for entity extraction. In Section 4, we cover statistical and rule-based techniques for relationship extraction. In Section 5, we discuss work on handling various performance and systems issues associated with creating an operational extraction system.

References

[1] 2004. ACE. Annotation guidelines for entity detection and tracking.

[2] E. Agichtein, "Extracting relations from large text collections," PhD thesis, Columbia University, 2005.

[3] E. Agichtein and V. Ganti, "Mining reference tables for automatic text segmentation," in *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Seattle, USA, 2004.

[4] E. Agichtein and L. Gravano, "Snowball: Extracting relations from large plain-text collections," in *Proceedings of the 5th ACM International Conference on Digital Libraries*, 2000.

[5] E. Agichtein and L. Gravano, "Querying text databases for efficient information extraction," in *ICDE*, 2003.

[6] R. Agrawal, H. Mannila, R. Srikant, H. Toivonen, and A. I. Verkamo, "Fast discovery of association rules," in *Advances in Knowledge Discovery and Data Mining*, (U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, eds.), ch. 12, pp. 307–328, AAAI/MIT Press, 1996.

[7] J. Aitken, "Learning information extraction rules: An inductive logic programming approach," in *Proceedings of the 15th European Conference on Artificial Intelligence*, pp. 355–359, 2002.

[8] R. Ananthakrishna, S. chaudhuri, and V. Ganti, "Eliminating fuzzy duplicates in data warehouses," in *International Conference on Very Large Databases (VLDB)*, 2002.

[9] R. Ando and T. Zhang, "A framework for learning predictive structures from multiple tasks and unlabeled data," *Journal of Machine Learning Research*, vol. 6, pp. 1817–1853, 2005.

[10] D. E. Appelt, J. R. Hobbs, J. Bear, D. J. Israel, and M. Tyson, "Fastus: A finite-state processor for information extraction from real-world text," in *IJCAI*, pp. 1172–1178, 1993.

[11] A. Arasu, H. Garcia-Molina, and S. University, "Extracting structured data from web pages," in *SIGMOD '03: Proceedings of the 2003 ACM SIGMOD International Conference on Management of Data*, pp. 337–348, 2003.

[12] S. Argamon-Engelson and I. Dagan, "Committee-based sample selection for probabilistic classifiers," *Journal of Artificial Intelligence Research*, vol. 11, pp. 335–360, 1999.

[13] M.-F. Balcan, A. Beygelzimer, and J. Langford, "Agnostic active learning," in *ICML*, pp. 65–72, 2006.

[14] M. Banko, M. J. Cafarella, S. Soderland, M. Broadhead, and O. Etzioni, "Open information extraction from the web," in *IJCAI*, pp. 2670–2676, 2007.

[15] N. Bansal, A. Blum, and S. Chawla, "Correlation clustering," in *FOCS '02: Proceedings of the 43rd Symposium on Foundations of Computer Science*, USA, Washington, DC: IEEE Computer Society, 2002.

[16] G. Barish, Y.-S. Chen, D. DiPasquo, C. A. Knoblock, S. Minton, I. Muslea, and C. Shahabi, "Theaterloc: Using information integration technology to rapidly build virtual applications," in *International Conference on Data Engineering (ICDE)*, pp. 681–682, 2000.

[17] R. Baumgartner, S. Flesca, and G. Gottlob, "Visual web information extraction with lixto," in *VLDB '01: Proceedings of the 27th International Conference on Very Large Data Bases*, pp. 119–128, USA, San Francisco, CA: Morgan Kaufmann Publishers Inc, 2001.

[18] M. Berland and E. Charniak, "Finding parts in very large corpora," in *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics*, pp. 57–64, 1999.

[19] M. Bhide, A. Gupta, R. Gupta, P. Roy, M. K. Mohania, and Z. Ichhaporia, "Liptus: Associating structured and unstructured information in a banking environment," in *SIGMOD Conference*, pp. 915–924, 2007.

[20] D. M. Bikel, S. Miller, R. Schwartz, and R. Weischedel, "Nymble: A high-performance learning name-finder," in *Proceedings of ANLP-97*, pp. 194–201, 1997.

[21] M. Bilenko, R. Mooney, W. Cohen, P. Ravikumar, and S. Fienberg, "Adaptive name-matching in information integration," *IEEE Intelligent Systems*, 2003.

[22] 2006. Biocreative — critical assessment for information extraction in biology. http://biocreative.sourceforge.net/.

[23] J. Blitzer, R. McDonald, and F. Pereira, "Domain adaptation with structural correspondence learning," in *Proceedings of the Empirical Methods in Natural Language Processing (EMNLP)*, 2006.

[24] A. Bordes, L. Bottou, P. Gallinari, and J. Weston, "Solving multiclass support vector machines with larank," in *ICML*, pp. 89–96, 2007.

[25] V. R. Borkar, K. Deshmukh, and S. Sarawagi, "Automatic text segmentation for extracting structured records," in *Proceedings of ACM SIGMOD International Conference on Management of Data*, Santa Barabara, USA, 2001.

[26] A. Borthwick, J. Sterling, E. Agichtein, and R. Grishman, "Exploiting diverse knowledge sources via maximum entropy in named entity recognition," in *Sixth Workshop on Very Large Corpora New Brunswick*, New Jersey, Association for Computational Linguistics, 1998.

[27] L. Bottou, "Stochastic learning," in *Advanced Lectures on Machine Learning, number LNAI 3176 in Lecture Notes in Artificial Intelligence*, (O. Bousquet and U. von Luxburg, eds.), pp. 146–168, Springer Verlag, 2004.

[28] J. Boulos, N. Dalvi, B. Mandhani, S. Mathur, C. Re, and D. Suciu, "Mystiq: A system for finding more answers by using probabilities," in *ACM SIGMOD*, 2005.

[29] A. Z. Broder, M. Fontoura, V. Josifovski, and L. Riedel, "A semantic approach to contextual advertising," in *SIGIR*, pp. 559–566, 2007.

[30] R. Bunescu and R. Mooney, "Learning to extract relations from the web using minimal supervision," in *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pp. 576–583, June 2007.

[31] R. Bunescu and R. J. Mooney, "Collective information extraction with relational markov networks," in *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, pp. 439–446, 2004.

[32] R. C. Bunescu, R. Ge, R. J. Kate, E. M. Marcotte, R. J. Mooney, A. K. Ramani, and Y. W. Wong, "Comparative experiments on learning information extractors for proteins and their interactions," *Artificial Intelligence in Medicine*, vol. 33, pp. 139–155, 2005.

[33] R. C. Bunescu and R. J. Mooney, "A shortest path dependency kernel for relation extraction," in *HLT '05: Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pp. 724–731, USA, Morristown, NJ: Association for Computational Linguistics, 2005.

[34] D. Burdick, P. M. Deshpande, T. S. Jayram, R. Ramakrishnan, and S. Vaithyanathan, "OLAP over uncertain and imprecise data," in *Proceedings of the 31st International Conference on Very Large Data Bases*, pp. 970–981, VLDB Endowment, 2005.

[35] D. Burdick, A. Doan, R. Ramakrishnan, and S. Vaithyanathan, "Olap over imprecise data with domain constraints," in *VLDB*, pp. 39–50, 2007.

[36] M. Cafarella, N. Khoussainova, D. Wang, E. Wu, Y. Zhang, and A. Halevy, "Uncovering the relational web," in *WebDB*, 2008.

[37] M. J. Cafarella, D. Downey, S. Soderland, and O. Etzioni, "KnowItNow: Fast, scalable information extraction from the web," in *Conference on Human Language Technologies (HLT/EMNLP)*, 2005.

[38] M. J. Cafarella and O. Etzioni, "A search engine for natural language applications," in *WWW*, pp. 442–452, 2005.

[39] M. J. Cafarella, C. Re, D. Suciu, and O. Etzioni, "Structured querying of web text data: A technical challenge," in *CIDR*, pp. 225–234, 2007.

[40] D. Cai, ShipengYu, Ji-RongWen, and W.-Y. Ma, "Vips: A vision based page segmentation algorithm," Technical Report MSR-TR-2003-79, Microsoft, 2004.

[41] Y. Cai, X. L. Dong, A. Y. Halevy, J. M. Liu, and J. Madhavan, "Personal information management with semex," in *SIGMOD Conference*, pp. 921–923, 2005.

[42] M. Califf and R. Mooney, *Bottom-up Relational Learning of Pattern Matching Rules for Information Extraction*, 2003.

[43] M. E. Califf and R. J. Mooney, "Relational learning of pattern-match rules for information extraction," in *Proceedings of the Sixteenth National Conference on Artificial Intelligence (AAAI-99)*, pp. 328–334, July 1999.

[44] V. T. Chakaravarthy, H. Gupta, P. Roy, and M. K. Mohania, "Efficiently linking text documents with relevant structured information," in *VLDB*, pp. 667–678, 2006.

[45] S. Chakrabarti, *Mining the Web: Discovering Knowledge from Hypertext Data*. Morgan-Kauffman, 2002.

[46] S. Chakrabarti, J. Mirchandani, and A. Nandi, "Spin: searching personal information networks," in *SIGIR*, p. 674, 2005.

[47] S. Chakrabarti, K. Punera, and M. Subramanyam, "Accelerated focused crawling through online relevance feedback," in *WWW*, Hawaii, ACM, May 2002.

[48] S. Chakrabarti, K. Puniyani, and S. Das, "Optimizing scoring functions and indexes for proximity search in type-annotated corpora," in *WWW*, pp. 717–726, 2006.

[49] A. Chandel, P. Nagesh, and S. Sarawagi, "Efficient batch top-k search for dictionary-based entity recognition," in *Proceedings of the 22nd IEEE International Conference on Data Engineering (ICDE)*, 2006.

[50] M. Charikar, V. Guruswami, and A. Wirth, "Clustering with qualitative information," *Journal of Computer and Systems Sciences*, vol. 71, pp. 360–383, 2005.

[51] S. Chaudhuri, K. Ganjam, V. Ganti, and R. Motwani, "Robust and efficient fuzzy match for online data cleaning," in *SIGMOD*, 2003.

[52] Chelba and Acero, "Adaptation of maximum entropy capitalizer: Little data can help a lot," in *EMNLP*, 2004.

[53] F. Chen, A. Doan, J. Yang, and R. Ramakrishnan, "Efficient information extraction over evolving text data," in *ICDE*, 2008.

[54] D. Cheng, R. Kannan, S. Vempala, and G. Wang, "A divide-and-merge methodology for clustering," *ACM Transactions on Database Systems*, vol. 31, pp. 1499–1525, 2006.

[55] R. Cheng, D. V. Kalashnikov, and S. Prabhakar, "Evaluating probabilistic queries over imprecise data," in *SIGMOD '03: Proceedings of the 2003 ACM SIGMOD International Conference on Management of Data*, pp. 551–562, USA, New York, NY: ACM Press, 2003.

[56] B. Chidlovskii, B. Roustant, and M. Brette, "Documentum eci self-repairing wrappers: Performance analysis," in *SIGMOD '06: Proceedings of the 2006 ACM SIGMOD International Conference on Management of Data*, pp. 708–717, USA, New York, NY: ACM, 2006.

[57] 1998. N. A. Chinchor, *Overview of MUC-7/MET-2*.

[58] J. Cho and S. Rajagopalan, "A fast regular expression indexing engine," in *ICDE*, pp. 419–430, 2002.

[59] Y. Choi, C. Cardie, E. Riloff, and S. Patwardhan, "Identifying sources of opinions with conditional random fields and extraction patterns," in *HLT/EMNLP*, 2005.

[60] F. Ciravegna, "Adaptive information extraction from text by rule induction and generalisation," in *Proceedings of the 17th International Joint Conference on Artificial Intelligence (IJCAI2001)*, 2001.

[61] W. Cohen and J. Richman, "Learning to match and cluster entity names," in *ACM SIGIR' 01 Workshop on Mathematical/Formal Methods in Information Retrieval*, 2001.

[62] W. W. Cohen, M. Hurst, and L. S. Jensen, "A flexible learning system for wrapping tables and lists in html documents," in *Proceedings of the 11th World Wide Web Conference (WWW2002)*, 2002.

[63] W. W. Cohen, E. Minkov, and A. Tomasic, "Learning to understand web site update requests," in *IJCAI*, pp. 1028–1033, 2005.

[64] W. W. Cohen, P. Ravikumar, and S. E. Fienberg, "A comparison of string distance metrics for name-matching tasks," in *Proceedings of the IJCAI-2003 Workshop on Information Integration on the Web (IIWeb-03)*, 2003. (To appear).

[65] W. W. Cohen and S. Sarawagi, "Exploiting dictionaries in named entity extraction: Combining semi-markov extraction processes and data integration methods," in *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2004.

[66] D. A. Cohn, Z. Ghahramani, and M. I. Jordan, "Active learning with statistical models," in *Advances in Neural Information Processing Systems*, (G. Tesauro, D. Touretzky, and T. Leen, eds.), pp. 705–712, The MIT Press, 1995.

[67] V. Crescenzi, G. Mecca, P. Merialdo, and P. Missier, "An automatic data grabber for large web sites," in *vldb'2004: Proceedings of the Thirtieth International Conference on Very Large Data Bases*, pp. 1321–1324, 2004.

[68] A. Culotta, T. T. Kristjansson, A. McCallum, and P. A. Viola, "Corrective feedback and persistent learning for information extraction," *Artificial Intelligence*, vol. 170, nos. 14–15, pp. 1101–1122, 2006.

[69] A. Culotta and J. Sorensen, "Dependency tree kernels for relation extraction," in *Proceedings of the 42nd Meeting of the Association for Computational Linguistics (ACL'04), Main Volume*, pp. 423–429, Barcelona, Spain, July 2004.

[70] C. Cumby and D. Roth, "Feature extraction languages for propositionalzed relational learning," in *Working Notes of the IJCAI-2003 Workshop on Learning Statistical Models from Relational Data (SRL-2003)*, (L. Getoor and D. Jensen, eds.), pp. 24–31, Acapulco, Mexico, August 11, 2003.

[71] H. Cunningham, "Information extraction, automatic," *Encyclopedia of Language and Linguistics*, 2005. second ed.

[72] H. Cunningham, D. Maynard, K. Bontcheva, and V. Tablan, "Gate: A framework and graphical development environment for robust nlp tools and applications," in *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics*, 2002.

[73] H. Cunningham, D. Maynard, K. Bontcheva, and V. Tablan, "GATE: A framework and graphical development environment for robust nlp tools and applications," in *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02)*, Philadelphia, 2002.

[74] E. Cutrell and S. T. Dumais, "Exploring personal information," *Communications on ACM*, vol. 49, pp. 50–51, 2006.

[75] N. N. Dalvi and D. Suciu, "Efficient query evaluation on probabilistic databases," in *VLDB*, pp. 864–875, 2004.

[76] S. Dasgupta, "Coarse sample complexity bounds for active learning," in *NIPS*, 2005.

[77] H. Daumé III, "Frustratingly easy domain adaptation," in *Conference of the Association for Computational Linguistics (ACL)*, Prague, Czech Republic, 2007.

[78] P. DeRose, W. Shen, F. C. 0002, Y. Lee, D. Burdick, A. Doan, and R. Ramakrishnan, "Dblife: A community information management platform for the database research community (demo)," in *CIDR*, pp. 169–172, 2007.

[79] T. Dietterich, "Machine learning for sequential data: A review," in *Structural, Syntactic and Statistical Pattern Recognition; Lecture Notes in Computer Science*, (T. Caelli, ed.), Vol. 2396, pp. 15–30, Springer-Verlag, 2002.

[80] P. Domingos, "Metacost: A general method for making classifiers cost-sensitive," in *Proceedings of the Fifth International Conference on Knowledge Discovery and Data Mining (KDD-99)*, 1999.

[81] D. Downey, M. Broadhead, and O. Etzioni, "Locating complex named entities in web text," in *IJCAI*, pp. 2733–2739, 2007.

[82] D. Downey, O. Etzioni, and S. Soderland, "A probabilistic model of redundancy in information extraction," in *IJCAI*, 2005.

[83] D. Downey, S. Schoenmackers, and O. Etzioni, "Sparse information extraction: Unsupervised language models to the rescue," in *ACL*, 2007.

[84] D. W. Embley, M. Hurst, D. P. Lopresti, and G. Nagy, "Table-processing paradigms: A research survey," *IJDAR*, vol. 8, nos. 2–3, pp. 66–86, 2006.

[85] D. W. Embley, Y. S. Jiang, and Y.-K. Ng, "Record-boundary discovery in web documents," in *SIGMOD 1999, Proceedings ACM SIGMOD International Conference on Management of Data, June 1–3, 1999*, pp. 467–478, Philadelphia, Pennsylvania, USA, 1999.

[86] O. Etzioni, M. Cafarella, D. Downey, S. Kok, A.-M. Popescu, T. Shaked, S. Soderland, D. S. Weld, and A. Yates, "Web-scale information extraction in KnowItAll: (preliminary results)," in *WWW*, pp. 100–110, 2004.

[87] O. Etzioni, B. Doorenbos, and D. Weld, "A scalable comparison shopping agent for the world-wide web," in *Proceedings of the International Conference on Autonomous Agents*, 1997.

[88] R. Fagin, A. Lotem, and M. Naor, "Optimal aggregation algorithms for middleware," *Journal of Computer and System Sciences*, vol. 66, nos. 614, 656, September 2001.

[89] R. Feldman, B. Rosenfeld, and M. Fresko, "Teg-a hybrid approach to information extraction," *Knowledge and Information Systems*, vol. 9, pp. 1–18, 2006.

[90] I. P. Fellegi and A. B. Sunter, "A theory for record linkage," *Journal of the American Statistical Society*, vol. 64, pp. 1183–1210, 1969.

[91] D. Ferrucci and A. Lally, "Uima: An architectural approach to unstructured information processing in the corporate research environment," *Natural Language Engineering*, vol. 10, nos. 3–4, pp. 327–348, 2004.

[92] J. R. Finkel, T. Grenager, and C. Manning, "Incorporating non-local information into information extraction systems by gibbs sampling," in *Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics (ACL 2005)*, 2005.

[93] J. R. Finkel, T. Grenager, and C. D. Manning, "Incorporating non-local information into information extraction systems by gibbs sampling," in *ACL*, 2005.

[94] G. W. Flake, E. J. Glover, S. Lawrence, and C. L. Giles, "Extracting query modifications from nonlinear svms," in *WWW*, pp. 317–324, 2002.

[95] Y. Freund, H. S. Seung, E. Shamir, and N. Tishby, "Selective sampling using the query by committee algorithm," *Machine Learning*, vol. 28, nos. 2–3, pp. 133–168, 1997.

[96] W. Gatterbauer, P. Bohunsky, M. Herzog, B. Krüpl, and B. Pollak, "Towards domain-independent information extraction from web tables," in *WWW '07: Proceedings of the 16th International Conference on World Wide Web*, pp. 71–80, ACM, 2007.

[97] R. Ghani, K. Probst, Y. Liu, M. Krema, and A. Fano, "Text mining for product attribute extraction," *SIGKDD Explorations Newsletter*, vol. 8, pp. 41–48, 2006.

[98] R. Grishman, "Information extraction: Techniques and challenges," in *SCIE*, 1997.

[99] R. Grishman, S. Huttunen, and R. Yangarber, "Information extraction for enhanced access to disease outbreak reports," *Journal of Biomedical Informatics*, vol. 35, pp. 236–246, 2002.

[100] R. Grishman and B. Sundheim, "Message understanding conference-6: A brief history," in *Proceedings of the 16th Conference on Computational Linguistics*, pp. 466–471, USA, Morristown, NJ: Association for Computational Linguistics, 1996.

[101] R. Gupta, A. A. Diwan, and S. Sarawagi, "Efficient inference with cardinality-based clique potentials," in *Proceedings of the 24th International Conference on Machine Learning (ICML)*, USA, 2007.

[102] R. Gupta and S. Sarawagi, "Curating probabilistic databases from information extraction models," in *Proceedings of the 32nd International Conference on Very Large Databases (VLDB)*, 2006.

[103] J. Hammer, H. Garcia-Molina, J. Cho, R. Aranha, and A. Crespo, "Extracting semistructure information from the web," in *Workshop on Mangement of Semistructured Data*, 1997.

[104] B. He, M. Patel, Z. Zhang, and K. C.-C. Chang, "Accessing the deep web," *Communications on ACM*, vol. 50, pp. 94–101, 2007.

[105] M. A. Hearst, "Automatic acquisition of hyponyms from large text corpora," in *Proceedings of the 14th Conference on Computational Linguistics*, pp. 539–545, 1992.

[106] C.-N. Hsu and M.-T. Dung, "Generating finite-state transducers for semistructured data extraction from the web," *Information Systems Special Issue on Semistructured Data*, vol. 23, 1998.

[107] J. Huang, T. Chen, A. Doan, and J. Naughton, *On the Provenance of Non-answers to Queries Over Extracted Data*.

[108] J. Huang, A. Smola, A. Gretton, K. Borgwardt, and B. Schölkopf, "Correcting sample selection bias by unlabeled data," in *Advances in Neural Information Processing Systems 20*, Cambridge, MA: MIT Press, 2007.

[109] M. Hurst, "The interpretation of tables in texts," PhD thesis, University of Edinburgh, School of Cognitive Science, Informatics, University of Edinburgh, 2000.

[110] P. G. Ipeirotis, E. Agichtein, P. Jain, and L. Gravano, "Towards a query optimizer for text-centric tasks," *ACM Transactions on Database Systems*, vol. 32, 2007.

[111] N. Ireson, F. Ciravegna, M. E. Califf, D. Freitag, N. Kushmerick, and A. Lavelli, "Evaluating machine learning for information extraction," in *ICML*, pp. 345–352, 2005.

[112] M. Jansche and S. P. Abney, "Information extraction from voicemail transcripts," in *EMNLP '02: Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing*, pp. 320–327, USA, Morristown, NJ: Association for Computational Linguistics, 2002.

[113] T. S. Jayram, R. Krishnamurthy, S. Raghavan, S. Vaithyanathan, and H. Zhu, "Avatar information extraction system," *IEEE Data Engineering Bulletin*, vol. 29, pp. 40–48, 2006.

[114] J. Jiang and C. Zhai, "A systematic exploration of the feature space for relation extraction," in *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pp. 113–120, 2007.

[115] N. Kambhatla, "Combining lexical, syntactic and semantic features with maximum entropy models for information extraction," in *The Companion Volume to the Proceedings of 42st Annual Meeting of the Association for Computational Linguistics*, pp. 178–181, Barcelona, Spain: Association for Computational Linguistics, July 2004.

[116] S. Khaitan, G. Ramakrishnan, S. Joshi, and A. Chalamalla, "Rad: A scalable framework for annotator development," in *ICDE*, pp. 1624–1627, 2008.

[117] M.-S. Kim, K.-Y. Whang, J.-G. Lee, and M.-J. Lee, "n-gram/2l: A space and time efficient two-level n-gram inverted index structure," in *VLDB '05: Proceedings of the 31st International Conference on Very Large Data Bases*, pp. 325–336, 2005.

[118] D. Klein and C. D. Manning, "Conditional structure versus conditional estimation in NLP models," in *Workshop on Empirical Methods in Natural Language Processing (EMNLP)*, 2002.

[119] D. Koller and N. Friedman, "Structured probabilistic models," Under preparation, 2007.

[120] V. Krishnan and C. D. Manning, "An effective two-stage model for exploiting non-local dependencies in named entity recognition," in *ACL-COLING*, 2006.

[121] N. Kushmerick, "Wrapper induction for information extraction," PhD thesis, University of Washington, 1997.

[122] N. Kushmerick, "Regression testing for wrapper maintenance," in *AAAI/IAAI*, pp. 74–79, 1999.

[123] N. Kushmerick, D. Weld, and R. Doorenbos, "Wrapper induction for information extraction," in *Proceedings of IJCAI*, 1997.

[124] S. R. Labeling 2008. http://www.lsi.upc.es/ srlconll/refs.html.

[125] J. Lafferty, A. McCallum, and F. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Proceedings of the International Conference on Machine Learning (ICML-2001)*, Williams, MA, 2001.

[126] S. Lawrence, C. L. Giles, and K. Bollacker, "Digital libraries and autonomous citation indexing," *IEEE Computer*, vol. 32, pp. 67–71, 1999.

[127] W. Lehnert, J. McCarthy, S. Soderland, E. Riloff, C. Cardie, J. Peterson, F. Feng, C. Dolan, and S. Goldman, "Umass/hughes: Description of the circus system used for tipster text," in *Proceedings of a Workshop on Held at Fredericksburg, Virginia*, pp. 241–256, USA, Morristown, NJ: Association for Computational Linguistics, 1993.

[128] K. Lerman, S. Minton, and C. A. Knoblock, "Wrapper maintenance: A machine learning approach," *Journal of Artificial Intellgence Research (JAIR)*, vol. 18, pp. 149–181, 2003.

[129] X. Li and J. Bilmes, "A bayesian divergence prior for classifier adaptation," *Eleventh International Conference on Artificial Intelligence and Statistics (AISTATS-2007)*, 2007.

[130] Y. Li and K. Bontcheva, "Hierarchical, perceptron-like learning for ontology-based information extraction," in *WWW '07: Proceedings of the 16th International Conference on World Wide Web*, pp. 777–786, ACM, 2007.

[131] B. Liu, M. Hu, and J. Cheng, "Opinion observer: Analyzing and comparing opinions on the web," in *WWW '05: Proceedings of the 14th international conference on World Wide Web*, pp. 342–351, 2005.

[132] D. C. Liu and J. Nocedal, "On the limited memory bfgs method for large-scale optimization," *Mathematic Programming*, vol. 45, pp. 503–528, 1989.

[133] L. Liu, C. Pu, and W. Han, "Xwrap: An xml-enabled wrapper construction system for web information sources," in *International Conference on Data Engineering (ICDE)*, pp. 611–621, 2000.

[134] Y. Liu, K. Bai, P. Mitra, and C. L. Giles, "Tableseer: Automatic table metadata extraction and searching in digital libraries," in *JCDL '07: Proceedings of the 2007 Conference on Digital Libraries*, pp. 91–100, USA, New York, NY: ACM, 2007.

[135] R. Malouf, "Markov models for language-independent named entity recognition," in *Proceedings of the Sixth Conference on Natural Language Learning (CoNLL-2002)*, 2002.

[136] R. Malouf, "A comparison of algorithms for maximum entropy parameter estimation," in *Proceedings of the Sixth Conference on Natural Language Learning (CoNLL-2002)*, pp. 49–55, 2002.

[137] C. D. Manning and H. Schütze, *Foundations of Statistical Natural Language Processing*. Cambridge, MA: The MIT Press, 1999.

[138] I. Mansuri and S. Sarawagi, "A system for integrating unstructured data into relational databases," in *Proceedings of the 22nd IEEE International Conference on Data Engineering (ICDE)*, 2006.

[139] S. Mao, A. Rosenfeld, and T. Kanungo, "Document structure analysis algorithms: A literature survey," *Document Recognition and Retrieval X*, vol. 5010, pp. 197–207, 2003.

[140] B. Marthi, B. Milch, and S. Russell, "First-order probabilistic models for information extraction," in *Working Notes of the IJCAI-2003 Workshop on Learning Statistical Models from Relational Data (SRL-2003)*, (L. Getoor and D. Jensen, eds.), pp. 71–78, Acapulco, Mexico, August 11 2003.

[141] D. Maynard, V. Tablan, C. Ursu, H. Cunningham, and Y. Wilks, "Named entity recognition from diverse text types," *Recent Advances in Natural Language Processing 2001 Conference*, Tzigov Chark, Bulgaria, 2001.

[142] A. McCallum, "Information extraction: Distilling structured data from unstructured text," *ACM Queue*, vol. 3, pp. 48–57, 2005.

[143] A. McCallum, D. Freitag, and F. Pereira, "Maximum entropy markov models for information extraction and segmentation," in *Proceedings of the International Conference on Machine Learning (ICML-2000)*, pp. 591–598, Palo Alto, CA, 2000.

[144] A. McCallum, K. Nigam, J. Reed, J. Rennie, and K. Seymore, *Cora: Computer Science Research Paper Search Engine*, http://cora.whizbang.com/, 2000.

[145] A. McCallum and B. Wellner, "Toward conditional models of identity uncertainty with application to proper noun coreference," in *Proceedings of the IJCAI-2003 Workshop on Information Integration on the Web*, pp. 79–86, Acapulco, Mexico, August 2003.

[146] A. K. McCallum, *Mallet: A Machine Learning for Language Toolkit*. http://mallet.cs.umass.edu, 2002.

[147] D. McDonald, H. Chen, H. Su, and B. Marshall, "Extracting gene pathway relations using a hybrid grammar: The arizona relation parser," *Bioinformatics*, vol. 20, pp. 3370–3378, 2004.

[148] R. McDonald, K. Crammer, and F. Pereira, "Flexible text segmentation with structured multilabel classification," in *HLT/EMNLP*, 2005.

[149] G. Mecca, P. Merialdo, and P. Atzeni, "Araneus in the era of xml," in *IEEE Data Engineering Bullettin, Special Issue on XML, IEEE*, September 1999.

[150] M. Michelson and C. A. Knoblock, "Semantic annotation of unstructured and ungrammatical text," in *Proceedings of the 19th International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 1091–1098, 2005.

[151] M. Michelson and C. A. Knoblock, "Creating relational data from unstructured and ungrammatical data sources," *Journal of Artificial Intelligence Research (JAIR)*, vol. 31, pp. 543–590, 2008.

[152] E. Minkov, R. C. Wang, and W. W. Cohen, "Extracting personal names from email: Applying named entity recognition to informal text," in *HLT/EMNLP*, 2005.

[153] R. J. Mooney and R. C. Bunescu, "Mining knowledge from text using information extraction," *SIGKDD Explorations*, vol. 7, pp. 3–10, 2005.

[154] I. Muslea, "Extraction patterns for information extraction tasks: A survey," in *The AAAI-99 Workshop on Machine Learning for Information Extraction*, 1999.

[155] I. Muslea, S. Minton, and C. Knoblock, "Selective sampling with redundant views," in *Proceedings of the Fifteenth National Conference on Artificial Intelligence, AAAI-2000*, pp. 621–626, 2000.

[156] I. Muslea, S. Minton, and C. A. Knoblock, "A hierarchical approach to wrapper induction," in *Proceedings of the Third International Conference on Autonomous Agents*, Seattle, WA, 1999.

[157] I. Muslea, S. Minton, and C. A. Knoblock, "Hierarchical wrapper induction for semistructured information sources," *Autonomous Agents and Multi-Agent Systems*, vol. 4, nos. 1/2, pp. 93–114, 2001.

[158] A. Niculescu-Mizil and R. Caruana, "Predicting good probabilities with supervised learning," in *ICML*, 2005.

[159] NIST. Automatic content extraction (ACE) program. 1998–present.

[160] B. Pang and L. Lee, "Opinion mining and sentiment analysis," *Foundations and Trends in Information Retrieval*, vol. 2, nos. 1–2, pp. 1–135, 2008.

[161] Parag and P. Domingos, "Multi-relational record linkage," in *Proceedings of 3rd Workshop on Multi-Relational Data Mining at ACM SIGKDD*, Seattle, WA, August 2004.

[162] H. Pasula, B. Marthi, B. Milch, S. Russell, and I. Shpitser, "Identity uncertainty and citation matching," in *Advances in Neural Processing Systems 15*, Vancouver, British Columbia: MIT Press, 2002.

[163] F. Peng and A. McCallum, "Accurate information extraction from research papers using conditional random fields," in *HLT-NAACL*, pp. 329–336, 2004.

[164] D. Pinto, A. McCallum, X. Wei, and W. B. Croft, "Table extraction using conditional random fields," in *SIGIR '03: Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval*, pp. 235–242, USA, New York, NY: ACM, 2003.

[165] A. Pivk, P. Cimiano, Y. Sure, M. Gams, V. Rajkovič, and R. Studer, "Transforming arbitrary tables into logical form with tartar," *Data Knowledge Engineering*, vol. 60, pp. 567–595, 2007.

[166] C. Plake, T. Schiemann, M. Pankalla, J. Hakenberg, and U. Leser, "Alibaba: Pubmed as a graph," *Bioinformatics*, vol. 22, pp. 2444–2445, 2006.

[167] A.-M. Popescu and O. Etzioni, "Extracting product features and opinions from reviews," in *HLT '05: Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pp. 339–346, 2005.

[168] F. Popowich, "Using text mining and natural language processing for health care claims processing," *SIGKDD Explorartion Newsletter*, vol. 7, pp. 59–66, 2005.

[169] K. Probst and R. Ghani, "Towards 'interactive' active learning in multi-view feature sets for information extraction," in *ECML*, pp. 683–690, 2007.

[170] J. R. Quinlan, "Learning logical definitions from examples," *Machine Learning*, vol. 5, 1990.

[171] L. Rabiner, "A tutorial on Hidden Markov Models and selected applications in speech recognition," in *Proceedings of the IEEE*, vol. 77, 1989.

[172] G. Ramakrishnan, S. Balakrishnan, and S. Joshi, "Entity annotation based on inverse index operations," in *EMNLP*, 2006.

[173] G. Ramakrishnan, S. Joshi, S. Balakrishnan, and A. Srinivasan, "Using ilp to construct features for information extraction from semi-structured text," in *ILP*, 2007.

[174] L. Ramaswamy, A. Iyengar, L. Liu, and F. Douglis, "Automatic fragment detection in dynamic web pages and its impact on caching," *IEEE Transactions on Knowledge and Data Engineering*, vol. 17, pp. 859–874, 2005.

[175] L. Ramaswamy, A. Iyengar, L. Liu, and F. Douglis, "Automatic detection of fragments in dynamically generated web pages," in *WWW*, pp. 443–454, 2004.

[176] J. Raposo, A. Pan, M. Álvarez, and N. Ángel Vira, "Automatic wrapper maintenance for semi-structured web sources using results from previous queries," in *SAC '05: Proceedings of the 2005 ACM Symposium on Applied Computing*, pp. 654–659, ACM, 2005.

[177] A. Ratnaparkhi, "Learning to parse natural language with maximum entropy models," *Machine Learning*, vol. 34, 1999.

[178] L. Reeve and H. Han, "Survey of semantic annotation platforms," in *SAC '05: Proceedings of the 2005 ACM Symposium on Applied Computing*, pp. 1634–1638, USA, New York, NY: ACM, 2005.

[179] F. Reiss, S. Raghavan, R. Krishnamurthy, H. Zhu, and S. Vaithyanathan, "An algebraic approach to rule-based information extraction," in *ICDE*, 2008.

[180] P. Resnik and A. Elkiss, "The linguist's search engine: An overview (demonstration)," in *ACL*, 2005.

[181] E. Riloff, "Automatically constructing a dictionary for information extraction tasks," in *AAAI*, pp. 811–816, 1993.

[182] B. Rosenfeld and R. Feldman, "Using corpus statistics on entities to improve semi-supervised relation extraction from the web," in *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pp. 600–607, June 2007.

[183] R. Ross, V. S. Subrahmanian, and J. Grant, "Aggregate operators in probabilistic databases," *Journal of ACM*, vol. 52, pp. 54–101, 2005.

[184] A. Sahuguet and F. Azavant, "Building light-weight wrappers for legacy web data-sources using w4f," in *International Conference on Very Large Databases (VLDB)*, 1999.

[185] S. Sarawagi, *The CRF Project: A Java Implementation*. http://crf.sourceforge.net, 2004.

[186] S. Sarawagi, "Efficient inference on sequence segmentation models," in *Proceedings of the 23rd International Conference on Machine Learning (ICML)*, Pittsburgh, PA, USA, 2006.

[187] S. Sarawagi and A. Bhamidipaty, "Interactive deduplication using active learning," in *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining(KDD-2002)*, Edmonton, Canada, July 2002.

[188] S. Satpal and S. Sarawagi, "Domain adaptation of conditional probability models via feature subsetting," in *ECML/PKDD*, 2007.

[189] K. Seymore, A. McCallum, and R. Rosenfeld, "Learning Hidden Markov Model structure for information extraction," in *Papers from the AAAI-99 Workshop on Machine Learning for Information Extraction*, pp. 37–42, 1999.

[190] W. Shen, A. Doan, J. F. Naughton, and R. Ramakrishnan, "Declarative information extraction using datalog with embedded extraction predicates," in *VLDB*, pp. 1033–1044, 2007.

[191] M. Shilman, P. Liang, and P. Viola, "Learning non-generative grammatical models for document analysis," *ICCV*, vol. 2, pp. 962–969, 2005.

[192] Y. Shinyama and S. Sekine, "Preemptive information extraction using unrestricted relation discovery," in *HLT-NAACL*, 2006.

[193] J. F. Silva, Z. Kozareva, V. Noncheva, and G. P. Lopes, "Extracting named entities. a statistical approach," in *Proceedings of the XIme Confrence sur le Traitement des Langues Naturelles — TALN, 19–22 Avril, Fez, Marroco*, (B. Bel and I. Merlien, eds.), pp. 347–351, ATALA — Association pour le Traitement Automatique des Langues, 04, 2004.

[194] P. Singla and P. Domingos, "Entity resolution with markov logic," in *ICDM*, pp. 572–582, 2006.

[195] S. Soderland, "Learning information extraction rules for semi-structured and free text," *Machine Learning*, vol. 34, 1999.

[196] F. M. Suchanek, G. Ifrim, and G. Weikum, "Combining linguistic and statistical analysis to extract relations from web documents," in *KDD '06: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 712–717, 2006.

[197] F. M. Suchanek, G. Kasneci, and G. Weikum, "Yago: A core of semantic knowledge," in *WWW '07: Proceedings of the 16th International Conference on World Wide Web*, pp. 697–706, 2007.

[198] B. M. Sundheim, "Overview of the third message understanding evaluation and conference," in *Proceedings of the Third Message Understanding Conference (MUC-3)*, pp. 3–16, San Diego, CA, 1991.

[199] C. Sutton and A. McCallum, "Collective segmentation and labeling of distant entities in information extraction," Technical Report TR # 04-49, University of Massachusetts Presented at ICML Workshop on Statistical Relational Learning and Its Connections to Other Fields, July 2004.

[200] K. Takeuchi and N. Collier, "Use of support vector machines in extended named entity recognition," in *The 6th Conference on Natural Language Learning (CoNLL)*, 2002.

[201] B. Taskar, "Learning structured prediction models: A large margin approach," PhD Thesis, Stanford University, 2004.

[202] B. Taskar, D. Klein, M. Collins, D. Koller, and C. Manning, "Max-margin parsing," in *EMNLP*, July 2004.

[203] B. Taskar, S. Lacoste-Julien, and M. I. Jordan, "Structured prediction, dual extragradient and bregman projections," *Journal on Machine Learning Research*, vol. 7, pp. 1627–1653, 2006.

[204] M. Theobald, G. Weikum, and R. Schenkel, "Top-k query evaluation with probabilistic guarantees," in *VLDB*, pp. 648–659, 2004.

[205] C. A. Thompson, M. E. Califf, and R. J. Mooney, "Active learning for natural language parsing and information extraction," in *Proceedings of 16th International Conference on Machine Learning*, pp. 406–414, Morgan Kaufmann, San Francisco, CA, 1999.

[206] E. F. Tjong Kim Sang and F. D. Meulder, "Introduction to the conll-2003 shared task: Language-independent named entity recognition," in *Seventh Conference on Natural Language Learning (CoNLL-03)*, (W. Daelemans and M. Osborne, eds.), pp. 142–147, Edmonton, Alberta, Canada: Association for Computational Linguistics, May 31–June 1, 2003. (In association with HLT-NAACL, 2003).

[207] A. Troussov, B. O'Donovan, S. Koskenniemi, and N. Glushnev, "Per-node optimization of finite-state mechanisms for natural language processing," in *CICLing*, pp. 221–224, 2003.

[208] I. Tsochantaridis, T. Joachims, T. Hofmann, and Y. Altun, "Large margin methods for structured and interdependent output variables," *Journal of Machine Learning Research (JMLR)*, vol. 6, pp. 1453–1484, September 2005.

[209] J. Turmo, A. Ageno, and N. Català, "Adaptive information extraction," *ACM Computer Services*, vol. 38, p. 4, 2006.

[210] P. D. Turney, "Cost-sensitive classification: Empirical evaluation of a hybrid genetic decision tree induction algorithm," *Journal of Artificial Intelligence Research*, pp. 369–409, 1995.

[211] P. D. Turney, "Expressing implicit semantic relations without supervision," in *ACL*, 2006.

[212] V. S. Uren, P. Cimiano, J. Iria, S. Handschuh, M. Vargas-Vera, E. Motta, and F. Ciravegna, "Semantic annotation for knowledge management: Requirements and a survey of the state of the art," *Journal of Web Semantics*, vol. 4, pp. 14–28, 2006.

[213] P. Viola and M. Narasimhan, "Learning to extract information from semi-structured text using a discriminative context free grammar," in *SIGIR '05: Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 330–337, USA, New York, NY: ACM, 2005.

[214] S. V. N. Vishwanathan, N. N. Schraudolph, M. W. Schmidt, and K. P. Murphy, "Accelerated training of conditional random fields with stochastic gradient methods," in *ICML*, pp. 969–976, 2006.

[215] M. Wang, "A re-examination of dependency path kernels for relation extraction," in *Proceedings of IJCNLP*, 2008.

[216] Y. Wang and J. Hu, "A machine learning based approach for table detection on the web," in *WWW '02: Proceedings of the 11th International Conference on World Wide Web*, pp. 242–250, ACM, 2002.

[217] B. Wellner, A. McCallum, F. Peng, and M. Hay, "An integrated, conditional model of information extraction and coreference with application to citation matching," in *Conference on Uncertainty in Artificial Intelligence (UAI)*, 2004.

[218] M. Wick, A. Culotta, and A. McCallum, "Learning field compatibilities to extract database records from unstructured text," in *Proceedings of the*

2006 Conference on Empirical Methods in Natural Language Processing, pp. 603–611, Sydney, Australia: Association for Computational Lingistics, July 2006.

[219] I. H. Witten, A. Moffat, and T. C. Bell, *Managing Gigabytes: Compressing and Indexing Documents and Images*. Morgan Kaufmann Publishing, San Francisco, 1999.

[220] F. Wu and D. S. Weld, "Autonomously semantifying wikipedia," in *CIKM*, pp. 41–50, 2007.

[221] B. Zadrozny and C. Elkan, "Learning and making decisions when costs and probabilities are both unknown," in *Proceedings of the Seventh International Conference on Knowledge Discovery and Data Mining (KDD)*, 2001.

[222] R. Zanibbi, D. Blostein, and R. Cordy, "A survey of table recognition: Models, observations, transformations, and inferences," *International Journal on Document Analysis and Recognition*, vol. 7, pp. 1–16, 2004.

[223] D. Zelenko, C. Aone, and A. Richardella, "Kernel methods for relation extraction," *Journal of Machine Learning Research*, vol. 3, pp. 1083–1106, 2003.

[224] M. Zhang, J. Zhang, J. Su, and G. Zhou, "A composite kernel to extract relations between entities with both flat and structured features," in *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pp. 825–832, Sydney, Australia: Association for Computational Linguistics, July 2006.

[225] S. Zhao and R. Grishman, "Extracting relations with integrated information using kernel methods," in *ACL '05: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pp. 419–426, 2005.

[226] G. Zhu, T. J. Bethea, and V. Krishna, "Extracting relevant named entities for automated expense reimbursement," in *KDD*, pp. 1004–1012, 2007.