Abstract

Recognising Specific Named Entities in a New Restricted Domain Using Conditional Random Fields

By Igal Gabbay

Named-entity recognition (NER) plays a vital role in information extraction, question answering and text mining. Classic NER research activity has focused on tagging instances of PERSON, LOCATION and ORGANISATION in the newswire domain. New fine-grained NER (FG-NER) covers subtypes of the classic NEs. The goal of this study was to investigate an FG-NER scenario with a set of new specific NEs (SNEs) typical to a new restricted journalistic domain. Reports on birth of animals in zoos were identified as such a productive domain. A 700-document corpus (241K tokens) named ZooBirth was compiled from a newspaper archive and annotated. It contained 2,811 instances of the ten most frequent numerical SNEs shortlisted from 43 candidates. Using Conditional Random Fields allowed testing positional and order-within-document features which were hypothesized to improve tagging SNEs. In support of positional features, analysis of distribution of SNEs within documents yielded SNE-specific patterns. The feature token position produced statistically significant but modest improvement in the case of two SNEs (82.2 to 84.4 strict precision, and 59.5 to 61.1 F-measure). Order-effect features improved with statistical significance the F-measure when tagging the weight at birth (from 68.4 to 71.1 strict, and from 75.5 to 80.6 lenient). In the final stage of the study a novel technique named subtractive tagging was introduced to enrich negative examples when training CRF. When tagging the newborn animal's date of birth and the age of its mother strict recall improved from 52.8 to 60.1 and 65.5 to 68.9, respectively, with statistical significance.

Acknowledgments

I would like to thank my PhD supervisor Dr Richard F. E. Sutcliffe for his guidance and advice.

Many thanks to the following people in the University of Limerick for their help during my research and writing-up: Dr Maurice Collins, Dr Angélica Rísquez, Dr Kieran White, Dr Darina Slattery, Dr Ita Richardson, Gerard Lyons, and Noreen O'Shea.

Finally, I would like to thank my family for their ongoing support.

I hereby declare that this thesis is entirely my own work and that it has not been submitted for any academic award.

Signature of Author

November 15, 2012

Date

Table of Contents

List of Figures

List of Tables

Chapter 1: Introduction

1.1 Outline

The first chapter presents the motivation for and the objectives of the project. A prefatory description of the experimental domain and related resources follows. The experiments and their results are then summarized. The introduction concludes with a guide to the rest of the thesis.

1.2 Motivation

Named-entity recognition (NER) plays a vital role in information extraction, question answering and text mining. Despite extensive research activity, particularly in the newswire domain within evaluation events such as the Message Understanding Conference (MUC), and the high performance reported, NER should not be considered a solved problem. There is a drop in performance when state-of-the-art systems are ported to different corpora, other genres or the web. More recently NER research has expanded to include tagging NEs in more specialised domains such as biomedicine and natural sciences. Within these domains, NEs (eg, protein, organism) are not necessarily of higher granularity than the ones in traditional newswire. Granularity is the focus of the relatively new branch of NER termed fine-grained NER (FG-NER) which mainly aims to subcategorise MUC entities. The limited research and resources in this area were the prime motivation behind this project. So far most research in the newswire domain has been conducted either as part of the evaluations activity mentioned above or using their legacy of tagged corpora for development and testing. The availability of these resources may have not encouraged the development of new restricted-domain collections, while recent novel NER seems to be moving away from newswire.

1.3 Objectives

The goal of the study was to investigate an FG-NER scenario with a set of new Specific NEs (SNEs) typical to a particular new restricted domain.

Once the corpus was compiled and the set of SNEs developed, the objective was to test features which were hypothesized to be effective in improving tagging subtypes of NEs: more specifically, features that capture the position and order of SNEs within documents.

1.4 Domain

The domain chosen for this project was news stories reporting the birth of animals in zoos. It offers a range of fine-grained NEs such as the date of birth of the newborn, the age of the mother and the gestation duration of the species. News of this type appears daily in the media around the world, and this regularity was deemed to make it suited for FG-NER. Another reason for choosing this zoological domain was the author's expertise in the field thanks to previous work in zoological centres.

1.5 Experiments

The initial step was to compile an appropriate corpus. This was done via the archived news service Nexus® (NexisLexis 2008). Initially, the corpus, named *ZooBirth*, comprised 500 documents (174,652 tokens) and later was expanded to 700 (240,848 tokens). Next, 43 types of SNEs were identified as appearing regularly in the news reports. These were then annotated manually.

Conditional Random Fields (CRF) was chosen as the machine learning method as it is

widely considered to be state-of-the-art in NER tasks and allows the use of rich sets of features that can be interdependent. The experimental design was 5×2 cross-validation. The experiments covered three main areas:

1.5.1 Positional Features

In the first stage the distribution of patterns of SNEs within the documents was explored to provide support to the use of positional features. In the second stage, various positional features were tested. These included token, sentence and paragraph positions.

1.5.2 Order Effects

SNEs which are subtypes of the same NE were tagged using features that take into account their relative order in the document. The two NEs that were tested were date and weight.

1.5.3 Subtractive Tagging

A novel procedure was developed in which the training set was stripped of the SNE to be tagged and of its context, followed by training on the original version of the training set and testing on its excised version. This results in tagging false positives by CRF which are then used in the 'real' testing stage. Essentially, it is a method to automatically enrich negative examples during learning. The method was demonstrated with the SNE date-of-birth of the newborn animal (DOBN) and tested on a further three.

1.6 Key Results

1.6.1 Corpus

A corpus of 241k tokens has been compiled and annotated by hand. Over 5,400 instances of numerical entities were marked up, of which over 2,800 were the ten most frequent numerical entities in the collection; these SNEs were the focus of the study. The corpus and its entire set of numerical entities could prove a productive resource for future research on information extraction in novel restricted domains.

1.6.2 Positional Features

The feature token position improved performance with statistical significance in the case of two SNEs: AGM (Age of Mother) and DOBN (Date Of Birth of Newborn). Strict precision of AGM tagging was only slightly higher than the baselines (82.8 and 82.4, respectively). Strict F-measure of DOBN tagging was 61.1 compared to a baseline value of 59.5. Otherwise the observed trends were mixed and suggest that positional features are weak on their own and perhaps only effective with certain SNEs.

1.6.3 Order Effects

In an experiment with weight SNEs, a simple count of the number of instances of weight units in the document, and a feature indicating if a weight unit instance was the first in the document, not first, or a single instance, improved with statistical significance the F-measure when tagging instances of WB (Weight of newborn at Birth) from 68.4 to 71.1 strict, and from 75.5 to 80.6 lenient and WEIGHT (any other type of weight) from 32.3 to 36.2 strict, and from 35 to 39.7 lenient).

1.6.4 Subtractive Tagging

When tagging DOBN the strict precision was reduced from 73.3 to 70.8 without statistical significance while strict recall improved from 52.8 to 60.1 with statistical significance. When tagging instances of AGM, strict precision declined slightly from 82.4 to 82.3 (without statistical significance), whereas recall improved from 65.5 to 68.9 and with statistical significance

1.7 Guide to Other Chapters

Chapter 2 is a literature review which covers information extraction, NER, fine-grained NER, Conditional Random Fields and text structure.

Chapter 3 describes the domain of animal births in zoos, the creation of the test corpus and contains an exhaustive list of SNEs illustrated with extracts from the corpus.

Chapter 4 covers the machine learning method used in this study and the experimental design (5×2 cross-validation).

Chapter 5 presents experiments in which positional features were used to train CRF to tag SNEs.

Chapter 6 reports on experiments which exploit the relative order of SNEs as features in CRF.

Chapter 7 introduces subtractive tagging, a suggested method to enrich negative examples when training CRF.

Chapter 8 sums up key findings and briefly discusses possible directions for further research.

Chapter 2: Background and Related Work

2.1 Outline

This chapter begins with a brief introduction to information extraction (IE), its history and main tasks. The following sections focus on named entity recognition (NER) and fine grained NER in particular. Next is an overview of the Conditional Random Field (CRF) model applied in this thesis' project. Finally, a background on the order of content in documents as a basis for information extraction is given.

2.2 Information extraction

2.2.1 A Real-World Information Extraction Task

When researchers in Boston University wanted to build a database which compares the sleep characteristics of 127 mammalian species (McNamara et al. 2008), they had to review 180 papers and manually extract data, such as values of daily sleep time, non-REM sleep, and sleep cycle length, as well as information on laboratory conditions (Figure 1). It is obvious that the scientists would have benefited from a system that could flag and label the relevant text snippets automatically. The following are excerpts from three papers which were analysed for the Phylogeny of Sleep Database (Boston University, 2007). The text in bold denotes information included in the database.

> The mean period of these REMS-SWS-REMS cycles was **11.8 min** (SEM . 2:4).
>
> (Nicol et al. 2000)

> The behaviour of **8 giraffes (*Giraffa camelopardalis*)**… was recorded continuously on two time-lapse video recorders (Panasonic AG-6720 A) **6–25 d** during the hours when the animals were in the animal house of a zoological garden
>
> (Tobler and Schwierin 1996)

> **Five male adult** specimens of the rodent ***Neotomodon alstoni alstoni***, weighing between 75 and 85 g, were used in this experimental work.
>
> (Ayala-Guerrero et al. 1998)

PHYLOGENY OF SLEEP DATABASE

Contact

HOMEPAGE SEARCH SUBMIT STUDY

« Return to Search Results

Nicol, S. C., Andersen, N. A., Phillips, N. H., & Berger, R. J. (2000). The Echidna Manifests Typical Characteristics Of Rapid Eye Movement Sleep. Neuroscience Letters, 283(1), 49-52.

STUDY

Species Name:	Tachyglossus Aculeatus	Common Name:	Echidna
Sex:	Mix	Age Class:	Adult & Juvenile
Number of Animals:	6	Animals in Mean:	6
Animals Sampled:	6	Age Average:	N/A

LAB CONDITIONS

EEG or ECoG Used:	Yes	Temperature:	No
Recording Length:	3 (24 hour recording period)	Light Conditions:	Normal light/dark conditions
Animal Allowed to Adapt:	Yes	Animal Restrained:	No
Tested in Wild:	No	Diet:	Normal
Behavioral Sleep Recorded:	1	Lab Condition Score:	11

SLEEP DATA (times in hours, unless otherwise specified)

Daily Sleep Time:	16.56	Daily Sleep Time Adjusted for Drowsiness:	N/A
Daily REM/Paradoxical Sleep Time:	1.032	Quiet/Non-REM Sleep Sleep Time:	15.528
Sleep Cycle Length in Minutes:	11.8	Monophasic or Polyphasic Sleep:	N/A

Figure 1 A screenshot of the web-based Phylogeny of Sleep Database (Boston University, 2007). The table summarizes the sleep characteristics of the echidna as extracted from the reference above the table (Nicol et al. 2000).

The scientific endeavour above illustrates the need to automate the task of information extraction (IE): to structure and combine selective data which is stated in natural language texts (Cowie and Wilks 2000). Once pre-specified information from this

unstructured source language populates a database, it can be interrogated and analysed by computers through queries or data mining for summarization (Mallett et al. 2004, White et al. 2001), translation (Aone et al. 1997, Son Bao et al. 2009), question answering (Srihari and Li 2000, Sutcliffe 2002, Sutcliffe et al. 2003, Sutcliffe et al. 2005) and text mining (Feldman and Sanger 2006, Mooney and Nahm 2003).

The Phylogeny of Sleep Database IE project may seem limited when compared to more recent attempts to extract open-domain information from the vast web corpora (Banko et al. 2007). However, the ambiguity and idiosyncrasy of natural language still pose a challenge in restricted domains. In addition, the ever growing information overload (Floridi 2009), regularly cited in introductions to IE papers, exists within almost every specialised field (Lok 2010) and motivates much of the research activity in IE.

When introducing IE, many authors make the distinction between IE and information retrieval (IR). IR preceded IE and its aim is to present to the user a subset of documents from a large collection in response to a query. IE systems complement IR by helping the user to find information within documents. Hence, IR is often the initial stage in IE systems (Gaizauskas and Wilks 1998).

2.2.2 The Development of Information Extraction

Cowie and Lehnart were the first to review IE in 1996 (Cowie and Lehnert 1996) but the roots of the field are in the 1960s (Gaizauskas and Wilks 1998). The work of Naomi Sager of the Linguistic String Project group at New York University is cited as one of the earliest IE projects. It was conducted in the medical domain and focused on deriving ‘information formats’ (i.e., templates) from radiology reports and hospital discharge summaries.

The Conceptual Dependency Theory (CDT) of Roger Schank at Yale University led to the development of IE tool prototypes in the 1970s (Moens 2006). This influential theory assumed an interlingual conceptual base to linguistic structures. Its aim was to parse texts into formal semantic representations. A CD script is meant to encompass all information about any event with predictable role players and sub events. Implementations of CDT used simplified forms named sketchy scripts. These only partially analysed the text and contained only the most relevant conceptualizations. The best known implementation of CD is Gerald DeJong's FRUMP (Fast Reading Understanding and Memory Program). The system employed 60 types of sketchy scripts to extract information from new stories (Gaizauskas and Wilks 1998). FRUMP matched every new story with the correct script on the basis of keywords and conceptual sentence analysis. Domain-specific expectations were used to instantiate descriptions of events based on the sketchy scripts.

In the early 1980s DaSilva and Dwiggins extracted satellite-flight information from global reports. The system was limited to single sentences and couldn't extract complete event descriptions (DaSilva and Dwiggins 1980).

Also in the 1980s, ATRANS was the first commercial IE system. Like FRUMP, It was based on scripts, and designed to automate processing of money transfer messages. The system identified script actors such as originating customer, originating bank, receiving bank to fill in a template that was used, after human inspection, to initiate money transfer (Lytinen and Gershman 1986). Other commercial systems that were developed in the 1980s included JASPER (Andersen et al. 1992) and SCISOR (Jacobs and Rau 1990). JASPER extracted information about earnings and dividends from company press releases to help journalists at Reuters validate and post-edit stories. SCISOR analysed corporate mergers and acquisitions. Two academic IE projects during that time were James Cowie's system to extract descriptions from wild flower guides

(Cowie 1983), and that of Gianpiero Zarri, to capture semantic biographical relations in French texts on historical figures (Zarri 1983).

In the mid-1980s the US Navy sponsored research to extract information from naval messages. The need to compare the performance of different systems developed for this task gave rise to the Message Understanding Conferences (MUC) which were sponsored by the American Defense Advanced Research Projects Agency (DARPA) (Grishman and Sundheim 1996). Seven MUCs altogether were held between 1987 and 1997 with the objective to establish a quantitative evaluation regime. They are considered to have been fruitful with significant contribution to the field of IE (Hobbs and Riloff 2010). For example, the annotated corpora they provided are still in use as standard test beds (Ireson et al. 2005). Table 1 summarizes key points from MUC-1 to MUC-7.

Table 1: a summary of the Message Understanding Conferences (MUC) 1 to 7.

Conference	Year	No. of Research Sites	Text	Domain	Task	Evaluation
MUC-1	1987	6	Naval reports	Naval operations	None defined	None / exploratory
MUC-2	1989	8	Naval reports	Naval operations	Fill a 10-slot template	Criteria developed but deemed inadequate
MUC-3	1991	15	News reports	Terror in Latin America	Fill an 18-slot template	Introduction of Precision and Recall from IR
MUC-4	1992	17	News reports	Terror in Latin America	Fill an 24-slot template	F measure introduced
MUC-5	1993	17	News reports	Joint ventures, microelectronics products both in Japanese and English	11 templates with 47 slots	Error Per Response introduced
MUC-6	1995	17	News reports	Corporate Management Succession	4 subtasks	Recall and Precision reinstated
MUC-7	1997	17	News reports	Aeroplane crashes, Missile Launches	6 subtasks	Same

The subtasks that were introduced in the MUC evaluations are now seen as typical IE tasks and will be discussed in the next section.

The Automatic Content Extraction programme (ACE) has been a successor to MUC since 1999 (Doddington et al. 2004) but differs in conflating a few of the MUC tasks, and increasing the tasks' complexity (Cunningham 2005). The reference corpus includes newswire, broadcast news and scans of newspapers.

Information extraction has been applied in a wide range of domains beyond military intelligence and finance (Hobbs and Riloff 2010), including bioinformatics (Blaschke et al. 2002, Gaizauskas et al. 2003, Humphreys et al. 2000, Krallinger et al. 2008, Ono et al. 2001, Shah et al. 2003, Skusa et al. 2005), law (Brüninghaus and Ashley 2001, Moens et al. 1999), clinical reports (de Bruijn et al. 2011, Demner-Fushman et al. 2009, Soysal et al. 2010, Xu et al. 2010), employment (Kessler et al. 2007, Loth et al. 2010, Wong et al. 2009), product reviews (Ghani et al. 2006, Hu and Liu 2004, Popescu and Etzioni 2005, Wong et al. 2008) and information extraction from the web (Chang et al. 2006, Etzioni et al. 2004, Ferrara et al. 2010), and in particular from informal, noisy text on social networks and blog pages (Bollen et al. 2011, Matsuo et al. 2007, Mendes et al. 2010, Moens 2009, Ritter et al. 2011), have been a very active field of research in recent years.

2.2.3 Typical Information Extraction Tasks

The MUC programme had split IE into the following tasks (Feldman and Sanger 2006):

1. Named Entity Recognition (NER)
2. Co-reference Resolution (CR)
3. Template Element construction (TE)
4. Template Relation construction (TR)
5. Scenario Template production (ST)

There are additional IE tasks such as time line recognition, but the above five are the most typical open domain tasks and the most extensively researched.

Table 2 describes the goal of each of the five tasks and demonstrates them using shortened examples from MUC-7 (Chinchor 2001). Figure 2 shows named entities marked up in a sentence from a document used in MUC-6.

The following sections will expatiate on the NER task, and in particular fine-grained NER, which is the focus of this thesis.

```
<s> The proposal to meet followed an announcement <TIMEX
TYPE="DATE">Wednesday</TIMEX> in which <ENAMEX
TYPE="PERSON">Philip Bakes</ENAMEX>, <ENAMEX
TYPE="ORGANIZATION">Eastern</ENAMEX>'s president, laid out pro-
posals to cut wages selectively an average of <NUMEX
TYPE="PERCENT">29%</NUMEX>. </s>
<s> The airline's three major labor unions, whose contracts
don't expire until year's end at the earliest, have vowed to re-
sist the cuts. </s>
```

Figure 2: Annotated sentence from MUC-6 Document No. 870123-0009 (Moens 2006)

Table 2 Typical IE tasks

Task	Goal	MUC Output Excerpt
NER	Identify mentions of proper names, dates and times, and quantities	The <ENAMEX TYPE="LOCATION">**U.K**.</ENAMEX> satellite television broadcaster said its subscriber base grew <NUMEX TYPE="PERCENT">**17.5 percent**</NUMEX> during <TIMEX TYPE="DATE">**the past year**</TIMEX> to 5.35 million
CR	Recognise identity between entities	***The U.K. satellite television broadcaster*** said ***its*** subscriber base grew 17.5 percent during the past year to 5.35 million
TE	Add descriptive information to NE results	ENT_NAME: "Dennis Gillespie" ENT_TYPE: **PERSON** ENT_DESCRIPTOR: "**Capt**." "the commander of Carrier Air Wing 11" ENT_CATEGORY: **PER_MIL**
TR	Identify relations between entities	<EMPLOYEE_OF-9602040136-5> := PERSON: <"Dennis Gillespie"> ORGANIZATION: <"NAVY">
ST	Fit TE and TR results into an event scenario	<LAUNCH_EVENT-9602140509-1> := VEHICLE_INFO: <VEHICLE_INFO-9602140509-1> PAYLOAD_INFO: <PAYLOAD_INFO-9602140509-1> LAUNCH_DATE: <TIME-9602140509-1> <TIME-9602140509-2> LAUNCH_SITE: <LOCATION-9602140509-1> MISSION_TYPE: CIVILIAN MISSION_FUNCTION: DEPLOY MISSION_STATUS: FAILED

2.3. Named Entity Recognition

2.3.1 What is meant by Named Entity?

Although the term ‘named entity’ had appeared sporadically as early as the 1970s in various fields, eg databases (DeRemer and Kron 1976, Patterson 1971) , in the context of information extraction it was coined in 1995 for the Sixth Message Understanding Conference (MUC-6) (Grishman and Sundheim 1995). The definition of the term followed the observation by the organisers of MUC that in order to perform information extraction it is essential to recognise information units such as names of people, organisations, locations, and numeric expressions (Nadeau and Sekine 2007). Early work prior to MUC-6 considered NER as the problem of recognizing proper names (Nadeau and Sekine 2007, Wolinski et al. 1995). In their review, Nadaeau and Sekine explain that the word ‘named’ restricts the task to entities with one or more rigid designators, as defined by the philosopher Saul Kripke (Kripke 1980). Rigid designators designate the same object wherever it exists and nothing else. Kripke argues that proper names are rigid designators. There is a general agreement among the NER community about the inclusion of temporal expressions and some numerical expressions. Some temporal expressions, such as a name of a month on its own, are invalid but for practical reasons the NE definition may be relaxed.

2.3.2 Brief outline of Named Entity Recognition research

As shown earlier, NER is a sub-problem of IE. And since IE is crucial to support higher level NLP tasks such as question answering, summarization, translation and text mining, NER’s role is fundamental. It can also augment tasks such as text classification (Armour et al. 2005).

One of the earliest papers in NER was presented by Lisa Rau at the seventh IEEE Conference on Artificial Intelligence Applications (Rau 1991) titled 'Extracting company names from text'. MUC-6 is considered a milestone in being the first high profile event dedicated to the task. Thereafter the rate of publication in the field accelerated significantly (Nadeau and Sekine 2007).

NER research has been driven by multiple evaluation forums in different languages. Other scientific events which followed MUC-6 included: Multilingual Entity Task Conference (MET) in Chinese and Japanese (Merchant et al. 1996), Special Interest Group on Chinese Language Processing (SIGHAN) (Zhu et al. 2003), evaluation contest for NER in Portuguese (HAREM) (Santos and Cardoso 2006), Information Retrieval and Extraction Exercise (IREX) in Japanese (Sekine and Eriguchi 2000), Conference on Computational Natural Language Learning (CoNLL) in Spanish, Dutch, German and English (Sang and Meulder 2003), TAC Knowledge Base Population Evaluation (TAC/KBP) (NIST 2011), and the Language Resources and Evaluation Conference (LREC) (ELRA 2011) which has been organizing workshops and tracks on the topic since 2000.

Apart from expanding the language coverage of NER beyond English, language-independent NER has also been investigated (Benajiba et al. 2009, Sang and Meulder 2003, Tjong Kim Sang and De Meulder 2003) as well as domain-independent NER .

The focus of NER research, as that of IE in general, has shifted to extraction of entities from the web (Cafarella et al. 2008, Carme et al. 2006, Ekbal and Bandyopadhyay 2008, Etzioni et al. 2004, Ferrara et al. 2010, Hidalgo et al. 2005, Moens 2009), with Google Inc. exploring NER on an unprecedented scale (Whitelaw et al. 2008). Working on web data also presents the challenge of dealing with poorly formatted and informal text (see next section) found on social networks and blogs (Finin et al. 2010).

On a practical level NER has been applied in both commercial and open source systems. Commercial systems which incorporate NER include Rosette Extractor (REX) (Basis-Technology 2011), ClearForest (Thomson-Reuters 2012), Inxight (SAP 2011), PolyAnalyst (Megaputer 2012), and SRA NetOwl (SRA 2012). Well known open source NER systems are Annie (AKT-Technologies 2012) and MinorThird (Cohen 2012).

So far there has been no mention of the main categories of methods (rule-based, supervised, unsupervised, hybrid) employed by NER systems. These will be reviewed in section 4 in relation to fine-grained entities.

2.3.3 The Named Entity Recognition problem

In MUC, systems were expected to identify expressions referring to organisations, locations, persons (`ENAMEX`), dates and times (`TIMEX`), and percentage and currency (`NUMEX`). The NER task therefore consists of two subtasks (Mansouri et al. 2008): Identifying proper names in the text and classifying the names into a set of predefined categories.

For humans NER looks easy. They have no problem with the lack of surface contextual cues often relied on by NER systems and can use semantics and predicate-argument selectional restrictions to recognize a vast set of entities (Vilain et al. 2007).

Initially it seems that a dictionary of proper names can solve the problem. However, new proper names keep appearing, so such a dictionary will have to be perpetually updated. Then there is the problem common to all NLP tasks, that of ambiguity. For example, *Darwin* and *Paris* can be names of persons but can also refer to a location;

White House can denote a location or an organisation; O_2 is a chemical element or a telecommunication provider; *Penguin* can refer to an organization (publisher), location (a town in Tasmania), or to a character in a comic book (Batman).

Vilain et al. (2007) challenge the preconception that NER is a 'boring solved problem'. By testing NER taggers that were developed to recognise entities in the journalistic newswire corpus of MUC-5 on business related documents compiled from reports and business websites, they demonstrated how poorly these systems performed when ported to a new domain which wasn't too dissimilar. The high performance of NER systems achieved in MUC (with F scores in the mid-90s) may belie their true brittle nature. The authors suggested that the drop in performance, as much as 30%, can be attributed to certain differences between journalistic and business texts: names of organisations are more prevalent in business texts and the most common type in such texts is less likely to be identified using gazetteers/dictionaries, a method which seems to be effective when recognising governmental or quasi-governmental organization names in the MUC corpus. A difference in editorial standards may also contribute to the decline in performance when transferring taggers across texts with different styles. For example, in business-related texts 'inc.' may be dropped from the names of companies. The well curated, pristine news texts may be atypically easy to tag.

Nadeau and Sekine (2007) also claim that the impact of the textual genre has been neglected in NER research. Poineau and Kosseim (2001) concluded that techniques developed for the newswire genre are generally not sufficient to deal with larger corpora containing texts that do not follow strict writing constraints (for example, technical e-mail messages, transcriptions of phone conversations).

The open web presents even more difficulties: the vast scale (millions of instances need to be recognized), the abundance of informal text, the impracticality of generating a

satisfactory training data set when using supervised learning and the difficulty of performance evaluation (Whitelaw et al. 2008).

One more aspect of the 'transfer gap' problem, when porting an NER system either to the web or to a corpus of a different genre, is the need to identify new types of named-entities, often subtypes of the ones defined in MUC. The problem should also exist when staying within the same genre yet attempting to recognise new, more specific types of named-entities. This was the focus of the project presented in this thesis and it is closely related to the somewhat neglected area of fine-grained NER which will be reviewed in the following section.

2.4. Fine-Grained NER (FG-NER)

2.4.1 Work on Fine-Grained Named Entities

It has become evident that the basic named-entity types are insufficient for the needs of NLP applications such as question answering, search engines and ontology population. For example, to be able to answer the question 'who was the US president in 1994?' a QA system would have to recognise a `PERSON` entity as the subtype 'president'. Similarly, when searching for the name of a person who won a Nobel Prize in Physics or received an MTV Music Video Award, it should look for a scientist and a singer, respectively (Kozareva et al. 2008). The NE granularity should be determined by the domain of application.

Sekine et al. (2002) developed an extended hierarchy of about 200 categories of named entities (see Figure 3). It attempted to capture the most frequent name types appearing in newspapers. Fleischman and Hovy (2002) automatically classified `PERSON` instances into eight finer-grained subcategories: athlete, politician/government, clergy, businessperson, entertainer/artist, lawyer, doctor/dentist, and police. Kozareva et al.

(2008) experimented with labelling a small set of names as either `PRESIDENT` or `SINGER`. White and Sutcliffe (2011) developed a method to determine the occupation of a person from syntactic data. They recognised 39 occupations, each held by at least 20 people mentioned in their test corpus.

Appendix: Extended NE hierarchy

```
TOP
 NAME
  PERSON                 # Bill Clinton, George W. Bush, Satoshi Sekine,
   LASTNAME              # Clinton, Bush, Sekine,
   MALE_FIRSTNAME        # Bill, George, Satoshi,
   FEMALE_FIRSTNAME      # Mary, Catherine, Ilene, Yoko

  ORGANIZATION           # United Nations, NATO
   COMPANY               # IBM, Microsoft
   COMPANY_GROUP         # Star Alliance, Tokyo-Mitsubishi Group
   MILITARY              # The U.S Navy
   INSTITUTE             # the National Football League, ACL
   MARKET                # New York Stock Exchange, NASDAQ
   POLITICAL_ORGANIZATION #
    GOVERNMENT           # Department of Education, Ministry of Finance
    POLITICAL_PARTY      # Republican Party, Democratic Party, GOP
    PUBLIC_INSTITUTION   # New York Post Office,
   GROUP                 # The Beatles, Boston Symphony Orchestra
    SPORTS_TEAM          # the Chicago Bulls, New York Mets
   ETHNIC_GROUP          # Han race, Hispanic
   NATIONALITY           # American, Japanese, Spanish

  LOCATION               # Times Square, Ground Zero
   GPE                   # Asia, Middle East, Palestine
    CITY                 # New York City, Los Angeles
    COUNTY               # Westchester
    PROVINCE             # State (US), Province (Canada), Prefecture (Japan)
    COUNTRY              # the United States of America, Japan, England
   REGION                # Scandinavia, North America, Asia, East coast
   GEOLOGICAL_REGION     # Altamira
    LANDFORM             # Rocky Mountains, Manzano Peak, Matterhorn
```

Figure 3: a section of the extended named-entity hierarchy devised by Sekine et al. (2002)

In their work on named-entity extraction from the web, Etzioni et al (2005) automatically extracted subclasses, for instance recognising biologists, chemists and physicists as subclasses of 'scientist'. The goal of Whitelaw et al. (2008) who conducted NER on the entire web, was to annotate all mentions of entities of hundreds of types. Their entity types fall into a hierarchy, with the highest level containing the

most MUC-like types and a longer list of less common but still important types such as titles of films, books, names of animals and plants, and cars. In their experimental work they selected a set of 32 labels. Ling and Weld (2012) introduced a set of 112 NE types derived from the collaborative knowledge base Freebase (see figure 4). This gave them broad coverage of entities and allowed tagging entities with multiple overlapping classes.

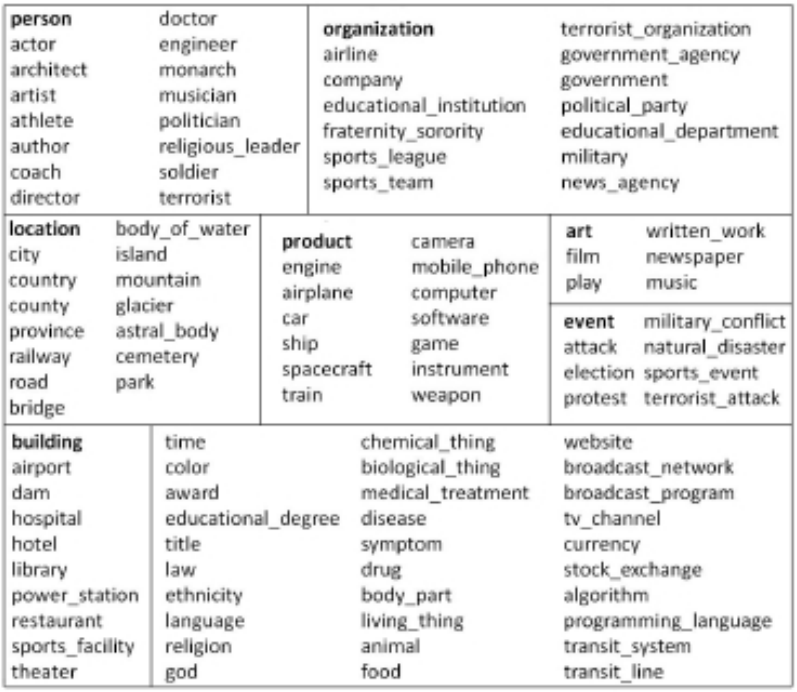

Figure 4: 112 NE tags used by Ling and Weld (2012). The set is derived from the knowledge base Freebase. The box at the bottom right corner contains mixed tags that are hard to categorise.

The examples above can be described as open-domain NER, either from the web or news corpora. There is by now a substantial amount of NER work in specialised, often scientific domains, including geography, chemistry (Corbett and Copestake 2008, Kolluru et al. 2011), geology (Sobhana et al. 2010), astronomy (Murphy et al. 2006) and in particular biomedicine (Kim and Yoon 2007, Lee et al. 2004, Song et al. 2005, Sung et al. 2006, Zhang et al. 2004, Zhou et al. 2004). NER in biomedical texts has become part of bioinformatics and has been driven by the availability of GENIA, the

largest semantically annotated corpus for bio text mining available to the public (Kim et al. 2003). This reflects a surge of interest in mining biomedical literature, e.g., MEDLINE abstracts (Chun et al. 2006, Perez-Iratxeta et al. 2001). The biomedical field has its own challenge evaluations such as BioCreative Gene Mention Recognition (Smith et al. 2008) and BioNLP (Kim et al. 2009). Work in this field is mainly trying to recognise instances of genes, proteins, gene products, organisms, drugs, diseases and chemical compounds. These are harder to recognise than traditional NEs (Smith et al. 2008, Tsai et al. 2006, Zhou et al. 2004) but in their granularity may not necessarily be called fine-grained considering the numerous subtypes each subsumes. At this point it is worth noting from a historical perspective that in the MUC events the entities at the level of `PERSON`, `LOCATION`, `ORGANISATION`/`PERCENT`, `MONEY`/`DATE`, `TIME` were referred to as NE subcategories (of `ENAMEX`, `NUMEX` and `TIMEX`, respectively).

More specific entities are appearing in the literature though. For example, Ananiadou et al. (2011) have developed tools for NER of four entities related to Type IV secretion systems: 1) bacteria names, 2) biological processes, 3) molecular functions, and 4) cellular components. These four entities are important to pathogenesis and virulence research. Biomedical NER is challenging as new terms are rapidly being introduced, while old ones are discarded. Biological names are complex and referenced by different communities with enormous variation such as synonyms, acronyms and morphological variants (Ananiadou et al. 2004).

Some of the research on semantic tagging overlaps FG-NER: examples include sub-categorisation of proper names in the Qur'an, such as synonyms of *Allah*, names of angels, prophets and their tribes (Sharaf and Atwell, 2009); extraction of thousands of SNOMED-CT (Systematized Nomenclature of Medicine—Clinical Terms) concepts from free text discharge summary reports (Hina et al. 2011); and a proposal to detect terrorist activities by extracting entities such as caller and recipient number, duration of

call, and suspicious words and phrases from phone tap transcripts (Brierley and Atwell 2011).

There has also been specialised NER work in connection with biodiversity informatics to automatically identify taxonomic names (Koning et al. 2005, Sautter and Böhm 2006). This could prove useful in projects such as the Biodiversity Heritage Library to scan millions of pages of taxonomic literature (Willis et al. 2010).

2.4.2 FG-NER Methods

Naturally, the methods used to automatically identify fine-grained named entities are rooted in 'coarse' NER. These can be divided into list-lookup, handcrafted rules, and machine learning (Kozareva et al. 2008, Sasaki et al. 2008). Machine learning methods can be supervised or unsupervised. In addition, there are hybrid systems which combine handcrafted rules and lists with statistical machine learning.

List lookup

In the context of NER, the terms 'dictionary', 'list', 'lexicon' and 'gazetteer' are often used interchangeably, though dictionaries and lexicons are expected to include information about each entity instance listed (Smith et al. 2008). Nadeau and Sekine (2007) identify in the literature three categories of lists used in NER:

1. General lists (general dictionary, stop words, capitalised nouns, common abbreviations);
2. Lists of entities (organisations, last names, countries, astral bodies);
3. Lists of entity cues (typical words in names of organisations, person titles, post

nominal letters, location-typical words)

1 and 2 are used in the simplest form of NER. A word needs to match an element in the list to be recognised as an entity. For instance, *Nairobi* in the text will be recognised as a name of a city if it is included in a pre-existing list of cities. One of the main drawbacks of this simple matching method is that the list may be incomplete or out of date: new names of people keep appearing in the news, new companies are founded, geographical names may change. For example, in the taxonomic domain the number of known species is over 1.5 million. Even if an exhaustive list could be created, the estimated number of species is much higher, in the range of 30-100 million (Koning et al. 2005). Moreover, new species names are continuously added, while previous nomenclature also gets revised. It is clear in this case that a list lookup can only offer a partial solution.

Entity cues can help recognise entities which are not on the list. For example, 'Dr' is likely to be followed by a name of a person, while 'corp' and 'inc.' are indicative of an organisation.

A common problem with lists is errors in recognising NE boundaries (Nadeau et al. 2006). This can occur when an NE consists of two or more words that are each listed separately or when one entity is a substring of another entity (eg, *Sydney Opera House*).

Another drawback is that often candidate words are required to exactly match at least one element of a list. However, for example, the town Southborough in Massachusetts is also often spelt Southboro, and in the same context both instances should be identified as referring to the same place. Stemming, lemmatization and fuzzy matching based on edit-distance can provide some flexibility (Nadeau and Sekine 2007).

The use of name lists does not solve the ever-present NLP problem of ambiguity. If for instance a list of cities is used to look up *Victoria*, the name will match every instance of *Victoria* in the text, even when it does not refer to a place. Furthermore, the use of the list on its own will not disambiguate between many possible geographical locations called Victoria, which are themselves in different subcategories of the NE `LOCATION`, including city, lake, island, port, street, dam and park.

In spite of the drawbacks of lists, lexical resources have become readily available on the web and make this NER approach easier to implement in various domains: In particular, the structure of Wikipedia is suited to function as a list of categorised fine-grained NEs (Bunescu and Pasca 2006, Nothman et al. 2008). For example, Beneti et al. (2006) exploited Wikipedia's hierarchy to classify NEs of the type `PERSON`, `LOCATION`, `ORGANISATION` into subcategories.

In a specialised task such as gene mention recognition (Smith et al. 2008) open systems use resources such as the Unified Medical Language System (UMLS). The UMLS is maintained by the US National Library of Medicine. It integrates over 2 million names for some 900 000 concepts from more than 60 families of biomedical vocabularies (Bodenreider 2004). The dictionary-based method is important in biomedical NER as biomedical thesauruses are needed to understand text in this domain (Liu et al. 2010).

Handcrafted Rules

There are two approaches to NER using rules: the first is based on internal structure and the second on context (Beneti et al. 2006, Narayanaswamy et al. 2003). In both approaches morphological, orthographic and syntactic features can help disambiguate certain categories. Examples for internal structure could be the capitalisation of taxonomic names according to Linnaean nomenclature (Koning et al. 2005), the

structure of an email address, which contains a @ sign and *.domain* at the end, *town/ville/land* suffix indicating a name of a location, and the prefix *tert-* designating a chemical name . Context could be a title preceding a person's name, strings that denote word boundaries of chemical names (Corbett and Copestake 2008) such as '-induced'; The presence of a location name can be indicated by 'Street', 'City', 'Avenue' following a contiguous sequence of capitalised word. Figure 4 shows a few examples of regular expressions used by Mikheev et al. (1999) to identify instances of NE in MUC-7 test materials.

Context Rule	Assign	Example
Xxxx+ is? a? JJ* PROF	PERS	Yuri Gromov, a former director
Xxxx+ is? a? JJ* REL	PERS	John White is beloved brother
Xxxx+ himself	PERS	White himself
Xxxx+, DD+,	PERS	White, 33,
shares in Xxxx+	ORG	shares in Trinity Motors
PROF of/at/with Xxxx+	ORG	director of Trinity Motors
Xxxx+ area	LOC	Beribidjan area

Figure 5: NER pattern rules: Xxxx+ is a sequence of capitalised words; DD is a digit; PROF is a profession; REL is a relative; J J* is a sequence of zero or more adjectives; LOC is a known location (Mikheev et al. 1999)

Rules can be combined with an NER lexical lookup component. The following example describes a method that was applied by Wang and Matthews (2008) to sub-categorise biomedical terms (proteins, genes) according to which species they are linked to. Their species tagger marked content words such as *human*, *murine* and *D. melanogaster*. In addition, grammar rules were used to identify species prefixes (eg, 'h' for *human*, 'm' for *mouse* in entities such as the protein *mSos-l*). Rules were used to assign a species. For instance, if the word preceding an entity was a species word, the species indicated by that word was assigned to the entity. Similarly, if a species word occurred left to the entity in the same sentence the entity was assigned the species indicated by that word.

The main weakness of using rules such as regular expressions is that they need to be written by hand and are unlikely to be portable to another domain. This can require significant human effort if a large number of fine-grained entity types need to be recognized (Sekine et al. 2002). Predetermined rules can miss instances of entities that match a pattern omitted by the system's developer. Just as lookup lists often cannot be exhaustive so it is with sets of pattern rules. Even distinctive entities such a taxonomic

names do not follow a restrictive syntax. For example, '*Prenolepis (Nylanderia) vividula Erin* subsp. *guatemalensis Forel var. itinerans Forel*' is a taxonomic name, and so is *Spheniscus humboldti* (Sautter and Böhm 2006*)*. In this case, regular expressions are unable to capture all taxonomic names and at the same time offer precision. Koning et al. (2005) used static dictionaries and regular expressions to recognise taxonomic names but omitted instances of names which lack a genus.

Machine Learning

The role of the rule-based approach in NER, despite its appealing transparency (Chiticariu et al. 2010) has diminished with the rise of machine learning techniques that now dominate the field with state-of-the-art results (Nadeau and Sekine 2007). Alpaydin (2004) defines machine learning as optimising a performance criterion using example data or past experience. This statistics-based learning optimises the parameters of a model using training data. The model can then be used to make predictions, or gain knowledge from data, or both. In the context of NER, machine learning most commonly refers to supervised learning. Its aim is 'to learn a mapping from the input to an output whose correct values are provided by a supervisor' (Alpaydin 2004). In practice a corpus of documents is annotated by hand to identify entities of interest that will serve as positive examples. Features such as literal words, pattern of orthography and parts of speech are then used to train a system to tag instances of NE in novel text (Freitag 2004). When developing systems the data is often split into training and test data. The training data is usually tuples of the form <data features-NE type> (Beneti et al. 2006). The system then assigns an NE type (or a non-NE tag) to every test instance.

The biggest disadvantage of supervised learning is the laborious effort required to annotate corpora for training (Lee and Lee 2007). The reliance on costly, low-yield expert annotation is thought to hinder the development of more adaptable, high-

performance NE taggers (Nothman et al. 2008). This is one of the reasons many studies have been availing of the tagged corpora created for NER evaluations (eg, MUC, CoNLL). However, these corpora are limited to the newswire domain and only tagged for coarse-grained entities such as `PERSON`, `LOCATION` and `ORGANISATION`, and are therefore not suited to more specialised NER. In biomedical NER researchers are able to use the publically available GENIA, the largest semantically annotated corpus for bio-text-mining (Kim et al. 2003). It consists of 2000 Medline abstracts in which 36 categories of biomedical NEs are annotated according to the Genia ontology. In other fields a new corpus needs to be annotated. For instance, to train a system to recognise chemical entities (compound, reaction, adjective, enzyme and prefix), Corbett and Copestake (2008) annotated a set of 42 chemical papers and 500 PubMed abstracts.

One way to reduce the effort of creating an annotated training set is active learning (Tsuruoka et al. 2008), in which samples are selected to be presented to a human annotator by a machine learning model interactively and iteratively.

More recently researchers have been recruiting annotators online through the crowd sourcing services Mechanical Turk (MTurk) and CrowdFlower (Finin et al. 2010, Lawson et al. 2010). Although the annotation tasks were performed at a reasonable level and cost effectively, creating specialised training corpora for NER still necessitates experts familiar with the domain. Consistency of annotators recruited from the general public could also become a problem when having to follow intricate guidelines to annotate general yet fine-grained NE.

The main techniques of supervised learning used in NER include Support Vector Machines (SVM) (Altincay et al. 2009), Bayesian classifiers and decision trees (Fleischman and Hovy 2002), Hidden Markov Models (HMM) (Collier et al. 2000),

Maximum Entropy (ME) (Murphy et al. 2006, Sutcliffe et al. 2010), and Conditional Random Fields (CRFs) (Lee et al. 2006). SVM and Bayes decision trees are classifier-based, whereas the rest are Markov model based. The latter excel at sequence labelling tasks (Tsai et al. 2006). Methods have been proposed in which the boundaries of named entities are learnt using a classifier such as SVM and then the classification of the NEs into predefined categories is treated as sequence labelling task by HMM or CRF (Lee et al. 2004, Shing-Kit and Wai 2007). More about CRF, the technique used in this project, can be found in section 5 of this chapter.

The scarcity of annotated training resources, in contrast to the vast amount of raw text now available on the web, and the effort involved in creating these resources, have led to the development of NER that exploits two alternative learning methods: semi-supervised learning (or weakly supervised) and unsupervised learning. The main technique of semi-supervised learning is bootstrapping (Collins and Singer 1999, Vlachos and Gasperin 2006). Bootstrapping relies on providing a small or relatively small number of seed instances that are then used to discover context patterns, which in turn are reapplied in iterative steps to extract further NE candidates and discover new contexts. For example, Lee and Lee (2007) used bootstrapping to recognise fine-grained geographic NEs (`COUNTRY`, `CITY`, `ISLAND`, `RIVER`, `MOUNTAIN`) in New-York Times articles. They first annotated a large raw corpus with unambiguous seed instances obtained from a gazetteer. From the initial annotation inter-phrasal and intra-phrasal contexts were learnt and reapplied to the corpus to obtain new candidates of each type. The work of Whitelaw et al. (2008) on web-scale fine-grained NER also relied on using a large seed set of entities extracted from web resources such as Wikipedia and IMDB (Internet Movie Data Base) in order to discover high precision simple context templates (eg, *and [drummer] on drums*) and to propagate name type information across web links. Their aim was to generate a training set for supervised machine learning. The seed set consisted of 5 million entity names of known type

which yielded, after filtering, 65 million matches. Their page/link propagation method increased this figure to 475 million trusted mentions. Similar web-scale work was reported earlier by Etzioni et al. (2005) with their Know-It-All system. Its goal was to extract lists of names of a given type (eg, names of politicians or scientists) by domain-specific pattern learning. The bootstrapping in that case was fully automatic through a seed set of domain-independent extraction patterns (eg, *NP "such as" NPList*). The task of creating lists of NEs is referred to as NE extraction, not recognition. It is not designed to disambiguate entities in a given document (Nadeau et al. 2006).

In unsupervised learning the learning is done without feedback. A typical technique is clustering (Freitag 2004). Unsupervised learning is not a popular approach for NER and NER systems that are described as unsupervised are usually not entirely unsupervised (Mansouri et al. 2008). Nadeau et al. (2006) also point out that the distinction between supervised and unsupervised systems is sometimes blurred with clever rules and heuristics replacing the human labour of annotating a training corpus. However, they argue that their own NER system can be described as unsupervised because of the system's 4-item seed lists, minimal use of domain knowledge and the availability of HTML markup. Elsner et al. (2009) describe an unsupervised system that clusters NEs using generative models. However, they assume that the NEs have already been correctly extracted and that they all fit into one of the three MUC-7 categories.

Other unsupervised systems described in the literature are named-entity **extraction** systems rather than NER: For example, Silva et al. (2004) extracted Multiword Lexical Units (MUWs) using n-grams without predefined categories. NE candidates were then filtered and clustered by just two attributes. Similarly, Zhang et al. (2010) used Automatic Term Recognition (ATR) techniques to determine domain specificity of words without domain-specific external knowledge. The terms were then used in the

form of clustered features to boost performance of supervised learning.

Addressing specifically FG-NER, Kozareva et al. (2008) categorised person names by using WordNet to calculate the domain distribution of the context surrounding the NE candidate. However, to select the set of names to be disambiguated they used regular expressions.

The work of Ling and Weld (2012), who introduced a set of 112 NE types derived from the collaborative knowledge base Freebase (see figure 4), can also be described as unsupervised. The use of Freebase allowed them to exploit broad coverage of entities and to tag entities with multiple overlapping classes. They then automatically labelled each text from Wikipedia, using the information encoded in anchor links to map it to a type in Freebase. They then proceeded to segment and classify the automatically generated training set using CRF and a perceptron classifier.

Hybrid Systems

Hybrid NER systems combine the main three methods reviewed above (Mansouri et al. 2008). NER systems that rely solely on one of the methods have become relatively rare. Even when relying on machine learning, more than one machine learning technique could be used (Ekbal and Bandyopadhyay 2009). Rules and lexicons are used in the preprocessing/pre-training stage of machine learning or in filtering/postprocessing its output. For example, McDonald and Pereira (2005) incorporated lexicon features when training a CRF system to recognise protein and gene mentions. The set of lexicons allowed them to remove false positives and recover false negatives. Sasaki et al. (2008) introduced a novel approach which they refer to as dictionary-based statistical NER. They first identified protein mentions with part-of-speech tagging based on both a general word dictionary and an NE dictionary. They

then trained a CRF system on the output of the tagger. Minkov et al. (2005) also integrated dictionary features when using CRF to extract personal names from emails. For instance, a word that is in the first-name dictionary and is not in the common words or last-name dictionaries is considered to be first-name for certain. Srihari et al. (2000) combined rules to constrain an HMM that generated the standard MUC tags of `PERSON`, `LOCATION`, `ORGANISATION`. They then used Maximum Entropy incorporated with gazetteers to derive subcategories such as airport or city from the basic tags. Leaman and Gonzalez (2008) applied two types of rule-based post-processing to the output of CRF trained to recognise biomedical entities: detecting when pairs of brackets and quotation marks were labelled differently, and resolving abbreviations to make sure that in instances such as *antilymphocyte globulin (ALG)* or *ALG (antilymphocyte globulin)* both *ALG* and *antilymphocyte globulin* are recognised.

The hybrid system of Nguyen and Cao (2008) detected proper names and linked them to their corresponding Wikipedia entries. In the first phase they used heuristics and patterns to narrow down candidates. In the second one a vector space model was used to rank ambiguous cases to select the right candidates.

It should be noted that domain portability remains a weakness with hybrid systems which are partly rule-based (Ekbal and Bandyopadhyay 2009, Mansouri et al. 2008).

2.4.3 Performance

Evaluation of NER has been usually in the context of information extraction competitions such as MUC. Due to lack of well established NER evaluation standards across competitions, comparison of performance is difficult (Krovetz et al. 2011). Comparison is even more elusive when FG-NER is concerned: although there have been efforts to draw up a comprehensive tag set of NE subcategories (Sekine et al.

2002), no consensus has been reached by the research community on standard NE subtypes (Ling and Weld 2012).

Performance of NER systems is measured by precision (P) and recall (R)—metrics adapted from information retrieval (Chinchor 1992):

$$Precision = \frac{number\ of\ correct\ responses}{number\ of\ responses}$$

$$Recall = \frac{number\ of\ correct\ responses}{number\ of\ correct\ answers\ in\ annotated\ text}$$

Realising that both measures of precision and recall are often important yet negatively correlated, the organisers of MUC-4 introduced van Rijsbergen's F-measure (Rijsbergen 1979) which combines precision and recall into a single measure as their harmonic average. The general formula for calculating this measure is:

$$F_\beta = \frac{(\beta^2 + 1.0) \times P \times R}{(\beta^2 \times P) + R}$$

β determines the relative importance given to recall over precision. If recall and precision are of equal weight β = 1.0. If recall is twice or half as important as precision, β = 2.0 or 0.5 respectively. The F-measure rewards values of recall and precision that are closer to the centre of the precision-recall graph and therefore balanced systems. For example, if β = 1.0, a system which has recall of 50% and precision of 50% would have a higher F-measure than a system which achieved recall of 30% and precision of 70% (Chinchor 1998).

Judgement of systems' responses can be strict or lenient. Under strict evaluation only an exact match would be considered correct, whereas lenient judgement would accommodate partial matches. In some domains of application, where the goal is to determine if particular sentences contain an NE, insisting on exact NE boundaries may be unnecessarily strict (Nadeau and Sekine 2007). Both strict and lenient values are often reported together in the literature. In MUC, each instance of NER was considered to consist of two tasks: boundary detection and type labelling. This differentiation meant that instances of NE labelled with the correct type regardless of exact detection of boundaries were still considered correct; and vice versa, cases of correct detection of boundaries regardless of type label were counted as correct. MUC's scoring system is not universal: at IREX and CoNLL only exact matches were counted as correct, whereas a complex algorithm for weighting entities differentially has been used at ACE (Doddington et al. 2004). This diversity of scoring systems hampers comparison.

The systems which participated in the NE task at MUC-6 and MUC-7 achieved high scores (see figure 6). In both MUC-6 and MUC-7 most systems scored more than 90% on the F-measure, with the top ones matching the performance of human annotators (Marsh and Perzanowski 1998, Sundheim 1995).

F-measure	Error	Recall	Precision
96.42	5	96	97
95.66	7	95	96
94.92	8	93	96
94.00	10	92	96
93.65	10	94	93
93.33	11	92	95
92.88	10	94	92
92.74	12	92	93
92.61	12	89	96
91.20	13	91	91
90.84	14	91	91
89.06	18	84	94
88.19	19	86	90
85.82	20	85	87
85.73	23	80	92
84.95	22	82	89
Annotators:			
96.68	6	95	98
93.18	11	92	95

Figure 6: Overall NER results in MUC-7. From (Marsh and Perzanowski 1998)

The performance of all systems was better when recognising person names than when recognising organisation or location names, and almost all systems scored higher on location names than organisations. This could be explained by the varied form of organisation names, their length and complexity such as internal punctuation. There were fewer errors in detecting the boundaries of ENAMEX entities (PERSON, ORGANISATION, LOCATION) than recognising their type, probably because NUMEX (PERCENT, MONEY) and TIMEX (DATE, TIME) had only two types each. When reporting the MUC NE results, Sundheim qualified the success in this task by highlighting the favourable evaluation conditions which should be kept in mind when these high scores are considered: Sundheim mentioned the uniform (journalistic) style of writing of the test set and its focus on certain topics, the small size of the test set (30 articles), and the lack of TIME expressions. For example, in the newswire domain correct usage of upper and lower case is to be expected. When in one experiment this reliable clue was removed by using an upper case version of the test set, the F-measure dropped by 10 points.

As discussed in section 3.3, the evaluation work by Vilain et al. (2007) demonstrated that commercial and research state-of-the-art systems, two of which were trained on MUC-6 data, perform more poorly than expected when tested on business-related texts (figure 7). Such texts seem, on the face on it, similar to journalistic stories in style and content. The drop in performance is equivalent to that observed when taggers developed for journalistic text are tested on informal language encountered in email messages, speech transcripts or more recently, tweets (Li et al. 2012).

	Pocahontas	Belle	Jasmine	Mulan	Ariel
SEC filings	R=58	R=28	R=50	R=50	R=71
	P=65	P=52	P=43	P=56	P=79
	F=61.1	F=36.4	F=42.7	F=52.6	F=74.5
SEC filings, *"the Corp."* optional	R=71	R=36	R=55	R=60	R=71
	P=65	P=52	P=40	P=56	P=79
	F=68.0	F=42.8	F=46.2	F=57.9	F=74.7
Business news	R=80 (82)	R=64 (69)	R=76	R=65	R=71 (75)
	P=80 (79)	P=86 (83)	P=63	P=74	P=74 (75)
	F=80.1 (81)	F=73.5 (75)	F=69.1	F=69.2	F=72.3 (75)
Current events (MUC-like)	R=94 (94)	R=59 (63)	R=79	R=79	R=89 (91)
	P=94 (93)	P=82 (80)	P=70	P=92	P=91 (92)
	F=94.3 (94)	F=68.5 (71)	F=74.5	F=84.9	F=90.4 (92)

Figure 7: scores of four undisclosed state-of-the-art NER systems and the author's rule-based SEC tuned system Ariel. Pocahontas and Belle are rule based; Jasmine and Mulan are statistical (HMM and CRF, respectively), MUC-trained. Scores are for ENAMEX unless in brackets (=all entities). SEC refers to financial reports filed by the Securities and Exchange Commission. "the Corp optional" refers to a run to isolate the contribution of the systems' failure to recognise rightwards shortenings of company names by excluding these cases in scores. Extracted from Vilain et al. (2007)

Organisation names are twice as common in business sources as in MUC-like data. As observed in MUC, scores for `ORGANISATION` are lower than scores for `PERSON` and `LOCATION`. This partly explains the poorer performance.

Krovetz et al. (2011) compared the agreement between three state-of-the-art NE taggers (Stanford, LBJ, Identifinder) when tagging the major three ENAMEX entities. They found that agreement between the taggers was 34%, 37%, 58% on `ORGANISATION`, `LOCATION`, `PERSON`, respectively. They also calculated the percentage of the ambiguous entities (having more than one classification across the corpus) co-occurring in a single document. Their analysis of the tagger run results showed that it is more than 40%. All these findings in relation to the more established MUC-type NER should be kept in mind when evaluating FG-NER.

To give the reader an idea about the performance levels of NER in specific domains and of FG-NER, table 3 presents a sample of results from biomedical/natural science NER. It is followed by table 4 which similarly summarises results from FG-NER work covered in section 2.4.1.

Table 3: a sample of state-of-the-art results reported for biomedical NER (P = Precision, R = Recall, F = F-measure).

Author	NE (best P/R/F)	Corpus	Method
(Leaman and Gonzalez 2008)	gene (DNA) (85/79/82)	BioCreative 2 GM training corpus	Machine learning: CRF
	DISEASE (69/45.5/55)	BioText disease/treatment corpus	
(Smith et al. 2008)	DNA (88.5/86/87)	BioCreative 2 GM training corpus	Semi-supervised/hybrid
(Shing-Kit and Wai 2007)	DNA (70/70/70) RNA (66/73/69) CELL_LINE (56/65/60)	Genia	Machine learning: CRF
(Narayanaswamy et al. 2003)	PROTEIN/DNA (96.5/62.5/75.9) CHEMICAL (93/86/91)	Medline abstracts	Rule-based
(Ananiadou et al. 2011)	BACTERIA (96/97/96) CELLULAR COMPONENT (74/62/68) BIOLOGICAL PROC. (87/81/84)	purpose-built corpus of bacterial Type IV secretion systems (T4SSs) documents	Dictionary/Machine learning
(Rocktäschel et al. 2012)	CHEMICAL (67/69/68)	SCAI (Corpora for Named Entity Recognition of Chemical Compounds)	Machine learning (CRF) + dictionary
(Sautter and Böhm 2006)	TAXON (99/99/99)	American Museum Novitates	Active learning + Rule-based

Table 4: a sample of best results of FG-NER/Classification (P = Precision, R = Recall, F = F-measure).

Author	NE (best P/R/F or accuracy)	Corpus	Method
(Fleischman and Hovy 2002)	8 subtypes of PERSON (70.5)	Person names compiled from TREC 9	WordNet + C4.5 decision tree classifier
(Kozareva et al. 2008)	2 subtypes of PERSON: PRESIDENT (72 /82.5/77) SINGER (79.5/67.5/73)	Self-compiled from the New York Times	WordNet context domain mapping
(Whitelaw and Patrick 2003)	35 labels, eg: COMPANY (42.5/62.5) FISH (44/62) PERSON (95/91) PLACE (86/86.5)	The web	Bootstrapping/lists/Machine learning (Perceptron)
(Ling and Weld 2012)	112 subtypes (see figure 4) Strict F-measure: 53 Lenient F-measure: 70	Reports from local newspaper	Unsupervised collection of training data/Machine learning (CRF/Perceptron)
(Lee et al. 2007)	147 subtypes (Korean) (83/74.5/78.5)	Korean reports	Machine learning: ME + CRF
(Sobhana et al. 2010)	17 geospatial and temporal NEs (77/77/76)	Scientific reports and papers on the geology of the Indian subcontinent	Machine learning: CRF

2.5. Why Choose Conditional Random Fields (CRFs) for NER?

Conditional Random Field is the probabilistic model used for the NER machine learning tasks in this project. CRF is considered by many to be a state-of-the-art technique for sequence tagging (Feldman and Sanger 2006, Finkel et al. 2005, Peng and McCallum 2004, Shing-Kit and Wai 2007, Klinger and Tomanek 2007). The technique has been proven to be useful in NER, especially in the biomedical domain (Leaman and Gonzalez 2008, McDonald and Pereira 2005, Shing-Kit and Wai 2007, Huang et al. 2009, Kou et al. 2005, Settles 2004, Struble et al. 2007, Suakkaphong et al. 2011, Vlachos 2007). More generally, CRF has been used to model linear sequence structures for natural language tasks such as part-of speech tagging and noun phrase chunking (Shah et al. 2003) and also has been applied in bioinformatics and computer vision (Sutton and McCallum 2006). CRF was introduced by Lafferty et al. (2001) who demonstrated the advantages of the new framework over Hidden Markov Models (HMMs) and Maximum Entropy Markov Models (MEMMs).

An HMM (Rabiner 1990) is a well known generative model. Its purpose is to maximise the joint probability of an observation sequence and its paired label /state sequence:

$$P(y, x) = \prod_{i=1}^{n} P(x_i \mid y_i) P(y_i \mid y_{i-1})$$

Where x = ($x_1, x_2, \ldots x_n$) is the observation sequence of words of length n and y = ($y_1, y_2, \ldots y_n$) is a sequence of labels. As can be seen, according to the Markov property each label y_i depends on the previous one y_{i-1} (transition probability), while each observation word x_i only depends on the current label (emission probability). Traditional HMMs therefore assume independence between words and their context or other features. This simplification does allow quick learning and global maximisation of the joint

probability over the whole observation and label sequences (Ponomareva et al. 2007). However, a generative model like HMM requires enumeration of all possible observation sequences. This means that the inference problem of an HMM model which represents interacting features or long-range dependencies would be intractable. An HMM models the observation sequence while in NER the goal is to optimise the *label* sequence.

In conditional/discriminative models no attempt is made to model the observation sequence. Such a model specifies the probabilities of label sequences given an observation sequence. This leads to a great reduction in the number of possible combinations between observation word features and their labels. Therefore the probabilities can depend on arbitrary, non-independent features of the observation sequence, adding more knowledge to the model. The probability of a transition between labels may depend also on past and future observations, not only on the current one.

Maximum Entropy Markov Models (MEMMs) are conditional models that confer these advantages. They are based on the *Principle of Maximum Entropy* (Berger et al. 1996) according to which, given a collection of facts, a model should be consistent with all the facts, but otherwise as uniform as possible. Uniformity of conditional distribution $p(y \mid x)$ is measured by the conditional entropy:

$$H(p) = -\sum_{x,y} p(y,x) \log p(y \mid x)$$

The goal of an ME model is to find $p(y \mid x)$ which has the largest H(p).

MEMMs (McCallum et al. 2000) replace the transition and observation functions of HMM with a single function $P(y' \mid y, x)$, which means it models the probability over the next state given the current state and the observations:

$$P(y' \mid y, x) = \frac{1}{\mathrm{Z}(y,x)} \left(\sum_k \lambda_k \, f_k(x, y, y') \right)$$

Where *f* is a feature and λ is a weight. Z is the per-state normalisation. A feature f_k is defined by k = <b,r>, where b is a binary feature of the current observation and r is a state value. For example, if $b(x_i)$ tests the capitalisation of the observed word and the word is tagged in the training corpus as part of a person's name (r = `PERSON`), then the feature will take the value 1. Otherwise it will be 0.

One weakness of MEMM is called the *label bias problem* (Lafferty et al. 2001): the bias is towards states/labels with fewer outgoing transitions. A state with a single outgoing transition could ignore an observation. Figure 8 shows an often quoted example which appeared originally in Bottou (1991).

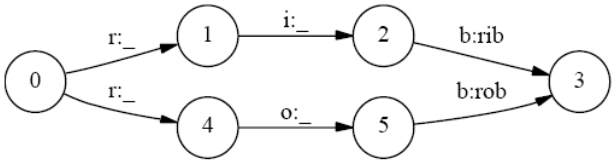

Figure 8: Label bias problem. Observation-label pairs *o : l* are placed on transition. '_' represents the null output label. (Lafferty et al. 2001)

The example depicts a finite-state model designed to distinguish between the words `rib` and `rob`. If the observation sequence is `rib`, `r` will match both transitions from the start state, and the probability mass is divided roughly equally between state 1 and 4. The next observation is i. State 1 has seen `i` often in training, while state 4 has almost never seen it. But both 1 and 4 have only one outgoing transition each, so state 4 must pass all its mass to that outgoing transition, as it is not generating the observation, just conditioned on it (and hence the absence of such a problem with generative HMM models).

The Conditional Random Fields undirected graphical model was proposed by Lafferty et al. (2001) as a way to solve this problem common to MEMM and other discriminative Markov based on directed graphs. They formally defined CRF:

> Let G = (V,E) be a graph such that $\mathbf{Y} = (\mathbf{Y}_v)_{v \in V}$, so that Y is indexed by the vertices of G. Then (**X**,**Y**) is a `conditional random field` in case, when conditioned on **X**, the random variables $\mathbf{Y_v}$ obey the Markov property with respect to the graph: $p(Y_v | X, Y_w, w \neq v) = p(Y_v | X, Y_w, w \sim v)$, where w ~ v means that w and v are neighbours in G.

The joint distribution over the label sequence Y given X is:

$$p_\theta(y \mid x) = \frac{1}{Z(x)} \exp\left(\sum_{e \in E,k} \lambda_k f_k(e, y|_e, x) + \sum_{v \in V,k} \mu_k g_k(v, y|_v, x) \right)$$

$$Z(x) = \sum_{y} \exp\left(\sum_{e \in E,k} \lambda_k f_k(e, y|_e, x) + \sum_{v \in V,k} \mu_k g_k(v, y|_v, x) \right)$$

Z(x) is a normalisation factor, *v* is a vertex from the vertex set, *V* is a set of label random variables, *e* is an edge from edge set *E* over *V*, $y|_e$ is the set of components of *y* defined by edge *e*, $y|_v$ is the set of components of *y* defined by vertex *v*, *k* is the number of features. In the simplest and most relevant example for modelling sequences of words, G is a linear chain.

λ_k and μ_k are learning weights of the feature functions f_k and $g_{k.}$ respectively. The weight parameters are similar to the logarithms of the HMM parameters p(y'|y) and p(x|y). The

feature functions are fixed and given. For example, a Boolean vertex feature g_k may be true if the word X_i ends with *land* and the tag Y_i is `COUNTRY`.

Unlike MEMM, which uses a per-state exponential model to predict the next state based on the current one, CRF presents a single exponential model for the joint probability of an entire sequence of labels, given the observation and allows certain transitions to 'vote' more strongly than others based on the corresponding observations. Figure 9 compares the graphical structures of HMM, MEMM and linear chain CRF.

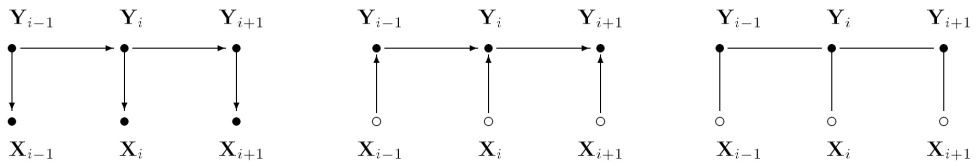

Figure 9: Graphical sequence structure of simple HMMs (left), MEMMs (centre), and the linear chain case of CRFs (right). An open circle indicates that the variable is not generated by the model Lafferty et al. (2001).

The training of CRF consists of weight evaluation in order to maximise conditional log likelihood of labelled sequences for a training data set (Ponomareva et al. 2007):

$$D = (x, y)^{(1)}, (x, y)^{(2)}, \ldots (x, y)^{(|D|)}$$

$$L(\theta) = \sum_{j=1}^{|D|} log\, P_\theta(y^{(j)} \mid x^{(j)})$$

Labelling a new unseen sequence of words (inference) is done through a Viterbi algorithm which finds the most likely label sequence according to the CRF model.

To summarise, CRF is reported to be a state-of-the-art model used in NER. Being

discriminative, it allows the incorporation of many correlated features found in real-world data, which adds knowledge to the model. Unlike MEMM, CRF is an undirected graphical model and avoids the label bias. This is particularly significant where there are many transitions that are nearly deterministic. In light of these key points, CRF was the approach chosen for the NER experiments of this project.

2.6 Could order patterns of content in documents support NER?

Order is a characteristic of natural language which distinguishes it from many other classification domains (Hachey and Grover 2005). In tasks such part-of-speech tagging or NER the order is that of words within sentences, but beyond the vocabulary, syntax and semantics of individual sentences, text too has been shown to have several structures (Moens et al. 1999). This is one of the subjects of the interdisciplinary study of discourse analysis, which goes beyond the sentence boundary and regards the text a whole grammatical unit. It examines the ways sentences are connected together ('cohesion') (Sinclair 1993) and the organisation of text. Discourse analysis identifies three main formal structures of texts:

1. schematic structure / superstructure
2. rhetorical structure
3. thematic structure

The *schematic structure* is conventional and specified in terms of the ordered parts the text is built of. A text type is made of parts of variable sizes (sentence(s)/paragraph(s)), some being optional, which occur in a fixed or partially fixed order. Such superstructures show the formulaic nature of many texts. Dillon (1991) showed that experienced readers of scientific journals possess a schema of this text type and

therefore can predict with high accuracy where information is located. Specifically, they were able to classify paragraphs into Introduction, Methods, Results, and Discussion with about 80% accuracy under time pressure.

Rhetorical structure refers to the organisation of coherent, continuous text and the rhetorical relations between its parts, such as succession, conditionality, motivation, circumstance and contrast. For instance, Teufel et al. (2009) applied Argumentative Zoning analysis to the argumentative and rhetorical structure of scientific papers. The analysis aims to model the stages of claiming ownership on a new piece of knowledge in scientific papers across disciplines. Mizuta et al. (2006) developed the model further and analysed the features of each zone in biology papers as a possible basis for information extraction.

Thematic structure is the overall organisation of topics in the text. The organisation is usually hierarchical: the theme of the whole text can be expressed in terms of increasingly more specific themes (sub-topics).

Superstructural and rhetorical relations may be signalled by surface linguistic forms and sometimes text layout. The themes of the text are closely linked to surface linguistics phenomena. There are additional markers which indicate topic shifts and locational cues such as the position of topic sentence.

An important feature of superstructure is the ways in which it controls thematic content order. For instance, in news discourse the headline and lead elements normally cover the more general themes of the report.

These observations led to attempts to exploit text structure for natural language processing tasks such as information retrieval, summarization and text ordering

(Barzilay and Lee 2004, Lapata 2003). Moens et al. (1999) illustrated how text grammar, which incorporates knowledge of discourse patterns, is used in a system that abstracts criminal cases. One fundamental task in discourse processing is text segmentation: locating the positions in which topics change in a stream of text. There has been a considerable amount of work in this area (Beeferman et al. 1999, Bestgen and Vonk 1995, Blei and Moreno 2001, Dias et al. 2005, Kan et al. 1998).

With information extraction tasks in mind, Shah et al. (2003) investigated the content extraction of keywords from the different standard sections (Abstract, Introduction, Methods, Results, Discussion) of scientific papers in the biomedical domains. Schuemie et al. (2004) described the distribution of information density of biomedical abstracts and full-text papers in relation to the abovementioned five sections. They found that within a single paper there were sections that contained more information than others, but a significant part of the information in any section was unique to it. Lin and Hovy (1997) developed a method of locating the likely positions of topic-bearing sentences by ranking sentences according to their yield of keywords shared with abstracts. They provided empirical validation of the Position Hypothesis that: (1) sentences which appear under certain headings are positively relevant; and (2) topic sentences tend to occur very early within a document. Experiments with positional features are reported in chapter 5.

2.7. Chapter Summary

The chapter began with an overview of information extraction (IE), its history and main tasks. The following sections reviewed named entity recognition (NER) and fine grained NER in particular. Next, the reasons for choosing the Conditional Random Field (CRF) model for this thesis' project were explained. Finally, a brief background on order of content/text structure as a possible basis for information extraction was provided. The next chapter will focus on the experimental domain of this project.

Chapter 3: Specific Named Entities and the ZooBirth Corpus

3.1 Outline

This chapter first provides background on the journalistic domain of animal birth events in zoos. Next, the compilation of the ZooBirth corpus resource is reported, followed by a detailed list of SNEs and related descriptive statistics, illustrated with extracts from the corpus.

3.2 Domain of Choice: News Reports on Animal Births in Zoos

As reviewed in the last chapter, most NER research in the newswire domain has been, for many years, conducted either as part of the evaluations activity of MUC, and later ACE and CoNLL or using their legacy of tagged corpora for development and testing. The availability of these resources may have not encouraged the development of new restricted-domain collections of journalistic reports due the effort of annotating a new corpus, while recent novel NER seems to be moving away from newswire. Currently no gold standard corpus with fine-grained annotation is available. Limited work has been reported on recognition of fine-grained named entities (FG-NER), and these named entities are still coarser than the numerical ones investigated here (see below).

News stories reporting the birth of animals in zoos offer a research opportunity in a novel restricted domain with a range of fine-grained NEs. Figures 10-12 are three typical examples of such news reports. Often, photos of the newborns, which are considered to be popular with readers, are a central element in the report (figure 11).

REUTERS

» Print

This copy is for your personal, non-commercial use only.

First panda born in Europe zoo after 25 years

Thu, Aug 23 2007

By Alexandra Zawadil

VIENNA (Reuters) - A giant panda gave birth to a cub in an Austrian zoo on Thursday, Europe's first such event in 25 years, officials said.

The cub was born in Vienna's Tiergarten Schoenbrunn zoo, 127 days after mother Yang Yang mated with male Long Hui, both on extended loan to Austria from China.

Caretakers spied the tiny cub, weighing just 100 grams (3.5 ounces) and measuring 10 cm (3.9 inches), on a surveillance camera after hearing tiny whimpering sounds in Yang Yang's den.

"The panda was born without artificial insemination and that is extremely rare. We're incredibly happy," zoo director Dagmar Schratter told a crowded news conference.

"In the next hours we'll keep our fingers crossed that the cub is healthy. But the first hurdle has been overcome -- the mother has accepted her baby," she said.

Giant pandas, one of the world's rarest and most endangered species, live in the wild only in China and are notoriously loathe to breed in captivity. Females typically ovulate just once a year for a few days.

Schratter said around 40 percent of baby pandas did not live beyond a year. The last panda born in a zoo in Europe was in 1982 in Madrid, she said.

The zoo released a photograph showing Yang Yang holding the hairless cub gently in her mouth. The cub's sex can only be determined after three months.

Yang Yang and Long Hui were loaned to Austria in 2003 and are expected to remain there for about 10 years. An estimated 1,600 wild giant pandas live in nature reserves in China's Sichuan, Gansu and Shaanxi provinces.

Figure 10: an example of a news article about a birth event of a zoo animal.

After several unsteady attempts, Titan learns to run during his first day on exhibit. Zoo Miami's newest baby giraffe, a 136-pound male named Titan, was let out on exhibit for the first time on Wednesday, July 11, 2012. He was born on June 28 and has been kept with his mother, Kita, in the holding area since then so that mother and baby could bond.
RON MAGILL / ZOO MIAMI

Figure 11: an example of a newborn photo in a news report about a birth in a zoo.

Baby white rhino makes his debut

A white rhino thought by keepers to be on the small side and sickly when born in March has fought his way back to full health.

The 55th white rhino born at Whipsnade Zoo in Bedfordshire since 1970 now weighs in at about 100kg (220lbs).

The 55th white rhino born at Whipsnade Zoo has no name

His mother Clara has fed the six-week-old since he was born and he has rapidly put on weight.

They made a first official appearance together this week but the baby calf has still to be named, the zoo said.

Figure 12: an online BBC News report on a new animal zoo birth.

The source of information reported is usually press releases issued by the zoos, who may view the event as a marketing opportunity. In an email to the author (April 2008), the press officer at the Zoological Society of London, Emma Kenly, explained that the facts included in the release are dictated purely by the demand for those facts from the press and intended to be as comprehensive as possible to pre-empt further questions: the press officers try to include information such as the birth date, name and age of the mother, whether it is her first offspring, and how many offspring were born. Sometimes there are tactical or political reasons for including or excluding facts. For example, a birth may not be announced until three months after the event because the zoo wants to be certain the animal will survive before making it public. In this instance, the birth date might be buried way down in the press release or possibly referred to very vaguely ('last month'). As with all press releases issued, zoo birth announcements try to cover the five Ws of journalism (Singer 2008): who, what, where, when and why. Some zoos such as London Zoo also try to raise their profile as a credible scientific conservation organisation, so would include information about the animal species' status in the wild and mention any conservation work in which the zoo has been involved.

Beyond the natural language processing context of this project, and despite sometimes seen as light-hearted news, reports on such events might be of serious interest to the zoo / animal welfare community. Studying one aspect of the media characterisation of zoos by analysing news reports was demonstrated by Hutchins (2006). He collected 148 articles focused on zoo and aquarium animal *deaths* and classified them into four categories (dispassionate, accusatory, sympathetic, balanced).

3.3 The ZooBirth Corpus

3.3.1 Document Retrieval and Pre-processing

Newspaper reports on zoo animal birth events were downloaded at Nexis® (NexisLexis 2008), an archived news service with more than 5 billion documents and records from over 34,000 sources. The current and archived news section offers access to local, national, and international newspapers, broadcast transcripts from major television and radio networks, wire services, magazines and trade journals.

To create ZooBirth, the search terms were set to `(HEADLINE (zoo) AND HEADLINE (birth OR born),` ie documents with a headline which contains the word 'zoo', and 'birth' or 'born'. The source was set to 'All English Language News'. Without a date limitation more than 1000 results were available. In such cases Nexis® expects the user to edit the search query in order to retrieve fewer records. The data limit selected was 'the previous 8 years' (up to October 13, 2006). This search yielded 965 documents. The collection included many duplicates—the same story, in identical or nearly identical form, was published by different newspapers as a consequence of sharing the same source, normally an international news agency such as AP, as well as inappropriate reports (for example, about the *death* of a captive born animal). Following manual elimination of such instances, the collection consisted of 513 documents which were then pruned to 500 (ZooBirth500). To increase the size of the

training and test sets, the collection was expanded to 700 documents on April 27, 2008 using the same download procedure and covering the complementary period before October 13, 1998 (486 documents) and after October 13, 2006 (281 documents). Again, duplicates and irrelevant results were weeded out. They were then added to ZooBirth500 to form ZooBirth700. ZooBirth500 was used early in the project in the experiments reported in chapter 6 about order effects. Figure 13 shows a section from one of the documents in the corpus.

Each downloaded raw text file was split into individual document files. Figure 13 shows a typical document section. All the lines which were not part of the body of the news report, such as ones beginning with COPYRIGHT, LOAD_DATE, BYLINE, DATELINE, HEADLINE, SECTION, LENGTH (see Appendix A for further examples) were removed using simple heuristics. However, copyright information, headline text and load-date were not entirely discarded but stored in a Prolog file with their matching document number. This was done in case these elements of metadata structure could be used as features in future experiments.

723 of 965 DOCUMENTS

Copyright 2000 The Kansas City Star Co.
Kansas City Star (Kansas & Missouri)

August 27, 2000, Sunday METROPOLITAN EDITION

SECTION: NATIONAL; Pg. A1

LENGTH: 540 words

HEADLINE: KC rhino gives birth to a 65-pound baby;
Both seem healthy, say officials at zoo

BYLINE: MATT CAMPBELL; The Kansas City Star

BODY:

A 65-pound bundle of baby rhinoceros arrived early Saturday at the Kansas City Zoo.

The female calf was born at 4:40 a.m., and within 50 minutes she was nursing and making high-pitched yeeps. Both baby and mother appeared healthy, zoo officials said.

"She follows mom around, and mom adjusts her height to allow the baby to nurse," said Conrad Schmitt, zoo curator. "She's doing exactly what we would want her to do."

The birth is of international significance because this subspecies, called the eastern black rhinoceros, is critically endangered. The main threats are poaching and loss of habitat. The International Rhino Foundation estimates there are just 660 eastern black rhinos left on the planet.

Now there are 661.

Figure 13: a document section from the ZooBirth corpus (see appendix A for a subset of the raw corpus).

3.3.2 Method of Annotation and Further Processing

Instances of Special Named Entities (see next section) were tagged manually and single-handedly by the author, who acquired expertise in the domain thanks to previous work in zoological centres. The annotation was carried out by adding special tags denoting all types of SNEs (see section 3.4.1) before and after each SNE; for instance `dobn September 29 dobn` is the date of birth of a newborn (September 29) in a news report, delimited by the tag `dobn` (see first document in Appendix A for a fully tagged example). The very simple enclosing tags resembled a mark-up language but did

not conform to any mark-up standard.

The tagged text of the documents' body was then tokenised, and split using Prolog into paragraphs and sentences using heuristics: the presence of non-breaking spaces in the case of paragraphs; punctuation patterns in the case of sentence boundaries. The next step was to create a database in which each token was stored as a Prolog clause with the following arguments: document number, paragraph number, sentence number, token number, SNE tag, and the token itself. An additional database was created just for SNEs (see Appendix B). Each clause of the predicate `sndb` consisted of two arguments: the document number and a list; each member of the list consisted of six arguments: the token, its SNE tag, paragraph number, sentence number, token number, and a tag indicating whether the token appeared at the beginning, the middle, or the end of a sentence. The last argument was originally a candidate feature but was never selected. The SNE database was mainly used to analyse and collect statistics about SNEs.

Preparing input files for CRF++ (see chapter 4) from the prolog clauses was straightforward: the exiting arguments such as the tokens and paragraph numbers were written out in the format accepted by the toolkit, and a shell command, again in Prolog, called CRF++ and various feature templates. In addition, as in the case of order-effects (chapter 6), the Prolog databases could be interrogated to extract additional features not explicitly represented by the clauses' arguments.

Table 5: descriptive statistics of ZooBirth500 and ZooBirth700

	ZooBirth500	ZooBirth700
Number of documents	500	700
Number of tokens	174,652	240,848
Number of sentences	7,427	10,091
Number of paragraphs	4,675	6,393
Average number of tokens per document/Median	349.3/272.5	354.2/275.5
Average number of sentences per document/Median	14.9/12	14.5/11
Average number of paragraphs per document/Median	9.3/8	9.2/8
Minimum number of tokens per document	38	37
Minimum number of sentences per document	1	1
Minimum number of paragraphs per document	1	1
Maximum number of tokens per document	1,689	2,134
Maximum number of sentences per document	80	115
Maximum number of paragraphs per document	52	52

3.4 Specific Named Entities (SNEs) in ZooBirth

The term Specific Named Entity (SNE) refers here to a subtype of an NE within the context of a restricted domain. In this study the entities are all numerical and relate to the domain of reports of animal birth events in zoos. Initially all instances of numerical entities (n = 5,313) in ZooBirth500 were tagged by hand to identify the recurring subtypes in the corpus. Eventually, the ten most frequent SNEs were selected as the test set for the machine learning experiments. All SNEs were subtypes of one of the following NEs: `NUMBER` (13), `DATE` (8), `TIME` (3), `WEIGHT (4)`, `LENGTH` (2), `DURATION` (or `LENGH-OF-TIME/LOT`) (12). All NE instances that could not be subcategorised were left with their respective NE tag. Figure 14 shows the distribution of these NEs in ZooBirth500. Section 3.4.1 lists the full initial set of 42 numerical SNE types identified manually in the corpus, with corresponding excerpts.

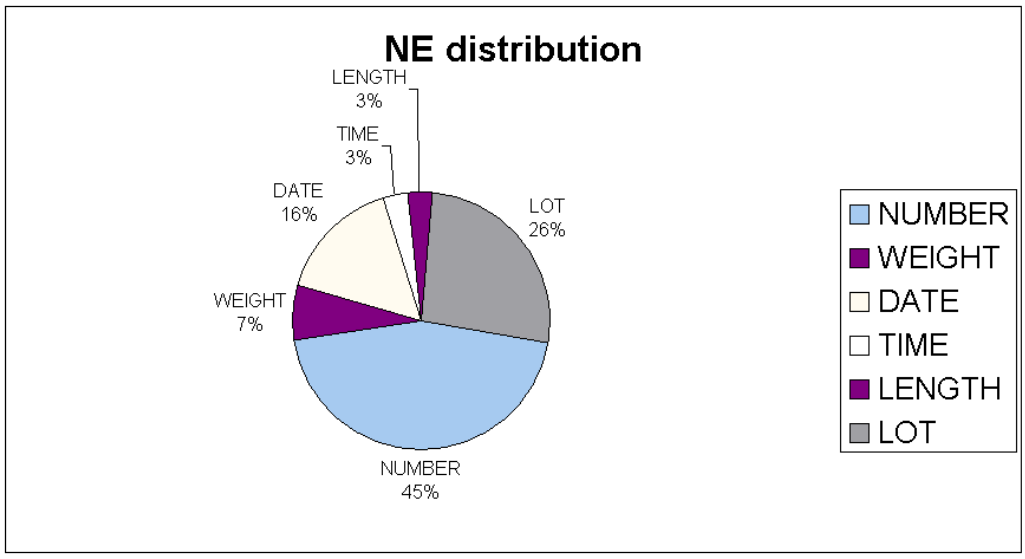

Figure 13: distribution of numerical NEs instances in the ZooBirth500 corpus.

3.4.1 Initial and Final Set of SNEs

NB (Number of newBorns):

The ***four*** cubs are only some of the many new arrivals being celebrated at the zoo.

Lion births in captivity are rare and big litters rarer still, so having ***three males and two females*** born to parents Henry and Vilas last week was a singular achievement for Madison's free zoo.

Winnipeg zoo officials are crediting the province's climate for their newest zoo babies - ***two*** exotic red pandas.

NO (Number of Offspring produced by the mother):

This is Maliku's **first** baby and she has settled into motherhood exceptionally well.

By day, zookeepers work with the ***first-time*** mother to prepare her for the birth and motherhood.

But zoo spokesman Paul Garcia conceded Friday morning that "time is definitely not on our side" for the ***second*** birth.

NOF (Number of Offspring sired by the Father):

Over the past decade, the zoo has had four successful births and its resident bull has sired a total of ***eight*** babies.

The 12-year-old red panda became a ***first-time*** father last year at the Rosamond Gifford Zoo at Burnet Park.

Bulwagi's calf in Toledo is his ***first*** offspring and only the second surviving African elephant born through artificial insemination.

NOL (Normal size Of Litter):

They only have ***one*** baby at a time.

A typical litter size is ***two to four*** cubs.

female duoc langurs give birth to ***one*** infant.

WP (size of Wild Population):

The Gelada population is now estimated at about ***100,000*** and the animal appears on some endangered lists, including Appendix II of the international CITES conservation list, which permits only monitored trade of the creatures between countries.

According to some estimates, the last ***15,000*** specimens survive in the wild now, including 10,000 in Namibia.

Fewer than 5,000 remain in their native Indonesia.

HWP (Historical size of Wild Population):

Myanmar's tiger population has plunged drastically to ***150*** or fewer in the wild in recent years from an estimated 3,000 nearly 25 years ago.

Lynx numbers have declined from ***100,000*** at the beginning of the 20th century to just 100 to 120 today, according to the the conservation group WWF.

Once numbering ***more than 200,000***, there are now an estimated 800 wild Nene in Hawaii.

WPR (size of Wild Population – Regional):

Fewer than 400 are believed to survive in the wild, about ***20*** of them in China and the rest in Russia.

According to some estimates, the last 15,000 specimens survive in the wild now, including ***10,000*** in Namibia.

The number in China is up from 30 to almost ***100***.

GCP (Global Captive Population):

Only ***175*** eastern black rhinos are in captivity around the world, with 72 in the North American captive population.

The 2-inch-high baby tamarins are part of a worldwide population that experts say numbers between ***200 and 500*** in captivity and 5,000 to 7,000 in the wild.

About ***200*** exist in captivity and 5,000 in the wild.

RCP (Regional Captive Population):

The plan began in 1982 and includes all ***44*** Asian rhinos in U.S. zoos.

“These births are particularly significant in terms of conservation because Asifa and Asal are two of only ***30*** Saudi goitered gazelles in North American zoos," said Ron Kagan, director of the Detroit Zoological Institute (DZI).

No South China tigers are known to survive in the wild, and there are only about ***40*** of them in zoos in China, according to the official Xinhua News Agency.

NBC (Number of Births in Captivity within a specific period):

There are just under ***100*** tigers in European zoos with around a dozen births of cubs a year.

Of the world's 140 captive pandas, this is only the ***fifth*** birth in 2005, following the arrival of two sets of twins at a panda conservation centre in China.

The population enjoyed ***12*** births in the past year, while it also saw 13 deaths.

ZS (number of Specimens at the Zoo belonging to the same species of the newborn):

The Zoo is home to *five* reticulated giraffes, the variety that has solid spots with defined edges as opposed to the more leafy pattern other giraffe varieties sport.

In addition to their sisters, Imara and Hatima, the National Zoo has ***nine*** cheetahs on exhibit - including the litter of five cubs born in April 2005.

This newborn is the ***fourth*** prehensile-tailed porcupine at the National Zoo.

ZSB (number of Specimens Born at the Zoo):

The Dvur Kralove zoo, in the north of the Czech Republic, on Wednesday announced the birth of a male black rhino, the ***29th*** of the endangered species born since breeding began at the zoo in 1971.

This was the ***third*** successful birth of an elephant in captivity for the Taiping Zoo, he said, adding that the zoo was fortunate to have a mature male elephant and a conducive environment for elephant breeding.

Fifteen gorillas have been born and raised at Zoo Atlanta since 1988.

PHON (Phone numbers):

Contact: Sarah Taylor, ***202/633-3081***; Peper Long, ***202/633-3082***

For more information call ***(248) 398-0900*** or visit http://www.detroitzoo.org/ .

NOTES: Zoo spokeswoman Lora LaMarca may be reached at ***(213) 666-4650, ext. 275***.

DOBN (Date of Birth of the Newborn):

The cheetahs, along with their two sisters, were born on ***Nov. 23, 2004*** , and were the Zoo's first litter of cheetahs in its 115-year history.

Polar bear "Rara" delivered her baby in ***December last year.***

Five-year old female cow Piroschka gave birth to her first calf on ***July 5***.

DOBM (Date of Birth of the Mother):

Abu was the first baby for its mother Sabi, born in Zimbabwe in ***1985***.

The mother was born in ***1991*** and hand-raised by keepers at the Cincinnati Zoo.

She was born in the wild around ***Jan. 1, 1976***, and orphaned when she was several months old.

DOBF (Date Of Birth of the Father):

The father, Triton was born at the Roger Williams Park Zoo in Providence, Rhode Island on **November 5, 1977** , and came to the Detroit Zoo in 1999.

The father, Garth, was born at the Henry Doorly Zoo through that process on **Nov. 5, 1991**.

Spike, who was born in **1992** at the Cincinnati Zoo, arrived at Cleveland Metroparks Zoo in August 1994.

DOP (Date of birth of Previous offspring):

Amanda's first-born, delivered here in ***1995***, died of malnutrition before attendants could determine that the infant had a cleft palate.

Kirina was born in ***June 1995***, while Tundi was born in ***July 1991***.

The father -- Pandu -- had been moved to the Philadelphia, Pennsylvania zoo to make room in the zoo's elephant house for his offspring and the new mother, who gave birth to a stillborn calf in ***July 1992***.

DOPZ (Date Of Previous births at the Zoo):

In summer ***2004*** ligress Zita was born in the zoo.

The others were born at the San Diego Zoo in **1999 and 2003.**

The first calf, named Kaisei, was born in ***late January*** , Crocker said.

DNOE (Date the Newborn will be put On Exhibit):

The mother and cubs are expected to stay off-exhibit until ***mid-June***.

Como Zoo anticipates the new calf will make her public debut on Mother's Day, ***May 8th***.

The piglets should be out for the public to see by the ***first weekend in May*** at the latest.

DOAM (Date Of Arrival of the Mother at the zoo):

Santi, a white tiger from India, was presented to the Surabaya Zoo in ***1992***.

They are Hua Mei's second set of cubs since she was sent from the San Diego Zoo to the Wolong Giant Panda Research Center in China in ***February 2004***.

Romina was brought to Bristol in ***2001*** as part of an international breeding program.

DOAF (Date Of Arrival of the Father at the zoo):

Tenang, now six, came as a two year old from Perth, Australia, in ***late 2002***.

The father is Michael, an 18-year-old male orangutan that came from Utah's Hogle Zoo in ***October 2000***.

He arrived here in ***March 2001*** from the Lincoln Park Zoo in Chicago, zoo officials said.

OOL (time of Onset of Labour):

At ***9:20*** the next morning, Emi began active labor.

'"We believe she had her first contraction ***between 1 and 2***," Doyle said.

Kathleen South, a spokeswoman for the zoo and aquarium in Tacoma, said the mother whale, Mauyak, went into labor about ***5 a.m***. yesterday

TOB (Time Of Birth):

The pair of un-named and un-sexed Sumatran cubs were born on Monday - one at ***2.44pm***, the second at ***4.15pm*** and are the first tiger cub births since 1988 .

Panda mom Bai Yun delivered her third cub just ***before 10 p.m.*** Tuesday.

A baby male giraffe was born ***shortly after midnight*** Wednesday, the first animal birth at the Erie Zoo this season.

ZOH (Zoo's Opening Hours):

The Toledo Zoo, voted fourth best zoo or aquarium in the Midwest by Family Fun magazine, is open daily from ***10 a.m. to 5 p.m.***

Como Zoo's baby Tamarins are on exhibit every day in the Primate Building. Como Zoo's hours are ***10 a.m. to 6 p.m.***

Visitors can see the young alpaca beginning ***10 a.m***. May 1, 1997 when the facility opens for the summer.

WB (Weight at Birth of the newborn):

"At birth, it was about ***three pounds*** in weight and the size of a football," Hubing said.

The unnamed calf, a healthy ***275 pounds*** at birth, is nursing, and mama Renee is attentive.

The ***344 gram (12 ounce)*** still-unnamed female was born at the zoo Sunday to 19-year-old mother Hecla.

WC (Current Weight of the newborn):

The baby bear, which weighed only 600 grams when born, now weighs ***16 kilograms***.

At her last weigh-in she was a healthy ***175 pounds***.

Now, it weighs ***28 ounces*** and is 40 inches long.

WBR (Range of Weight at Birth):

A panda cub typically weighs from ***3 ounces to 5 ounces*** and is no larger than a stick of butter.

A baby weighs an average of ***100 to 125 kg***, with its mother gaining about 400 kg during pregnancy.

Newborn orangutans generally are ***3 1/2 to 4 pounds***, which is Schmidt's best guess for this baby.

WA (Weight of the Adult):

A full-grown panther weighs around ***60 to 70kg***.

The monkeys average ***500-600 grammes*** and live in rain forest trees along South America's Atlantic coast line.

A grown-up male weighs up to ***a tonne***.

LOB (Length Of newborn at Birth):

The calf weighed in at 135 pounds and stood nearly ***5 1/2 feet*** tall on arrival.

The ***4-centimetre*** long newborn turtle weighs 12 grammes.

The female baby giraffe, or calf, weighed 151 pounds and was ***5'9***" at birth.

LOA (Length Of Adult):

The adults, which grow to a length of ***12 inches***, have lobes around the neck that can be raised to deter predators or rivals - hence the name.

Small and fuzzy-faced, the animals are expected to begin eating meat in their third month and eventually grow to ***more than two metres*** in length.

The baby river otter is one of the largest species in the zoo, growing up to ***1.5 metres***.

AGM (AGe of the Mother):

Zoo Atlanta is celebrating the rare birth of twin gorillas, delivered overnight by Kuchi, ***21***, and fathered by Taz, a 16-year-old silverback.

He's the first offspring for Ellie, a ***20-year-old*** African elephant.

The previous kittens of the mother, ***six-year-old*** lioness Nelly, were all brought up on artificial nutrition.

AGF (AGe of the Father):

An ***eight-year-old*** tigress and a ***seven-year-old*** tiger gave birth to the cubs more than a month ago.

Yvonne, 13, and her male counterpart, Saba, 12, lived for about 10 years together at the zoo without any interest in each other.

In addition to Bai Yun, 12, and her cub, Hua Mei, Gao Gao, a male believed to be about ***11 to 13 years old***, is in residence.

G (gestation duration):

Giraffe pregnancies last ***15 months***.

Sloth bears usually mate in early summer and females give birth ***six to seven months*** later.

Born on Tuesday - about a month earlier than the usual ***22-month*** gestation period for elephants - the calf and the mother, 29-year-old African elephant Kubwa, were in good health, zoo officials said.

LS (Life Span):

An estimated 12,000 to 15,000 cheetahs survive in the wild, where they typically live ***eight to 10 years***.

Mothers are only able to bear four or five young during their lifetime and can live to their ***late 50s*** in captivity.

They can live for up to ***35 years***.

LDOP (Length of time since the birth Of Previous Offspring of the same mother):

Bai Yun, who proved herself to be an adept mother with her first cub, Hua Mei, born ***four years ago*** today, is a little more casual with this youngster, Lindburg noted.

Three years ago, Mandara gave birth to her first baby, a male -- the National Zoo's first gorilla birth in 19 years.

By all accounts, the new baby is doing very well and has already started sparring with his ***2-year-old*** brother, who outweighs him by about 2,000 pounds.

LDOPZ (length of time since last birth of an animal of the same species at the zoo):

It's been ***10 years*** since the zoo had Siberian tiger cubs.

Five months after a cheetah gave birth at the National Zoo for the first time in its history, another one of the speedy cats has had a litter.

Another scarce rhinoceros has been born at the National Zoo, the second one in ***six weeks***.

NOE (length of time until the newborn is put On Exhibit):

It will be at least ***four to six weeks*** before the trio make a public appearance and are expected to be a big draw for visitors.

"The two cubs will be moved to another cage that can be shown by visitors when they are ***four months*** old," he said.

While the cub won't be on exhibit for about ***three months***, the zoo's Web site will be updated weekly with current pictures.

LOLR (Length Of Labour):

The cub spent most of yesterday sleeping atop its mother, who endured ***three-hour*** labor.

The 13-year-old Bai Yun gave birth to the 112 gram (four-ounce) cub on Tuesday night after ***three hours*** of labour, said Don Lind, panda exhibit team leader at the San Diego Zoo.

The elephant, not yet named, was born to "Pang Pha," who weathered a ***nine-hour*** delivery after carrying her daughter for 668 days.

LDOBN (Length of time since the recent birth):

The baby bison were born ***two weeks ago*** to two different mothers on the same day.

An eight-year-old tigress and a seven-year-old tiger gave birth to the cubs ***more than a month ago***.

BABY May-Tagu shelters from the rain under her mother, just ***two weeks*** after making her extraordinary entrance into the world.

LDOAM (Length of time since the Arrival Of the Mother at the zoo):

Two white tigers, a rare species, came to the Moscow zoo ***several years ago***, the tigress from Sweden and the tiger from the Netherlands.

The mating pair were brought to Tulsa ***more than two years ago***, she said.

Sheena and her sister, Dhari, a white tiger, were brought to the zoo ***less than a month ago***.

LDOAF (Length of time since the Arrival Of the Father at the zoo):

Their father, Doni, was brought from Minnesota ***last year*** to breed with them.

The father, Raymond, is from the Philadelphia Zoo and has lived in Madison's zoo for ***three years***.

Miri's father, Batu, was born at Allwetter Zoo in Munster, Germany and arrived at Twycross ***nine years ago.***

LDEP (Length of time the newborn is expected to be DEpendant on the mother):

Those babies can be dependent on their mothers for the ***better part of 10 years*** in the wild," Dr. Shumaker said.

After the cubs are weaned at ***six months*** the zoo may decide to send any of its five tigers to other accredited facilities to maintain genetic diversity within the captive breeding population.

Staggs said the baby will live with his mother for another ***10 months*** or so, until he reaches sexual maturity.

NUM (all other numerical instances):

There are only ***four*** patches of forest in Sumatra that have viable populations.

Dvur Kralove is one of the main European zoos specialising in African wildlife, with many animals able to roam free on its ***64*** -hectare site.

About ***15,000*** people visit the zoo every month.

DATE (all other date instances):

The zoo has housed zebras since ***1966*** but acquired the Grant's species just a few years ago.

Zoo veterinarians hoped to capture a pregnancy on a sonogram, but the panda had not sat still for one minute since ***June 20***.

Romina, a western lowland gorilla, underwent two operations to restore her sight in ***April 2002*** and ***September 2003***

TIME (all other instances of time):

It was about ***1 a.m***. ***EDT*** when the volunteer on duty at the Panda House noticed on camera that Mei Xiang seemed restless and unable to settle down.

Twycross is to host a presentation called Conservation of the Orang-utan - Securing Their Future, on October 30 at ***6.15pm***.

The zoo did not release the news of the birth until after ***4 p.m***., but some visitors were aware of it.

WEIGHT (all other instances of weight):

In the wild, young rhinos gain weight at the rate of ***50 pounds*** per month and spend the first three years with their mother.

Each will eat ***10 to 15 pounds*** per day.

A baby weighs an average of 100 to 125 kg , with its mother gaining about ***400 kg*** during pregnancy.

LENGTH (all other instances of length):

For now, the unnamed female has gray hair and miniature versions of her parents' nearly ***two-foot***-long horns.

There are thought to be only around 5,000 snow leopards left in the wild, living in the mountains of central Asia, roaming at altitudes of up to ***6,000 metres (20,000 feet)*** in China, Nepal, Mongolia, India, Afghanistan and Russia.

LOT (any other instance of Length Of Time):

Tenang, now six, came as a ***two year*** old from Perth, Australia, in late 2002.

Senior zookeeper Mat Shah Mahadon said the cubs were fed milk from a bottle every ***three hours***.

The cub will not be given a name until it is ***100 days*** old.

3.4.2 SNEs Chosen For The Study

Table 6: most common SNEs that were selected as the test set in the study.

SNE	Total number of instances in ZooBirth500	Total number of instances in ZooBirth700
NB (number of newborns)	344	432
WB (newborn weight)	274	365
AGM (age of the mother)	273	350
NO (number of offspring by the same mother)	238	279
DOBN (birth date of newborn)	226	307
WP (population size in the wild)	181	240
ZS (number of specimens at the zoo)	171	204
WA (adult weight)	129	165
G (gestation duration)	126	158
AGF (age of the father)	115	144

Table 7: distribution of the ten SNEs in ZooBirth500

Number of SNEs in document	Number of documents
1	169
2	146
3	91
4	59
5	23
6	7
7	4
8	0
9	1

The figures in tables 6 and 7 suggest that the ten SNEs provide good coverage of the domain. In other words, it would be highly unlikely to find a report about animal births in zoos that does not feature at least one of these SNEs. This probably reflects the journalistic conventions in reporting such stories as described in section 3.2.

3.4.3 Notes About Marking Up Studied SNEs

WB (Weight of newborn at Birth): sometimes the actual weight at birth was not reported but instead the normal newborn weight in that species was given. When it was felt that this was meant to replace the reporting of the actual weight, it was marked up. When the species' weight at birth was reported alongside general facts about the species it was not marked up.

WP (Wild Population): when several figures were reported, i.e., for a certain

country/region or a subspecies alongside the world population in the wild, only the figure for the global wild population was marked.

NO (Number of Offspring): any entity was marked which indicated unambiguously the number of former pregnancies, litters, or offspring produced previously by the same mother.

G (Gestation): when both actual gestation and the normal value for the species were reported, only the actual duration was marked up.

AGM (Age of the Mother): in most cases the age of the mother was reported in the form of X-year-old. Dates of birth were not marked up. They were rare.

3.5 Chapter Summary

The chapter introduced the domain chosen for this project. Next, the compilation and annotation of the ZooBirth corpus was described, followed by a full list of Specific Named Entities (SNEs) with related descriptive statistics and illustrative excerpts from the corpus. The next chapter will describe the use of CRF in this study and the method employed for cross validation in the experimental design.

Chapter 4: CRF setup and Evaluation

4.1 Outline

The chapter explains how a CRF tool was used to train the system to recognise SNEs, with details on the list of baseline features. This is followed by a report on evaluation, the cross-validation method of choice (5×2) and statistical testing.

4.2 CRF++

CRF++ (Kudo 2007) is the open source general purpose toolkit used in this study. It implements Conditional Random Fields for data segmentation/labelling following Lafferty et al. (2001). The user has to specify the feature template in advance to be used in training. During testing (decoding) CRF++ uses the model file generated in training. The training and test/validation files must consist of multiple tokens and a fixed but unlimited number of columns. Each token to be tagged must be represented by one line, with the columns separated by white spaces. The input sequence unit is a sentence. The boundary between sentences is denoted by an empty line in the file. The last column in the training file is the true answer tag (figure 15). Template files have to be prepared separately: special macros in the format `%x[row,col]` specify a token in the input data. `row` specifies the relative position from the current token up to four tokens down/upstream, and `col` specifies the absolute position of the column. For example, in the training data in figure 15, if the current token is 'offspring', the template `%x[0,1]` will be expanded to the feature 'token' and %x[-1,1] will be expanded to the feature 'number'. CRF++ automatically generates a set of feature functions such as `func1 = if (output = ag and feature="U01:number") return 1 else return 0`. 'U01' is a unique identifier.

The	token	0
baby	token	0
is	token	0
the	token	0
17th	number	0
offspring	token	0
of	token	0
the	token	0
37	number	ag
-	token	ag
year	token	ag
-	token	ag
old	token	ag
Coco	token	0
,	token	0
Pate	token	0
said	token	0
.	token	0

Figure 14: a simplified example of the training input format expected by CRF++. Here the first column (column 0) is a sentence sequence of tokens from ZooBirth and the second (column 1) a basic tag (see baseline features). The last column is the answer tag: in this case '0' or the SNE tag `AG` (age of mother).

CRF++ offers the choice between unigram and bigram template modes. Here the terms 'unigram' and 'bigram' relate to the **output tags**. In a bigram mode, a combination of the current output token and previous output token is automatically generated. The bigram mode generates a total of (L×L×N) distinct features, where L is the number of output classes and N is the number of unique strings expanded from a given template. This can be inefficient compared to the unigram mode where the number of distinct features is only L×N.

When CRF++ is applied to large set of data the number of unique features can amount to millions. In this case, the `-f NUM` parameter can set a cut-off threshold so only features that occur no fewer than `NUM` times in the given training data are used. The

default value is 1. In the experiments reported here the value was set to 3.
The test output of CRF++ is the test data file in the same format used in training but with the inferred answer tags in the last column. Probability is reported for each tag and for the entire sentence output.

Answer tags used in this study only indicated SNE type (or '0' for a non-SNE token). There are more elaborate labelling models in which tags indicate the position of the token **within** entity (Leaman and Gonzalez 2008): the IO model (Inside/Outside), IOB (Inside/Outside/Beginning) and the most complex, IOBEW (Inside/Outside/Beginning/End/one-Word entity). However, using any of these models would have increased the number of output classes considerably and made running CRF experiments to recognise multiple SNE types impractical due to the computational cost.

The pre-processing and feature extraction of the ZooBirth data was implemented in Prolog (see Appendix B). The list clauses storing the data were written out to training and test files compliant with the CRF++ format.

4.3 Baseline Features

To test the hypotheses and approaches proposed in this study, the baseline machine learning setup incorporated only minimal external knowledge in the form of rudimentary tagging. Each token from the ZooBirth corpus was tagged in the training file as a `weight` (unit), `month`, `year`, `number`, `ordinal number`, or otherwise `token` based on list lookup. Another, binary column in the file indicated if the token was a capitalised word or not. The CRF++ baseline template (figure 16) took into account the current token, its basic tag and capitalisation, the preceding and following

four tokens and **their** basic tag and capitalisation binary value. The template was set to binary mode (see section 4.2). The example presented in figure 17 is of a short sentence extracted from the corpus. The leftmost column is column 0 under which the sentence tokens are listed, each on a separate row. The following column (column 1) is the basic tags assigned to each token. The values under column 2 are binary (0 or 1) and indicate if the current token is capitalised. Column 3 (the fourth column) is the answer tag.

```
U01:%x[-4,0]
U02:%x[-3,0]
U03:%x[-2,0]
U04:%x[-1,0]
U05:%x[4,0]
U06:%x[3,0]
U07:%x[2,0]
U08:%x[1,0]
U09:%x[0,0]
U10:%x[0,1]
U11%x[-1,1]
U12%x[-2,1]
U13%x[-3,1]
U14%x[-4,1]
U15%x[1,1]
U16%x[2,1]
U17%x[3,1]
U18%x[4,1]
U19%x[0,2]
U20%x[1,2]
U21%x[2,2]
U22%x[3,2]
U23%x[4,2]
U24%x[-1,2]
U25%x[-2,2]
U26%x[-3,2]
U27%x[-4,2]

B
```

Figure 15: the CRF++ template used in baseline runs. U numbers are unique identifiers followed by %x[row,column]. Row 0 refers to the current token. Column 1 is its basic tag (token, weight, number, ordinal number, month, year). Column 2 is the binary, indicating if the token is a capitalised word or not. B denotes bigram mode.

Ripley	token	1	0
,	token	0	0
who	token	0	0
weighs	token	0	0
only	token	0	0
a	token	0	wb
kilogram	weight	0	wb
or	token	0	wb
two	number	0	wb
,	token	0	0
was	token	0	0
a	token	0	0
surprise	token	0	0
birth	token	0	0
.	token	0	0

Figure 16: an example of a short sentence extracted from ZooBirth with its baseline features in training-ready format for CRF++ .

In the example shown in figure 17, the input features of the token 'two' are the token itself, 'two', its tag 'number', its capitalisation binary value 0 **and** the same features in each row of the four-token window around 'two' (ie, 'only a kilogram or', and ', was a surprise'). As explained earlier, CRF++ creates a feature function during training. For instance, in this case the following feature function will be generated: `func1 = if (output = wb and feature="U12:weight") return 1 else return 0.`

4.4 Evaluation

The different feature configurations were evaluated by dividing the ZooBirth corpus into training and testing sets, specifically following a 5×2 cross-validation method (see

below). The judgment of the answer tag output of CRF++ and the computing of precision, recall and F-measure were completed automatically.

4.4.1 Judgement of SNE tagging

The CRF++ output answer tags were judged at the SNE, **not** the token level. A contiguous sequence of tokens labelled with the same SNE type was considered to be an instance of an SNE identified by the system. If the system tagged only part of the SNE sequence (eg, only the `a kilogram` in `a kilogram or two`) the answer was judged as partially correct. Such partial matching can also occur when the SNE answer sequence contains superfluous tokens. As was discussed in section 4.2.3 about performance, partial answers can still be valuable for information extraction tasks and lenient measures which include them are often reported in the literature. In this study both lenient and strict versions of precision, recall and F-measure were calculated.

4.4.2 Choosing 5 × 2 Cross Validation

When creating training/validation set pairs from a dataset, the sets should be as large as possible to provide robust error estimates; In addition, the overlap between them should be minimal. Also, the respective proportion of the classes (SNE types in this case) should be preserved in the held out data subsets.

One option is to use K-fold cross-validation in which the dataset is repeatedly divided randomly into K equal-sized parts. One part is used for validation, while the remaining K – 1 parts are combined to form the training set. The problem is that the validation/testing set is small. Furthermore, the training sets overlap to a great degree (any two training sets share K – 2 parts).

The considerable effort that was required to single-handedly annotate ZooBirth limited the size of corpus that could be generated. The relatively small size of ZooBirth (n = 500 and later n = 700) meant that it could not be divided randomly into K parts, with each divided into training and validation halves. In cases where machine learning algorithms need to be assessed using small datasets it is common to use cross-validation (Alpaydin 2004) in which the same data is used repeatedly but split differently each time. The drawback is that the error percentages are not independent because the different sets share data.

The 5×2 cross-validation method was proposed by Dietterich (1998). The difference from K fold cross-validation is that the training and validation sets are of equal size. The dataset is divided randomly into two halves: one is used for training and the other for validation. Then the roles of the halves are reversed: the first half now becomes the validation set and the second used for training. This is the first fold. To create five folds the process is repeated four more times with the dataset shuffled randomly before every two-way split. The result is ten training and validation sets. More than five folds can be created, but beyond five the overlap is too great and the error rates too dependent and hence uninformative.

4.4.4 Statistical Testing

The statistical test used to compare the performance of various feature configurations in this study is the **5×2 cv paired *F* test** (Alpaydin 2004) where F refers to the Fisher test, not the F-measure encountered in this thesis. The test is an extension of the 5×2 cv paired *t* test which was proposed by Dietterich (1998) . The two are intended to test the null hypothesis that two classification algorithms have the same error rate.

In the context of this study $P_i^{(j)}$ is the difference between the error rates of two CRF experiment runs on fold $j = 1,2$ of replication $i = 1,\ldots,5$.

The average on replication i is $\overline{p_i} = (p_i^{(1)} + p_i^{(2)})/2$ and the estimated variance is $s_i^2 = (p_i^{(1)} - \overline{p_i})^2 + (p_i^{(2)} - \overline{p_i})^2$. If $p_i^{(1)}$ and $p_i^{(2)}$ are assumed to be independent normals with unknown variance σ^2 (which is not strictly true because their training and validation tests are not independent), then s_i^2/σ^2 has a chi-square distribution with one degree of freedom. If each of the s_i^2 is assumed to be independent, their sum is chi-square with five degrees of freedom:

$$M = \frac{\sum_{i=1}^{5} s_i^2}{\sigma^2} \sim \chi_5^2$$

and

$$t = \frac{p_1^{(1)}}{\sqrt{M/5}} = \frac{p_1^{(1)}}{\sqrt{\sum_{i=1}^{5} s_i^2 / 5}} \sim t_5$$

giving a t statistic with five degrees of freedom. The 5×2 cv paired F test replaces the arbitrary numerator with one which combines the ten possible statistics:

$$N = \frac{\sum_{i=1}^{5} \sum_{j=1}^{2} (p_i^{(j)})^2}{\sigma^2} \sim \chi_{10}^2$$

The result is the ratio of two chi-square distributed random variables, and F distributed with ten and five degrees of freedom:

$$f = \frac{N/10}{M/5} = \frac{\sum_{i=1}^{5} \sum_{j=1}^{2} (p_i^{(j)})^2}{2\sum_{i=1}^{5} s_i^2} \sim F_{10,5}$$

The null hypothesis is accepted at significance level α if this value is less than $F_{\alpha,10,5}$ ($F_{0.05,10,5} = 4.74$ was used to evaluate the results in this study).

4.5 Chapter Summary

This chapter explained how CRF was used in practice to recognise SNEs, with details on the baseline feature setup. Next, evaluation methods were discussed, concluding with an introduction to 5×2 cross-validation, and the 5×2 cv paired *F* test used to infer statistical significance of results produced by this cv method. The next chapter opens the experimental work part of this thesis, presenting experiments with positional features.

Chapter 5: Experiments with Positional Features

5.1 Outline

The chapter covers experiments in which positional features were used to train CRF to recognise the ZooBirth SNEs. It begins with related statistics and is followed by a detailed description of the various experiments. The last section discusses the results.

5.2 Positional Features

Positional features denote the position of every token in the document at the token, sentence and paragraph levels. To allow generalisation when training, the ordinal numbers of tokens, sentences and paragraphs within the document were normalised by dividing them by the total number of tokens, sentences and paragraphs in the document, respectively. The value of the ratio was discretised by assigning it one of five letters denoting five equal intervals within the range of 0 to 1 (see table 8).

Table 8: positional features used in training CRF++.

Positional feature	Feature label in CRF++ training file
Token position (TokPos)	$NormalisedTokPos = \frac{\text{Token's ordinal number in document}}{\text{Total number of tokens in document}}$ $a = 0 < NormalisedTokPos < 0.2$ $b = 0.2 \leq NormalisedTokPos < 0.4$ $c = 0.4 \leq NormalisedTokPos < 0.6$ $d = 0.6 \leq NormalisedTokPos < 0.8$ $e = 0.8 \leq NormalisedTokPos \leq 1.0$
Sentence position (SenPos)	$NormalisedSenPos = \frac{\text{Sentence's ordinal number in document}}{\text{Total number of sentences in document}}$ $a = 0 \leq NormalisedSenPos < 0.2$ $b = 0.2 \leq NormalisedSenPos < 0.4$ $c = 0.4 \leq NormalisedSenPos < 0.6$ $d = 0.6 \leq NormalisedSenPos < 0.8$ $e = 0.8 \leq NormalisedSenPos \leq 1.0$
Paragraph position (ParagPos)	$NormalisedParagPos = \frac{\text{Paragraph's ordinal number in document}}{\text{Total number of paragraphs in document}}$ $a = 0 \leq NormalisedParagPos < 0.2$ $b = 0.2 \leq NormalisedParagPos < 0.4$ $c = 0.4 \leq NormalisedParagPos < 0.6$ $d = 0.6 \leq NormalisedParagPos < 0.8$ $e = 0.8 \leq NormalisedParagPos \leq 1.0$

5.3 Exploration of SNE Positional Distribution within Documents

The rationale behind the idea to use positional features to improve the performance of SNE recognition is that the restricted domain of animal births in zoos seems to follow thematic structures. It was hypothesised that such underlying structures would mean that an SNE is likely to appear earlier or later in the news report, and that these differences would become significant when disambiguating SNEs subsumed in the same NE.

The histograms in figures 18–21 show the positional distribution of time, date and weight SNEs within a document. In figure 18 it can be seen that the SNE `time-of-birth` (TOB) appears earlier in the news report whereas a reverse trend can be observed in relation to the `zoo's opening hours` (ZOH). This is most noticeable in the histogram's leftmost column. It clearly shows that if an instance of the NE time is detected in one of the first sentences in a ZooBirth report, it is most likely to be the time of birth, rather than the zoo's opening hours or other type of time mention which has not been predefined. The time of birth relates specifically to the subject of the story, while the zoo's opening hours, though useful, are not newsworthy information. Similarly, figure 19 shows a growing proportion of non story-specific instances of date towards the end of the report. Therefore, the date of birth of the newborn, who is the subject of the news report, is more likely to appear early in the report; the date of birth of a previous sibling (DOP) never appears in the first interval (column a); and the date of arrival of the mother to the zoo (DOAM) can appear in small proportion throughout the document but with slightly larger frequency in the middle of the news story. In figure 20, `weight-at-birth` (WB) occurs with greater frequency earlier in the document, whereas the `weight-of-adult` (WA) appears with highest frequency towards the document's end (column d). These exploratory graphic findings indicate

that positional features should be useful for SNE recognition.

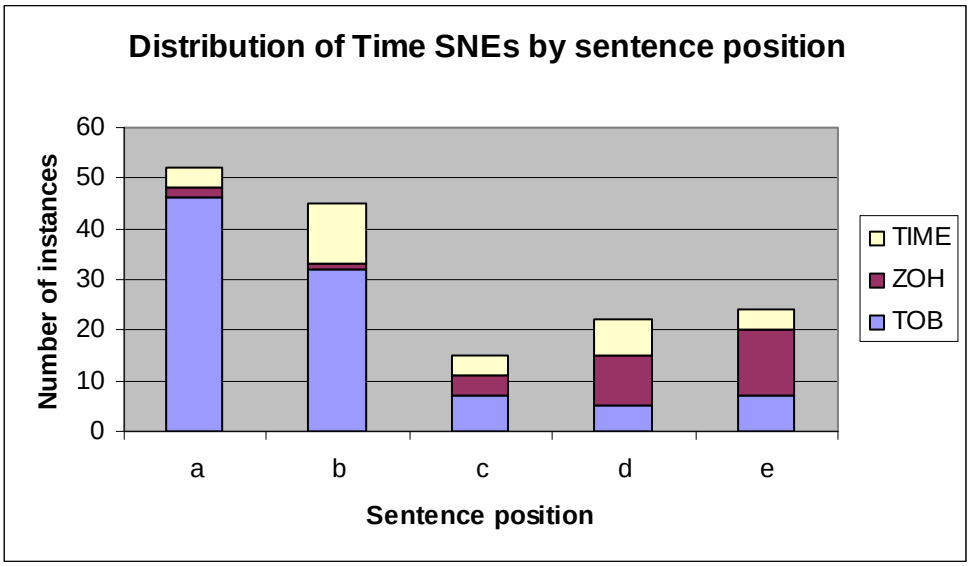

Figure 17: Positional distribution of time SNEs. TOB = Time Of Birth, ZOH = Zoo's opening hours, TIME = any other instance of time.

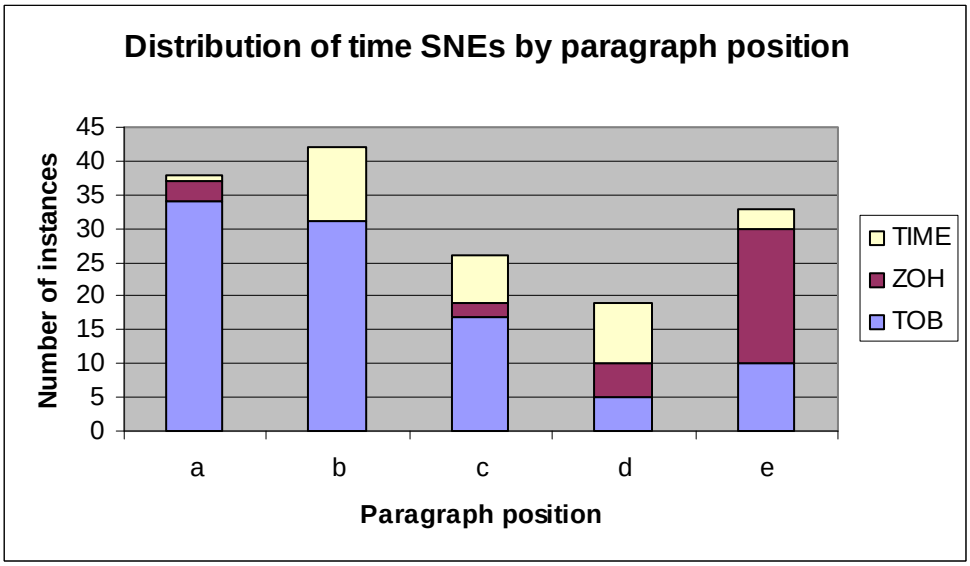

Figure 18: Positional Distribution of Time SNEs. TOB = Time Of Birth, ZOH = Zoo's opening hours, TIME = any other instance of time.

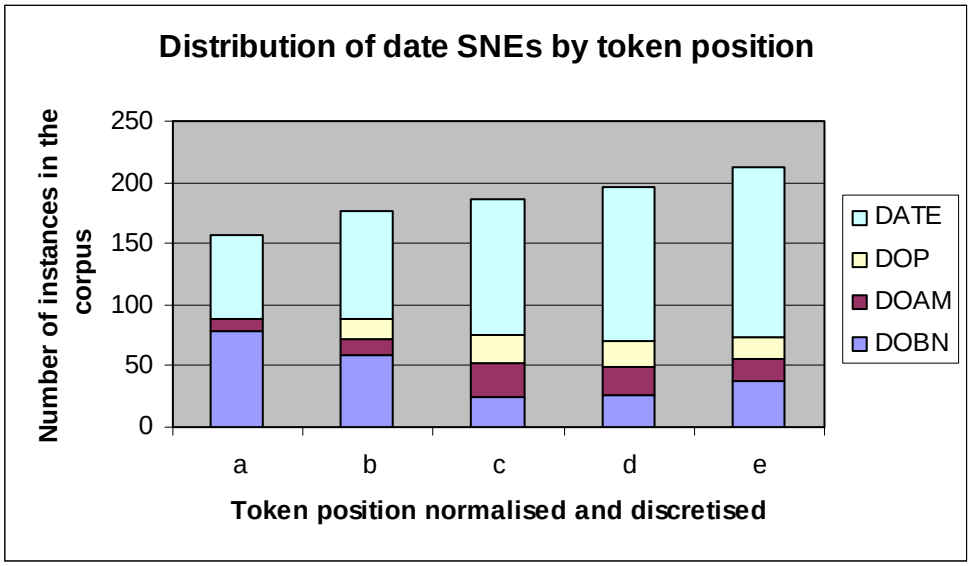

Figure 19: Positional distribution of date SNEs. DOBN = Date Of Birth of Newborn, DOAM = Date Of Arrival of the Mother, DOP = Date Of Birth of Previous sibling, DATE = any other instance of date.

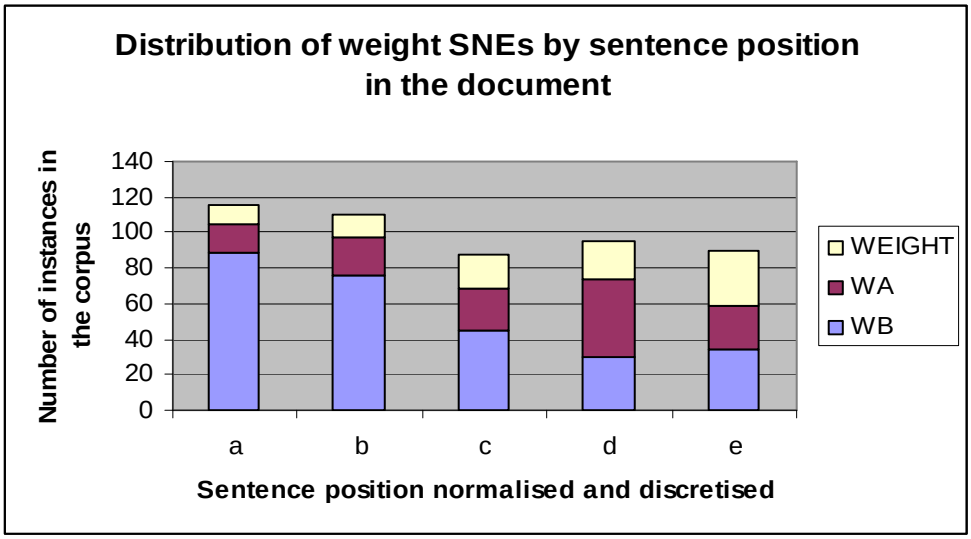

Figure 20: Positional distribution of weight SNEs. WB = Weight at Birth, WA = Weight of Adult, WEIGHT = any other instance of weight.

5.4 Results of CRF Runs Using Positional Features

The possible benefit of positional features was evaluated by training and testing the system on the ZooBirth700 collection. The experimental design and statistical testing followed the 5×2 cross-validation procedure (see Section 4.4). Each run tested the ability of the system to recognise one type of SNE. Recognising more than one SNE simultaneously is very time consuming. More importantly, labelling more than one type of SNE, especially if they belong to the same NE, affects performance due to possible complex interactions (see next chapter). Therefore, in the following experiments the task of the system was to simply label each token as belonging to an SNE instance or not. Tables 9–18 present the results for each of the ten SNEs. Each SNE experiment consisted of five runs:

1. baseline (baseline features as described in section 4.3)
2. baseline + paragraph position feature
3. baseline + sentence position feature
4. baseline + token position feature
5. baseline + all three position features

Table 9: Recognising the SNE ZS (number of specimens at the zoo) in the ZooBirth700 corpus using positional features.

5 by 2	Precision (S)	Precision (L)	Recall (S)	Recall (L)	F (S)	F (L)
Base line	77.2	84.2	26.3	28.9	39.2	42.8
Paragraph position	74	80.5	26.1	28.5	38.3	41.9
Sentence position	76.4	82.6	26	28.1	38.7	41.8
Token position	74.8	81.7	25.9	28.3	38.4	41.9
paragraph + Sentence + Token positions	74.4	81	27.1	29.4	39.7	43

Table 10: Recognising the SNE WA (adult's weight) in the ZooBirth700 corpus using positional features

5 by 2	Precision (S)	Precision (L)	Recall (S)	Recall (L)	F (S)	F (L)
Base line	70.1	73.7	26.1	44.3	37.7	55.1
Paragraph position	70.1	73.2	27	43.4	38.6	54
Sentence position	70.3	73.8	26.3	44	37.7	54.7
Token position	68.5	71.4	25.5	42.9	36.7	53.2
paragraph + Sentence + Token positions	67.2	70.5	26.8	43.7	37.8	53.4

Table 11: Recognising the SNE WB (weight at birth) in the ZooBirth700 corpus using positional features.

5 by 2	Precision (S)	Precision (L)	Recall (S)	Recall (L)	F (S)	F (L)
Base line	78.6	82.4	62.3	69.3	69.4	75.1
Paragraph position	78.9	83.1	61.5	68.6	69	75
Sentence position	78.9	82.9	61	68.2	68.7	74.7
Token position	78.5	82.6	61.5	67.9	68.9	74.3
+ paragraph + Sentence + Token positions	79.1	82.9	61	68	68.8	74.5

Table 12: Recognising the SNE WP (population in the wild) in the ZooBirth700 corpus using positional features.

5 by 2	Precision (S)	Precision (L)	Recall (S)	Recall (L)	F (S)	F (L)
Base line	78.2	82.2	42.8	70.1	55.2	75.6
Paragraph position	77.9	81.6	42.7	70.4	55	75.5
Sentence position	78.2	82.1	42.6	70.5	55	75.7
+ Token position	79.1	82.6	43.2	70.9	55.6	76.2
+ paragraph + Sentence + Token positions	77.6	81.6	43.2	70.8	55.3	75.7

Table 13: Recognising the SNE AGM (age of the mother) in the ZooBirth700 corpus using positional features. A star denotes statistical difference from the corresponding value in table 2. (5 x 2 cv paired F test, $F_{0.05,10,5} = 4.74$).

5 by 2	Precision (S)	Precision (L)	Recall (S)	Recall (L)	F (S)	F (L)
Base line	82.4	83.3	65.5	66.6	72.9	73.9
Paragraph position	81.8	83.1	66.1	67.3	73	74.3
Sentence position	82.1	83.3	67.5	68.8	74	75.2
Token position	82.8*	84*	66.9	68.4	73.9	75.3
+ paragraph + Sentence + Token positions	82.4	83.6	66.8	68.3	73.8	75.1

Table 14: Recognising the SNE AGF (age of the father) in the ZooBirth700 corpus using positional features. A star denotes statistical difference from the corresponding value in table 2. (5 x 2 cv paired F test, $F_{0.05,10,5} = 4.74$).

5 by 2	Precision (S)	Precision (L)	Recall (S)	Recall (L)	F (S)	F (L)
Base line	75.1	81.2	36.7	36.9	48.9	50.3
Paragraph position	70.8	75.8	35.9	36.1	47.2	48.5
Sentence position	72.2	77.3	37	37.3	48.8	50
+ Token position	73.1	77.5	36	36.3	48	49.1
+ paragraph + Sentence + Token positions	68.8*	73	35.6	35.9	46.6	47.8

Table 15: Recognising the SNE G (gestation duration) the ZooBirth700 corpus using positional features.

5 by 2	Precision (S)	Precision (L)	Recall (S)	Recall (L)	F (S)	F (L)
Base line	94.1	97.7	53.6	60.9	68.1	74.7
+ **Paragraph position**	93.9	96.7	53.7	60.8	68	74.3
+ **Sentence position**	93.1	96.3	52.9	59.6	67.2	73.4
+ **Token position**	93.6	96.9	53.1	60.2	67.4	73.8
+ **paragraph** + **Sentence** + **Token positions**	93.3	96.6	52.7	59.6	67	73.4

Table 16: Recognising the SNE NB (number of newborns) in the ZooBirth700 corpus using positional features.

5 by 2	Precision (S)	Precision (L)	Recall (S)	Recall (L)	F (S)	F (L)
Base line	72.4	73.1	35.7	37.1	47.7	49.2
Paragraph position	71.2	72.3	35.6	36.9	47.4	48.8
Sentence position	72.5	73.7	35.6	37.1	47.7	49.3
Token position	72.7	73.8	36.7	38.3	48.7	50.3
+ paragraph + Sentence + Token positions	72.4	73.5	37.1	38.7	49	50.6

Table 17: Recognising the SNE NO (number of offspring produced in ZooBirth700 using positional features.

5 by 2	Precision (S)	Precision (L)	Recall (S)	Recall (L)	F (S)	F (L)
Base line	82.7	84.1	34.3	35	48.3	49.2
Paragraph position	83.7	85.1	34.7	35.3	48.9	49.7
Sentence position	83.6	85.4	35.1	35.7	49.2	50.1
Token position	84.3	85.8	35.4	36	49.6	50.5
+ paragraph + Sentence + Token positions	83.1	84.8	35.4	36	49.4	50.3

Table 18: Recognising the SNE DOBN (date of birth of newborn) in the ZooBirth700 corpus using positional features. A star denotes statistical difference from the corresponding value in table 2. (5 x 2 cv paired F test, $F_{0.05,10,5} = 4.74$).

5 by 2	Precision (S)	Precision (L)	Recall (S)	Recall (L)	F (S)	F (L)
Base line	72.7	74.1	51.9	54.7	60.3	62.7
Paragraph position	72.5	73.6	50.0	53.3	58.9	61.6
Sentence position	72.6	73.7	50.9	54.1	59.5	62.2
Token position	74.4	76.1	52.1	55.4	61.1*	64*
+ paragraph + Sentence + Token positions	72.4	73.9	52.4	55.7	60.6	63.4

5.5 Discussion

Statistically significant positive results were observed in the case of two SNEs (AGM, DOBN). In both cases these positive results were achieved when token position was used as a feature. In a single instance a negative result was statistically significant (AGF, strict precision measure of paragraph + sentence + token position run, Table 14). This may be related to AGF being the rarest of the ten SNEs (see table 6), hence a small sample. Non-significant mixed trends were observed across the SNEs: For example, positional features had a negative effect on the recognition of G (table 15), whereas a positive effect was seen in the NO run (table 17). This mixed picture suggests that positional features can improve performance but only when certain SNEs are concerned. Although in both cases where statistical significance of positive results was observed the feature used was token position, the sample (ie, number of SNEs with such results) is too small to conclude that token position is a universally superior positional feature. It is worth noting that in both these token position runs all performance measures of precision and recall showed improvement over the baseline even though not all were statistically significant.

Exploration of SNE distribution as presented in section 5.3 suggests that different SNEs can be typified by positional features. However, they appear to be weak features that can only make statistically significant contribution in certain scenarios.

5.6 Chapter Summary

The chapter described experiments in which positional features were used to train CRF to recognise ten SNEs in ZooBirth700. First, the distribution nature of SNEs within documents was explored. This was followed by the experimental results. The chapter concluded with a short discussion on the effectiveness of positional features.

Chapter 6: Experiments with Order Effects

6.1 Outline

The chapter reports on experiments which exploit the relative order of SNEs as features in CRF. The first section describes the effect of concurrent recognition of SNEs on performance. The second and main section focuses on testing features which capture the order of SNEs belonging to the same NE (weight, date).

6.2 Recognising Multiple SNEs Concurrently

As noted in chapter 5, recognising more than one type of SNE in a single run can affect performance due to possible interactions between the features of the SNEs. Compare the baseline performance of the system on ZooBirth500 when recognising each of the ten SNEs on its own to a run where all of them were recognised simultaneously (tables 19 and 20, respectively): concurrent recognition seems to improve recall and lower precision. The effect was found to be significant in the case of AGF and WB, both with NE counterparts (AGM and WB).

A similar effect on precision and recall can be seen when comparing a run in which DOBN was recognised on its own to one where it was labelled alongside other date entities (Table 21). When CRF was trained on text which was labelled as either DOBN, DATE or 0, recall of DOBN rose to 60.4% (strict) and 60.9% (lenient) while precision declined to 71.5% (strict) and 72.8%* (lenient). When DOBN was recognised with DOAM (Date of Arrival of the Mother at the zoo), DOP (Date of birth of Previous siblings), and DATE, recall of DOBN improved and precision declined further.

It should be noted that when CRF is trained to label any instance of date without discriminating between date SNEs using the standard base line features used

throughout this project, the performance is relatively high (table 23). These figures are within the range reported in the literature. For example, Ahn et al. (2005) achieved 85.5% strict precision, 74.8% strict recall and an F measure of 79.8% when training CRF with extra features to recognise temporal expressions (which are broader than explicit month/year dates. Without the extra features their results were 79.8, 68.5 and 73.7 respectively.

Table 19: baseline performance of CRF++ when each of the ten SNEs is recognized in ZooBirth500 independently, in a separate run. Precision and recall values are averages based on five replications of twofold cross- validation: each average reported here is of the five twofold averages. Features used in a baseline run are current token, four previous and four following tokens, pre-tag (number, weight, month, year, animal) of current token, and capitalization.

SNE	Precision (s)	Precision (L)	Recall (s)	Recall (L)
AGM	81.8	82.7	64.1	66.1
AGF	75	79.6	31.1	31.8
G	93.6	97.3	50.2	56.5
NB	73.6	74.8	25.8	26.4
WA	74.5	78.7	35.8	38
WB	73.7	80	58.6	65.2
DOBN	72.9	75.3	54.7	55.8
WP	87.8	91.7	57.5	61.5
NO	80.7	82.3	29.7	31
ZS	71.4	79.7	21.5	24

Table 20: baseline performance of CRF++ when each of the ten SNEs is recognized in ZooBirth500 alongside the remaining nine in a single run. Precision and recall values are averages based on five replications of twofold cross- validation: each average reported here is of the five twofold averages. Features used in the baseline runs are current token, four previous and four following tokens, pre-tag (number, weight, month, year, animal) of current token, and capitalization. A star denotes statistical difference from the corresponding value in table 2. (5 x 2 cv paired F test, $F_{0.05,10,5} = 4.74$).

SNE	Precision (s)	Precision (L)	Recall (s)	Recall (L)
AGM	77	79.1	67.5	70.6
AGF	70.6	73	38.3*	39.3*
G	92.6	97	48.9	57.4
NB	66.1	72.6	28.9	31*
WA	63.3*	73.3	44.8*	49.3
WB	70.2	76.1	65.3	72.3*
DOBN	74.6	76.8	55	55.9
WP	87.5	91.1	59.8	63.7
NO	77.2	78.9	29.8	31
ZS	66.9	74.7	23	27

Table 21: Recognising DOBN instances with or without other date SNEs.

Run	P(S)	P(L)	R(S)	R(L)
DOBN on its own	72.9	75.3	54.7	55.8
DOBN with DATE	71.5	72.8	60.4*	60.9*
DOBN with DOAM, DOP and DATE	69.5	70.4*	62.9*	63.2

Table 22: Recognising all instances of the NE date.

P(S)	P(L)	R(S)	R(L)	F(S)	F(L)
88.8	90.3	80.5	82.9	84.4	86.5

6.3 Order Effects When Recognising Related SNEs

Section 5.3 of the previous chapter showed that SNEs subsumed in the same NE differ in their positional distribution within documents. However, these differences were aggregated across documents. For example, although the weight of the newborn (WB) is more likely to appear in the beginning of the document, whereas the weight of the adult tends to appear towards the end of the document, those observations could come from separate news reports containing only one of either SNE. This section examines the relation between related SNEs within the same document, and how it can be exploited for NER.

6.3.1 Weight SNEs

To remind the reader, four main types of weight entities were identified in the corpus. Table 23 shows their frequency in ZooBirth500, the corpus used in these experiments.

Since WC and WBR are rare, and training CRF to identify more than 3-4 SNEs simultaneously is time consuming, the task was simplified to labelling WA, WB, and WEIGHT. The WEIGHT label subsumes all instances of WC and WBR.

Table 23: Weight SNEs

SNE	Number of instances in the ZooBirth500 corpus
WB (weight of newborn at birth)	273
WA (weight of adult)	129
WC (current weight of newborn)	50
WBR (normal range of weight at birth)	24
WEIGHT (any other instance of weight)	21

Three features were introduced that attempt to capture the relation between weight SNEs when tagging them:

1. **Simple Count (SC)**: The absolute number of instances of weight measurement units in the document.

2. **First/Not First (FT)**: This feature can take the value `a` (first instance in the document) or `b` (any subsequent instance of weight unit in the document). It takes the value `s` when there is a single instance of weight unit in the document.

3. **Order (O)**: The ordinal number of the weight unit instance in the sequence of weight units in the document. In documents containing a single instance the value of this feature is `s`.

Four settings of these features were selected for the experiment: SC, FT + SC, O, O + SC. The settings FT and FT + O were not tested because O already encapsulates the information of FT. The results are given in three tables, one for each weight SNE. They are analysed in section 6.3.3.

Table 24: Recognising WB alongside WA and WEIGHT using the features Simple Count (SC), First/Not First (FT), and Order (O). A star indicates statistical significance (5 x 2 cv paired F test, $F_{0.05,10,5} = 4.74$). Best F measure is in bold.

Run	P(S)	P(L)	R(S)	R(L)	F(S)	F(L)
Baseline	68.6	74.8	68.1	76.1	68.4	75.5
SC	70.4	77.5	69.0	77.8	69.7	77.6
FT + SC	74.2	81.4	70.1	79.8	**72.1***	**80.6***
O	72.9	79.9	70.2	79.3	71.5	79.6
O + SC	73.4	80.6	70.1	79.7	71.7	80.2

Table 25: Recognising WA alongside WB and WEIGHT using the features Simple Count (SC), First/Not First (FT), and Order (O). A star indicates statistical significance (5 x 2 cv paired F test, $F_{0.05,10,5} = 4.74$). Best F measure is in bold.

Run	P(S)	P(L)	R(S)	R(L)	F(S)	F(L)
Baseline	66.2	70.6	50.8	55.4	57.5	62.1
SC	66.1	71.3	54.6	60.8	59.8	65.6
FT + SC	66.5	71.5	56.2	62.8	60.9	**66.8**
O	67.8	72.7	55.8	61.3	**61.2**	66.5
O + SC	67.2	72.4	55.2	61.8	60.6	66.7

Table 26: : Recognising WEIGHT alongside WB and WA using the features Simple Count (SC), First/Not First (FT), and Order (O). A star indicates statistical significance (5 x 2 cv paired F test, F0.05,10,5 = 4.74). Best F measure is in bold.

Run	P(S)	P(L)	R(S)	R(L)	F(S)	F(L)
Baseline	48.3	50.8	24.3	26.7	32.3	35.0
SC	44.1	47.1	25.4	27.6	32.2	34.8
FT + SC	48.7	51.3	28.8	32.3	**36.2***	**39.7***
O	47.8	50.0	26.8	30.0	34.3	37.5
O + SC	48.6	50.7	27.7	31.2	35.3	38.7

6.3.2 Date SNEs

A procedure identical to the one used to recognise weight SNEs in section 6.3.1 was followed when labelling date SNEs. The SNEs recognised were DOBN, DOAM (Date Of Arrival of the Mother at the zoo), DOP (Date Of Birth of Previous sibling) and DATE (any other instance of date). The results are presented in tables 27–30.

Table 27: Recognising DOBN alongside DOAM, DOP and DATE using the features Simple Count (SC), First/Not First (FT), and Order (O). A star indicates statistical significance (5 x 2 cv paired F test, F0.05,10,5 = 4.74). Best F measure is in bold.

Run	P (S)	P (L)	R (S)	R (L)	F (S)	F (L)
Base line	69.5	70.6	62.6	63.2	65.9	66.7
SC	70.5	71.9	62.9	63.6	66.5	**67.5**
SC + FT	69.7	70.8	63.7	64.2	**66.5**	67.4
O	69.9	71.2	63.4	64.0	66.5	67.4
SC + O	69.5	70.4	62.9	63.2	66.1	66.6

Table 28: Recognising DOP alongside DOBN, DOAM and DATE using the features Simple Count (SC), First/Not First (FT), and Order (O). A star indicates statistical significance (5 x 2 cv paired F test, F0.05,10,5 = 4.74). Best F measure is in bold.

Run	P (S)	P (L)	R (S)	R (L)	F (S)	F (L)
Base line	64.1	64.1	22.1	22.6	32.9	33.4
SC	60.1	60.1	24.1	24.7	**34.4**	**35.0**
SC + FT	65.4	65.4	22.6	23.0	33.6	34.0
O	59.9	59.9	22.7	23.1	32.9	33.4
SC + O	69.8	69.8	21.7	22.1	33.1	33.6

Table 29: Recognising DATE alongside DOBN, DOAM, DOP and DATE using the features Simple Count (SC), First/Not First (FT), and Order (O). A star indicates statistical significance (5 x 2 cv paired F test, F0.05,10,5 = 4.74). Best F measure is in bold.

Run	P (S)	P (L)	R (S)	R (L)	F (S)	F (L)
Base line	60.5	62.2	59.5	62.4	60.0	62.3
SC	60.6	62.3	59.7	62.5	60.1	62.4
SC + FT	60.5	62.2	59.9	62.5	**60.2**	**62.4**
O	60.6	62.5	60.2	63.3	60.4	62.9
SC + O	60.0	62.0	59.0	61.0	59.5	61.5

6.3.3 Discussion

In the case of the weight SNEs the use of simple count of measurement units in the document (SC) in combination with a feature that indicates whether the current measurement unit is the first instance in the document (FT) produced statistically significant improvement in the F measure. In addition, a positive trend was observed across all measures of precision and recall, though not statistically significant.

The picture is less clear when the date SNEs are concerned. Firstly, no statistically significant results were attained. Secondly, the results were mixed with no particular setting emerging as consistently beneficial across the SNEs. Only in the case of DOAM, the best F measure involved the feature O. With all three other date SNES, the best F measures were achieved with either SC or SC + FT.

In certain situations it is possible that whenever there is a single instance of weight or date they are likely to be a particular SNE. For example, in ZooBirth500, of the 153 documents that contain a single instance of date, 54 contain a DOBN (35%) and 52 (34%) the generic DATE.

The first mention of an SNE is normally refers to the subject of the report, ie the newborn animal so it is not surprising that a feature which discriminates between a first mention and any other mention should be effective. For example, normally WB precedes WA. For example:

> The calf was, born March 24, stands about 2 feet tall and weighs approximately **30 pounds**. When fully grown, the alpaca will stand about four feet high and weigh approximately **200 pounds**.

However, here is an opposite example, where WB follows WA:

> An **8,300-pound** elephant and her newborn, who was **280 pounds** at birth, are doing well at a breeding compound and retirement center for elephants in this southeastern Oklahoma city.

The contribution of the SC feature might be thanks to the structure of a news story, that is often described as an inverted pyramid, with the most essential information at the top (Gabbay and Sutcliffe 2004). This means that in a detailed report with multiple related SNEs the first mention may be more likely to be about the main subject of the report.

6.4 Chapter Summary

The chapter reported on experiments which exploit the relative order of SNEs as features in CRF. The first section introduced the effect of concurrent recognition of SNEs: reduced precision and improved recall. The second and main section described the use of features meant to capture order effects within a document. In these experiments date and weight SNEs were recognised, with statistically significant results when labelling weight SNEs.

Chapter 7: Subtractive Tagging

7.1 Outline

The chapter introduces a novel training method to enrich negative examples when training CRF. The next section reports the results of testing the method's performance in tagging instances of four of the SNEs followed by a discussion.

7.2 The Subtractive Tagging Method

So far, negative examples provided when training CRF to label SNEs were simply tokens labelled as 0s or tags of other SNEs. The method introduced here is meant to generate an additional category of negative examples automatically and is demonstrated with the SNE DOBN.

Step 1: ZooBirth700 corpus was partitioned randomly five times to P_A and P_B (figure 22). Each of the two partitions contains exactly 250 documents. This is identical to the way the corpus has been processed in order to run 5×2 cross-validations. All instances of DOBN are then annotated in these documents.

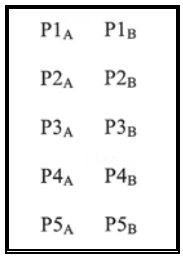

Figure 21: random fivefold partitioning of the corpus (step 1 of subtractive tagging).

Step 2 (figure 23): All annotated instances of DOBN and their context, a window of up to 4 tokens on each side, as set in CRF++, are subtracted from the corpus partitions. The two cleared partitions in each of the five folds are referred to here as A' and B'.

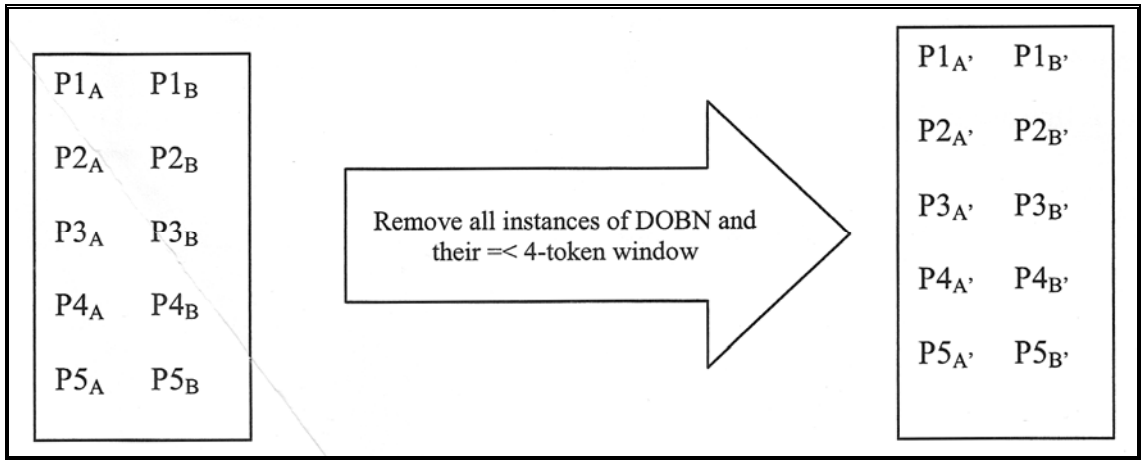

Figure 22: step 2 of subtractive tagging.

Step 3 (figure 24): CRF is trained on each original partition and tested on the DOBN-free version of the partition.

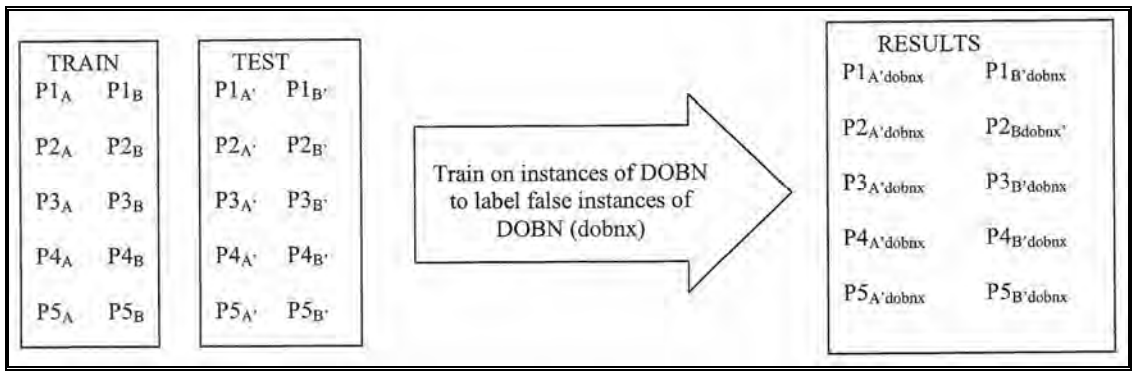

Figure 23: step 3 of subtractive tagging.

Step 4 (figure 25): The tags of false DOBN (`dobnx`) are added to their respective original partitions (A/B).

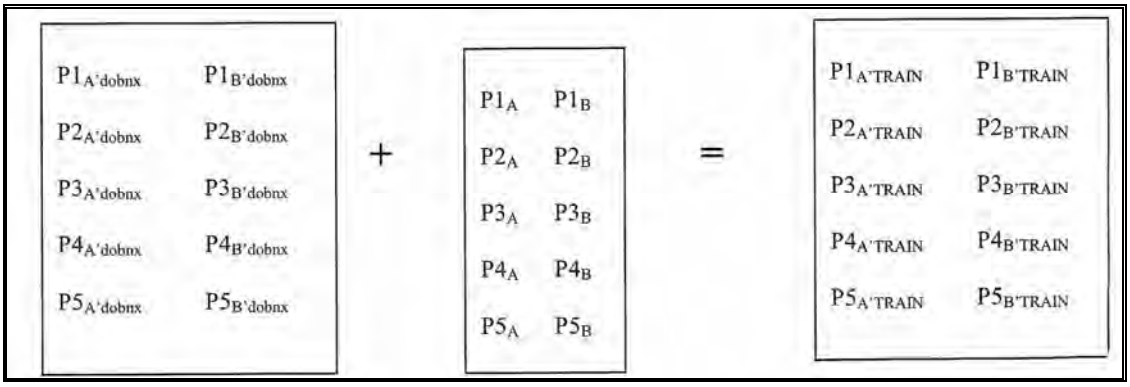

Figure 24: step 4 of subtractive tagging.

In the **final step**, in each of the five folds, CRF is first trained on the A'_{TRAIN} partition and tested on the original version of the B partition, then trained on B'_{TRAIN} and tested on A. CRF labels tokens as 0, `dobn`, or `dobnx`.

Initially, the standard base line features that have been used throughout the thesis (see section 4.3) were also used in step 3. However, this set of features guaranteed that CRF assigned no `dobn` tags to tokens in the DOBN-free version. To overcome this problem, the feature set was reduced to `weight` (unit), `month`, `year`, `number`, `ordinal number` and capitalisation, ignoring the surface text.

7.3 Results

Subtractive tagging improved recall, both strict and lenient, with statistical significance when tagging instances of DOBN. The slight decline in precision was not statistically significant. Table 30 shows the results.

Table 30: Performance of CRF when labelling instances of DOBN using the subtractive tagging method.

Run	P(S)	P(L)	R(S)	R(L)	F(S)	F(L)
Baseline	73.3	74.4	52.8	54	61.1	62.4
Subtractive tagging	70.8	72.2	60.1*	61.7*	64.7	66.3

The following are examples of false instances of DOBN which were tagged in step 3 of the method (see last section):

> The zoo is holding a contest to name the cubs. Called “Dub the Cubs,” it is being conducted starting Monday through **July 24**.
>
> She’s been under night watch since **March 1**, and the zoo staff is weary but prepared.
>
> The zoo has launched a competition to name the newborn rhino with a deadline of **February 28** for suggestions.
>
> So the zoo on Aug. 1 started intensive, round-the-clock care and put her on a special diet starting **Aug. 1**.

Table 31: Performance of CRF when labelling instances of AGM, NO and NB using the subtractive tagging method.

SNE	RUN	P(S)	P(L)	R(S)	R(L)	F (S)	F(L)
AGM	Baseline	82.4	83.3	65.5	66.6	72.9	73.9
	Sub Tag	82.2	83.1	68.9*	69.8*	73.8	75.1
NO	Baseline	82.7	84.1	34.3	35.0	48.3	49.1
	Sub Tag	82.0	83.5	35.8	36.5	49.5	50.5
NB	Baseline	72.3	73	35.7	37.1	47.7	49.2
	Sub Tag	71.3	72.3	34.4	36.2	46.3	48.1

Table 31 shows the tagging results of three additional SNEs. AGM's results are similar to DOBN's: improved recall, both strict and lenient, with statistical significance and a slight decline in precision which was not statistically significant. Tagging NO confirms this trend but in the case of NO, without statistical significance. Subtractive tagging deteriorated performance when recognising instances of NB, without statistical significance.

7.4 Discussion

Improving recall is highly desirable as CRF's precision has been consistently higher than recall across the various experiments reported in this thesis. The subtractive tagging method which uses the output of one CRF run as features in the testing stage is similar in concept to the two-stage approach suggested by Krishnan and Manning (2006) to handle non-local dependencies in NER. However, subtractive tagging is different from other two-pass NER because it is used for pre-processing of the *training* set, as opposed to using the output from a first pass over the testing set.

Fine-tuning the reduction of the feature set that will eventually be used in the final testing stage, in order to tag false positives for training could be a crucial step to capture weaknesses in the system.

It might be expected that presenting specific negative examples to CRF should improve precision, not recall. However, blanket tagging in the training set of negative examples (ie, a `O` tag), may present it with too many instances that are similar to true positives but are assigned the `O` tag, making it 'err on the safe side' when tested. This could lead to low recall.

7.5 Chapter Summary

This chapter introduced a novel method called subtractive tagging to label false positives in the training set. Statistically significant improvement in recall was observed when testing the method with CRF tagging instances of the SNEs DOBN and AGM.

Chapter 8: Conclusions

8.1 Outline

This final brief chapter first presents the key findings of the project followed by suggestions for further research.

8.2 Key Findings

1. A novel named-entity corpus was compiled from a news archive. Manual annotation, using personal domain expertise, resulted in marking up over 5,400 instances of numerical entities of which 2,811 belonged to the ten most frequent subtypes, the focus of this study.

2. An extensive and rich set of numerical entities specific to a new restricted domain was identified. Of these, the most common SNEs proved useful in NER experimentation.

3. SNEs differed in their typical positional distribution within a given document. This observation highlighted the potential of various positional features.

4. Positional features (token position) improved performance with statistical significance in the case of two SNEs: DOBN and AGM. Strict precision of AGM tagging was only slightly higher than the baselines (82.8 and 82.4, respectively). Strict F-measure of DOBN tagging was 61.1 compared to a baseline value of 59.5. The observed trends were mixed and suggest that positional features are only effective when used to tag particular SNEs.

5. Tagging multiple SNEs concurrently lowered precision but improved recall, in some cases with statistical significance of the recall and F measures: Strict and lenient recall of AGF tagging improved from 31.1 to 38.3 and from 31.8 to 39.3, respectively; The strict F measure of WA rose from 35.8 to 44.8; Lenient F measure of NB improved from 26.4 to 31, and that of WB from 65.2 to 72.3.

6. The order of related SNEs (ie, subtypes of the same NE) within a document can be used as a feature in CRF when the NE can be readily recognised using simple heuristics. In an experiment with weight SNEs, a simple count of the number of instances of weight units in the document, and a feature indicating if a weight unit instance was the first in the document, not first, or a single instance improved with statistical significance the F-measure when tagging WB (from 68.4 to 71.1 strict, and from 75.5 to 80.6 lenient) and WEIGHT (from 32.3 to 36.2 strict, and from 35 to 39.7 lenient)

7. A novel technique to automatically tag false positives with CRF by self training it on an SNE-free version of the training set, and then using the output in the testing phase, proved effective in improving recall. The method was tested on four SNEs. In the case of DOBN the strict precision was reduced from 73.3 to 70.8 without statistical significance while strict recall improved from 52.8 to 60.1 with statistical significance. When tagging instances of AGM, strict precision declined slightly from 82.4 to 82.3 whereas recall improved from 65.5 to 68.9 and with statistical significance.

8. CRF performs well out-of-the box at the NE level but across all the various experiments showed low recall. However, with minimal external and linguistic knowledge and only short distance context, state-of-the-art performance was not

expected, and in that sense it can be said that CRF performed reasonably well. It is also a useful tool to test various hypotheses. One drawback of CRF is that the way it works is often opaque (compared to hand-written rules) especially when used to learn a complex interaction between features.

8.3 Further Research

ZooBirth could be published and made available for re-use by the information extraction research community. Technically, if the annotated corpus is made public, it would be desirable to convert the nonstandard SNE tagging to XML format, more widely used in corpus annotation. The Prolog clause database could be transformed to XML using an SGML Prolog library and additional XML writers. However, distributing the corpus outside the University of Limerick by uploading the annotated documents online depends on obtaining permission from the copyright holders. In response to a query about securing such permission, a licensing manager at LexisNexis suggested contacting each news source used. An alternative to distributing the corpus itself is to provide enough data (ie, document identifiers and a corresponding SNE database) to colleagues; this should allow them to reconstruct ZooBirth on their own.

The work on this project started before crowd-sourcing annotation (Lawson et al. 2010, Finin et al. 2010) became available. These services may now allow researchers to compile a larger SNE corpus while not having to devote a considerable amount of time to manual annotation, as long as the domain is not too specialised (eg, bioinformatics). Creating new resources would be essential to test FG-NER in various strict domains. It would be interesting to repeat some of the experiments reported here on a significantly larger corpus in the zoo birth domain, with many more instances of the rarer SNEs, or port the methods to a whole new restricted domain. For example, the positional distribution of unknown SNEs could be aid unsupervised NER. Positional distribution

patterns of SNEs could also be used to profile new genres of text.

Subtractive tagging should be investigated further with different permutations to find the optimal settings of features in each of the method's two stages.

Improved computational power should make multi-SNE tagging by CRF more wieldy and therefore improve recall.

8.4 Chapter Summary

This chapter summed up the key findings of the experimental work and made a few suggestions for further research.

References

Ahn, D., Adafre, S. F. and Rijke, M. d. (2005) 'Extracting Temporal Information from Open Domain Text: A Comparative Exploration', *Journal of Digital Information Management*, 3(1), 14–20.

AKT-Technologies (2012) 'ANNIE - Open Source Information Extraction from The University of Sheffield', [online], available: http://www.aktors.org/technologies/annie/ [accessed 2 January, 2012].

Alpaydin, E. (2004) *Introduction to Machine Learning* Cambridge, MA: MIT Press.

Altincay, H., Dimililer, N. and Varoglu, E. (2009) 'Classifier subset selection for biomedical named entity recognition', *Applied Intelligence*, 31(3), 267-282.

Ananiadou, S., Friedman, C. and Tsujii, J. (2004) 'Introduction: named entity recognition in biomedicine', *Journal of Biomedical Informatics*, 37(6), 393-395.

Ananiadou, S. A., S., Sullivan, D., Black, W., Levow, G. A., Gillespie, J. J., Mao, C. H., Pyysalo, S., Kolluru, B., Tsujii, J. and Sobral, B. (2011) 'Named Entity Recognition for Bacterial Type IV Secretion Systems', *Plos One*, 6(3).

Andersen, P. M., Hayes, P. J., Huettner, A. K., Schmandt, L. M., Nirenburg, I. B. and Weinstein, S. P. (1992) 'Automatic extraction of facts from press releases to generate news stories', in *Proceedings of the third conference on Applied natural language processing*, Trento, Italy, 974531: Association for Computational Linguistics, 170-177.

Aone, C., Charocopos, N. and Gorlinsky, J. (1997) 'An intelligent multilingual information browsing and retrieval system using information extraction', in *Proceedings of the fifth conference on Applied natural language processing*, Washington, DC, 974606: Association for Computational Linguistics, 332-339.

Armour, Q., Japkowicz, N. and Matwin, S. (2005) 'The Role of Named Entities in Text Classification ', *CLiNE 05: 3rd Computational Linguistics in the North-East Workshop*, available: [accessed 23 Aug, 2008].

Ayala-Guerrero, F., L. Vargas-Reynaa, Ramos, J. I. and Mexicanoa, G. (1998) 'Sleep patterns of the volcano mouse (Neotomodon alstoni alstoni)', *Physiology & Behavior*, 64(4), 577-580.

Banko, M., Cafarella, M., Soderland, S., Broadhead, M. and Etzioni, O. (2007) 'Open Information Extraction from the Web', in Veloso, M. M., ed. *Proceedings of the 20th International Joint Conference on Artificial Intelligence*, Hyderabad, India, 6–12 Jan, 2007, Menlo Park, CA, USA: AAAI Press, 2670-2676.

References

Barzilay, R. and Lee, L. (2004) 'Catching the Drift: Probabilistic Content Models, with Applications to Generation and Summarization', in *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, Boston, Massachusetts, USA, 2-7 May, 2004, East Strousburg, PA, USA: Association for Computational Linguistics, 113-120.

Basis-Technology (2011) 'Rosette Entity Extractor', 2011(December 9, 2011)available: http://www.basistech.com/datasheets/Rosette-Entity-Extractor-EN.pdf [accessed 2 January, 2012].

Beeferman, D., Berger, A. and Lafferty, J. (1999) 'Statistical Models for Text Segmentation', *Machine Learning*, 34(1-3), 177-210.

Benajiba, Y., Diab, M. and Rosso, P. (2009) 'Using Language Independent and Language Specific Features to Enhance Arabic Named Entity Recognition', *International Arab Journal of Information Technology*, 6(5), 464-472.

Beneti, A., Hammoumi, W., Hielscher, E., Mller, M. and Persons, D. (2006) *Automatic Generation of Fine-Grained Named Entity Classications*, technical report, Amsterdam: University of Amsterdam.

Berger, A. L., Pietra, V. J. D. and Pietra, S. A. D. (1996) 'A maximum entropy approach to natural language processing', *Comput. Linguist.*, 22(1), 39-71.

Bestgen, Y. and Vonk, W. (1995) 'The role of temporal segmentation markers in discourse processing', *Discourse Processes*, 19, 385–406.

Blaschke, C., Hirschman, L. and Valencia, A. (2002) 'Information extraction in molecular biology', *Briefings in Bioinformatics*, 3(2), 154-166.

Blei, D. M. and Moreno, P. J. (2001) 'Topic segmentation with an aspect hidden Markov model', in *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, New Orleans, Louisiana, United States, September 9-12, 2001, ACM, 343 - 348.

Bodenreider, O. (2004) 'The Unified Medical Language System (UMLS): integrating biomedical terminology', *Nucleic Acids Research*, 32(suppl 1), D267-D270.

Bollen, J., Mao, H. and Zeng, X. (2011) 'Twitter mood predicts the stock market', *Journal of Computational Science*, 2(1), 1-8.

Boston University (2007) 'Phylogeny of Sleep Database', [online], available: http://www.bu.edu/phpbin/sleep/search/ [accessed April 30, 2008].

References

Bottou, L. (1991) *Une Approche théorique de l'Apprentissage Connexionniste: Applications à la Reconnaissance de la Parole*, unpublished thesis (PhD thesis), Université de Paris XI.

Brierley, C. and Atwell, E. (2011) 'Terrorist Activities: Making Sense', in *Proceedings of International Crime and Intelligence Analysis Conference 2011*, Manchester.

Brüninghaus, S. and Ashley, K. D. (2001) 'Improving the representation of legal case texts with information extraction methods', in *Proceedings of the 8th international conference on Artificial intelligence and law*, St. Louis, Missouri, United States, ACM, 42-51.

Bunescu, R. and Pasca, M. (2006) 'Using Encyclopedic Knowledge for Named Entity Disambiguation', in McCarthy, D., Wintner, S. and (chairs), eds., *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, Trento, Italy, East Stroudsburg, PA, USA: The Association for Computer Linguistics 9-16.

Cafarella, M. J., Madhavan, J. and Halevy, A. (2008) 'Web-Scale Extraction of Structured Data', *Sigmod Record*, 37(4), 55-61.

Carme, J., Ceresna, M., Frolich, O., Gottlob, G., Hassan, T., Herzog, M., Holzinger, W. and Krupl, B. (2006) 'The Lixto project: Exploring new frontiers of web data extraction', *Flexible and Efficient Information Handling*, 4042, 1-15.

Chang, C.-H., Kayed, M., Girgis, M. R. and Shaalan, K. F. (2006) 'A Survey of Web Information Extraction Systems', *IEEE Trans. on Knowl. and Data Eng.*, 18(10), 1411-1428.

Chinchor, N. (1992) 'MUC-4 evaluation metrics', in *Proceedings of the 4th conference on Message understanding*, McLean, Virginia, 1072067: Association for Computational Linguistics, 22-29.

Chinchor, N. (1998) 'MUC-7 named entity task definition', *Proceedings of the Seventh Message Understanding Conference (MUC-7); April 1998; Fairfax, Virginia*.

Chinchor, N. A. (2001) 'Overview of MUC-7/MET-2', [online], available: http://www-nlpir.nist.gov/related_projects/muc/proceedings/muc_7_proceedings/overview.html [accessed 2 January, 2012].

Chiticariu, L., Krishnamurthy, R., Li, Y., Reiss, F. and Vaithyanathan, S. (2010) 'Domain adaptation of rule-based annotators for named-entity recognition tasks', in *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, Cambridge, Massachusetts, 1870756: Association for Computational Linguistics, 1002-

1012.

Chun, H. W., Tsuruoka, Y., Kim, J. D., Shiba, R., Nagata, N., Hishiki, T. and Tsujii, J. (2006) 'Extraction of gene-disease relations from Medline using domain dictionaries and machine learning', *Pac Symp Biocomput*, 4 - 15.

Cohen, W. (2012) 'MinorThird', [online], available: http://sourceforge.net/ apps/ trac/ minorthird/wiki [accessed 3 May, 2010].

Collier, N., Nobata, C. and Tsujii, J. (2000) 'Extracting the names of genes and gene products with a Hidden Markov Model', *COLING'2000*, 201 - 207.

Collins, M. and Singer, Y. (1999) 'Unsupervised Models for Named Entity Classification', in Fung, P. and Zhou, J., eds., *Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, University of Maryland, College Park, MD, USA, June 21-22, 1999, Association for Computational Linguistics, 100-110.

Corbett, P. and Copestake, A. (2008) 'Cascaded Classifiers for Confidence-Based Chemical Named Entity Recognition', in Demner-Fushman, D., ed. *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing*, Columbus, Ohio, USA, 19 Jun, 2008, East Stroudsburg, PA, USA: Association for Computational Linguistics (ACL), 54-62.

Cowie, J. R. (1983) 'Automatic analysis of descriptive texts', in *Proceedings of the first conference on Applied natural language processing*, Santa Monica, California, 974218: Association for Computational Linguistics, 117-123.

Cowie, J. and Lehnert, W. (1996) 'Information Extraction', *Communications of the ACM*, 39(1), 80-91.

Cowie, J. and Wilks, Y. (2000) 'Information Extraction' in Dale, R., Moisl, H. and Somers, H. L., eds., *Handbook of Natural Language Processing*, New York : Marcel Dekker, 241-260.

Cunningham, H. (2005) 'Information Extraction, Automatic' in Brown, K., ed. *Encyclopedia of Language and Linguistics*, 2nd ed., Amsterdam, the Netherlands: Elsevier Science Publishers,, 665-677.

DaSilva, G. and Dwiggins, D. (1980) 'Towards a PROLOG Text Grammar.', *ACM SIGART Newsletter*, 20-25.

de Bruijn, B., Cherry, C., Kiritchenko, S., Martin, J. and Zhu, X. (2011) 'Machine-learned solutions for three stages of clinical information extraction: the state of the art at i2b2 2010', *Journal of the American Medical Informatics Association*.

References

Demner-Fushman, D., Chapman, W. W. and McDonald, C. J. (2009) 'What can natural language processing do for clinical decision support?', *Journal of Biomedical Informatics*, 42(5), 760-772.

DeRemer, F. and Kron, H. H. (1976) 'Programming-in-the-Large Versus Programming-in-the-Small', *Software Engineering, IEEE Transactions on*, SE-2(2), 80-86.

Dias, G., Alves, E. and Nunes, C. (2005) 'Topic Segmentation: How Much Can We Do By Counting Words And Sequences of Words', *Pliska Studia Mathematica Bulgarica Journal*, 17, 39-70.

Dietterich, T. G. (1998) 'Approximate statistical tests for comparing supervised classification learning algorithms', *Neural Computation*, 10(7), 1895-1923.

Dillon, A. (1991) 'Reader's models of text structures: the case of academic articles', *Int. J. Man-Mach. Stud.*, 35(6), 913-925.

Doddington, G., Mitchell, A., Przybocki, M., Ramshaw, L., Strassel, S. and Weischedel, R. (2004) 'The Automatic Content Extraction (ACE) Program--Tasks, Data, and Evaluation', in *The International Conference on Language Resources and Evaluation (LREC)*, Lisbon, Portugal, 24-30 May, 2004 837--840.

Ekbal, A. and Bandyopadhyay, S. (2008) 'A web-based Bengali news corpus for named entity recognition', *Language Resources and Evaluation*, 42(2), 173-182.

Ekbal, A. and Bandyopadhyay, S. (2009) 'Voted NER system using appropriate unlabeled data', in *Proceedings of the 2009 Named Entities Workshop: Shared Task on Transliteration*, Suntec, Singapore, 1699749: Association for Computational Linguistics, 202-210.

ELRA (2011) 'LREC Conferences ', [online], available: http://www.lrec-conf.org/ [accessed 12 October, 2011].

Elsner, M., Charniak, E. and Johnson, M. (2009) 'Structured generative models for unsupervised named-entity clustering', in *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, Boulder, Colorado, 1620778: Association for Computational Linguistics, 164-172.

Etzioni, O., Cafarella, M., Downey, D., Popescu, A.-M., Shaked, T., Soderland, S., Weld, D. S. and Yates, A. (2004) 'Methods for domain-independent information extraction from the web: an experimental comparison', in *Proceedings of the 19th national conference on Artifical intelligence*, San Jose, California, 1597213: AAAI Press, 391-398.

Etzioni, O., Cafarella, M., Downey, D., Popescu, A. M., Shaked, T., Soderland, S., Weld, D. S. and Yates, A. (2005) 'Unsupervised named-entity extraction from the Web: An experimental study', *Artificial Intelligence*, 165(1), 91-134.

References

Feldman, R. and Sanger, J. (2006) *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*, Cambridge: Cambridge University Press.

Ferrara, E., Fiumara, G. and Baumgartner, R. (2010) 'Web Data Extraction, Applications and Techniques: A Survey', *ACM Transactions on Computational Logic*, V(June 2010), 1-20.

Finin, T., Murnane, W., Karandikar, A., Keller, N., Martineau, J. and Dredze, M. (2010) 'Annotating named entities in Twitter data with crowdsourcing', in *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, Los Angeles, California, 1866709: Association for Computational Linguistics, 80-88.

Finkel, J. R., Grenager, T. and Manning, C. (2005) 'Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling', in Night, K., ed. *ACL '05 Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, Ann Arbor, Michigan, USA, East Stroudsburg, PA, USA: Association for Computational Linguistics, 363-370.

Fleischman, M. and Hovy, E. (2002) Fine Grained Classification of Named Entities', in *Proceedings of the 19th international Conference on Computational Linguistics Taipei*, Taiwan, 24 Aug - 1 Sept, 2002, Morristown, NJ: Morristown, NJ: Association for Computational Linguistics, 1-7.

Floridi, L. (2009) 'The Information Society and Its Philosophy: Introduction to the Special Issue on "The Philosophy of Information, Its Nature, and Future Developments"', *InformationSociety*, 25(3), 153-158.

Freitag, D. (2004) 'Trained Named Entity Recognition Using Distributional Clusters', in Lin, D. and (chairs), D. W., eds., *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP 2004)*, Barcelona, Spain, 25-26 Jul, 2004i, East Stroudsburg, PA, USA: Association for Computational Linguistics, 262-269.

Gabbay, I. and Sutcliffe, R. F. E. (2004) 'A Qualitative Comparison of Scientific and Journalistic Texts from the Perspective of Extracting Definitions', in Aliod, D. M. and Vicedo, J. L., eds., *ACL 2004: Question Answering in Restricted Domains*, Barcelona, Spain, July 2004, Association for Computational Linguistics, 16-22.

Gaizauskas, R., Demetriou, G., Artymiuk, P. J. and Willett, P. (2003) 'Protein Structures and Information Extraction from Biological Texts: The PASTA System', *Bioinformatics*, 19(1), 135-143.

Gaizauskas, R. and Wilks, Y. (1998) 'Information Extraction: Beyond Document Retrieval', *Journal of Documentation* 54(1), 70-105.

References

Ghani, R., Probst, K., Liu, Y., Krema, M. and Fano, A. (2006) 'Text mining for product attribute extraction', *SIGKDD Explor. Newsl.*, 8(1), 41-48.

Grishman, R. and Sundheim, B. (1995) 'Design of the MUC-6 evaluation', in *Proceedings of the 6th conference on Message understanding*, Columbia, Maryland, 1072401: Association for Computational Linguistics, 1-11.

Grishman, R. and Sundheim, B. (1996) 'Message Understanding Conference-6: a brief history', in Tsujii, J., ed. *Proceedings of the 16th conference on Computational linguistics - Volume 1*, Copenhagen, Denmark, 5-9 Aug, 1996, Morristown, NJ: Association for Computational Linguistics, 466-471.

Hachey, B. and Grover, C. (2005) 'Sequence Modelling for Sentence Classification in a Legal Summarisation System', in Liebrock, L. M., ed. *Proceedings of the 2005 ACM Symposium on Applied Computing*, Santa Fe, New Mexico, USA, 13 -17 Mar, 2005, New York, NY, USA: ACM, 292-296

Hidalgo, J. M. G., Garcia, F. C. and Sanz, E. P. (2005) 'Named entity recognition for Web content filtering', *Natural Language Processing and Information Systems, Proceedings*, 3513, 286-297.

Hina, S., Atwell, E. and Johnson, O. (2011) 'Semantic Tagging of Medical Narratives with Top Level Concepts from SNOMED CT Healthcare Data Standard', *International Journal of Intelligent Computing Research (IJICR)*, Volume 2,(1/2/3/4), 204-210.

Hobbs, J. R. and Riloff, E. (2010) 'Information Extraction' in Indurkhya, N. and Damerau, F. J., eds., *Handbook of Natural Language Processing*, Boca Raton, FL: Chapman and Hall/CRC, 511-532.

Hu, M. and Liu, B. (2004) 'Mining and summarizing customer reviews', in *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, Seattle, WA, USA, 1014073: ACM, 168-177.

Huang, D. G., Li, L. S. and Zhou, R. P. (2009) 'Two-phase biomedical named entity recognition using CRFs', *Computational Biology and Chemistry*, 33(4), 334-338.

Humphreys, K., Demetriou, G. and Gaizauskas, R. (2000) 'Two Applications of Information Extraction to Biological Science Journal Articles: Enzyme Interactions and Protein Structures', in Altman, R. B., Dunker, A. K., Hunter, L., Klein, T. E. and chairs, eds., *Proceedings of the 7th Pacific Symposium on Biocomputing (PSB 2002)*, Lihue, Hawaii, USA, 3-7 Jan, 2002., Hackensack, NJ, USA: World Scientific Publishing, 505-516.

Hutchins, M. (2006) 'Death at the Zoo: The Media, Science, and Reality', *Zoo Biology*, 25(2), 101-115.

Ireson, N., Ciravegna, F., Califf, M. E., Freitag, D., Kushmerick, N. and Lavelli, A. (2005) 'Evaluating Machine Learning for Information Extraction', in (chair), S. D., ed. *Proceedings of the 22nd International Conference on Machine Learnin*g, Bonn, Germany, 7-11 Aug, 2005, New York, NY: ACM, 345-352.

Jacobs, P. S. and Rau, L. F. (1990) 'SCISOR: extracting information from on-line news', *Commun. ACM,* 33(11), 88-97.

Kan, M., Klavans, J. and McKeown, K. (1998) 'Linear Segmentation and Segment Significance', in Charniak, E., ed. *Proceedings of the 6th Workshop on Very Large Corpora (WVLC-98)*, Montreal, Canada, 15- 16 Aug, 1998, East Sroudsburg, PA, USA: Association for Computational Lingustics, 197-205.

Kessler, R., Torres-Moreno, J. M. and El-Bèze, M. (2007) 'E-Gen: automatic job offer processing system for human resources', in *Proceedings of the artificial intelligence 6th Mexican international conference on Advances in artificial intelligence*, Aguascalientes, Mexico, 1776068: Springer-Verlag, 985-995.

Kim, J.-D., Ohta, T., Pyysalo, S., Kano, Y. and Tsujii, J. i. (2009) 'Overview of BioNLP'09 shared task on event extraction', in *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing: Shared Task*, Boulder, Colorado, 1572342: Association for Computational Linguistics, 1-9.

Kim, J.-D., Ohta, T., Tateisi, Y. and Tsujii, J. (2003) 'GENIA corpus—a semantically annotated corpus for bio-textmining', *Bioinformatics,* 19(suppl 1), i180-i182.

Kim, S. and Yoon, J. (2007) 'Experimental study on a two phase method for biomedical named entity recognition', *Ieice Transactions on Information and Systems*, E90d(7), 1103-1110.

Klinger, R. and Tomanek, K. (2007) *Classical Probabilistic Models and Conditional Random Fields*, Dortmund, Germany: Faculty of Computer Science, Dortmund Univiversity of Tech nology.

Kolluru, B., Hawizy, L., Murray-Rust, P., Tsujii, J. and Ananiadou, S. (2011) 'Using Workflows to Explore and Optimise Named Entity Recognition for Chemistry', *Plos One,* 6(5).

Koning, D., Sarkar, I. N. and Moritz, T. (2005) 'TaxonGrab: Extracting Taxonomic Names From Text', *Biodiversity Informatics,* 2, 79-82.

Kou, Z., Cohen, W. W. and Murphy, R. F. (2005) 'High-recall protein entity recognition using a dictionary', *Bioinformatics*, 21(suppl_1), i266-273.

Kozareva, Z., Vazquez, S. and Montoyo, A. (2008) Domain Information for Fine-Grained Person Name Categorization', in Gelbukh, A., ed. *Computational Linguistics and*

Intelligent Text Processing 9th International Conference, CICLing 2008, Haifa, Israel, 17-23 Feb, 2008, Berlin Heidelberg: Springer-Verlag, 311-321.

Krallinger, M., Valencia, A. and Hirschman, L. (2008) 'Linking genes to literature: text mining, information extraction, and retrieval applications for biology', *Genome Biology*, 9.

Kripke, S. A. (1980) *Naming and necessity*, Cambridge, Mass.: Harvard University Press.

Krishnan, V. and Manning, C. D. (2006) An Effective Two-Stage Model for Exploiting Non-Local Dependencies in Named Entity Recognition', in Calzolari, N., ed. *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the ACL*, Sydney, Australia, 2006, 17-18 Jul, 2006, East Stroudsburg, PA, USA: Association for Computational Linguistics, 1121-1128.

Krovetz, R., Deane, P. and Madnani, N. (2011) 'The web is not a person, Berners-Lee is not an organization, and African-Americans are not locations: an analysis of the performance of named-entity recognition', in *Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World*, Portland, Oregon, 2021135: Association for Computational Linguistics, 57-64.

Kudo, T. (2007) 'CRF++: Yet Another CRF toolkit', [online], available: http://crfpp.sourceforge.net/ [accessed 5 Jan, 2008].

Lafferty, J., McCallum, A. and Pereira, F. (2001) 'Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data', *Proc. 18th International Conf. on Machine Learning*, 282 - 289.

Lapata, M. (2003) 'Probabilistic Text Structuring: Experiments with Sentence Ordering', in Dignum, F., ed. *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*, Sapporo, Japan, July 7-12, 2003, East Stroudsburg, PA, USA: Association for Computational Linguistics, 545-552.

Lawson, N., Eustice, K., Perkowitz, M. and Yetisgen-Yildiz, M. (2010) 'Annotating large email datasets for named entity recognition with Mechanical Turk', in *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, Los Angeles, California, 1866708: Association for Computational Linguistics, 71-79.

Leaman, R. and Gonzalez, G. (2008) 'BANNER: An Executable Survey of Advances in Biomedical Named Entity Recognition', in Altman, R. B., Dunker, A. K., Hunter, L., Murray, T. and Klein, T. E., eds., *The Thirteenth Pacific Symposium on Biocomputing (PSB)*, Hawaii, USA, 4-8 Jan, 2008, Hackensack, NJ, USA: World Scientific Publishing, 652-663.

Lee, C., Hwang, Y.-G. and Jang, M.-G. (2007) 'Fine-Grained Named Entity Recognition and Relation Extraction for Question Answering', in Kraaij, W. and Vries, A. P. d., eds.,

Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Amsterdam, The Netherlands, 23-27 Jul, 2007, New York, NY, USA: Association for Computing Machinery 799-800.

Lee, C., Hwang, Y. G., Oh, H. J., Lim, S., Heo, J., Lee, C. H., Kim, H. J., Wang, J. H. and Jang, M. G. (2006) 'Fine-grained Named Entity Recognition using Conditional Random Fields for Question Answering', *Information Retrieval Technololgy, Proceedings*, 4182, 581-587.

Lee, K. J., Hwang, Y. S., Kim, S. and Rim, H. C. (2004) 'Biomedical named entity recognition using two-phase model based on SVMs', *Journal of Biomedical Informatics*, 37(6), 436-447.

Lee, S. and Lee, G. G. (2007) 'Exploring phrasal context and error correction heuristics in bootstrapping for geographic named entity annotation', *Information Systems*, 32(4), 575-592.

Li, C., Weng, J., He, Q., Yao, Y., Datta, A., Sun, A. and Lee, B.-S. (2012) 'TwiNER: Named Entity Recognition in Targeted Twitter Stream', in *The 35th Annual International ACM SIGIR Conference (SIGIR'12)*, Portland, Oregon, USA, August 12–16, 2012.

Lin, C.-Y. and Hovy, E. (1997) 'Identifying Topics by Position', in *Proceedings of the Fifth Conference on Applied Natural Language Processing*, Washington, DC, 31 Mar - 3 Apr, 1997, San Francisco, CA: Morgan Kaufmann Publishers Inc., 283-290

Ling, X. and Weld, D. S. (2012) 'Fine-Grained Entity Recognition Xiao Ling and Daniel S. Weld', available: http://www.cs.washington.edu/ai/pubs/ling-aaai12.pdf [accessed May 25, 2012].

Liu, J., Huang, M. and Zhu, X. (2010) 'Recognizing biomedical named entities using skip-chain conditional random fields', in *Proceedings of the 2010 Workshop on Biomedical Natural Language Processing*, Uppsala, Sweden, 1869963: Association for Computational Linguistics, 10-18.

Lok, C. (2010) 'Literature mining: Speed reading', *Nature*, 463(7280), 416-418.

Loth, R., Battistelli, D., Chaumartin, F.-R., Mazancourt, H. d., Minel, J.-L. and Vinckx, A. (2010) 'Linguistic information extraction for job ads (SIRE project)', in *Adaptivity, Personalization and Fusion of Heterogeneous Information*, Paris, France, 1937114: LE CENTRE DE HAUTES ETUDES INTERNATIONALES D'INFORMATIQUE DOCUMENTAIRE, 222-224.

Lytinen, S. L. and Gershman, A. (1986) 'ATRANS Automatic Processing of Money Transfer Messages', in *Proceedings of the Fifth National Conference on Artificial Intelligence*, Philadelphia, August 11–15, 1986, 1089-1095.

Mallett, D., Elding, J. and Nascimento, M. A. (2004) 'Information-Content Based Sentence Extraction for Text Summarization', in *Proceedings of the International Conference on Information Technology: Coding and Computing (ITCC'04) Volume 2 - Volume 2*, Las Vegas, Nevada, USA, 5-7 Apr, 2004,, Washington, DC, USA: IEEE Computer Society 214-217.

Mansouri, A., Affendey, L. S. and Mamat, A. (2008) 'Named Entity Recognition Approaches', *International Journal of Computer Science and Network Security,*, 8(2), 339-344.

Marsh, E. and Perzanowski, D. (1998) 'MUC-7 EVALUATION OF IE TECHNOLOGY: Overview of Results', available: http://www.itl.nist.gov/iaui/894.02/related_projects/muc/proceedings/muc_7_proceedings/marsh_slides.pdf [accessed 3 May, 2010].

Matsuo, Y., Mori, J., Hamasaki, M., Nishimura, T., Takeda, H., Hasida, K. and Ishizuka, M. (2007) 'POLYPHONET: An advanced social network extraction system from the Web', *Web Semantics: Science, Services and Agents on the World Wide Web*, 5(4), 262-278.

McCallum, A., Freitag, D. and Pereira, F. C. N. (2000) 'Maximum Entropy Markov Models for Information Extraction and Segmentation', in *Proceedings of the Seventeenth International Conference on Machine Learning*, Stanford , California, June 29-July 2, Morgan Kaufmann Publishers Inc.

McDonald, R. and Pereira, F. (2005) 'Identifying Gene and Protein Mentions in Text Using Conditional Random Fields', *BMC Bioinformatics*, 6 (Suppl 1)(Suppl 1), S6, available: [accessed 29 Sept, 2008].

McNamara, P., Capellini, I., Harris, E., Nunn, C. L., Barton, R. A. and Preston, B. (2008) 'The Phylogeny of Sleep Database: A New Resource for Sleep Scientists', *The Open Sleep Journal*, (1), 11-14.

Megaputer (2012) 'PolyAnalyst, Data Mining. Text Mining. All in a single, intuitive package.', [online], available: http://www.megaputer.com/polyanalyst.php [accessed 2 January, 2012].

Mendes, P. N., Passant, A. and Kapanipathi, P. (2010) 'Twarql: tapping into the wisdom of the crowd', in *Proceedings of the 6th International Conference on Semantic Systems*, Graz, Austria, 1839762: ACM, 1-3.

Merchant, R., Okurowski, M. E. and Chinchor, N. (1996) 'The multilingual entity task (MET) overview', in *Proceedings of a workshop on held at Vienna, Virginia: May 6-8, 1996*, Vienna, Virginia, 1119075: Association for Computational Linguistics, 445-447.

Mikheev, A., Moens, M. and Grover, C. (1999) 'Named Entity recognition without gazetteers', in *Proceedings of the ninth conference on European chapter of the Association for Computational Linguistics*, Bergen, Norway, 977037: Association for Computational Linguistics, 1-8.

Minkov, E., Wang, R. C. and Cohen, W. W. (2005) 'Extracting Personal Names from Email: Applying Named Entity Recognition to Informal Text', in Mooney, R. J. and chair, eds., *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing (HLT/EMNLP)*, Vancouver, British Columbia, Canada, 6-8 Oct, 2005, East Stroudsburg, PA, USA: Association for Computational Linguistics.

Mizuta, Y., Korhonen, A., Mullen, T. and Collier, N. (2006) 'Zone Analysis in Biology Articles as a Basis for Information Extraction', *International Journal of Medical Informatics*, 75(6)(6), 468-487.

Moens, M.-F. (2006) *Information extraction : algorithms and prospects in a retrieval context, Information retrieval series 21*, Dordrecht: Springer.

Moens, M.-F. (2009) 'Information extraction from blogs' in Jansen, B. J., Spink, A. and Taksa, I., eds., *Handbook of Research on Web Log Analysis*, Hershey PA IGI Global, 469-487.

Moens, M.-F., Uyttendaele, C. and Dumortier, J. (1999) 'Information Extraction from Legal Texts: The Potential of Discourse Analysis', *International Journal of Human-Computer Studies*, 51(6)(6), 1155-1171.

Mooney, R. J. and Nahm, U. Y. (2003) 'Text Mining with Information Extraction', in Daelmans, W., du Plessis, T., Snyman, C. and Teck, L., eds., Multilingualism and Electronic Language Management: *Proceedings of the 4th International MIDP Colloquium*, Bloemfontein, South Africa, September 22-23, Van Schaik Pub., 141-160.

Murphy, T., McIntosh, T. and Curran, J. R. (2006) Named Entity Recognition for Astronomy Literature', in *Proceedings of the 2006 Australasian Language Technology Workshop (ALTW2006)*, Sydney, Australia, November 30 - December 1, 2006, Australasian Language Technology Association, 59–66.

Nadeau, D. and Sekine, S. (2007) 'A Survey of Named Entity Recognition and Classification', *Linguisticae Investigationes*, 30(1), 3-26.

Nadeau, D., Turney, P. and Matwin, S. (2006) 'Unsupervised Named-Entity Recognition: Generating Gazetteers and Resolving Ambiguity (AI 2006)', in Lamontagne, L. and Marchand, M., eds., *19th Canadian Conference on Artificial Intelligence*, Québec City, Québec, Canada, 7 Jun, 2006., Heidelberg-Berlin, Germany: Springer.

Narayanaswamy, M., Ravikumar, K. E. and Vijay-Shanker, K. (2003) 'A Biological Named Entity Recognizer', in Altman, R. B., Dunker, A. K., Hunter, L. and Klein, T. E., eds., *Proceedings of the 8th Pacific Symposium on Biocomputing (PSB 2003*), Lihue, Hawaii, USA, January 3-7, 2003,, Hackensack, NJ, USA: World Scientific Publishing, 427-438.

References

NexisLexis (2008) 'Nexis', [online], available: http://web.lexis-nexis.com/professional/ [accessed April 24, 2008].

Nguyen, H. and Cao, T. (2008) 'Named Entity Disambiguation: A Hybrid Statistical and Rule-Based Incremental Approach' in Domingue, J. and Anutariya, C., eds., *The Semantic Web*, Springer Berlin Heidelberg, 420-433.

Nicol, S. C., Andersen, N. A., Nathan H. Phillips and Berger, R. J. (2000) 'The echidna manifests typical characteristics of rapid eye movement sleep', *Neuroscience Letters* 283 49-52.

NIST (2011) 'The Text Analysis Conference (TAC) ', [online], available: http://www.nist.gov/tac/ [accessed 19 July, 2012].

Nothman, J., Curran, J. R. and Murphy, T. (2008) 'Transforming Wikipedia into Named Entity Training Data', in *Proceedings of the Australasian Language Technology Workshop*, Hobart, Tasmania December 8-10, 2008, 124–132.

Ono, T., Hishigaki, H., Tanigami, A. and Takagi, T. (2001) 'Automated extraction of information on protein-protein interactions from the biological literature', *Bioinformatics*, 17(2), 155 - 161.

Patterson, A. C. (1971) 'Requirements for a generalized data base management system', in *Proceedings of the November 16-18, 1971, fall joint computer conference*, Las Vegas, Nevada, 1479157: ACM, 515-522.

Peng, F. and McCallum, A. (2004) 'Accurate Information Extraction from Research Papers Using Conditional Random Fields', in Hirschberg, J. and chair, eds., *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, Boston, Massachusetts, USA, 2-7 May, 2004, New York, NY, USA: Association for Computational Linguistics, 197-207.

Perez-Iratxeta, C., Bork, P. and Andrade, M. A. (2001) 'XplorMed: a tool for exploring MEDLINE abstracts', *Trends Biochem Sci*, 26, 573 - 575.

Poibeau, T. and Kosseim, L. (2001) 'Proper Name Extraction from Non-Journalistic Texts', *Language and Computers*, 144-157.

Ponomareva, N., Rosso, P., Pla, F. and Molina, A. (2007) 'Conditional Random Fields vs. Hidden Markov Models in a biomedical Named Entity Recognition task', in Mitkov, R., ed., *Recent Advances in Natural Language Processing, RANLP-2007*, Borovets, Bulgaria, 27-29 Sept, 2007, 479-483.

Popescu, A.-M. and Etzioni, O. (2005) 'Extracting product features and opinions from reviews',

in *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, Vancouver, British Columbia, Canada, 1220618: Association for Computational Linguistics, 339-346.

Rabiner, L. R. (1990) 'A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition' in Waibel, A. and Lee, K.-F., eds., *Readings in speech recognition*, San Francisco, USA: Morgan Kaufmann Publishers Inc., 267-296.

Rau, L. F. (1991) 'Extracting company names from text', in *Proceedings of the Seventh IEEE Conference on Artificial Intelligence Applications*, Miami Florida, 24-28 Feb 1991, 29-32.

Rijsbergen, C. J. V. (1979) *Information Retrieval*, Butterworth-Heinemann.

Ritter, A., Sam Clark, Mausam and Etzioni, O. (2011) 'Named Entity Recognition in Tweets: An Experimental Study', in *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, Edinburgh, Scotland, UK, July 27–31, 2011, Association for Computational Linguistics, 1524–1534.

Rocktäschel, T., Weidlich, M. and Leser, U. (2012) 'ChemSpot: A Hybrid System for Chemical Named Entity Recognition', *Bioinformatics*.

Sang, E. F. T. K. and Meulder, F. D. (2003) 'Introduction to the CoNLL-2003 shared task: language-independent named entity recognition', in *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003 - Volume 4*, Edmonton, Canada, 1119195: Association for Computational Linguistics, 142-147.

Santos, D. and Cardoso, N. (2006) 'A golden resource for named entity recognition in Portuguese', *Computational Processing of the Portuguese Language, Proceedings*, 3960, 69-79.

SAP (2011) 'SAP BUSINESSOBJECTS DATA SERVICES', [online], available: http://sap.com/solutions/sapbusinessobjects [accessed 2 January, 2012].

Sasaki, Y., Tsuruoka, Y., McNaught, J. and Ananiadou, S. (2008) 'How to Make the Most of NE Dictionaries in Statistical NER', in Demner-Fushman, D., ed., *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing*, Columbus, Ohio, USA, 19 Jun, 2008, East Stroudsburg, PA, USA: Association for Computational Linguistics (ACL), 63-70.

Sautter, G. and Böhm, K. (2006) 'The Difficulties of Taxonomic Name Extraction and a Solution', in Verspoor, K., Cohen, K. B., Goertzel, B. and Mani, I., eds., *Proceedings of the BioNLP Workshop on Linking Natural Lanaguge Processing and Biology at HLT-NAACL 06*, New York City, 8 Jun, 2006, East Stroudsburg, PA, USA: Association for Computational Linguistics, 126-133.

Schuemie, M. J., Weeber, M., Schijvenaars, B. J. A., van Mulligen, E. M., van der Eijk, C. C., Jelier, R., Mons, B. and Kors, J. A. (2004) 'Distribution of Information in Biomedical Abstracts and Full-Text Publications', *Bioinformatics*, 20(16)(16), 2597-2604.

Sekine, S. and Eriguchi, Y. (2000) 'Japanese named entity extraction evaluation: analysis of results', in *Proceedings of the 18th conference on Computational linguistics - Volume 2*, Saarbr\&\#252;cken, Germany, 992814: Association for Computational Linguistics, 1106-1110.

Sekine, S., Sudo, K. and Nobata, C. (2002) 'Extended Named Entity Hierarchy', in *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC-2002)*, Las Palmas de Gran Canaria, Spain, 29-31 May, 2002, 1818-1824.

Settles, B. (2004) 'Biomedical Named Entity Recognition Using Conditional Random Fields', *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications (NLPBA)*.

Shah, P., Perez-Iratxeta, C., Bork, P. and Andrade, M. (2003) 'Information Extraction from Full Text Scientific Articles: Where are the Keywords?', *BMC Bioinformatics*, 4(1)(1), 20, available: http://www.biomedcentral.com/1471-2105/4/20 [accessed 20 Aug, 2008].

Sharaf, A.-B. M. and Atwell., E. (2009) *The Qur'an Annotation for Text Mining* [online]: available: http://www.comp.leeds.ac.uk/scsams/transfer/TransferReport-Sharaf.pdf [accessed 24 Feb, 2013].

Shing-Kit, C. and Wai, L. (2007) 'Efficient Methods for Biomedical Named Entity Recognition', in BIBE 2007: *Proceedings of the 7th IEEE International Conference on Bioinformatics and Bioengineering*, Boston, Massachusetts, USA, 14-17 Oct, 2007, New York, NY, USA: IEEE, 729-735.

Silva, J. F. d., Kozareva, Z., Noncheva, V. and Lopes, G. P. (2004) 'Extracting Named Entities. A Statistical Approach', in Bel, B. and Merlien, I., eds., *Proceedings of the XIème Conférence sur le Traitement des Langues Naturelles (TALN)*, Fez, Morocco, 19 - 22 Apr, 2004, Paris, France: ATALA - Association pour le Traitement Automatique des Langues 347-351.

Sinclair, J. M. (1993) 'Written Discourse Structure' in Sinclair, J. M., Hoey, M. and Fox., G., eds., *Techniques of Description: Spoken and Written Discourse : a Festschrift for Malcolm Coulthard*, London: London, UK: Routledge, 6-31.

Singer, J. B. (2008) 'Five Ws and an H: Digital Challenges in Newspaper Newsrooms and Boardrooms', *International Journal on Media Management*, 10(3), 122-129.

Skusa, A., Ruegg, A. and Kohler, J. (2005) 'Extraction of Biological Interaction Networks from Scientific Literature', *Briefings in Bioinformatics*, 6(3)(3), 263-276.

Smith, L., Tanabe, L., Ando, R., Kuo, C.-J., Chung, I. F., Hsu, C.-N., Lin, Y.-S., Klinger, R., Friedrich, C., Ganchev, K., Torii, M., Liu, H., Haddow, B., Struble, C., Povinelli, R., Vlachos, A., Baumgartner, W., Hunter, L., Carpenter, B., Tsai, R., Dai, H.-J., Liu, F., Chen, Y., Sun, C., Katrenko, S., Adriaans, P., Blaschke, C., Torres, R., Neves, M., Nakov, P., Divoli, A., Mana-Lopez, M., Mata, J. and Wilbur, W. J. (2008) 'Overview of BioCreative II gene mention recognition', *The BioCreative II - Critical Assessment for Information Extraction in Biology Challenge*, 9(Suppl 2)available: http://genomebiology.com/2008/9/S2/S2 [accessed 28 Sept, 2008].

Sobhana, N. V., Mitra, P. and S. K, G. (2010) 'Conditional Random Field Based Named Entity Recognition in Geological text', *Foundation of Computer Science*, 1(3), 119–125.

Son Bao, P., Giang Binh, T., Dang Duc, P., Kien Chi, P. and Kien Trung, N. (2009) 'An Information Extraction Approach to English-Vietnamese Weather Bulletins Machine Translation', in *First Asian Conference on Intelligent Information and Database Systems, 2009 (ACIIDS 2009)*, Dong hoi, Quang binh, Vietnam, 1-3 April 2009, IEEE Computer Society, 161-166.

Song, Y., Kim, E., Lee, G. G. and Yi, B. K. (2005) 'POSBIOTM-NER: a trainable biomedical named-entity recognition system', *Bioinformatics*, 21(11), 2794-2796.

Soysal, E., Cicekli, I. and Baykal, N. (2010) 'Design and evaluation of an ontology based information extraction system for radiological reports', *Computers in Biology and Medicine*, 40(11-12), 900-911.

SRA (2012) 'Text and Entity Analytics for Big Data', [online], available: http://www.sra.com/netowl/ [accessed 2 January, 2012].

Srihari, R. and Li, W. (2000) 'A question answering system supported by information extraction', in *Proceedings of the sixth conference on Applied natural language processing*, Seattle, Washington, 974170: Association for Computational Linguistics, 166-172.

Srihari, R., Niu, C. and Li, W. (2000) 'A hybrid approach for named entity and sub-type tagging', in *Proceedings of the sixth conference on Applied natural language processing*, Seattle, Washington, 974181: Association for Computational Linguistics, 247-254.

Struble, C. A., Povinelli, R. J., Johnson, M. T., Berchanskiy, D., Tao, J. and Trawicki, M. (2007) 'Combined conditional random fields and n-gram language models for gene mention recognition', *Proceedings of the Second BioCreative Challenge Evaluation Workshop; 23 to 25 April 2007; Madrid, Spain*, 81 - 83.

Suakkaphong, N., Zhang, Z. and Chen, H. C. (2011) 'Disease Named Entity Recognition Using Semisupervised Learning and Conditional Random Fields', *Journal of the American Society for Information Science and Technology*, 62(4), 727-737.

Sundheim, B. M. (1995) 'Overview of results of the MUC-6 evaluation', in *Proceedings of the 6th conference on Message understanding*, Columbia, Maryland, 1072402: Association for Computational Linguistics, 13-31.

Sung, T. Y., Tsai, R. T. H., Wu, S. H., Chou, W. C., Lin, Y. C., He, D., Hsiang, J. and Hsu, W. L. (2006) 'Various criteria in the evaluation of biomedical named entity recognition', *Bmc Bioinformatics*, 7, 1-8.

Sutcliffe, R. F. E. (2002) 'Question Answering using the DLT system at TREC 2002', in Voorhees, E. M. and Buckland, L. P., eds., *Proceedings of the Eleventh Text REtrieval Conference (TREC 2002)*, Guthersburg, Maryland, November 19-22, 2002, NIST, 677-685.

Sutcliffe, R. F. E., Gabbay, I., Mulcahy, M., & White, K. (2003). Question Answering using the DLT System at TREC 2003. In E. M. Voorhees and L. P. Buckland (Eds) *Notebook of the Twelfth Text REtrieval Conference (TREC 2003)*, Gaithersburg, Maryland, November 18-21, 2003, 638-644.

Sutcliffe, R. F. E., Mulcahy, M., White, K., Gabbay, I. and O'Gorman, A. (2005) 'Question Answering Using the DLT System at TREC 2005', *Proceedings of the Fourteenth Text REtrieval Conference (TREC 2005)*, available: http://trec.nist.gov/pubs/trec14/papers/ulimerick.qa.pdf [accessed May 15, 2011].

Sutcliffe, R. F. E., White, K., & Kruschwitz (2010) 'Named Entity Recognition in an Intranet Query Log', in Sutcliffe, R. F. E., Kruschwitz, U. and Mandl, T., eds., *Proceedings of the Workshop on Web Logs and Question Answering (WLQA2010)* Valetta, Malta, May 22, 2010, ELRA, 43-49.

Sutton, C. and McCallum, A. (2006) 'An Introduction to Conditional Random Fields for Relational Learning' in Getoor, L. and Taskar, B., eds., *Introduction to Statistical Relational Learning*, Cambridge, MA, USA: MIT Press, 93-128.

Teufel, S., Siddharthan, A. and Batchelor, C. (2009) 'Towards discipline-independent argumentative zoning: evidence from chemistry and computational linguistics', in *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3 - Volume 3*, Singapore, 1699696: Association for Computational Linguistics, 1493-1502.

Thomson-Reuters (2012) 'ClearForest', [online], available: http://www.clearforest.com/ [accessed.

Tjong Kim Sang, E. F. and De Meulder, F. (2003) 'Introduction to the CoNLL-2003 shared task: language-independent named entity recognition', in *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003 - Volume 4*, Edmonton, Canada, 1119195: Association for Computational Linguistics, 142-147.

Tobler, I. and Schwierin, B. (1996) 'Behavioural sleep in the giraffe in a zoological garden', *Journal of Sleep Research*, 5(1), 21-32.

Tsai, T.-h., Chou, W.-C., Wu, S.-H., Sung, T.-Y., Hsiang, J. and Hsu, W.-L. (2006) 'Integrating linguistic knowledge into a conditional random fieldframework to identify biomedical named entities', *Expert Systems with Applications*, 30(1)(1), 117-128.

Tsuruoka, Y., Tsujii, J. i. and Ananiadou, S. (2008) Accelerating the Annotation of Sparse Named Entities by Dynamic Sentence Selection', in Demner-Fushman, D., ed. *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing*, Columbus, Ohio, USA, 19 Jun, 2008, East Stroudsburg, PA, USA: Association for Computational Linguistics (ACL), 30-37.

Vilain, M., Su, J. and Lubar, S. (2007) 'Entity Extraction is a Boring Solved Problem---Or is it?', in Sidner, C., Schultz, T., Stone, M. and Zhai, C., eds., *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics;* , April 22-27, Rochester, New York: Association for Computational Linguistics, 181-184.

Vlachos, A. (2007) 'Tackling the BioCreative2 gene mention task with conditional random fields and syntactic parsing', *Proceedings of the Second BioCreative Challenge Evaluation Workshop; 23 to 25 April 2007; Madrid, Spain*, 85 - 87.

Vlachos, A. and Gasperin, C. (2006) 'Bootstrapping and Evaluating Named Entity Recognition in the. Biomedical Domain', in *Proceedings of BioNLP in HLT-NAACL*, New York City, USA,, June 8, 138–145.

Wang, X. and Matthews, M. (2008) 'Species Disambiguation for Biomedical Term Identification', in Demner-Fushman, D., ed. *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing*, Columbus, Ohio, USA, 19 Jun, 2008, East Stroudsburg, PA, USA: Association for Computational Linguistics (ACL), 71-79..

White, K. and Sutcliffe, R. F. E. (2011) 'Butcher, baker, or candlestick maker? Predicting occupations using predicate–argument relations', *Journal of the American Society for Information Science and Technology*, 62(7), 1325-1344.

White, M., Korelsky, T., Cardie, C., Ng, V., Pierce, D. and Wagstaff, K. (2001) 'Multidocument summarization via information extraction', in *Proceedings of the first international conference on Human language technology research*, San Diego, 1072206: Association for Computational Linguistics, 1-7.

Whitelaw, C., Kehlenbeck, A., Petrovic, N. and Ungar, L. (2008) 'Web-scale named entity recognition', in *Proceedings of the 17th ACM conference on Information and knowledge management*, Napa Valley, California, USA, 1458102: ACM, 123-132.

Whitelaw, C. and Patrick, J. (2003) 'Evaluating corpora for named entity recognition using character-level features', *Ai 2003: Advances in Artificial Intelligence*, 2903, 910-921.

Willis, A., King, D., Morse, D., Dil, A., Lyal, Chris and Roberts, D. (2010) 'From XML to XML: The why and how of making the biodiversity literature accessible to researchers', in *Language Resources and Evaluation Conference (LREC)*, Malta, May 19-20, 2010.

Wolinski, F., Vichot, F. and Dillet, B. (1995) 'Automatic processing of proper names in texts', in *Proceedings of the seventh conference on European chapter of the Association for Computational Linguistics*, Dublin, Ireland, 976978: Morgan Kaufmann Publishers Inc., 23-30.

Wong, T.-L., Lam, W. and Chen, B. (2009) 'Mining employment market via text block detection and adaptive cross-domain information extraction', in *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, Boston, MA, USA, 1571991: ACM, 283-290.

Wong, T.-L., Lam, W. and Wong, T.-S. (2008) 'An unsupervised framework for extracting and normalizing product attributes from multiple web sites', in *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, Singapore, Singapore, 1390343: ACM, 35-42.

Xu, H., Stenner, S. P., Doan, S., Johnson, K. B., Waitman, L. R. and Denny, J. C. (2010) 'MedEx: a medication information extraction system for clinical narratives', *Journal of the American Medical Informatics Association*, 17(1), 19-24.

Zarri, G. P. (1983) 'Automatic representation of the semantic relationships corresponding to a French surface expression', in *Proceedings of the first conference on Applied natural language processing*, Santa Monica, California, 974222: Association for Computational Linguistics, 143-147.

Zhang, J., Shen, D., Zhou, G. D., Su, J. and Tan, C. L. (2004) 'Enhancing HMM-based biomedical named entity recognition by studying special phenomena', *Journal of Biomedical Informatics*, 37(6), 411-422.

Zhang, Z., Iria, J. and Ciravegna, F. (2010) 'Improving Domain-specific Entity Recognition with Automatic Term Recognition and Feature Extraction', in Calzolari, N., Choukri, K., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., Rosner, M. and Tapias, D., eds., *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10*, Valletta, Malta, May 19-21, 2010, European Language Resources Association (ELRA).

Zhou, G., Zhang, J., Su, J., Shen, D. and Tan, C. (2004) 'Recognizing Names in Biomedical Texts: A Machine Learning Approach', *Bioinformatics*, 20(7), 1178-1190.

Zhu, X., Li, M., Gao, J. and Huang, C.-N. (2003) 'Single character Chinese named entity recognition', in *Proceedings of the second SIGHAN workshop on Chinese language processing - Volume 17*, Sapporo, Japan, 1119268: Association for Computational Linguistics, 125-132.

Appendix A: ZooBirth700 Sample

St. Louis Post-Dispatch (Missouri)

April 19, 1998, Sunday, THREE STAR EDITION

RARE BABY PANDA BORN IN CHINESE ZOO AND FEARED TOO SICK TO LIVE HAS GAINED ITS HEALTH

BYLINE: REUTERS NEWS SERVICE

SECTION: METRO, Pg. C8

LENGTH: 96 words

DATELINE: BEIJING

A rare baby panda born in the Beijing Zoo in **dobn** July **dobn** and feared too sick to survive has been given a clean bill of health, the Xinhua news agency says.

The panda, Ben Ben, weighed **wb** 5.6 ounces **wb** at birth and failed to meet health standards at **lot** 1 month **lot** old, Xinhua reported.

"Ben Ben now weighs **wc** 73 pounds **wc**, and its health is normal," Xinhua said, lauding "the panda keepers' efforts" for the dramatic turnaround.

Ben Ben's mother, Yong Yong, was **agm** 16 years old **agm** and very ill when she gave birth to the panda and its sibling, which was too weak to survive, Xinhua said.

LOAD-DATE: April 20, 1998

Appendix A: ZooBirth700 Sample

AAP NEWSFEED

May 24, 1998, Sunday

US: RARE ALBINO KOALA BIRTH KEPT A SECRET BY SAN DIEGO ZOO

BYLINE: By Dale Paget

SECTION: Nationwide General News; Overseas News

LENGTH: 171 words

SAN DIEGO, May 23 AAP - The only albino koala in captivity in the world has been a secret of the San Diego Zoo in the United States for the last few months.

The koala was given the aboriginal name "Onya-Birri" which means "Ghost Boy".

Zoo keepers discovered they had an extremely rare newborn when Onya-Birri appeared from his mothers' pouch in March. He has a completely white coat and a pink nose.

News of Onya-Birri's arrival was a secret until a San Diego newspaper reported the birth today.

The zoo was home to another albino koala, Goolara, but it died of cancer in 1992 at seven years of age. At the time Goolara was the only albino koala in captivity.

Goolara and Onya-Birri were not believed to be related.

The San Diego Zoo was home to the largest colony of koalas outside Australia with 65 koalas, many were on loan to other zoos in the US and elsewhere.

The San Diego Zoo donates half of the proceeds from the koala loan program to koala habitat preservation efforts in Australia.

LOAD-DATE: May 23, 1998

Appendix A: ZooBirth700 Sample

Buffalo News (New York)

February 8, 1993, Monday, City Edition

FOURTH GORILLA IN THREE YEARS IS BORN AT ZOO

BYLINE: By TOM BUCKHAM, News Staff Reporter

SECTION: LOCAL

LENGTH: 307 words

The Buffalo Zoo's gorilla breeding program, which for a decade failed to produce any offspring, appears to be in full bloom.

The fourth baby lowland gorilla born in the last three years arrived Friday and appears to be in good health, the zoo announced today. Zoo officials believe the baby is a female, which would be another first for the program.

It was the third baby for Becky, an 11-year-old female, and Omega, the 31-year-old silverback. They cooperated to produce B.K., the zoo's first baby lowland gorilla, born in March 1991, and the second, Samson, born last April. Omega also is the father of a third male, still unnamed, born in August to 9-year-old Zira.

The latest baby and its parents will be allowed to move freely between the gorilla habitat and the privacy of an adjacent holding area for about a week while they adjust to one another. Other members of the collection will remain off exhibit during this period.

Zoo officials hope Becky, who eventually rejected her first two infants, which had to be hand-reared by a nurse, will have learned enough from those experiences to be a nurturing mother this time.

The Buffalo Zoo, starting with B.K., has produced more baby gorillas than any other U.S. zoo. It is contemplating the addition of another young adult female to the family in the spring or summer, said Executive Director Minot H. Ortolani.

NATURAL BORN GORILLAS;
Gorilla called Frala shows off her new baby at Taronga Zoo

SECTION: Page 21

LENGTH: 159 words

He isn't the most beautiful baby in the world - but mum thinks he's just wonderful.

And so do zoologists around the world, for this little fellow really is rather special.

The two-week-old youngster is the first gorilla born naturally in captivity in the southern hemisphere.

Keepers at Sydney's Taronga Zoo only discovered yesterday he was a boy – it was the first time proud mum Frala had let them close enough to establish the baby's sex.

But the tot's doting dad - a huge 400lb silverback called Kibabu - was standing close by to make sure no harm came to him.

Female gorillas produce one child every four years on average and 17-year-old Frala, who was transferred from a Dutch zoo a year ago, isn't the first to give birth in Australia.

But the others were either the result of IVF treatment or were delivered by caesarean section.

So far, the new baby doesn't have a name. They're just calling him The Wonder from Down Under.

LOAD-DATE: March 20, 1998

Appendix A: ZooBirth700 Sample

Belle Isle Zoo Celebrates Season Opening with Alpaca Birth

SECTION: State and Regional News

LENGTH: 269 words

DATELINE: ROYAL OAK, Mich., May 1

The Belle Isle Zoo (BIZ) is pleased to announce the birth of a female alpaca. Visitors can see the young alpaca beginning 10 a.m., May 1, 1997 when the facility opens for the summer.

The calf was, born March 24, stands about 2 feet tall and weighs approximately 30 pounds. When fully grown, the alpaca will stand about four feet high and weigh approximately 200 pounds. She is black with white markings on her eyes and lips and has white "socks" on her feet. The birth brings the BIZ alpaca population to nine.

Alpacas (Llama glama pacos) are, native to the Andes Mountains in South America. They are a close relative of the llama and are bred primarily for their fleece, which is longer and silkier than that of other llama species.

Other popular exhibits at the BIZ include lion-tailed macaques, a maned wolf, Egyptian geese, trumpeter swans, Sumatran tigers, African lions, red-crowned cranes, Bactrian deer and Grant's zebras.

Of the 47 species at the Belle Isle Zoo, 12 are officially listed as endangered or threatened. The Belle Isle Zoo is a Detroit Zoological Institute (DZI) facility.

In 1980, the Belle Isle Zoo was expanded from 3 acres to 13 acres and re-designed with a 3/4 mile-long, elevated boardwalk which offers visitors a rare perspective of the animals. It is located between Central and Tanglewood streets on Belle Isle and is open from 10 a.m. to 5 p.m. daily through November 1. Admission is $3 for ages 13 and older; $2 for seniors ages 62 and older: $1 for ages 2 to 12 and free for children under 2.

SOURCE Detroit Zoological Institute

Appendix A: ZooBirth700 Sample

The Post-Standard (Syracuse, NY)

March 1, 1997 Saturday Metro Edition

MOTHER, DAUGHTER DOING WELL AT ZOO SYRACUSE'S BURNET PARK IS HOME TO A BABY ELEPHANT, MALI, BORN FRIDAY MORNING.

BYLINE: LILLIAN ABBOTT PFOHL The Post-Standard, Jeff Stage contributed to this report.

SECTION: LOCAL NEWS; Pg. A7

LENGTH: 409 words

It took awhile, but Burnet Park Zoo's newest baby elephant got the knack of nursing.

When you've got a trunk constantly in your way, that's quite a trick to get down in your first day of life.

Mali arrived at 4:25 a.m. Friday. She's the first daughter born to Targa, a 14-year-old Asian elephant who came to Syracuse in 1990 from Busch Gardens in Florida.

It took Mali a little while Friday to learn to nurse, but by the end of the day she was nursing more regularly, said Chuck Doyle, the zoo's curator of mammals.

"Sometimes, when the baby goes to nurse, Targa gets a little nervous, and the handlers go in to calm her," Doyle said. "With first-time moms, it's not unusual to have a little bit of difficulty settling down."

Right now, Mali "is apparently healthy, and she's getting plenty of milk," Doyle said. "The first few days of life are tenuous at best, but we're very pleased with how well everything went and we're cautiously optimistic the baby will survive."

Mali is the first second-generation captive-born elephant at the zoo. At birth, she stood 38 inches high and weighed 293 pounds.

Mali and Targa are separated from five other elephants at the zoo. Baker

said they hope to get the herd back together as soon as possible. All other elephants at the zoo have seen the baby.

"It looks like Targa is going to be a very, very protective mother," zoo Director Anne Baker said. "We want her to get real comfortable with the baby and let them tell us when they're ready" to rejoin the herd.

Doyle said the weather, Mali's health and the mother-daughter bonding process will dictate when the new calf is introduced to the public.

"We would like to get her out as soon as possible, and we'll have some sort of coming-out party for her," Doyle said.

The zoo will announce when Mali makes her debut, which should be sometime in the next month, he said. Mali is the fourth elephant born in the 1990s in Syracuse.

The baby's father is the zoo's Indy, who sired all four calves born in Syracuse and two born at other zoos.

Tundi and Kirina - both of whom have the same mother, Romani - still reside at the Burnet Park Zoo. Kirina was born in June 1995, while Tundi was born in July 1991.

It's unclear whether Mali will stay at the zoo permanently, Doyle said.

"She'll definitely stay here for a long while," he said. "But it may be that it's better for the species as a whole for her to go somewhere else to breed later on."

LOAD-DATE: January 30, 2003

Appendix A: ZooBirth700 Sample

PR Newswire

June 25, 1996, Tuesday - 10:17 Eastern Time

TWIN SAUDI GOITERED GAZELLES BORN AT THE DETROIT ZOO

SECTION: State and Regional News

LENGTH: 436 words

After 5 1/2 months of anticipation, twin Saudi goitered gazelles were born at the Detroit Zoo on May 25. "Asifa," an Arabic word meaning tempest, and "Asal," which means honey, are on exhibit now.

Asifa and Asal, both females, weigh four and three pounds, respectively. They have a sandy-brown coat with tiny, delicate legs. Two-year-old father Muaddib and 4-year-old mother Jazar are also on exhibit. The Detroit Zoo now has a total of eight Saudi goitered gazelles.

ROYAL OAK, Mich., June 25

Saudi goitered gazelles are an endangered species native to the deserts and plateaus of Saudi Arabia. They are named for a large, visible larynx in the mid-throat region that gives the appearance of a goiter. The Detroit Zoo has nicknamed this species "disappearing desert dweller" because of their camouflage coloring, which is useful in the desert, and their endangered status.

When fully-grown, the gazelles stand 2 1/2 feet tall and weigh 60 pounds. Males develop large, lyrate horns while females' horns grow long, slender and fragile.

Due largely to hunting, the population of Saudi goitered gazelles is now estimated at less than 1,200 in the wild.

"It's always exciting when a birth occurs at the Zoo. These births are particularly significant in terms of conservation because Asifa and Asal are two of only 30 Saudi goitered gazelles in North American zoos," said Ron Kagan, director of the Detroit Zoological Institute (DZI).

The Detroit, San Diego and San Antonio zoos are the only North American parks that exhibit Saudi goitered gazelles and are cooperating to preserve

these animals.

Scott Carter, curator of animals for the DZI, is also the studbook keeper for the North American population of Saudi goitered gazelles. He developed a population management plan that uses a series of breeding coefficients to determine the best breeding pairs for the captive U.S. population. With the next five to 10 years, Carter hopes to grow the captive population to 100 animals.

Opened in 1928, the 125-acre Detroit Zoo was the first zoo in the United States to use barless exhibits extensively. It is a natural habitat for more than 1,250 animals and 700 varieties of trees, shrubbery and flowering plants. Of the 48 mammal species at the zoo, 34 are officially listed as endangered, threatened or extinct in the wild. The Detroit Zoo opens at 10 a.m. 362 days a year. It is located at the intersection of 10 Mile Road and Woodward Avenue, off of I-696, Royal Oak. CONTACT: Lisa Viselli or Michele Scott of Hermanoff & Associates, 313-964-6644

LOAD-DATE: June 26, 1996

Appendix A: ZooBirth700 Sample

New Straits Times (Malaysia)

July 19, 1996

First serow born in Malacca Zoo

SECTION: National; Pg. 17

LENGTH: 201 words

MALACCA, Thurs. - Malacca Zoo scored a first when a serow (kambing gurun) was born under its captive breeding programme.

A five-year-old female named Anip gave birth to the kid in April. It is fathered by seven-year-old Kheow which was acquired from the Thailand Zoo through an animal exchange programme.

The male kid, weighing 3kg at birth, is now 8kg and has become the zoo's pride.

Malacca Zoo veterinary surgeon Dr Razeem Mazlan Abdullah said it was difficult to breed serow, a protected and endangered species, in captivity.

"The zoo did it somehow and it now has five serows - a kid, an adult male and three females," he said.

The young serow is one of the eight animals born under the zoo's captive breeding programme this year. A female springbok was born two days ago.

Two weeks ago three Malayan tiger cubs were born but only one survived. The 3.2kg cuddly cub will be hand-fed until it is five months' old. It is expected to make its first public appearance at the zoo's pets' corner on Aug 15.

Malacca Zoo has 11 tigers from which 30 cubs were born. Many of these cubs have been sent to other zoos in the country and abroad under animal exchange programmes.

LOAD-DATE: March 17, 1999

Appendix A: ZooBirth700 Sample

The Associated Press State & Local Wire

May 4, 2007 Friday 3:01 PM GMT

Elephant born at zoo

SECTION: STATE AND REGIONAL

LENGTH: 213 words

DATELINE: HUGO Okla.

An 8,300-pound elephant and her newborn, who was 280 pounds at birth, are doing well at a breeding compound and retirement center for elephants in this southeastern Oklahoma city.

Val was born to Whimpy after nearly 12 hours of labor on April 27 at the Endangered Ark Foundation. This is the third birth at the center, which was founded by D.R. Miller, whose family started the first Hugo-based circus.

"We have them in the barn right now. The baby is doing well. She's nursing, and her mom is taking good care of her," said Tim Friscia, manager of the foundation.

Although Miller died in 1999, his family has continued to work toward his vision of having a breeding center for Asian elephants. Asian elephants are an endangered species and can't be imported.

The gestation period for elephants is about 22 months. Once the baby elephant is born, it will nurse for 18 to 24 months, Friscia said.

The father is Tommy, a 12-year-old who also lives at the compound.

Whimpy and Val will stay in the facility's barns for a few more days before being moved outside to a special pen, Friscia said.

Male elephants have little to do with the caretaking of their offspring. Most of the rearing is done by the mother and other female elephants.

Appendix B: SNE Prolog Database

```
sndb( 2, [ ( '5', lob, 1, 1, 6, mid_sen ), ( '1', lob, 1, 1, 7, mid_sen ), (
/, lob, 1, 1, 8, mid_sen ), ( '2', lob, 1, 1, 9, mid_sen ), ( feet, lob, 1, 1,
10, mid_sen ) ] ).
sndb( 2, [ ( '138', wb, 1, 1, 13, mid_sen ), ( pounds, wb, 1, 1, 14, mid_sen )
] ).
sndb( 2, [ ( five, zs, 1, 5, 127, mid_sen ) ] ).
sndb( 2, [ ( four, no, 3, 6, 175, mid_sen ) ] ).
sndb( 2, [ ( '15', g, 2, 7, 204, mid_sen ), ( months, g, 2, 7, 205, mid_sen )
] ).
sndb( 2, [ ( first, num, 2, 9, 276, mid_sen ) ] ).
sndb( 2, [ ( second, num, 2, 9, 278, mid_sen ) ] ).
sndb( 2, [ ( two, lot, 1, 10, 297, mid_sen ), ( years, lot, 1, 10, 298,
mid_sen ) ] ).
sndb( 2, [ ( two, num, 1, 11, 334, mid_sen ) ] ).
sndb( 2, [ ( '8', zoh, 2, 11, 351, mid_sen ), ( a, zoh, 2, 11, 352, mid_sen ),
( '.', zoh, 2, 11, 353, mid_sen ), ( m, zoh, 2, 11, 354, mid_sen ), ( '.',
zoh, 2, 11, 355, mid_sen ), ( to, zoh, 2, 11, 356, mid_sen ), ( '7', zoh, 2,
11, 357, mid_sen ), ( p, zoh, 2, 11, 358, mid_sen ), ( '.', zoh, 2, 11, 359,
mid_sen ), ( m, zoh, 2, 11, 360, mid_sen ), ( '.', zoh, 2, 11, 361, mid_sen )
] ).
sndb( 2, [ ( four, no, 6, 11, 422, mid_sen ) ] ).
sndb( 6, [ ( double, nb, 1, 1, 9, mid_sen ) ] ).
sndb( 6, [ ( pair, nb, 1, 2, 16, mid_sen ) ] ).
sndb( 6, [ ( one, num, 1, 2, 32, mid_sen ) ] ).
sndb( 6, [ ( '2', tob, 1, 2, 34, mid_sen ), ( '.', tob, 1, 2, 35, mid_sen ), (
'44pm', tob, 1, 2, 36, mid_sen ) ] ).
sndb( 6, [ ( second, num, 1, 2, 39, mid_sen ) ] ).
sndb( 6, [ ( '4', tob, 1, 2, 41, mid_sen ), ( '.', tob, 1, 2, 42, mid_sen ), (
'15pm', tob, 1, 2, 43, mid_sen ) ] ).
sndb( 6, [ ( first, num, 1, 2, 47, mid_sen ) ] ).
sndb( 6, [ ( '1988', dopz, 1, 2, 52, mid_sen ) ] ).
sndb( 6, [ ( four, noe, 1, 4, 90, mid_sen ), ( to, noe, 1, 4, 91, mid_sen ), (
six, noe, 1, 4, 92, mid_sen ), ( weeks, noe, 1, 4, 93, mid_sen ) ] ).
sndb( 6, [ ( trio, num, 1, 4, 96, mid_sen ) ] ).
sndb( 6, [ ( '100', rcp, 1, 7, 190, mid_sen ) ] ).
sndb( 6, [ ( dozen, nbc, 1, 7, 198, mid_sen ) ] ).
sndb( 6, [ ( '1990s', date, 1, 8, 249, mid_sen ) ] ).
sndb( 6, [ ( eight, lot, 1, 12, 353, mid_sen ), ( or, lot, 1, 12, 354, mid_sen
), ( nine, lot, 1, 12, 355, mid_sen ), ( year, lot, 1, 12, 356, mid_sen ) ] ).
sndb( 6, [ ( six, agf, 1, 15, 416, mid_sen ) ] ).
```

```
sndb( 6, [ ( two, lot, 1, 15, 421, mid_sen ), ( year, lot, 1, 15, 422, mid_sen
) ] ).
sndb( 6, [ ( late, doaf, 1, 15, 430, mid_sen ), ( '2002', doaf, 1, 15, 431,
mid_sen ) ] ).
sndb( 6, [ ( four, ag, 1, 16, 436, mid_sen ) ] ).
sndb( 6, [ ( two, ldoam, 1, 16, 441, mid_sen ), ( years, ldoam, 1, 16, 442,
mid_sen ), ( ago, ldoam, 1, 16, 443, mid_sen ) ] ).
sndb( 6, [ ( '2', loa, 1, 17, 458, mid_sen ), ( '.', loa, 1, 17, 459, mid_sen
), ( '5', loa, 1, 17, 460, mid_sen ), ( metres, loa, 1, 17, 461, mid_sen ) ]
).
sndb( 6, [ ( '140', wa, 1, 17, 468, mid_sen ), ( kilos, wa, 1, 17, 469,
mid_sen ) ] ).
sndb( 6, [ ( '2', loa, 1, 17, 477, mid_sen ), ( '.', loa, 1, 17, 478, mid_sen
), ( '3', loa, 1, 17, 479, mid_sen ), ( metres, loa, 1, 17, 480, mid_sen ) ]
).
sndb( 6, [ ( '100', wa, 1, 17, 483, mid_sen ), ( kilos, wa, 1, 17, 484,
mid_sen ) ] ).
sndb( 6, [ ( one, num, 1, 19, 523, mid_sen ) ] ).
sndb( 6, [ ( one, num, 1, 22, 612, mid_sen ) ] ).
sndb( 6, [ ( first, num, 1, 22, 615, mid_sen ) ] ).
sndb( 7, [ ( two, num, 1, 2, 15, mid_sen ) ] ).
sndb( 7, [ ( two, num, 2, 2, 56, mid_sen ) ] ).
sndb( 7, [ ( 'Nov', dobn, 2, 2, 62, mid_sen ), ( '.', dobn, 2, 2, 63, mid_sen
), ( '23', dobn, 2, 2, 64, mid_sen ), ( ',', dobn, 2, 2, 65, mid_sen ), (
'2004', dobn, 2, 2, 66, mid_sen ) ] ).
sndb( 7, [ ( first, zsp, 2, 2, 74, mid_sen ) ] ).
sndb( 7, [ ( '115', lot, 2, 2, 80, mid_sen ), ( -, lot, 2, 2, 81, mid_sen ), (
year, lot, 2, 2, 82, mid_sen ) ] ).
sndb( 7, [ ( two, num, 1, 3, 143, mid_sen ) ] ).
sndb( 7, [ ( '10', noe, 1, 3, 169, mid_sen ), ( weeks, noe, 1, 3, 170, mid_sen
), ( of, noe, 1, 3, 171, mid_sen ), ( age, noe, 1, 3, 172, mid_sen ) ] ).
sndb( 7, [ ( nine, zs, 3, 3, 217, mid_sen ) ] ).
sndb( 7, [ ( five, nb, 3, 3, 226, mid_sen ) ] ).
sndb( 7, [ ( 'April', dobn, 3, 3, 230, mid_sen ), ( '2005', dobn, 3, 3, 231,
mid_sen ) ] ).
sndb( 7, [ ( '60', num, 1, 4, 244, mid_sen ) ] ).
sndb( 7, [ ( '12', wp, 3, 4, 280, mid_sen ), ( ',', wp, 3, 4, 281, mid_sen ),
( '000', wp, 3, 4, 282, mid_sen ), ( to, wp, 3, 4, 283, mid_sen ), ( '15', wp,
3, 4, 284, mid_sen ), ( ',', wp, 3, 4, 285, mid_sen ), ( '000', wp, 3, 4, 286,
mid_sen ) ] ).
sndb( 7, [ ( eight, e, 3, 4, 297, mid_sen ), ( to, e, 3, 4, 298, mid_sen ), (
'10', e, 3, 4, 299, mid_sen ), ( years, e, 3, 4, 300, mid_sen ) ] ).
sndb( 7, [ ( '202', phon, 1, 5, 307, mid_sen ), ( /, phon, 1, 5, 308, mid_sen
```

```
), ( '633', phon, 1, 5, 309, mid_sen ), ( -, phon, 1, 5, 310, mid_sen ), (
'3081', phon, 1, 5, 311, mid_sen ) ] ).
sndb( 7, [ ( '202', phon, 1, 5, 316, mid_sen ), ( /, phon, 1, 5, 317, mid_sen
), ( '633', phon, 1, 5, 318, mid_sen ), ( -, phon, 1, 5, 319, mid_sen ), (
'3082', phon, 1, 5, 320, mid_sen ) ] ).
sndb( 9, [ ( 'Three', nb, 1, 1, 1, mid_sen ) ] ).
sndb( 9, [ ( third, zsp, 1, 1, 11, mid_sen ) ] ).
sndb( 9, [ ( 'September', date, 1, 1, 15, mid_sen ) ] ).
sndb( 9, [ ( 'April', dobn, 1, 2, 29, mid_sen ), ( '29', dobn, 1, 2, 30,
mid_sen ) ] ).
sndb( 9, [ ( 'September', dop, 1, 3, 99, mid_sen ) ] ).
sndb( 9, [ ( three, num, 1, 3, 101, mid_sen ) ] ).
sndb( 9, [ ( one, num, 1, 3, 104, mid_sen ) ] ).
sndb( 9, [ ( a, lot, 1, 3, 110, mid_sen ), ( month, lot, 1, 3, 111, mid_sen ),
( ago, lot, 1, 3, 112, mid_sen ) ] ).
sndb( 9, [ ( two, num, 1, 5, 124, mid_sen ) ] ).
sndb( 9, [ ( 'September', date, 1, 5, 135, mid_sen ) ] ).
sndb( 9, [ ( seven, zs, 1, 5, 142, mid_sen ) ] ).
sndb( 9, [ ( two, num, 1, 6, 171, mid_sen ) ] ).
sndb( 9, [ ( '150', wp, 1, 7, 188, mid_sen ) ] ).
sndb( 9, [ ( '3', hwp, 1, 7, 200, mid_sen ), ( ',', hwp, 1, 7, 201, mid_sen ),
( '000', hwp, 1, 7, 202, mid_sen ) ] ).
sndb( 9, [ ( '25', lot, 1, 7, 204, mid_sen ), ( years, lot, 1, 7, 205, mid_sen
), ( ago, lot, 1, 7, 206, mid_sen ) ] ).
sndb( 10, [ ( eight, zs, 1, 1, 12, mid_sen ) ] ).
sndb( 10, [ ( two, nb, 1, 1, 19, mid_sen ) ] ).
sndb( 10, [ ( five, ag, 1, 2, 29, mid_sen ), ( -, ag, 1, 2, 30, mid_sen ), (
year, ag, 1, 2, 31, mid_sen ), ( -, ag, 1, 2, 32, mid_sen ), ( old, ag, 1, 2,
33, mid_sen ) ] ).
sndb( 10, [ ( two, nb, 1, 3, 41, mid_sen ) ] ).
sndb( 10, [ ( '9', tob, 1, 3, 52, mid_sen ), ( '.', tob, 1, 3, 53, mid_sen ),
( '30', tob, 1, 3, 54, mid_sen ), ( a, tob, 1, 3, 55, mid_sen ), ( '.', tob,
1, 3, 56, mid_sen ), ( m, tob, 1, 3, 57, mid_sen ), ( '.', tob, 1, 3, 58,
mid_sen ) ] ).
sndb( 10, [ ( '2', tob, 1, 3, 60, mid_sen ), ( '.', tob, 1, 3, 61, mid_sen ),
( '50', tob, 1, 3, 62, mid_sen ), ( p, tob, 1, 3, 63, mid_sen ), ( '.', tob,
1, 3, 64, mid_sen ), ( m, tob, 1, 3, 65, mid_sen ), ( '.', tob, 1, 3, 66,
mid_sen ) ] ).
sndb( 10, [ ( two, nb, 1, 4, 79, mid_sen ) ] ).
sndb( 10, [ ( two, lot, 1, 4, 100, mid_sen ), ( months, lot, 1, 4, 101,
mid_sen ) ] ).
sndb( 10, [ ( two, nb, 1, 5, 114, mid_sen ) ] ).
```