Resources and linguistic processors

Deliverable D4.1

Version DRAFT

Authors: Rodrigo Agerri[1], Itziar Aldabe[1], Egoitz Laparra[1], German Rigau[1], Beñat Zapirain[1], Marieke van Erp[2], Sara Tonelli[3], Marco Rospocher[3] and Piek Vossen[2]

Affiliation: (1) EHU, (2) VUA, (3) FBK

Building structured event indexes of large volumes of financial and economic data for decision making

ICT 316404

Grant Agreement No.	316404
Project Acronym	NEWSREADER
Project Full Title	Building structured event indexes of large volumes of financial and economic data for decision making.
Funding Scheme	FP7-ICT-2011-8
Project Website	http://www.newsreader-project.eu/
Project Coordinator	Prof. dr. Piek T.J.M. Vossen VU University Amsterdam Tel. + 31 (0) 20 5986466 Fax. + 31 (0) 20 5986500 Email: piek.vossen@vu.nl
Document Number	Deliverable D4.1
Status & Version	DRAFT
Contractual Date of Delivery	June 2013
Actual Date of Delivery	July 23, 2013
Type	Report
Security (distribution level)	Public
Number of Pages	127
WP Contributing to the Deliverable	WP4
WP Responsible	EHU
EC Project Officer	Sophie Reig

Authors: Rodrigo Agerri[1], Itziar Aldabe[1], Egoitz Laparra[1], German Rigau[1], Beñat Zapirain[1], Marieke van Erp[2], Sara Tonelli[3], Marco Rospocher[3] and Piek Vossen[2]

Keywords: existing tools, existing datasets, event detection

Abstract: The research activities conducted within the NewsReader project strongly rely on the automatic detection of events. Events are the core information unit underlying news and WP04 addresses the development of text processing modules. The modules detect mentions of events, participants, their roles and the time and place expressions in the four project languages. This deliverable consists of an in-depth survey of the current state of the art, data sources, tools and technology related to event detection for English, Dutch, Spanish and Italian.

Table of Revisions

Version	Date	Description and reason	By	Affected sections
0.1	April 2013	Structure of the deliverable set	Itziar Aldabe	All
0.2	May 2013	First draft of the deliverable	Rodrigo Agerri, Itziar Aldabe, Egoitz Laparra, German Rigau, Beñat Zapirain	All
0.3	08 June 2013	Revision of Introduction, Processing Events in Text, Text Classification, NERC, Coref., NED, WSD, SRL	Rodrigo Agerri, Itziar Aldabe, Egoitz Laparra, German Rigau, Beñat Zapirain	Sections 1-7,9
0.4	19 June 2013	Revision of Sentiment Analysis, Discourse Analysis. Split factuality and time section	Marieke van Erp	Sections 8, 10-12
0.5	20 June 2013	Revision of Event Relations	Sara Tonelli	Section 15
0.6	21 June 2013	Revision of WSD	German Rigau	Section 7
1.0	July 2013	Draft version of D4.1	Itziar Aldabe, Egoitz Laparra, German Rigau, Marieke van Erp, Piek Vossen	All
1.1	13 July 2013	Internal Review of Sections 1-7,9,16-18	Marieke van Erp	Sections 1-7,9,16-1
1.2	18 July 2013	Revision according to the internal review comments	Itziar Aldabe, German Rigau	Sections 1-7,9,16-18

Executive Summary

This document presents a review of the current state-of-the-art in event detection from text and the components available to the NewsReader project, taking into account licensing issues. Therefore the main outcome of the deliverable is a collection of these components including its description, accessibility, availability, etc. This report is split into two parts: the list of the identified sources and data models, and the main components to analyze it and to provide the functionality needed by the project.

Contents

List of Tables

1 Introduction

This deliverable consists of an in-depth survey of the current state of the art, data sources, tools and technology related to **Event Detection** for English, Dutch, Spanish and Italian. The research activities conducted within the NewsReader project strongly rely on the automatic managing of events, which are considered as the core information unit underlying news and therefore any decision making process that depends on news. The research focuses on four challenging aspects: event detection (addressed in WP04 -Event Detection-), event processing (addressed in WP05 -Event Modelling-), storage and reasoning over events (addressed in WP06 -Knowledge Store-), and scalling to large textual streams (addressed in WP2 -System Design-). Given that one of the main project goals is the extraction of event structures from large streams of documents and their manipulation, a thorough analysis of what is an event, how its participants are characterized and how events are related to each other is of paramount importance.

WP04 (Event Detection) addresses the development of text processing modules that detect mentions of events, participants, their roles and the time and place expressions in the four project languages. Another objective is to classify textual information on the factuality of the events and to derive the authority and trust of the source.

NewsReader plans to use an open architecture for Natural Language Processing (NLP) as a starting point. The system plans to use an extension of KAF [Bosma *et al.*, 2009] as a layered annotation format for text that can be shared across languages and that can be extended with more layers when needed. Separate modules will be developed to add interpretation layers using the output of previous layers. We plan to develop new modules to perform event detection and to combine separate event representations. When necessary, new modules will be developed using the gold standards and training data developed in WP03 (Benchmarking). Specific input and output wrappers need to be developed or adapted to work with the new formats and APIs defined in WP02 (System Design). For that, NewsReader plans to exploit a variety of knowledge-rich and machine-learning approaches. All modules will work on all the languages in NewsReader: English, Dutch, Spanish and Italian. Additionally, NewsReader plans to provide an abstraction layer for large-scale distributed computations, separating the "what" from the "how" of computation and isolating NLP developers from the details of concurrent programming.

Text-processing requires basic and generic NLP steps, such as tokenization, lemmatization, part-of-speech tagging, parsing, word sense disambiguation, named entity and semantic role recognition for all the languages in NewsReader. Furthermore, named entities are as much as possible linked to possible Wikipedia pages as external sources (Wikification) and entity identifiers. We plan to use existing state-of-the-art technology and resources for this. Technology and resources will be selected for quality, efficiency, availability and extendability to other languages. NewsReader will provide (1) wide-coverage linguistic processors adapted to the financial domain and (2) new techniques for achieving interoperable semantic interpretation of English, Dutch, Spanish and Italian.

The semantic interpretation of the text is directed towards the detection of event mentions and those named entities that play a role in these events, including time and location expressions. This implies covering all expressions (verbal, nominal and lexical) and meanings that can refer to events, their participating named entities, time and place expressions but also resolving any co-reference relations for these named entities and explicit (causal) relations between different event mentions. The analysis results in an augmentation of the text with semantic concepts and identifiers. This allows us to access lexical resources and ontologies that provide for each word and expression 1) the possible semantic type (e.g. to what type of event or participant can it refer), 2) the probability that it refers to that type (as scored by the word sense disambiguation and named entity recognition), 3) what types of participants are expected for each event (using background knowledge resources) and 4) what semantic roles are expected for each event (also using background knowledge resources. Such constraints can be used in rule-based, knowledge-rich and hybrid machine-learning systems to determine the actual events structures in texts.

We also plan to develop classifiers (e.g. on the basis of textual and structural markers such as *not*, *failed*, *succeeded*, *might*, *should*, *will*, *probably*, etc.) that provide a factuality score which indicates the likelihood that an event took place. Authority and trust can be based on the metadata available on each source, the number of times the same information is expressed by different sources (possibly combined with the type of source), but also on stylistic properties of the text (formal or informal, use of references, use of direct and indirect speech) and richness and coherence of the information that is given. For each unique event, we also derive a trust and authority score based on the source data and a factuality score based on the textual properties. This information can easily be added to the layered annotation format in separate layers connected to each event, without complicating the current representations.

The textual sources defined in WP01 (User Requirements) by the industrial partners come in various formats. In WP02 (System Design), we are defining the RDF formats to represent the information of these sources. In WP04, we will process the textual information to compatible RDF formats and make them available for subsequent NewsReader modules.

Finally, following T02.4 ("scaling requirements"), NewsReader will provide an abstraction layer for large-scale distributed computations, separating the "what" from the "how" of computation and isolating NLP developers from the details of concurrent programming. The different modules and the resources that they need to access or load will be adapted to be used in such a format and to provide optimal performance.

The remainder of the document consists of the following sections. Section 2 presents the event detection task. Sections 3 to 15 presents current academic and industrial systems and data sources for all the subtasks which are part of event detection. Some conclusions of this deliverable will be discussed in Section 16. The languages for which every data source and module are available are explicitly listed. In order to make this document as self-contained as possible, every section will start by offering a description of each of the tasks. Moreover, it also contains a table of available data sources and technology modules in order to obtain a general overview of current availability of technology relevant for WP04 in NewsReader.

2 Processing Events in Text

This section introduces the main tasks to process events across documents in four different languages: English, Dutch, Spanish and Italian. This process involves the identification of event mentions, event participants, the temporal constraints and, if relevant, the location. Furthermore, it also implies the detection of expressions of factuality of event mentions and the authority of the source of each event mention.

2.1 Event detection from multilingual textual sources

One of the main research objectives of NewsReader is the identification of event mentions across documents in four different languages: English, Dutch, Spanish and Italian. In addition, we will extract information about the event participants, the temporal constraints and the location. Furthermore, we also need to detect expressions of factuality of event mentions and the authority of the source of each event mention. The former is derived from expressions that indicate whether an event took place or is speculative. The latter can be based on textual properties (subjectivity of the text and style) and on the meta-data related to the source.

In NewsReader, event detection will be performed mainly in WP04 (Event Detection), and we plan to explore both supervised and unsupervised approaches. Specifically, we will take advantage of existing resources with TimeML[1] annotation in English, Italian and Spanish to train the event detection module, while for Dutch additional annotation and/or unsupervised techniques will be required. Furthermore, novel approaches will be investigated to relate participants information to event mentions by extending the TimeML framework. Event detection will be evaluated by comparing the module coverage and precision against existing benchmarks, such as TimeBank[2] which also includes annotations for Italian and Spanish, and the data sets developed within the TempEval-2 evaluation campaign[3]. A portion of these benchmarks can be manually enriched, within WP03 (Benchmarking), with participant information following a new version of the KYOTO annotation format. The software should show progress on the current state-of-the-art with respect to gold-standards currently employed in evaluation tracks and developed in the project and it should show comparable results across the four languages.

Thus, NewsReader will develop (1) wide-coverage linguistic processors adapted to the financial domain and (2) new techniques for achieving interoperable Semantic Interpretation of English, Dutch, Spanish and Italian. The main goal is to reduce ambiguity to allow improvement of performance. Morphologic, syntactic and semantic processors should be adapted thus defining a methodology for tool customization according to the new domain.

[1] http://timeml.org/site/index.html
[2] http://www.timeml.org/site/timebank/timebank.html
[3] http://www.timeml.org/tempeval2/

2.2 Progress on semantic processing

Although there have been many relevant advances in the research field, Natural Language Processing (NLP) is still far from achieving full natural language understanding, since it demands for a complex analysis of different semantic components, from the detection and classification of named entities to semantic role labelling for the identification of participants. Several semantic tasks are needed for allowing sentences to produce full meaning representations. For instance, in order to create consistent event chains and to identify event mentions that describe the same action, the analysis of the event participants is necessary. On the one hand, at the lexical level, a good performance is needed for detecting and classifying named entities and word sense interpretation. At the sentence level, semantic role labeling is crucial for eventually construing full sentence representations. The semantic tasks needed to accomplish this goal are described below in more detail.

In order to allow interoperable semantic interpretation of texts, we plan to exploit existing wordnets (such as those integrated in the Multilingual Central Repository[4] and MultiWordNet[5]) and Word Sense Disambiguation technology. **Word Sense Disambiguation** (WSD) stands for labelling every word in a text with its appropriate meaning or sense depending on its context [Agirre and Edmonds, 2006]. State-of-the-art WSD systems obtain around 60-70% precision for fine-grained senses and 80-90% for coarser meaning distinctions [Izquierdo *et al.*, 2009]. Lately, graph-based WSD systems are gaining growing attention [Agirre and Soroa, 2009; Laparra *et al.*, 2010]. These methods are language independent since only requires a local wordnet connected to the Princeton WordNet. For instance, using UKB[6], KYOTO developed knowledge-based WSD modules for English, Spanish, Basque, Italian, Dutch, Chinese and Japanese.

First, named entities need to be recognized in running text via **Named Entity Recognition**. Current state-of-the-art processors achieve high performance in recognition and classification of general categories such as people, places, dates or organisations [Nadeau and Sekine, 2007; Singh *et al.*, 2010]. This task also requires to identify of which expressions in a sentence or document refer to the same named entity [Bryl *et al.*, 2010], also known as **co-reference resolution**. The best performing system in the task is a multi-pass sieve co-reference resolution system [Lee *et al.*, 2011]. Current performance rates of around 80% can be improved by using a common platform and drawing information from multiple languages resources at the same time (see for instance some of the tools and resources developed by JRC[7] for Europe Media Monitor[8]). Named entities are very common in financial news and in NewsReader they will be identified and resolved across documents in different languages. In a multilingual setting, the knowledge captured for a particular named entity in one language can be ported to another once converted to a language-neutral representation, likewise balancing resources and technological ad-

[4] `http://adimen.si.ehu.es/web/MCR`
[5] `http://multiwordnet.fbk.eu`
[6] `http://ixa2.si.ehu.es/ukb/`
[7] `http://langtech.jrc.ec.europa.eu/JRC-Names.html`
[8] `http://emm.newsbrief.eu/overview.html`

vances across languages [Steinberger and Pouliquen, 2007]. In NewsReader, we will build a multilingual extension of the cross-document coreference system developed within the LiveMemories project [Poesio and Magnini, 2009] and successfully evaluated in the Evalita 2011 evaluation campaign for Italian.

Furthermore, once the named entities have been recognized, they can be identified with respect to an existing catalogue. Wikipedia has become the de facto standard catalogue for named entity disambiguation, and may be particularly relevant to the creation of background event models because it provides additional information related to event participants, thus allowing to define explicit links among them. **Wikification** is then the process of automatic linking of the named entities occurring in free text to their corresponding Wikipedia articles. This task is typically regarded as a word sense disambiguation problem, where Wikipedia provides both the dictionary and training examples. For instance, DBpedia Spotlight[9]) have achieved good classification accuracy also in multilingual settings and it shows a better coverage of named entities compared to disambiguation models trained on WordNet [Mendes *et al.*, 2011]. Existing architectures are already multilingual, and can be applied to the four languages of the project after training the model on language-specific Wikipedia dumps. In NewsReader, we will also have the option of linking to entities already stored in the Knowledge Store as defined in WP06.

The creation of a web-based large-scale repository of named entities has already been implemented in the Okkam project[10], whose current repository contains 7.5 million entities. In NewsReader we plan to build upon the findings of Okkam by identifying Named Entities participating in the same events and by integrating them into the extracted narrative schemas. For this, we will need to associate the Named entities in a text with the semantic arguments of the predicates denoting specific events. This task, which is usually called **Semantic Role Labeling** (SRL) relies on the role repository encoded in the domain-specific background models, such as those appearing in FrameNet [Baker *et al.*, 1998a] or PropBank [Kingsbury and Palmer, 2002a]. As an alternative, we plan to explore the possibility to use more generic roles, such as Agent, Patient, Instrument or Location. Such quite general and widely-recognized labels are used in building corpora and other linguistic resources [Kipper *et al.*, 2006]. SRL is a crucial task for establishing "Who does What, Where, When and Why" and it is a key technology for applications involving any level of semantic interpretation [Gildea and Jurafsky, 2002a; Carreras and Màrquez, 2005a; Zapirain *et al.*, 2008]. There are only few systems performing semantic role labelling on unrestricted domains and mainly on English. For instance, Mate-tools[11] [Björkelund *et al.*, 2009a] and SEMAFOR[12] [Chen *et al.*, 2010a].

SRL focuses on the extraction of explicit propositional meaning within a sentence boundary. Propositional meaning makes assertions about the world that can be true or false. Non-propositional meaning conveys aspects of meaning that do not have a truth-value (attitudes, sentiment, opinion) or that change the propositional meaning (nega-

[9] http://spotlight.dbpedia.org/
[10] http://www.okkam.org/
[11] http://code.google.com/p/mate-tools/
[12] http://code.google.com/p/semafor-semantic-parser/

tion). Research on **modality and negation** have been focused on two main tasks, the detection of various forms of modality and negation, and the resolution of the scope of modality and negation cues. Several rule and pattern-based [Chapman *et al.*, 2001; Mutalik *et al.*, 2001; Huang and Lowe, 2007; Rokach *et al.*, 2008] and machine learning [Goldin and Chapman, 2003] systems have been developed to detect nagated entities and events in texts, as well as to detect the scope of negation cues [Morante and Daelemans, 2009]. **Modality** allows to express aspects related to the attitude of the speaker towards its own statements in terms of degree of factuality [Saurí and Pustejovsky, 2009a], subjectivity [Wiebe *et al.*, 2004], certainty [Rubin *et al.*, 2006], evidentiality [Aikhenvald, 2004], hedging [Hyland, 1998], comitted belief [Diab *et al.*, 2009], etc. **Scope resolution** is concerned with determining at a sentence level which tokens are affected by negation and modality [Morante and Daelemans, 2009; Özgür and Radev, 2009; Øvrelid *et al.*, 2010]. Despite the progress in recent works, the performance of scope resolvers is low and their capabilities does not include determining exactly which entity or event is negated or speculated; finding uncertainty should be performed at a proposition level, instead of at a sentence level, since a sentence can contain more than one proposition and not all of them need to be uncertain; there are no modality taggers that can tag different types of modality; Finally, existing work focuses mostly on English.

Traditionally, SRL systems have focused in searching the fillers of those explicit roles appearing within sentence boundaries [Gildea and Jurafsky, 2000; Gildea and Jurafsky, 2002b; Carreras and Màrquez, 2005b; Surdeanu *et al.*, 2008; Hajič *et al.*, 2009]. These systems limited their search space to the elements that share a syntactical relation with the predicate. However, when the participants of a predicate are implicit this approach obtains incomplete predicative structures with null arguments. Early works addressing **implicit SRL** cast this task as a special case of anaphora or coreference resolution [Palmer *et al.*, 1986; Whittemore *et al.*, 1991; Tetreault, 2002]. Recently, the task has been taken up again around two different proposals. On the one hand, [Ruppenhofer *et al.*, 2010] presented a task in SemEval-2010 that included an implicit argument identification challenge based on FrameNet [Baker *et al.*, 1998b]. Besides the two systems presented to the task, some other systems have used the same dataset and evaluation metrics to explore alternative linguistic and semantic strategies [Ruppenhofer *et al.*, 2011], [Laparra and Rigau, 2012], [Gorinski *et al.*, 2013] and [Laparra and Rigau, 2013c]. On the other hand, [Gerber and Chai, 2010; Gerber and Chai, 2012] studied the implicit argument resolution on NomBank. All these works agree that implicit arguments must be modeled as a particular case of coreference together with features that include lexical-semantic information, to build selectional preferences. Another common point is the fact that these works try to solve each instance of the implicit arguments independently, without taking into account the previous realizations of the same implicit argument in the document. [Laparra and Rigau, 2013a] propose that these realizations, together with the explicit ones, must maintain a certain coherence along the document and, in consequence, the filler of an argument remains the same along the following instances of that argument until a stronger evidence indicates a change.

Semantic parsing is considerably more complex than Semantic Role Labeling (SRL). In fact, there are not many semantic interpretation systems for unrestricted domains. For instance, Lingo/LKB [Copestake, 2002] or Boxer [Bos, 2008] are not easy to adapt to other languages. For NewsReader, we will not need the full complexity of semantic parsing systems. We can restrict ourselves to more robust and local structures from which we will build up more complex structures in so far they are relevant and fit the general application constraints. Likewise, we will keep the system scalable and robust.

Parsing **discourse** [Kamp and Reyle, 1993] consist of finding binary discourse relations in text. Discourse conective such as *but, although, however, etc.* are considered to be the anchors of discourse relations such as *cause, contrast, conditional, etc.* that relate prepositions, beliefs, facts or eventualities. Several discourse parsers are available for English. Moreover, the analysis of discourse structure of news genre have been also previously studied [Bell, 1991; Bell, 1998].

Furthermore, all linguistic processors developed by this project will be adapted to financial domain. The main goal is to reduce ambiguity to allow the improvement of performance. Morphologic, syntactic and semantic processors should be adapted thus defining a methodology for tool customization according to the new domain [Agirre *et al.*, 2009a; Zapirain *et al.*, 2008].

However, full natural language understanding demands for complete identification of every concepts in the text as well as every relation occurring between them. Moreover, natural language understanding requires knowledge and processing abilities which are far beyond simple word processing. That is, a large set of knowledge expressed implicitly by the linguistic surface elements are needed to be considered by the NLP processors. Thus, natural language understanding involves much more than performing syntactic parsing and looking for words in a dictionary. Real language understanding largely relies in a large amount of semantic and general world knowledge as well as the capability to apply contextual knowledge (pragmatics) to fill gaps and to disambiguate meanings – a routine for speakers but a big challenge for machines. Moreover, current databases are still far from universal coverage, therefore most of non-trivial inferences usually can not be achieved.

2.3 Progress on event-detection

Existing semantic paradigms such as VerbNet[13] [Kipper *et al.*, 2006], FrameNet[14] [Baker *et al.*, 1998a] and TimeML [Pustejovsky *et al.*, 2010] are built upon specifications of events that often contradict each other, and no unitary framework for the analysis of events, relations and event participants over time has been applied to document processing so far. NewsReader aims at filling this gap by developing an architecture that detects, processes, stores and manipulates events in a interoperable multilingual setting.

Event detection has recently become an active area of research with many dedicated workshops (e.g. at LREC 2002, TERQAS 2002, TANGO 2003, ACL 2006, NAACL 2013[15]) and specific evaluation campaigns (i.e. TempEval-1 and TempEval-2). In this context,

[13] `http://verbs.colorado.edu/~mpalmer/projects/verbnet.html`
[14] `https://framenet.icsi.berkeley.edu/fndrupal/`
[15] `https://sites.google.com/site/cfpwsevents/`

the specification language called TimeML has been developed, and consolidated as an ISO standard for the annotation of events, temporal expressions and the anchoring and ordering relations between them [Pustejovsky *et al.*, 2010]. With respect to other existing annotation schemes, ISO-TimeML presents a unifying approach to event and temporal identification:

- it extends the TIDES-TIMEX2 standard [Ferro *et al.*, 2007] for a more detailed annotation of temporal expressions
- it identifies all the textual elements which explicitly express the relations between temporal expressions and events
- it identifies a wide range of linguistic expressions realizing events (including nominalizations and event naming)
- it creates various kinds of dependencies between events and/or temporal expressions allowing the temporal anchoring and ordering of events.

However, ISO-TimeML does not include the identification of event arguments. The definition of the argument structure is essential to perform deep reasoning and full inference over events within texts. For this reason, we plan to adopt the ISO-TimeML specifications in NewsReader, considering the possibility of defining the appropriate argumenthood within event markup, taking Pustejovsky's proposal as a starting point. For the creation of the gold standard we plan to extend the functionalities developed in CAT, the CELCT Annotation Tool[16], that has already been used for the manual annotation of a corpus following the ISO-TimeML standard.

3 Text Classification

Automatic Text Classification involves assigning a text document to a set of pre-defined classes automatically [Aggarwal and Zhai, 2012]. In the research community, the dominant approach to this problem is based on machine learning techniques: a general inductive process automatically builds a classifier by learning, from a set of preclassified documents, the characteristics of the categories. The advantages of this approach over the knowledge engineering approach (consisting in the manual definition of a classifier by domain experts) are a very good effectiveness, considerable savings in terms of expert labor power, and straightforward portability to different domains. [Sebastiani, 2002] discusses the main approaches to text categorization that fall within the machine learning paradigm. An evaluation of different kinds of text classification methods can be found in [Yang and Liu, 1999]. A number of the techniques discussed in this deliverable have also been converted into software packages and are publicly available through multiple toolkits such as the

[16] `http://www.celct.it/projects/CAT.php`

BOW toolkit[17] [McCallum, 1996], Mallet[18] [McCallum, 2002], WEKA[19] and LingPipe[20]. However, there are also successful knowledge-based approaches to text classification such as JEX [21] [Steinberger *et al.*, 2012].

While numerous studied text categorization in the past, good test collections are by far less abundant. This scarcity is mainly due to the huge manual effort required to collect a sufficiently large body of text, categorize it, and ultimately produce it in machine-readable format. Most studies use the Reuters-21578[22] collection as the primary benchmark. Others use 20 Newsgroups[23] and OHSUMED[24], while TREC filtering experiments often use the data from the TIPSTER[25] or AP[26] corpus. TechTC[27] - Technion Repository of Text Categorization Datasets provides a large number of diverse test collections for use in text categorization research. The PASCAL Large Scale Hierarchical Text Classification[28] (LSHTC) Challenge is a hierarchical text classification competition, using large datasets. The challenge is based on two large datasets: one created from the ODP web directory (DMOZ) and one from Wikipedia. The datasets are multi-class, multi-label and hierarchical. The number of categories range between 13,000 and 325,000 roughly and the number of the documents between 380,000 and 2,400,000.

3.1 Tools

3.1.1 JEX

JEX[29] [Steinberger *et al.*, 2012] is multi-label classification software that automatically assigns a ranked list of the over six thousand descriptors (classes) from the controlled vocabulary of the EuroVoc thesesaurus[30] to new texts. JEX has been trained for twenty-two EU languages. The software allows users to re-train the system with their own documents, or with a combination of their own documents and the data provided together with the software. JEX can also be trained using classification schemes other than EuroVoc.

[17] `http://www.cs.cmu.edu/~mccallum/bow/`
[18] `http://mallet.cs.umass.edu/`
[19] `http://www.cs.waikato.ac.nz/ml/weka/`
[20] `http://alias-i.com/lingpipe/demos/tutorial/classify/read-me.html`
[21] `http://ipsc.jrc.ec.europa.eu/index.php?id=60`
[22] `http://trec.nist.gov/data/reuters/reuters.html`
[23] `http://kdd.ics.uci.edu/databases/20newsgroups/20newsgroups.html`
[24] `ftp://medir.ohsu.edu/pub/ohsumed/`
[25] `http://trec.nist.gov/data.html`
[26] `http://www.daviddlewis.com/resources/testcollections/trecap/`
[27] `http://techtc.cs.technion.ac.il/`
[28] `http://lshtc.iit.demokritos.gr/`
[29] `http://ipsc.jrc.ec.europa.eu/index.php?id=60`
[30] `http://eurovoc.europa.eu/`

3.1.2 Mahout

Mahout[31] is a toolbox for clustering, classification and batch based collaborative filtering implemented on top of Apache Hadoop[32] using the map/reduce paradigm. However the library does not restrict contributions to Hadoop based implementations. The library can run on a single node or on a non-Hadoop cluster as well. The core libraries are highly optimized to allow for good performance also for non-distributed algorithms and to calable to process reasonably large data sets. Currently Mahout supports mainly four use cases:

- Recommendation mining takes users' behavior and from that tries to find items users might like.
- Clustering takes e.g. text documents and groups them into groups of topically related documents.
- Classification learns from existing categorized documents what documents of a specific category look like and is able to assign unlabelled documents to the (hopefully) correct category.
- Frequent itemset mining takes a set of item groups (terms in a query session, shopping cart content) and identifies, which individual items usually appear together.

Mahout is licensed under Apache 2.0 License.

3.1.3 OpenNLP Document Categorizer

The OpenNLP Document Categorizer[33] can classify text into pre-defined categories. It is based on maximum entropy framework.

3.1.4 Classifier4j

Classifier4j[34] is a Java library designed to do text classification. It comes with an implementation of a Bayesian classifier, and now has some other features, including a text summary facility.

3.1.5 jTCat

jTCat[35] (java Text Categorization) is a tool for Text Categorization. It is based on a supervised machine learning approach. In particular, jTCat uses a combination of kernel functions to embed the original feature space in a low dimensional one. jTCat requires only shallow linguistic processing, such as tokenization, part-of-speech tagging (optional) tagging and lemmatization (optional). jTCat is freely available for research purposes.

[31] http://mahout.apache.org/
[32] http://hadoop.apache.org/
[33] http://opennlp.apache.org/
[34] http://classifier4j.sourceforge.net/
[35] http://hlt.fbk.eu/en/technology/jTCat

3.1.6 RTextTools

RTextTools[36] is a free, open source machine learning package for automatic text classification that makes it simple for both novice and advanced users to get started with supervised learning. The package includes nine algorithms for ensemble classification (svm, slda, boosting, bagging, random forests, glmnet, decision trees, neural networks, maximum entropy), comprehensive analytics, and thorough documentation. The license of the package is GPL-3.

3.1.7 TCatNG

TCatNG Toolkit[37] is a Java package that you can use to apply N-Gram analysis techniques to the process of categorizing text files. TCatNG is a Java package that implement the classification technique described in [Cavnar and Trenkle, 1994]. The central idea is to calculate a "fingerprint" of a document with an unknown category, and compare this with the fingerprints of a number of documents for which the categories are known. The categories of the closest matches are output as the classification. A fingerprint is a list of the most frequent n-grams occurring in a document, ordered by frequency. Fingerprints are compared with a simple "out-of-place" metric.

This package also implements some extentions to the original proposal. Among other things, the software offers support for Good-Turing smoothing and new fingerprint comparison methods based on the similarity metrics proposed by [Lin, 1998; Jiang and Conrath, 1997]. Other classification methods besides nearest neighbour are also implemented, such as Support Vector Machines or Bayesian Logistic Regression. TCatNG is released under the BSD License.

3.1.8 libTextCat

libTextCat[38] is a library with functions that also implement the classification technique described ij [Cavnar and Trenkle, 1994]. It was primarily developed for language guessing, a task on which it is known to perform with near-perfect accuracy. The library is released under the BSD License.

3.1.9 TexLexAn

TexLexAn[39] is an open source text analyser for Linux, able to estimate the readability and reading time, to classify and summarize texts. It has some learning abilities and accepts html, doc, pdf, ppt, odt and txt documents. Written in C and Python. The license of the package is GPLv2.

[36] `http://www.rtexttools.com/about-the-project.html`
[37] `http://tcatng.sourceforge.net/`
[38] `http://software.wise-guys.nl/libtextcat/`
[39] `http://sourceforge.net/projects/texlexan/`

3.1.10 Mallet

MALLET[40] [McCallum, 2002] is a Java-based package for statistical natural language processing, document classification, clustering, topic modeling, information extraction, and other machine learning applications to text. MALLET includes sophisticated tools for document classification: efficient routines for converting text to "features", a wide variety of algorithms (including Naïve Bayes, Maximum Entropy, and Decision Trees), and code for evaluating classifier performance using several commonly used metrics. The toolkit is Open Source Software, and is released under the Common Public License.

4 Named Entity Recognition and Classification

The term "Named Entity", now widely used in Natural Language Processing, was coined for the Sixth Message Understanding Conference (MUC-6) [Grishman and Sundheim, 1996]. At that time, MUC was focusing on Information Extraction (IE) tasks where structured information of company activities and defense related activities is extracted from unstructured text, such as newspaper articles. In defining the task, people noticed that it is essential to recognize information units such as names, including person, organization and location names, and numeric expressions including time, date, money and percent expressions. Identifying references to these entities in text was recognized as one of the important sub-tasks of IE and was called "Named Entity Recognition and Classification (NERC)".

The NERC field can perhaps be tracked from 1991 to present days, although the NERC task has been partially superseded by the Named Entity Disambiguation via Wikification or Entity Linking tasks since around 2007 [Mihalcea and Csomai, 2007]. While early systems were making use of handcrafted rule-based algorithms, modern systems most often resort to machine learning techniques. It was indeed concluded in an influential conference that the choice of features is at least as important as the choice of technique for obtaining a good NERC system [Tjong Kim Sang and De Meulder, 2003]. Moreover, the way NERC systems are evaluated and compared is essential to the progress in the field.

A good proportion of work in NERC research is devoted to the study of English but a possibly larger proportion addresses language independence and multilingualism. With respect to the languages involved in NewsReader. Spanish and Dutch are strongly represented, boosted by a major devoted conference: CoNLL-2002[41]. Similarly, there have been numerous studies for Italian [Black *et al.*, 1998; Cucchiarelli and Velardi, 2001];

Overall, the most studied types are three specializations of "proper names": names of "**persons**", "**locations**" and "**organizations**". These types are collectively known as "**enamex**" since the MUC-6 competition. The type "location" can in turn be divided into multiple subtypes of "fine- grained locations": city, state, country, etc. [Fleischman and Hovy, 2002]. Similarly, "fine-grained person" sub-categories like "politician" and "enter-

[40] `http://mallet.cs.umass.edu/`
[41] `http://www.clips.ua.ac.be/conll2002/ner/`

tainer" appear in the aforementioned work [Fleischman and Hovy, 2002]. In the ACE[42] program, the type "facility" subsumes entities of the types "location" and "organization", and the type "GPE" is used to represent a location which has a government, such as a city or a country.

The type "miscellaneous" is used in the CoNLL conferences and includes proper names falling outside the classic "enamex". The class is also sometimes augmented with the type "product" [Bick, 2004]. The "timex" (also coined in MUC) types "date" and "time" and the "numex" types "money" and "percent" are also quite predominant in the literature. Since 2003, a community named TIMEX2 proposes an elaborated standard for the annotation and normalization of temporal expressions[43]. Finally, marginal types are sometime handled for specific needs: "film" and "scientist" [Etzioni *et al.*, 2005], "email address" and "phone number" [Witten *et al.*, 1999; Maynard *et al.*, 2001], "brand" [Bick, 2004].

Other work does not limit the possible types to extract and is referred as "open domain" NERC [Alfonseca and Manandhar, 2002; Evans and Street, 2004]. For example, a named entity hierarchy has been defined that includes many fine grained subcategories, such as museum, river or airport, and adds a wide range of categories, such as product and event, as well as substance, animal, religion or color. The hierarchy tries to cover most frequent name types and rigid designators appearing in a newspaper, and the number of categories is about 200 [Sekine and Nobata, 2004].

Most approaches rely on manually annotated newswire corpora, namely, in the MUC 6 and 7 [Grishman and Sundheim, 1996; Chinchor, 1998] conference, in the CoNLL 2002 and 2003 shared tasks mentioned below, and later detailed NE annotations were added to the Penn Treebank [Marcus *et al.*, 1993] by the BBN Pronoun Co-reference and Entity Type Corpus [Weischedel and Brunstein, 2005].

With a well-defined evaluation methodology in MUC and CoNLL tasks and the manually annotated corpora, most of the NERC systems consisted of language independent systems based on automatic learning of statistical models (for technical details of these approaches see [Nadeau and Sekine, 2007]). However, the reliance on expensively manually annotated data hinders the creation of NERC systems for most languages and domains. This has been a major impediment to adaptation of existing NERC systems to other domains, such as the scientific or the biomedical domain [Ciaramita and Altun, 2005].

Some works started to use external knowledge to reduce the dependence on quality manually annotated data. Most of these approaches incorporated knowledge in the form of *gazetteers*, namely, lists of categorized names or common words extracted from the Web [Etzioni *et al.*, 2005] or knowledge resources such as Wikipedia [Toral and Munoz, 2006]. However, this does not necessarily correspond to better results in NERC performance [Mikheev *et al.*, 1999], the bottom line being that *gazetteers* will never be exhaustive and contain all naming variations for every named entity, or free of ambiguity.

As a consequence, the use of external knowledge for NERC has moved on towards semi-supervised approaches and low-cost annotation (in the form of silver standard corpora) as

[42] http://www.itl.nist.gov/iad/mig/tests/ace/
[43] http://www.timexportal.info/system:page-tags/tag/timex2

opposed to supervised approaches highly dependent on large amounts of manually annotated data (gold standard). A crucial role has been the rise to prominence of Wikipedia. Wikipedia provides a large source world knowledge which can be potentially a source of silver-standard data for NE annotations [Richman and Schone, 2008; Mika *et al.*, 2008; Nothman *et al.*, 2008; Nothman *et al.*, 2012].

In section 4.1, the main existing data sources currently available for the development (both in industrial and academic environments) and evaluation of NERC systems are described. Generally, since MUC and CoNLL shared tasks, these data sources consisted of manually annotated data which served as training machine learning models for NERC classification. The performance of these systems is usually evaluated using the F-measure: the harmonic mean of precision and recall. As previously mentioned, more recent trends aim at building automatic silver-standard and gold-standard datasets from existing large knowledge resources such as Wikipedia [Mika *et al.*, 2008; Nothman *et al.*, 2012]. The tools and services for NERC described in section 4.2 are mostly based on supervised machine learning approaches, although some systems make use of knowledge resources such as gazetteers.

4.1 Data Sources

Table 1 lists the data sources, available for the 4 languages included in the project (English, Dutch, Italian and Spanish), in the form of annotated corpora for training and evaluation of NERC systems. Specific details about them are also included. The meaning of the individual columns of Table 1 is as follows:

- Data Entity: name or identification of the data resource, namely, LDC Ontonotes version 4.0.
- Type of data: the type of data which is gathered, i.e. main stream news / blogs / twitter / Facebook /...
- How it is provided: method and availability of the data. For example, API, WS, files, databases, etc.
- Stored as: A brief description of the data format in which it is stored, plain text, XML, ontology, Linked Open Data.
- Amount: size of data.
- Language: Language in which the data is available.
- License: identifies whether the data is only available for the project purposes (PR) or it is also publicly available (PU). When applicable, the license in which the data is release is also listed.
- Web site URL: address of the web site which includes the documentation and information of the data source.

Data Entity	Type of data	How it is provided	Stored as	Amount	Language	License	Website
CoNLL 2002	Newswire articles made available by the Spanish EFE News Agency, May 2000	Source files available at CoNLL 2002	Plain text CoNLL format	369171 annotated tokens for dev/train/test	Spanish	Free for research purposes	http://www.clips.ua.ac.be/conll2002/ner/
CoNLL 2002	Newswire articles from Belgian newspaper "De Morgen" of 2000	Source files available at CoNLL 2002	Plain text CoNLL format	303450 annotated tokens for dev/train/test	Dutch	Free for research purposes	http://www.clips.ua.ac.be/conll2002/ner/
CoNLL 2003 datasets	Newswire from Reuters corpus	Annotations available at CoNLL 2003, need to access Reuters corpus at NIST to build the complete dataset.	Plain text CoNLL format.	301418 annotated tokens for dev/train/test	English	Free for research purposes.	http://www.clips.ua.ac.be/conll2003/ner/; Reuter corpus at http://trec.nist.gov/data/reuters/reuters.html
JRC Names	Analysis of hundreds of millions of news articles from the Europe Media Monitor since 2004 until 2011.	Recognized names available at http://langtech.jrc.it/JRC-Names.html	Database of lists of names	205,000 distinct known entities and its variants	20+ languages, including NewsReader languages	Free for research purposes. See license	http://langtech.jrc.it/JRC-Names.html
Ancora Corpus	Newswire, web text	Downloadable as files from http://clic.ub.edu/corpus/ancora	Sentences with semantic, syntactic and named entity annotations	500K words	Spanish	Public	http://clic.ub.edu/corpus/ancora
I-CAB	News text	Available at http://ontotext.fbk.eu/icab.html	Corpora with semantic annotation	180K words	Italian	Available for research purposes	http://ontotext.fbk.eu/icab.html

Table 1: Resources for Named Entity Recognition and Classification

4.1.1 CoNLL 2002 datasets

The CoNLL 2002 shared task was focused on language independent NERC based on machine learning techniques for person names, organizations, locations and miscellaneous names that do not belong to the previous three groups. The languages available for this task were Spanish and Dutch. The data consisted of two columns separated by a single space. The first item on each line is a word and the second the named entity tag. For example:

Wolff	B-PER
,	O
currently	O
a	O
journalist	O
in	O
Argentina	B-LOC
,	O
played	O
with	O
Del	B-PER
Bosque	I-PER
in	O
the	O
final	O
years	O
of	O
the	O
seventies	O
in	O
Real	B-ORG
Madrid	I-ORG
.	O

The Spanish data is a collection of news wire articles made available by the Spanish EFE News Agency from May 2000. The annotation was carried out by the TALP Research Center[44] of the Technical University of Catalonia (UPC) and the Center of Language and Computation (CLiC)[45] of the University of Barcelona (UB).

The Dutch data consist of four editions of the Belgian newspaper "De Morgen" of 2000 (June 2, July 1, August 1 and September 1). The data was annotated as a part of the Atranos[46] project at the University of Antwerp.

[44] http://www.talp.upc.es/
[45] http://clic.fil.ub.es/
[46] http://atranos.esat.kuleuven.ac.be/

4.1.2 CoNLL 2003 datasets

The shared task of CoNLL-2003[47] was also focused on language-independent named entity recognition for four types of named entities: **persons, locations, organizations and names of miscellaneous entities** that do not belong to the previous three groups. The participants of the shared task were offered training and test data for English and German and their objective was to build a NERC system based on machine learning techniques.

The data files consist of four columns separated by a single space. Each word is put on a separate line and there is an empty line after each sentence. The first item on each line is a word, the second a part-of-speech (POS) tag, the third a syntactic chunk tag and the fourth the named entity tag. The chunk tags and the named entity tags have the format I-TYPE which means that the word is inside a phrase of type TYPE. Only if two phrases of the same type immediately follow each other the first word of the second phrase will have tag B-TYPE to show that it starts a new phrase. A word with tag O is not part of a phrase. For example:

U.N.	NNP	I-NP	I-ORG
official	NN	I-NP	O
Ekeus	NNP	I-NP	I-PER
heads	VBZ	I-VP	O
for	IN	I-PP	O
Baghdad	NNP	I-NP	I-LOC
.	.	O	O

The English data is a collection of news wire articles from the Reuters Corpus[48]. Due to copyright issues only the annotations were made available at CoNLL and to build the complete datasets it is necessary to access the Reuters Corpus, which can be obtained from NIST for research purposes. The annotations for English and German were done by researchers at the University of Antwerp.

4.1.3 JRC Names

JRC-Names[49] is a highly multilingual named entity resource for person and organization names. It consists of large lists of names and their many spelling variants (up to hundreds for a single person), including across scripts (Latin, Greek, Arabic, Cyrillic, Japanese, Chinese, etc.). JRC Names contains the most important names of the EMM name database[50], namely, those names that were found frequently or that were verified manually or found on Wikipedia.

The first release of JRC Names (September 2011) contains the names of about 205,000 distinct known entities, plus about the same amount of variant spellings for these entities. Additionally, it contains a number of morphologically inflected variants of these names.

[47] `http://www.clips.ua.ac.be/conll2003/`
[48] `http://trec.nist.gov/data/reuters/reuters.html`
[49] `http://langtech.jrc.it/JRC-Names.html`
[50] `http://emm.newsexplorer.eu`

The **resource grows** by about 230 new entities and an additional 430 new name variants per week.

4.1.4 Ancora

AnCora consist of a Catalan corpus (AnCora-CA) and a Spanish corpus (AnCora-ES), each of them of 500,000 words. The following six named entity types are annotated: **Person, Organization, Location, Date, Numerical expression, and Others**.

4.1.5 Italian Content Annotation Bank (I-CAB)

I-CAB is an annotated corpus consisting of 525 news stories taken from the local newspaper "L'Adige", for a total of around 180,000 words. It is annotated with semantic information at different levels: temporal expressions, entities such as persons, organizations, locations; relations between entities such as the affiliation relation connecting a person to an organization. This annotation has been realized in conjunction with CELCT and the current version contains temporal expressions and entities.

I-CAB is accessible through the I-CAB Web Browser, a dedicated web interface. A version of the Ontotext portal for ICAB is also available. I-CAB is freely available for research purposes upon acceptance of a license agreement. It has been used in the following tasks at EVALITA:

- Entity Recognition at EVALITA 2009 (Local Entity Detection and Recognition and Named Entity Recognition subtasks)
- Temporal Expression Normalization and Recognition at EVALITA 2007
- Named Entity Recognition at EVALITA 2007

Web Site: `http://ontotext.fbk.eu/icab.html`

4.2 Tools

Table 2 lists the services and available downloadable systems and tools to perform NERC for the 4 languages relevant to NewsReader. The services and modules are also described in more detail. The meaning of the individual columns of Table 2 is as follows:

- System/Service: Name or identification of the Service or System (e.g., OpenCalais)
- Sources availability: Type of availability of the source code yes/no/partly
- How it is provided: The type of accessibility, namely, library, Web services, etc.
- Programming Language: The type of language used by the components: Java, C++, etc.

- License: The type of license i.e. GNU/GPL, Creative Commons licenses, proprietary, etc.
- Web site URL: address of the web site which includes the documentation and information of the service/system.

System/Service	Languages	Sources availability	How it is provided	Programming Language	License	URL
Open Calais	English, Spanish	No	Web service	Java, PHP, RDF	CC-SA	`http://www.opencalais.com`
Stanford CoreNLP	English	Yes	Library	Java	GNU GPLv2 or later	`http://nlp.stanford.edu/software/corenlp.shtml`
Freeling	English, Spanish	Yes	Library	C++, APIs also in Java, Perl, Python	GNU GPLv3	`http://nlp.lsi.upc.edu/freeling/`
Illinois Named Entity Tagger	English	Yes	Jar	Java	Research purposes	`http://cogcomp.cs.illinois.edu/page/download_view/NETagger`
OpenNLP	English, Spanish, Dutch	Yes	Library	Java	Apache license v.2	`http://opennlp.apache.org/`
TextPro	English, Italian	No	Executable binary	Java,C++	Free for research, propetary otherwise	`http://textpro.fbk.eu/`

Table 2: Tools for Named Entity Recognition and Classification

4.2.1 OpenCalais

The OpenCalais Web Service automatically creates rich semantic metadata for unstructured documents. Based on machine learning and other methods, it not only analyses the documents to find the entities, but it also provides with the facts and events hidden within the text.

Entities are things such as people, places, companies are geographies. Facts are relationships like *John Doe* is *the CEO of Acme Corporation*. Events are things that happened: *there was a natural disaster of type landslide in place Chula Vista*.

The web service is an API that accepts unstructured text (such as news articles, blog postings, etc.), processes them and returns RDF formatted entities, facts and events. It is possible to send four transactions per second and 50,000 per day free of cost, although commercial and service support is available. It is available for its use in commercial and non-commercial applications, the former at a cost. A number of Web applications using OpenCalais are listed in this URL: `http://www.opencalais.com/showcase`.

4.2.2 Stanford CoreNLP

Stanford CoreNLP includes a module for NERC. Stanford CoreNLP is a general NLP suite that provides a set of natural language analysis tools. The tools take raw English language as text input and they give, in a wide variety of output formats, different information: forms of words, parts of speech, named entities, normalize dates, times, and numeric quantities. The tools alos mark up the structure of sentences in terms of phrases and word dependencies, and indicate which noun phrases refer to the same entities. Stanford CoreNLP is an integrated framework that allows the analysis of a piece of text at different levels.

The Stanford CoreNLP code is written in Java and licensed under the GNU General Public License[51] (v2 or later). Source is included. It requires at least 4GB to run. The general suite is available for English. The Stanford NERC module for English includes a 4 class model trained for CoNLL, a 7-class model trained for MUC, and a 3-class model trained on both data sets for the intersection of those class sets.

4.2.3 Illinois Named Entity Tagger

This is a state of the art NER tagger [Ratinov and Roth, 2009] that tags plain text with named entities (people / organizations / locations / miscellaneous). It uses gazetteers extracted from Wikipedia, word class model derived from unlabeled text and expressive non-local features. The best performance is 90.8 F1 on the CoNLL03 shared task data for English. The software is licensed for academic purposes only.

4.2.4 Freeling

Freeling [Carreras *et al.*, 2004] is an open-source C++ library of language analyzers for building end-to-end NLP pipelines. The Freeling NERC module is based on their participation in the CoNLL shared tasks [Carreras *et al.*, 2003]. NERC is available in Freeling for English and Spanish. Freeling is licensed under the GPL. Each module requires about 2GB to run.

4.2.5 OpenNLP

OpenNLP is a general suite of NLP processing part of the Apache Software Foundation. The NERC module provides pre-trained models for English, Spanish and Dutch based on the CoNLL datasets. It is developed in Java and distributed under the Apache license v.2.

4.2.6 TextPro

TextPro is a flexible, customizable, integratable and easy-to-use NLP tool, which has a set of modules to process raw or customized text and perform NLP tasks such as: web page

[51] `http://www.gnu.org/licenses/gpl-2.0.html`

cleaning, tokenization, sentence detection, morphological analysis, pos-tagging, lemmatization, chunking and named-entity recognition. The current version, TextPro 2.0, supports English and Italian languages.

In TextPro there is the possibility to add dynamically new/customized processor, without affecting the flow of the pipeline. A Java interface class is available, which allows to deal with the input/output of the module. The “tab” format (table format) is used as interchange format between them. Each processor adds its specific information on a different column of the table. The IOB labelling format allows the system to annotate a span of token in a single column. All components are developed by researchers at FBK under a single license and ensuring more simplicity, modularity and portability. Distributions for Linux, Mac are available, for both research and commercial purposes. Also a web-service version of the system is availble.

The main modules of TextPro are:

1. HTML cleaner, CleanPro: it removes mark-up tags and irrelevant text (i.e. words used as navigation menu, common header and footer, etc.) from HTML pages.

2. Tokenizer, TokePro: it is a rule based splitter to tokenize raw text, using some predefined rules specific for each language and producing one token per line. TokenPro provides also:

   - UTF8 normalization of the token;
   - the char position of the token inside the input text;
   - sentence splitting.

3. Postagger, TagPro: it comes with two language models, Italian and English. The Italian model is trained on a corpus using a subset of the ELRA tagset. The English model is trained using the BNC tagset. TagPro processes the tokens to assign them their part of speech.

4. Morphological analyzer, MorphoPro: it processes the tokens to produce all the possible morphological analyses of a token. It has an Italian dictionary with 1,878,285 analyses for 149,372 lemmas, while there are 222,579 analyses for 78,721 English lemmas.

5. Lemmatizer, LemmaPro: it provides the lemma and the compatible morphological analysis of a token.

6. Named entity recognizer, EntityPro: it discovers the named entities in a text and classifies them. The available categories are person (PER), organization (ORG), geopolical entity (GPE) and location (LOC).

7. Chunker, ChunkPro: it assigns the Italian tokens to one of these 2 categories: NP (noun phrase) or VX (verb phrase). For English, there is a larger number of categories: ADJP (adjectival phrase), ADVP (adverbal phrase), CONJP (conjunction

phrase), INTJ (interjection), LST (list marker, includes surrounding punctuation), NP (noun phrase), PP (prepositional phrase), PRT (particle), B-SBAR (clause introduced by a, possibly empty, subordinating conjunction), UCP (unlike coordinated phrase), VP (verb phrase).

8. Keywords extractor, KX: it extracts the most important keywords from the document. For each keyword, it indicated its relevance and the number of occurrences in the text.

9. Recognizer of temporal expressions, TimePro: it identifies the tokens corresponding to temporal expressions in English and Italian and assignes them to one of the 4 Timex classes defined in ISO-TimeML.

5 Coreference Resolution

Coreference resolution is the task of linking noun phrases to the entities that they refer to. This problem has been widely studied in the literature. The first attempts to solve coreference were based on knwoledge and modeled and applied some lingusitic theories [Hobbs, 1977; Lappin and Leass, 1994; Grosz *et al.*, 1995], later approaches got some improvement applying machine learning and data mining techniques, both supervisded and unsupervised . However, recent works have recovered deterministic models with great success [Raghunathan *et al.*, 2010; Lee *et al.*, 2011].

Over the last fifteen years, various competitions have been run to promote research in the field of coreference resolution. The first competition of this kind was MUC, which in its sixth edition (MUC-6, 1995) added a coreference resolution task. The experiment was repeated in the seventh and final edition (MUC-7, 1997). Later, a coreference resolution task was added to ACE from 2002 to the most current competitions. After a few years without competition in this area, nowadays there is a new wave of interest thanks to the SemEval-2010[52] [Recasens *et al.*, 2010] and CoNLL-2011[53] [Pradhan *et al.*, 2011] tasks. These last two tasks incorporate all known measures (except ACE- value) and have much larger corpora. In addition, the corpora and participants' output can be downloaded for future comparison. On the one hand, the main goal of SemEval-2010 task on Coreference Resolution in Multiple Languages was to evaluate and compare automatic coreference resolution systems for six different languages (Catalan, Dutch, English, German, Italian, and Spanish). On the other hand, the coreference resolution task of CoNLL-2011 use the English language portion of the OntoNotes data, which consists of a little over one million words. The main goal was to automatically identify coreferring entities and events given predicted information on the other layers.

Automatic evaluation measures are crucial for coreference system development and comparison. Unfortunately, there is no agreement at present on a standard measure for

[52]`http://stel.ub.edu/semeval2010-coref`
[53]`http://conll.bbn.com`

coreference resolution evaluation. First, there are two metrics associated with international coreference resolution contests: the MUC scorer [Vilain *et al.*, 1995a] and the ACE value (Nist). Second, two commonly used measures, B3 [Bagga and Baldwin, 1998a] and CEAF [Luo, 2005a], are also used. Finally, an alternative metric called BLANC was presented [Recasens and Hovy, 2011]. B3 and CEAF are mention-based, whereas MUC and BLANC are link-based.

5.1 Data Sources

5.1.1 MUC

MUC The Message Understanding Conferences (MUC) were initiated in 1987 by DARPA [Grishman and Sundheim, 1996; Chinchor, 1998] as competitions in information extraction. The goal was to encourage the development of new and better methods for many tasks related to information extraction. Many research teams competed against one another, and coreference resolution was included in the competition in MUC-6 (1995) and MUC-7 (1997). Annotated corpora in English for coreference are copyrighted by the Linguistic Data Consortium[54]. MUC-6 used 30 text documents with 4381 mentions for training, and another 30 documents with 4,565 mentions for testing. MUC-7 consisted of 30 text documents with 5,270 mentions for training, and 20 documents with 3,558 mentions for testing.

5.1.2 ACE

ACE Automatic Content Extraction (ACE)[55] [Strassel *et al.*, 2008] is a program that supports the automatic processing of human language in text form (NIST, 2003). Promoted by the National Institute of Standards and Technology (NIST), it was originally devoted to the three source types of newswires, broadcast news (with text derived from ASR), and newspapers (with text derived from OCR). The most recent versions of ACE may have different source types. In addition, texts are available in Chinese, Arabic, and English. ACE annotations include information about the entities (for instance, their semantic class) and their relations that is used in other fields of information extraction. There are many ACE corpora, dating from 2002 until the present, and each one has a different size. The corpus is commonly divided into three parts according to documents of diverse nature: Broadcast News (bnews), Newspaper (npaper), and Newswire (nwire). Each of these parts is further divided into training and development/test sets. Documents in npaper are, on average, larger than the others. While an npaper document has between 200 and 300 mentions, a document in bnews or nwire has about 100 mentions.

The main differences between MUC and ACE are found in three different levels: syntactic, semantic, and task understanding, and are described as follows [Stoyanov *et al.*, 2010]. First, at the syntactic level, the MUC annotated mentions do not include nested named

[54]`http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2001T02`
[55]`http://www.nist.gov/speech/tests/ace`

entities, such as "Washington" in the named entity "University of Washington," relative pronouns, and gerunds, but do allow nested nouns. On the contrary, ACE annotations include gerunds and relative pronouns, but exclude nested nouns that are not themselves NPs, and allow some geopolitical nested named entities such as "U.S." in "U.S. officials." Second, ACE restricts mentions to a limited set of semantic classes: person, organization, geopolitical, location, facility, vehicle, and weapon. MUC has no limitations on entity semantic classes.And third, MUC does not include singletons. A singleton is a mention not coreferring to any other in the document. For instance, the named entity "San Sebastián" in a document is annotated as a mention only if there is another mention referring to the same city, such as another occurrence of "San Sebastián" or "the city."

5.1.3 OntoNotes

The **OntoNotes** project has created a corpus of large-scale, accurate, and integrated annotations of multiple levels of the shallow semantic structure in text. The idea is that this rich, integrated annotation covering many linguistic layers will allow for richer, cross-layer models enabling significantly better automatic semantic analysis. In addition to coreferences, this data is also tagged with syntactic trees, high-coverage verbs, and some noun propositions, verb and noun word senses, and 18 named entity types [Pradhan *et al.*, 2007b]. Moreover, OntoNotes 2.0 was used in SemEval Task 1 [Recasens *et al.*, 2010] and OntoNotes 4.0 (the fourth version of annotations) has been used in the CoNLL 2011 shared task on coreference resolution [Pradhan *et al.*, 2011].

The English corpora annotated with all the layers contains about 1.3M words. It comprises 450,000 words from newswires, 150,000 from magazine articles, 200,000 from broadcast news, 200,000 from broadcast conversations, and 200,000 web data. Note that this corpus is considerably larger than MUC and ACE.

5.1.4 AnCora-Co

AnCora-CO [Recasens and Martí, 2010] is a corpus in Catalan and Spanish that contains coreference annotations of entities composed of pronouns and full noun phrases (including named entities), plus several annotation layers of syntactic and semantic information: lemmas, parts-of-speech, morphological features, dependency parsing, named entities, predicates, and semantic roles. Most of these annotation layers are dually provided as gold standard and predicted, namely, manually annotated versus predicted by automatic linguistic analyzers. The coreference annotation also includes singletons. AnCora-CO was used in SemEval Shared Task 1: Coreference resolution in multiple languages [Recasens *et al.*, 2010]. The size of AnCora-CO is about 350,000 words of Catalan and a similar quantity in Spanish.

Data Entity	Type of data	How it is provided	Stored as	Amount	Language	License	Website
MUC [Grishman and Sundheim, 1996; Chinchor, 1998]	Newswire		50 documents with 9K mentions		English	Linguistic Data Consortium	`http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2001T02`
ACE [Strassel *et al.*, 2008]	Broadcast News, Newspaper, and Newswire	Availabe upon request `http://projects.ldc.upenn.edu/ace/data/`	Corpus	English: 260K words; Chinese: 205K words; Arabic: 100K words	Chinese, Arabic, and English	Linguistic Data Consortium	`http://www.itl.nist.gov/iad/mig//tests/ace/`
OntoNotes [Pradhan *et al.*, 2007b]	Newswire and web text	Available at `http://www.ldc.upenn.edu/Catalog/catalogEntry.jsp?catalogId=LDC2011T03`	Treebank sentences with named entity and coreference information	1M words	English	Private	`http://www.ldc.upenn.edu/Catalog/catalogEntry.jsp?catalogId=LDC2011T03`
AnCora-CO [Recasens and Martí, 2010]	Newswire, web text	Downloadable as files from `http://http://clic.ub.edu/corpus/ancora`	Sentences with semantic, syntactic and named entity annotations	500K words	Spanish	Public	`http://http://clic.ub.edu/corpus/ancora`

Table 3: Resources for Coreference resolution

5.2 Tools

5.2.1 GUITAR

GUITAR[56] [Steinberger *et al.*, 2007], is a freely available tool designed to be modular and usable as an off-the-shelf component of a NLP pipeline. The system resolves pronouns, definite descriptions and proper nouns in coreference chains.

The anaphora resolution proper part of guitar is designed to take XML input, in a special format called MAS-XML, and produce an output in the same format, but which additionally contains anaphoric annotation. The system can therefore work with a variety of pre-processing methods, ranging from a simple part-of-speech tagger to a chunker to a full parser, provided that appropriate conversion routines into MAS-XML are implemented. The version used for these experiments uses Charniak's parser [Charniak, 2000].

The latest version includes an implementation of the MARS pronoun resolution algorithm [Mitkov *et al.*, 2002] to resolve personal and possessive pronouns. This system resolves definite descriptions using a partial implementation of the algorithm proposed in [Vieira and Poesio, 2000], augmented with a statistical discourse new classifier. Finally, it also includes an implementation of the shallow algorithm for resolving coreference with proper names proposed by [Bontcheva *et al.*, 2002]. The evaluation of GUITAR has been carried out on the GNOME corpus[57], consisting of a variety of texts from different domains– and 37 texts from the CAST corpus[58] [Orăsan *et al.*, 2003] consisting of news articles, mostly from the Reuters corpus. [Steinberger *et al.*, 2007] report a precision of 70.2, a recall of 72.5 and an F1 of 71.3 . On the CAST corpus the results were much modest: precision of 55.2, recall of 45.8 and F1 of 50.1.

5.2.2 BART

The BART[59] toolkit [Versley *et al.*, 2008] has been developed as a tool to explore the integration of knowledge-rich features into a coreference system at the Johns Hopkins Summer Workshop 2007. It is based on code and ideas from the system of [Ponzetto and Strube, 2006b], but also includes some ideas from GUITAR [Steinberger *et al.*, 2007] and other coreference systems. BART is a modular toolkit for coreference resolution that supports state-of-the-art statistical approaches to the task and enables efficient feature engineering. BART has originally been created and tested for English, but its flexible modular architecture ensures its portability to other languages and domains. Given a corpus in a new language, one can re-train BART to obtain baseline results. Such a language-agnostic system, however, is only used as a starting point: substantial improvements can be achieved by incorporating language-specific information with the help of the Language Plugin. This design provides effective separation between linguistic and machine learning aspects of the problem. The BART toolkit has five main components: pre-processing pipeline, mention

[56] http://cswww.essex.ac.uk/Research/nle/GuiTAR/gtarNew.html
[57] http://cswww.essex.ac.uk/Research/nle/corpora/GNOME
[58] http://clg.wlv.ac.uk/projects/CAST/corpus/index.php
[59] http://www.bart-coref.org/

factory, feature extraction module, decoder and encoder. In addition, an independent LanguagePlugin module handles all the language specific information and is accessible from any component. The pre-processing pipeline converts an input document into a set of linguistic layers, represented as separate XML files. The mention factory uses these layers to extract mentions and assign their basic properties (number, gender etc). The feature extraction module describes pairs of mentions as a set of features. The decoder generates training examples through a process of sample selection and learns a pairwise classifier. Finally, the encoder generates testing examples through a (possibly distinct) process of sample selection, runs the classifier and partitions the mentions into coreference chains. The BART toolkit supports several models of coreference (pairwise modeling, rankers, semantic trees), as well as different machine learning algorithms. In SemEval-2010 Task 1 on Coreference Resolution, BART shown reliable performance for English, German and Italian.

5.2.3 Illinois Coreference Package

This Illinois Package[60] contains a Coreference Resolver, along with a collection of coreference related features [Bengtson and Roth, 2008]. The system presents a rather simple pairwise classification model for coreference resolution, developed with a well-designed set of features. These features include gender and number match, WordNet relations including synonym, hypernym, and antonym, and ACE entity types (e.g. semantic classes such as person, organization, and geopolitical entity). These features also include an anaphoricity classifier trained using machine learning techniques. This collection of features is a key ingredient in the performance of the included coreference classifier. To train the coreference classifier, an annotated training data such as the LDC's ACE 2004 corpus is needed. Both the source files and a compiled and trained jar distribution of the Illinois coreference system can be downloaded.

5.2.4 ARKref

ARKref[61] is a Noun Phrase Coreference System. It is a Java implementation of a syntactically rich, rule-based within-document coreference system very similar to the syntactic components of [Haghighi and Klein, 2009]. It is useful as a starting point for incorporating coreference into larger information extraction and natural language processing systems. For example, by tweaking the gazetteers, customzing mention identification, turning the syntactic rules into log-linear features, etc. It performs about as well as [Haghighi and Klein, 2009] system on the development data set (they do not provide evaluation results on the test dataset). Its F-score is slightly higher, and the precision/recall tradeoff is different. Note that there is no semantic compatibility subsystem ("+SEM-COMPAT") and that they use the supersense tagger [Ciaramita and Altun, 2006a] rather than a named

[60] `http://cogcomp.cs.illinois.edu/page/software_view/18`
[61] `http://www.ark.cs.cmu.edu/ARKref`

entity recognizer. It depends on having a phrase structure parser. They use the Stanford Parser and include it in the download package. ARKref also makes heavy use of the Stanford Tregex library for implementation of syntactic rules.

5.2.5 Reconcile

Reconcile[62] [Stoyanov *et al.*, 2010] is an automatic coreference resolution system that was developed to provide a stable test-bed for researchers to implement new ideas quickly and reliably. It achieves roughly state of the art performance on many of the most common coreference resolution test sets, such as MUC-6, MUC-7, and ACE. Reconcile comes ready out of the box to train and test on these common data sets (though the data sets are not provided) as well as the ability to run on unlabeled texts. Reconcile utilizes supervised machine learning classifiers from the Weka toolkit, as well as other language processing tools such as the Berkeley Parser and Stanford Named Entity Recognition system. The source language is Java, and it is freely available under the GPL.

5.2.6 MARS

MARS[63][64] (Mitkov's Anaphora Resolution System) [Mitkov *et al.*, 2002] indicates the antecedent of each 3rd person NP-anaphoric pronoun. A table is printed under each pronoun, listing all candidates considered as its potential antecedents. The weights assigned to each candidate by different salience factors are also printed. A full description of salience factors and their weights appears in [Mitkov *et al.*, 2002]. MARS uses the Connexor FDG Parser to perform syntactic analysis.

5.2.7 CherryPicker

CherryPicker[65] [Rahman and Ng, 2009] is a coreference resolution tool that implements a cluster-ranking model as well as two existing learning-based coreference models (the mention-pair model and the mention-ranking model). Cluster rankers aim to address the major weaknesses of the widely-investigated mention-pair model.

All coreference models included in CherryPicker employ linguistic features that are largely motivated by those described in [Ng and Cardie, 2002a], and were trained using SVMlight on the English portion of the ACE 2005 multilingual training corpus. Since ACE 2005 restricts coreference to noun phrases that belong to one of seven semantic classes (namely, person, organization, GPE (geo-political entity), facility, location, vehicle, and weapon), the resulting coreference models will generate coreference chains only for noun phrases belonging to these semantic classes.

CherryPicker also includes a mention detector that was trained using CRF++ on the same training data to identify noun phrases that belong to these seven semantic classes,

[62] `http://www.cs.utah.edu/nlp/reconcile`
[63] `http://clg.wlv.ac.uk/demos/MARS/index.php`
[64] `http://clg.wlv.ac.uk/demos/MARS/mars2.tar.gz`
[65] `http://www.hlt.utdallas.edu/~altaf/cherrypicker.html`

so there is no need for the user to provide noun phrases as input. For feature generation, CherryPicker relies on the following NLP tools:

1. The Stanford Log-linear Part-Of-Speech Tagger
2. The Stanford Named Entity Recognizer (NER)
3. The Charniak Statistical Syntactic Parser
4. The MINIPAR Parser

All these software tools, as well as SVMlight and CRF++, are included as part of our software package. CherryPicker only assumes as input a text that is sentence-delimited, with one sentence per line, and produces coreference chains in the MUC format.

CherryPicker may be freely downloaded and used for all educational and research activities, but may not be used for commerical or for-profit purposes.

The current version of CherryPicker has only been tested on Unix/Linux machines. Since some of the software tools on which it relies run on Unix/Linux machines only, we do not expect CherryPicker to be able to run on other platforms.

5.2.8 Stanford CoreNLP

The Stanford coreference resolution system is a module integrated into the Stanford CoreNLP.[66] Stanford CoreNLP is an integrated framework that allows the analysis of a piece of text at different levels. The Stanford CoreNLP code is written in Java and licensed under the GNU General Public License (v2 or later). Source is included. Note that this is the full GPL, which allows many free uses, but not its use in distributed proprietary software. The download is 259 MB and requires Java 1.6+.

The Stanford multi-pass sieve coreference resolution (or anaphora resolution) system is described in [Lee *et al.*, 2011] and [Raghunathan *et al.*, 2010]. The approach applies tiers of coreference models one at a time from highest to lowest precision. Each tier builds on the entity clusters constructed by previous models in the sieve, guaranteeing that stronger features are given precedence over weaker ones. Furthermore, each model's decisions are richly informed by sharing attributes across the mentions clustered in earlier tiers. This ensures that each decision uses all of the information available at the time. They implemented all components using only deterministic models. All these components are unsupervised, in the sense that they do not require training on gold coreference links. Furthermore, this framework can be easily extended with arbitrary models, including statistical or supervised models.

This system was the top ranked system at the CoNLL-2011 shared task. The score is higher than that in EMNLP 2010 paper because of additional sieves and better rules (see [Lee *et al.*, 2011] for details). Mention detection is included in the package.

[66] `http://nlp.stanford.edu/software/corenlp.shtml`

System/Service	Languages	Sources availability	How it is provided	Programming Language	License	URL
GUITAR	English	YES	jar	Java	GPL	`http://cswww.essex.ac.uk/Research/nle/GuiTAR/gtarNew.html`
BART	English,Italian	YES	jar	Java	Apache v2.0, GPL	`http://www.bart-coref.org`
Ilinois	English	YES	jar	Java	Research only	`http://cogcomp.cs.illinois.edu/page/software_view/18`
ARKref	English	YES	jar	Java	GPL	`http://www.ark.cs.cmu.edu/ARKref`
Reconcile	English	YES	jar	Java	GPL	`http://www.cs.utah.edu/nlp/reconcile`
MARS	English	YES	tar	Perl, C++		`http://clg.wlv.ac.uk/demos/MARS/index.php`
CherryPicker	English	YES	jar	Java	Research only	`http://www.hlt.utdallas.edu/~altaf/cherrypicker.html`
Stanford	English	YES	jar	Java	GPL	`http://nlp.stanford.edu/software/corenlp.shtml`
RelaxCor	English, Spanish	YES	tar	Perl, C++	GPL	`http://nlp.lsi.upc.edu/relaxcor`
JavaRAP	English	YES	jar	Java	GPL (but contact developer)	`http://wing.comp.nus.edu.sg/~qiu/NLPTools/JavaRAP.html`

Table 4: Tools for Coreference resolution

5.2.9 RelaxCor

Relaxcor[67] [Sapena *et al.*, 2011] is a coreference resolution system based on constraint satisfaction. It represents the problem as a graph connecting any pair of candidate coreferent mentions and applies relaxation labeling, over a set of constraints, to decide the set of most compatible coreference relations. Decisions are taken considering the entire set of mentions, which ensures consistency and avoids local classification decisions.

The Relaxcor implementation is 90% Perl and 10% C++. The performances of Relaxcor are in the state-of-the-art, achieving the second position at CONLL-2011 Shared Task [Pradhan *et al.*, 2011].

The main advantages of using Relaxcor are the language adaptation and the possibility to incorporate handwritten constraints, or constraints acquired from other sources. Regarding the languages, Relaxcor is ready to work on English, Spanish and Catalan, and apparently, the incorporation of new languages requires minimal changes in the software.

5.2.10 JavaRAP

JavaRAP[68] [Qiu *et al.*, 2004] is an implementation of the classic Resolution of Anaphora Procedure (RAP) given by (Lappin and Leass 1994). It process English texts and resolves third person pronouns, lexical anaphors, and identifies pleonastic pronouns. It is written in Java and requires the Charniak parser. Evaluation on the MUC-6 coreference task shows that JavaRAP has an accuracy of 57.9%.

6 Named Entity Disambiguation

As explained in section 4, Named Entity Recognition and Classification (NERC) deals with the detection and identification of specific entities in running text. Current state-of-the-art processors achieve high performance in recognition and classification of general categories such as people, places, dates or organisations [Nadeau and Sekine, 2007].

Once the named entities are recognized they can be identified with respect to an existing catalogue. Wikipedia has become the de facto standard as such a named entity catalogue. Wikification [Mihalcea and Csomai, 2007] is the process of automatic linking of the named entities occurring in free text to their corresponding Wikipedia articles. This task is typically regarded as a Word Sense Disambiguation (WSD) problem [Agirre and Edmonds, 2006], where Wikipedia provides both the dictionary and training examples. Public demos of systems which exploit Wikification (only for English) are Spotlight[69], CiceroLite from

[67] `http://nlp.lsi.upc.edu/relaxcor`
[68] `http://wing.comp.nus.edu.sg/~qiu/NLPTools/JavaRAP.html`
[69] `http://spotlight.dbpedia.org/demo/index.html`

LCC[70] and, Zemanta[71], TAGME[72] or The Wiki Machine[73].

Automatic text wikification implies solutions for named entity disambiguation [Mihalcea and Csomai, 2007]. For unambiguous terms it is not a problem, but in other cases words sense disambiguation must be performed.

For example, the Wikipedia disambiguation page lists many different articles that the term **BMW** might refer to (the German manufacturer Bayerische Motoren Werke AG, a Jamaican reggae band[74]).

The following sentence provides an example of BMW with the corresponding Wikipedia links:

> **BMW**[75] produces motorcycles under **BMW Motorrad**[76]. In 2010, the BMW group produced 1,481,253 **automobiles**[77] and 112,271 motorcycles across all its brands.

The named entity ambiguity problem has been formulated in two different ways. Within computational linguistics, the problem was first conceptualised as an extension of the coreference resolution problem [Bagga and Baldwin, 1998c]. The Wikification approach later used Wikipedia as a word sense disambiguation data set by attempting to reproduce the links between pages, as linked text is often ambiguous [Mihalcea and Csomai, 2007]. Finally, using Wikipedia as in the Wikification approach, NERC was included as a preprocessing step and a link or NIL was required for all identified mentions [Bunescu and Pasca, 2006]. This means that, as opposed to Wikification, links are provided only for named entities. The resulting terminology of these various approaches is cross-document coreference resolution (CDCR), Wikification, and Named Entity Linking (NEL). The term Named Entity Disambiguation (NED) will be used to refer to any of these three tasks indistinctly [Hachey *et al.*, 2013].

NewsReader will extract the appropriate semantic knowledge and properties concerning the named entities of interest. The same approach can be extended to languages other than English. Current performance rates can be improved by focusing on the named entities only, thus, avoiding the annotation of the remainder of the text. In a multilingual setting, once in a language-neutral representation, the knowledge captured for a particular NE in one language can be ported to another, balancing resources and technological advances across languages [Steinberger and Pouliquen, 2007].

This section describes the relevant data sources and tools for Named Entity Disambiguation (NED). The data sources are mainly either text corpora developed for NLP

[70] http://demo.languagecomputer.com/cicerolite/
[71] http://www.zemanta.com
[72] http://tagme.di.unipi.it/
[73] http://thewikimachine.fbk.eu/html/index.html
[74] http://en.wikipedia.org/wiki/BMW_(disambiguation)
[75] http://en.wikipedia.org/wiki/BMW
[76] http://en.wikipedia.org/wiki/BMW_Motorrad
[77] http://en.wikipedia.org/wiki/Automobiles

applications or Linked Data as part of the Linked Data[78] initiative. Most of the research on NED systems has been undertaken on text corpora, although, as we will see in section 6.2, some systems are already using Linked Data datasets such as DBpedia[79].

6.1 Data Sources

The data sources and systems described in this section will be those relevant to cross-document coreference resolution (CDCR), Wikification, and Named Entity Linking (NEL). The term Named Entity Disambiguation (NED) will be used to refer to any of these three tasks indistinctly [Hachey *et al.*, 2013].

Most CDCR datasets are collected by searching a set of canonical entity names, ignoring non-canonical coreferent forms, as it is shown by the datasets collected by the Web People Search WePS shared evaluation tasks [Mann and Yarowsky, 2003; Artiles *et al.*, 2007; Artiles *et al.*, 2009; Artiles *et al.*, 2010].

With the rise to prominence of Wikipedia, the Wikification task was sorted [Mihalcea and Csomai, 2007]. Instead of clustering entities, as in CDCR, mentions of important concepts in the text were to be linked to its corresponding Wikipedia article. Crucially, the Wikification task differs from Named Entity Linking (NEL) in that the concepts to be disambiguated are not necessarily named entities and in assuming that the knowledge base is complete.

The first large datasets on NEL were created by the Text Analysis Conference (TAC) for the Knowledge Base Population (KBP) track. The goal of KBP is to promote research in automated systems that discover information about named entities as found in a large corpus and incorporate this information into a knowledge base. TAC 2013 fields tasks in three areas, all aimed at improving the ability to automatically populate knowledge bases from text. For our purposes the Entity-Linking task is the most relevant:

> "The entity linking task is to link name mentions of entities in a document collection to entities in a reference KB, or to new named entities discovered in the collection. The document collection will comprise a combination of newswire articles and posts to blogs, newsgroups, and discussion fora. Given a query that consists of a document with a specified name mention of an entity, the task is to determine the correct node in the reference KB for the entity, adding a new node for the entity if it is not already in the reference KB. Entities can be of type PER (person), ORG (organization), or GPE (geopolitical entity). In addition to monolingual English entity linking, cross-lingual entity linking tasks will be offered in Chinese (Chinese and English documents, English reference KB) and Spanish (Spanish and English documents, English reference KB)".[80]

So far there have been 5 editions since 2009. Originally the datasets sorted were only for English but the 2012 and 2013 editions include documents in Spanish. In addition to

[78] `http://linkeddata.org/`
[79] `http://dbpedia.org`
[80] `http://www.nist.gov/tac/2013/KBP/EntityLinking/index.html`

the KBP datasets, several others have been created [Cucerzan, 2007; Fader *et al.*, 2009]. Furthermore, there is some work on integrating NEL annotation with existing NERC datasets such as the CONLL 2003 datasets [Hoffart *et al.*, 2011].

Other valuable datasets listed in table 5 for NED are those related with Linked Data. Linked Data is defined as "about using the Web to connect related data that wasn't previously linked, or using the Web to lower the barriers to linking data currently linked using other methods". More specifically, Wikipedia defines Linked Data as "a term used to describe a recommended best practice for exposing, sharing, and connecting pieces of data, information, and knowledge on the Semantic Web using URIs and RDF." Of course, the data to be linked can consist of any type of named entity currently available in the Web. Well known and large linked data resources in the NLP community are DBpedia, Freebase[81] and Yago[82], but there are many others including those supported by large organizations such as the BBC, the British Government, NASA, CIA, Yahoo, etc. Current count in the list of Linked Data datasets is more than 300.

Data Entity	**Type of data**	**How it is provided**	**Stored as**	**Amount**	**Language**	**License**	**Website**
Cucerzan 2007	Newswire	Text corpora	Source documents and gold standard for evaluation	756 surface forms of entities	English	Public	`http://research.microsoft.com/en-us/um/people/silviu/WebAssistant/TestData/`
KBP 2009	Newswire	Text corpora available from Linguistic Data Consortium (LDC)	Annotated files for development and evaluation	3904 instances	English	Private	`http://apl.jhu.edu/~paulmac/kbp.html`
KBP 2010	News, Blogs, Web data	Datasets available from LDC	Annotated files for development and evaluation	3750 instances	English	Private	LDC
KBP 2011	News, Web data	Datasets available from LDC	Annotated files for development and evaluation	6000 instances for development, training and evaluation	English	Private	LDC
KBP 2012	News, Web data	Datasets available from LDC	Annotated files for development and evaluation		English, Spanish	Private	LDC
KBP 2013	News, Web data	TBA	TBA	TBA	English, Spanish	TBA	TBA
Fader 2009	News	Datasets available on request to the author	Annotated files evaluation	500 instances for evaluation	English		`http://www.cs.washington.edu/homes/afader/`

[81] `http://www.freebase.com`

[82] `http://www.mpi-inf.mpg.de/yago-naga/yago/`

Dredze 2010	News	Available on request to the author	Annotated files for training	1496 instances	English	Private	http://www.cs.jhu.edu/~mdredze/
ACEtoWiki	News, Web, Transcripts	Available as text corpora, distributed by LDC	Annotated files with truth links for evaluation	16851 instances	English	Free for research purposes during duration of project	http://www.celct.it/resources.php?id_page=acewiki2010
AIDA CoNLL YAGO	Newswire	Available as text corpora	Annotated files	34596 annotated mentions	English	CC-BY 3.0 license, PU	http://www.mpi-inf.mpg.de/yago-naga/aida/downloads.html
Illinois Wikifier Data	Wikipedia, new	Text corpora	Annotated files	928 annotated instances	English	Public	http://cogcomp.cs.illinois.edu/page/resources/data
Wikipedia Miner	Wikipedia, news	Text corpora	Annotated files	727 annotated instances	English	Public	http://www.nzdl.org/wikification
Dbpedia	Wikipedia articles	API, dump	Linked Data	3.77 million named entities	Multilingual, including English, Spanish, Dutch, Italian	CC-BY-SA license	http://dbpedia.org
Freebase	Web pages	API, dump	Linked Data	23 million of named entities	Multilingual, including English, Spanish, Dutch, Italian	CC-BY 3.0 license, PU	http://www.freebase.com
YAGO2	Web pages, Wikipedia	API, dump	Linked Data	10 million of named entities	Multilingual, including English, Spanish, Dutch, Italian		http://www.mpi-inf.mpg.de/yago-naga/yago/
GeoNames	Web	Web services, dump, premium dump	Linked Data	8 million geographic entities	Multilingual, including English, Spanish, Dutch, Italian	CC-BY 3.0 license, PU	http://www.geonames.org/
LinkedGeoData	Web	Web service, API, dump	Linked Data	6 million location instances	Multilingual, including English, Spanish, Dutch, Italian	CC-BY-SA license	http://linkedgeodata.org

Table 5: Resources for Named Entity Disambiguation

6.1.1 KBP at TAC

The TAC KBP 2009 edition distributed **a knowledge base extracted from a 2008 dump of Wikipedia and a test set of 3,904 queries**. Each query consists of an ID that identified a document within a set of Reuters news articles, a mention string that occurrs at least once within that document, and a node ID within the knowledge base. Each knowledge base node contains the Wikipedia article title, Wikipedia article text, a predicted entity type (person, organization, location or misc), and a key-value list of information extracted from the article's infobox. Only articles with infoboxes that are predicted to correspond to a named entity are included in the knowledge base. The annotators favour mentions that are likely to be ambiguous, in order to provide a more challenging evaluation. If the entity referred to does not occur in the knowledge base, it is labelled NIL. A high percentage of queries in the 2009 test set does not map to any nodes in the knowledge base: the gold standard answer for 2,229 of the 3,904 queries is NIL.

In the 2010 challenge the same configuration as the 2009 challenge is used with the same knowledge base. In this edition, however, a training set of 1,500 queries is provided, with a test set of 2,250 queries. In the 2010 training set, only 28.4% of the queries are NIL, compared to 57.1% in the 2009 test data and 54.6% in the 2010 test data. This mismatch between the training and test data show the importance of the NIL queries and it is argued that it may have harmed performance for some systems because it can be quite difficult to determine whether a candidate that seems to weakly match the query should be discarded, in favour of guessing NIL. The most successful strategy to deal with these issue in the 2009 challenge is augmenting the knowledge base with extra articles from a recent Wikipedia dump. If a strong match against articles that do not have any corresponding node in the knowledge base is obtained, then NIL is return for these matches.

In the KBP 2012 and 2013 editions, the reference KB is derived from English Wikipedia, while source documents come from a variety of languages, including English, Chinese, and Spanish.

6.1.2 Cucerzan 2007

Cucerzan [Cucerzan, 2007] manually linked all entities from 20 MSNBC news articles to a 2006 Wikipedia dump, for a total of 756 links, with 127 resolving to NIL. This data set is particularly interesting because mentions were linked exhaustively over articles, unlike the KBP data, where mentions were selected for annotation if the annotators regarded them as interesting. The Cucerzan dataset thus gives a better indication of how a real-world system might perform.

6.1.3 Fader 2009

The authors evaluated their NED system against 500 predicate-argument relations extracted by TextRunner from a corpus of 500 million Web pages, covering various topics and genres. Considering only relations where one argument was a proper noun, the authors manually identified the Wikipedia page corresponding to the first argument, assigning NIL if there is no corresponding page. Overall, 160 of the 500 mentions resolved to NIL [Fader *et al.*, 2009].

6.1.4 Dredze 2010

In order to general additional training data, the authors performed manual annotation using a similar methodology to the KBP challenges. They linked 1,496 mentions from news text to the KBP knowledge base, of which 270 resolved to NIL [Dredze *et al.*, 2010]. As it can be noted, this is a substantially lower percentage of NIL linked queries than the 2009 and 2010 KBP datasets.

6.1.5 ACEtoWIKI

ACEtoWIKI is the result of a joint effort between FBK[83] and CELCT[84]. The resource has been created by adding a manual annotation layer connecting the English ACE-2005 Corpus to Wikipedia.

[83] `http://www.fbk.eu/`
[84] `http://www.celct.it`

ACEtoWiki has been produced by manually annotating the non-pronominal mentions, namely, the named (NAM) and nominal (NOM) mentions contained in the English ACE 2005 corpus with links to appropriate Wikipedia articles.

Each mention of type NAM is annotated with a link to a Wikipedia page describing the referred entity. For instance, "George Bush" is annotated with a link to the Wikipedia page *George_W._Bush*. NOM mentions are annotated with a link to the Wikipedia page which provides a description of its appropriate sense. Note that the object of linking is the textual description of an entity, and not the entity itself.

Moreover, mentions of type NOM can often be linked to more than one Wikipedia page. In such cases, links are sorted in order of relevance, where the first link corresponds to the most specific sense for that term in its context. For instance, for the NOM mention "President" which in the context identifies the United States President George Bush the following links are selected as appropriate: *President_of_the_ United_States* and *President*.

6.1.6 AIDA CoNLL Yago

This corpus contains assignments of entities to the mentions of named entities annotated for the original CoNLL 2003 entity recognition task. The entities are identified by YAGO2 entity name, by Wikipedia URL, or by Freebase mid[85]. The CoNLL 2003 dataset is required to create the corpus.

6.1.7 Illinois Wikifier Datasets

These datasets were created for the evaluation of the paper from which originated the Illinois Wikifier system [Ratinov *et al.*, 2011].

[Ratinov *et al.*, 2011] constructed two data sets. The first is a subset of the ACE coreference data set, which has the advantage that mentions and their types are given, and the coreference is resolved. Using Amazon's Mechanical Turk annotators linked the first nominal mention of each coreference chain to Wikipedia, if possible. Finding the accuracy of a majority vote of these annotations to be approximately 85%, they manually corrected the annotations to obtain ground truth for their experiments.

The second data set is a sample of paragraphs from Wikipedia pages. Mentions in this data set correspond to existing hyperlinks in the Wikipedia text. Because Wikipedia editors explicitly link mentions to Wikipedia pages, their anchor text tends to match the title of the linked-to page. As a result, in the overwhelming majority of got correct serve as positive examples, the disambiguation task is trivial. The ACE based corpus contains 257 mentions whereas the Wikipedia-based consists of 928 mentions.

6.1.8 Wikipedia Miner

The Wikipedia Miner system was mainly tested on Wikipedia articles, by taking the links out and trying to put them back in automatically. In addition, the system was also tested

[85]http://wiki.freebase.com/wiki/Machine_ID

on news stories from the AQUAINT corpus, to see if it would work as well "in the wild" as it did on Wikipedia. The stories were automatically wikified, and then inspected by human evaluators. This dataset contains the news stories of the AQUAINT corpus.

6.1.9 DBpedia

Dbpedia is the Linked Data version of Wikipedia. The DBpedia data set currently provides information about more than 1.95 million "things", including at least 80,000 persons, 70,000 places, 35,000 music albums, 12,000 films classified in a consistent ontology. In total it contains almost 4 million entities. It also provides descriptions in 12 different languages. Altogether, the DBpedia data set consists of (more than) 103 million RDF triples.

The data set is interlinked with many other data sources from various domains (life sciences, media, geographic government, publications, etc.), including the aforementioned Freebase and YAGO, among many others[86].

6.1.10 Freebase

Freebase has information about approximately 20 million topics or entities. Each one of them has a unique identifier, which can help distinguish multiple entities which have similar names (named entity synonymy) such as 'Henry Ford', which can refer to the industrialist or the footballer (e.g,. see `http://en.wikipedia.org/wiki/Henry_Ford_disambiguation`).

Most of their topics are associated with one or more named entity type (such as people, places, books, films, etc) and may have additional properties like "date of birth" for a person or latitude and longitude for a location. Freebase is created using information from many other Web pages[87].

6.1.11 YAGO2

YAGO2 is a large semantic knowledge base, derived from Wikipedia, WordNet and GeoNames. Currently, YAGO2 has knowledge of more than 10 million entities (like persons, organizations, cities, etc.) and contains more than 120 million facts about these entities. The accuracy of YAGO2 has been manually evaluated, claiming an accuracy of 95%. Every relation is annotated with its confidence value. YAGO2 is an ontology that is anchored in time and space. YAGO2 attaches a temporal dimension and a spacial dimension to many of its facts and entities. YAGO2 is particularly suited for disambiguation purposes, as it contains a large number of names for entities. It also knows the gender of people. YAGO2 is part of the Linked Data cloud and is directly linked to DBpedia.

6.1.12 GeoNames

GeoNames contains over 10 million geographical names and consists of over 8 million unique features whereof 2.8 million populated places and 5.5 million alternate names. All

[86] `http://wiki.dbpedia.org/Datasets`
[87] `http://wiki.freebase.com/wiki/Freebase_data`

features are categorized into one out of nine feature classes and further subcategorized into one out of 645 feature codes. GeoNames is integrating geographical data such as names of places in various languages, elevation, population and others from various sources. All lat/long coordinates are in WGS84 (World Geodetic System1984). The data is accessible free of charge through a number of Web services and a daily database export. GeoNames is serving up to over 30 million web service requests per day.

6.1.13 LinkedGeoData

LinkedGeoData uses the comprehensive OpenStreetMap[88] spatial data collection to create a large spatial knowledge base. It consists of more than 1 billion nodes and 100 million ways and the resulting RDF data comprises approximately 20 billion triples. The data is available according to the Linked Data principles and interlinked with DBpedia and GeoNames.

6.2 Tools

Most of the currently available systems have been developed as a result of the popularity of the Wikification and KBP tasks. Furthermore, the rise of Linked Data datasets have also contributed to the development of industrial NED systems. Most systems either perform Wikification (every concept is linked) or NEL (only named entities are disambiguated) and some others perform also coreference resolution, the third aspect needed for Named Entity Resolution. As in previous sections, table 6 lists the available systems and services for NED and thereafter some details of each system are provided.

6.2.1 OKKAM

The overall goal of the OKKAM project[89] was to enable the Web of Entities, a global digital space for publishing and managing information about entities, where every entity is uniquely identified, and links between entities can be explicitly specified and exploited in a variety of scenarios. Compared to the WWW, the main differences are that the domain of entities is extended beyond the realm of digital resources to include objects in other realms like products, organizations, associations, countries, events, publications, hotels or people; and that links between entities are extended beyond hyperlinks to include virtually any type of relation. They developed the **Entity Name System** (ENS) as a NED system. In order to feed the system for NED, they harvested entities (together with an automatically created profile) from some popular public data sources like Wikipedia/DBpedia, GeoNames, UNIProt, etc. They were aiming at a repository of about 10 million entities by the end of the project. There is a public demo of the ENS and the tools are available to download[90].

[88] `http://openstreetmap.org/`
[89] `http://www.okkam.org`
[90] `http://community.okkam.org/`

System/Service	Languages	Sources availability	How it is provided	Programming Language	License	URL
Zemanta	English	NO	Browser add-on, API	Multiple	Free for non-commercial uses	http://www.zemanta.com
OKKAM	Multilingual	YES	Java Library	Java	Apache v2.0	http://www.okkam.org
The Wiki Machine	English, Italian	Yes	Library			http://thewikimachine.fbk.eu
Illinois Wikifier	English	Yes	Jar, Library	Java	Public	http://cogcomp.cs.illinois.edu/page/software_view/Wikifier
DBpedia Spotlight	English	Yes	API, library, source code	Java	Apache 2.0, part of the code uses LingPipe Royalty Free license	http://dbpedia-spotlight.github.com/
TAGME	English, Italian	NO	Restful API			http://tagme.di.unipi.it/
WikiMiner	English	Yes	Jar, library	Java	GNU GPLv3	http://wikipedia-miner.cms.waikato.ac.nz/

Table 6: Tools for Named Entity Disambiguation

6.2.2 The Wiki Machine

The Wiki Machine[91] is a Wikification system developed at the FBK in Trento, Italy. In addition to machine learning techniques, they use Linked Data to offer multilingual (English, Portuguese and Italian) wikification via DBpedia and Freebase. They also offer a public available demo in which you can compare their results with respect to AlchemyAPI, Zemanta and OpenCalais.

6.2.3 Zemanta

Zemanta is a service for bloggers that helps to blog better, easier and faster. By suggesting related articles, pictures, relevant in-text links and tags you can enrich your posts in a way to get more traffic, more clicks, more recommendations and to make your posts look more attractive. They have several tools to enrich your blogs as you write, providing related articles, image suggestions, and tag suggestions for your blog. Crucially, they also provide what they call in-text links which is basically a Wikification system to automatically provide the users with relevant links to the most important concepts of the blog, including named entities. The links use a variety of sources from the Web. Zemanta ltd. operates the Zemanta service. There is a basic free service, and they also offer paid upgrades for advanced features such as customization and guaranteed service levels. In principle, it is not available for commercial applications.

6.2.4 Illinois Wikifier

The Illinois Wikifier system is developed at Cognitive Computation Group at the of the University of Illinois at Urbana Champaign[92]. They present a Wikification system [Ratinov *et al.*, 2011] using both local and global features. The results reported claim to outperform previous systems [Milne and Witten, 2008]. It should be noted, however, that are not many approaches to NED who have evaluated their results with the same datasets. The KBP participants being the general exception.

6.2.5 DBpedia Spotlight

DBpedia Spotlight is a Wikification tool for automatically annotating mentions of DBpedia resources in text, providing a solution for linking unstructured information sources to the Linked Open Data cloud through DBpedia. DBpedia Spotlight recognizes that names of concepts or entities have been mentioned (e.g. "Michael Jordan"), and subsequently matches these names to unique identifiers (e.g. dbpedia:Michael_I._Jordan[93], the machine learning professor or dbpedia:Michael_Jordan[94] the basketball player).

[91] `http://thewikimachine.fbk.eu/html/index.html`
[92] `http://cogcomp.cs.illinois.edu/`
[93] `http://dbpedia.org/page/Michael_I._Jordan`
[94] `http://dbpedia.org/page/Michael_Jordan`

DBpedia Spotlight can used through their Web Application or Web Service endpoints. The Web Application is a user interface that allows to enter text in a form and generates an HTML annotated version of the text with links to DBpedia. The Web Service endpoints provide programmatic access to the demo, allowing to retrieve data also in XML or JSON. DBpedia is released under the Apache License 2.0.

6.2.6 WikiMiner

Wikipedia Miner is a wikification system developed by the University of Waikato, New Zealand [Milne and Witten, 2008]. The Wikipedia Miner can be used as a Web service or as a library via a Java API. The system uses machine learning and graph-based approaches to detect and disambiguate and link terms in running text to their Wikipedia articles. The system was the first publicly available tool for Wikification and many works still have it as a reference to evaluate their performance. Wikipedia Miner provided several benefits over previous Wikification work [Mihalcea and Csomai, 2007], by: (I) Identifying in the input text of a set C of so-called context pages, namely, pages linked by spots that are not ambiguous because they only link to one article; (ii) calculating a relatedness measure between two articles based on the overlap between their in-linking pages in Wikipedia; and (iii) defining a notion of coherence with other context pages in the set C. These three main components of the system allowed them to obtain around 75% F measure over long and richly linked Wikipedia articles.

6.2.7 TAGME

TAGME is a Wikification system developed by the University of Pisa, Italy. In principle they are particularly interested in short texts and they use the TAGME datasets, which partially consist of tweets to train their system [Ferragina and Scaiella, 2010]. Their aim is to obtain good performance annotating texts which are poorly written or formed, such as tweets, search engine snippets, etc. TAGME is inspired by previous systems such as Wikipedia Miner but they try to address the problem of having a very small context C available for training their machine learning models by using ranking algorithms. They report better results on short and long articles than previous approaches such as Wikipedia Miner.

7 Word Sense Disambiguation

Word sense disambiguation (WSD) stands for labelling every word in a text with its appropriate meaning or sense depending on its context. WSD is a very relevant research topic in NLP. General NLP books dedicate separate chapters to WSD [Manning and Schütze, 1998; Fox *et al.*, 1999]. There are also special issues on WSD in NLP journals [Ide and Véronis, 1998; Edmonds and Kilgarriff, 2002] and surveys [Navigli, 2009]; and books focussing to this issue [Ravin and Leacock, 2000; Stevenson, 2003; Agirre and Edmonds, 2006]. Despite the work devoted to the task, it can be said that no large-scale broad-coverage and accurate WSD system has been built up to date. State-of-the-art WSD systems obtain around 60-70% precision for fine-grained senses and 80-90% for coarser meaning distinctions [Izquierdo *et al.*, 2009]. Such a level of performance allows for improving tasks such as Machine Translation [Chan *et al.*, 2007], syntactic parsing [Agirre *et al.*, 2008], Information Retrieval [Stokoe *et al.*, 2003; Liu *et al.*, 2005b] and Cross-Linguistic Information Retrieval [Clough and Stevenson, 2004; Vossen *et al.*, 2006]. Lately, graph-based WSD systems are gaining growing attention [Agirre and Soroa, 2009; Laparra *et al.*, 2010]. These methods are language independent since only requires a local wordnet connected to the Princeton WordNet. For instance, using UKB[95], KYOTO developed knowledge-based WSD modules for English, Spanish, Basque, Italian, Dutch, Chinese and Japanese. This type of algorithms are also useful to compute semantic similarity of words and sentences [Agirre *et al.*, 2010a].

Deep approaches to WSD presume access to a large amount and comprehensive body of knowledge (both linguistic and world knowledge), which is used to determine the sense for words in the text. These approaches are very challenging in practice, mainly because such a body of knowledge is very hard to encode in computer-readable format, outside limited domains or without a very large investment. This is the case of Cyc [Lenat, 1995] which compiles a complex knowledge base with a vast quantity of world knowledge, including facts, terms, rules and axioms.

However, WSD systems have traditionally used a shallow approach. Shallow approaches do not try to perform complete understanding of the text. Usually, they only consider simple heuristics to determine the meaning of a word in a particular context. For instance, by testing the presence of a particular word in the surrounding context as in the rules "if *bass* has words *sea* or *fishing* nearby, it is probably the fish sense; if *bass* has the words *music* or *song* nearby, it is probably the music sense". These rules can be automatically derived by machine learning techniques, using a training corpus of word examples tagged with their corresponding word senses. However, such simple heuristics can confuse the correct sense of *bark* in "The dogs bark at the tree", which contains the word bark near both tree and dogs.

WSD systems are usually classified as *supervised* or *unsupervised*. However, nowadays it is difficult to establish a strict classification, since there are methods using different degrees of supervision. In order to avoid any confusion we will call unsupervised methods those which are "not supervised" at all. Supervised methods are those using *machine learning* methods to learn classifiers from sense-annotated corpora. On the other hand, approaches such as graph-based methods using WordNet glosses annotated with word senses like [Agirre and Soroa, 2009] will be considered unsupervised. Additionally, there are systems that combine both approaches to benefit from their advantages [Rigau *et al.*, 1997] or [Montoyo *et al.*, 2005].

Supervised approaches [Màrquez *et al.*, 2006], include Probabilistic methods (as Naive Bayes or Maximum Entropy), similarity methods (as Vector Space Models or K-Nearst Neighbours), those based on discriminating rules (as Decision Lists or Decision Trees) or

[95] `http://ixa2.si.ehu.es/ukb`

those margin based methods (Support Vector Machines), etc.

Machine learning (ML) classifiers are undeniable effective. However, in order to achieve high performance, supervised approaches require large training sets where instances (target words in context) are hand-annotated with the most appropriate word senses [Gale *et al.*, 1992b]. Due to this knowledge acquisition bottleneck problem, they will not be feasible until having reliable methods for acquiring large sets of training examples with a minimum human annotation effort.

There are several challenges that limit the performance of supervised WSD systems to around 70% accuracy [Martinez, 2004]. WSD depends on the characteristics of the used sense inventory such as granularity, coverage and richness of the encoded information. Also, the most usual feature sets consisting in bigrams, trigrams, and "bags of words" are too limited for modelling the contexts of the target words. Thus, some researchers have enriched the feature representation by including more sophisticated features such as syntactic dependencies [Chen and Palmer, 2009] or semantic classes [Izquierdo *et al.*, 2010].

Moreover, it also seems that existing corpora manually annotated with word senses is not large enough for improving the current state-of-the-art supervised WSD systems. Obviously, high-quality manually annotated data is very difficult and costly to obtain. Inter-annotator agreement (ITA) can be used to measure the consistency of the manually annotated data. Producing this kind of knowledge is extremely costly: the annotation rate is estimated to be about one word sense per minute [Edmonds and Cotton, 2001]. Furthermore, it is also worth mentioning that usually the most frequent sense baseline is extremely hard to improve upon even slightly [Gale *et al.*, 1992a].

For instance, [Ng, 1997] estimates that to obtain a high accuracy domain-independent system for English, about 1,000 occurrences of each of at least 3,200 words should be tagged. The necessary effort for constructing such a training corpus is estimated to be 16 person-years per language, according to the experience of [Ng and Lee, 1996]. However [Ng, 1997] suggests that active learning methods, described later in this section, could reduce the required effort significantly.

In order to overcome this problem, a number of research lines are being pursued. For instance, by using automatic methods for acquiring Sense Examples from the web by using WordNet as a knowledge base to characterize word-sense queries [Leacock *et al.*, 1998; Mihalcea and Moldovan, 1999; Agirre and Martínez, 2000; Agirre and Lopez de Lacalle, 2004; Cuadros and Rigau, 2008]. Recently, [Mihalcea, 2007] describes a method for generating sense-tagged data using Wikipedia as a source of sense annotations showing that Wikipedia-based sense annotations are reliable enough to construct accurate sense classifiers.

Additionally, WSD systems trained on general corpora are known to perform worse when moved to specific domains. Previous work [Escudero *et al.*, 2000; Martínez and Agirre, 2000] has shown that there is a large loss of performance when training on one corpora and testing on a different one. Recently, [Izquierdo *et al.*, 2010] presents a system that achieves results over the most-frequent-sense baseline in environmental domain [Agirre *et al.*, 2010b]. The system uses semantic class classifiers instead of word classifiers, and monosemous examples obtained from a background set of documents from the same domain.

Traditionally, *unsupervised* approaches are grouped as:

- **Knowledge Based methods:** These methods use the explicit information gathered from an existing lexicon or knowledge base. The lexicon may be a machine readable dictionary such as LDOCE [Procter, 1987], WordNet [Fellbaum, 1998] or a thesaurus such as Roget's [Roget, 1911].

  One of the first knowledge based approaches to WSD, is the Lesk algorithm [Lesk, 1986]. Given a word to disambiguate, the dictionary definition or gloss of each of its sense is compared to the glosses (or definition) of every other word in the context. A sense whose gloss shares the largest number of words in common with the glosses of the words in context is assigned.

  [Brockmann and Lapata, 2003] give a detailed analysis of these approaches, while [Agirre and Martinez, 2001] report a comparative evaluation of some of these approaches. A whole overview of the impact of the knowlege sources applied to Word Sense Disambiguation is summarized in [Agirre and Stevenson, 2005].

- **Corpus Based methods:** These methods perform WSD using information gathered from corpora. Corpus based unsupervised algorithms use non-annotated corpora to induce their models.

  [Pedersen, 2006] provides a complete overview of unsupervised corpus based methods.

- **Graph based methods:** Lately, graph-based methods for knowledge based WSD have gained much attention in the NLP community [Navigli and Velardi, 2005; Sinha and Mihalcea, 2007a; Navigli and Lapata, 2007; Mihalcea, 2005; Agirre and Soroa, 2009]. These methods use well-known graph based techniques to find and exploit the structural properties of the graph underlying a particular knowledge base, for instance WordNet. Graph based WSD methods manage to exploit the interrelations among the senses in the given context.

  Graph based methods have great advantages. Firstly, no training corpora is required. Furthermore, these methods are language independent since they only need a knowledge base for the target language, or multilingual connections to the graph. Finally, they also obtain good results when they are applied to a set of closely related words.

- **Hybrid and semi-supervised methods:** These methods use a mixture of corpus data and knowledge from an explicit knowledge base. Most of the unsupervised approaches fall in this category.

  For instance, [Yarowsky, 1992] proposed an unsupervised method that disambiguate words using statistical models inferred from raw, untagged text by using the Roget's Thesaurus [Roget, 1911].

As empirically demonstrated by the last SensEval and SemEval exercises[96], despite the wide range of approaches investigated and the large effort devoted to tackle this problem,

[96] `http://www.senseval.org`

assigning the appropriate meaning to words in context has resisted all attempts to be fully successfully addressed.

However, still with low performance, WSD has been proved to be useful to improve tasks such as in parsing [Agirre *et al.*, 2008], information retrieval (IR) [Agirre *et al.*, 2009c], machine translation [Carpuat and Wu, 2007] or information extraction [Chai and Biermann, 1999].

Albeit its inherent drawbacks, supervised corpus-based methods obtain better performance results than unsupervised methods. The achieved performance varies depending on the number of sense-tagged examples to train, the domain, the sense repository, etc., but considering the all-words task as the most realistic scenario, state-of-the-art performance is between 50% and 80% of accuracy. For instance, in the last SemEval exercise [Izquierdo *et al.*, 2010] achieved 51% recall on a specific domain. However, [Chen and Palmer, 2009] presented in SemEval 2007 a supervised WSD system for English verbs (usually more difficult than nouns) that using linguistically motivated features obtained accuracy rates over 90%.

However, unsupervised methods and, in particular, graph-based methods present very appealing advantages. They are not dependent on a manually labelled corpus for training. In comparison, graph-based methods obtain better results when applied to a set of closely related words than when applied to running text [Navigli and Velardi, 2005; Niemann and Gurevych, 2011].

When addressing WSD in particular domains, supervised methods perform worse compared with they performance in general domain [Escudero *et al.*, 2000; Martínez and Agirre, 2000]. Following this direction [Agirre *et al.*, 2009b; Agirre and Lopez deLacalle, 2009] study the problem of domain WSD using different knowledge based and machine learning techniques. The best performing methods seem to be graph based.

7.1 Data Sources

7.1.1 SemCor

SemCor [Miller *et al.*, 1993] is a subset of the Brown Corpus [Kučera and Francis, 1967] whose content words have been manually annotated with part-of-speech tags, lemmas, and word senses from the WordNet inventory. SemCor is composed of 352 texts: in 186 texts all the open-class words (nouns, verbs, adjectives, and adverbs) are annotated with these information, while in the remaining 166 texts only verbs are semantically annotated with word senses.

Overall, SemCor comprises a sample of around 234,000 semantically annotated words, thus constituting the largest manually sense-tagged corpus for training sense classifiers in supervised disambiguation settings. The original SemCor was annotated according to WordNet 1.5. However, mappings exist to more recent versions (e.g., 3.0, etc.)[97].

Based on SemCor, a bilingual corpus was created by [Bentivogli and Pianta, 2005]: **MultiSemCor** is an English/Italian parallel corpus aligned at the word level which pro-

[97] `http://www.cse.unt.edu/~rada/downloads.html#semcor`

vides for each word its part of speech, its lemma, and a sense from the English and Italian versions of WordNet (namely, MultiWordNet [Pianta *et al.*, 2002]). The corpus was built by aligning the Italian translation of SemCor at the word level. The original word sense tags from SemCor were then transferred to the aligned Italian words.

7.1.2 OntoNotes

OntoNotes Release 4.0[98] [Hovy *et al.*, 2006], was developed as part of the OntoNotes project, a collaborative effort between BBN Technologies, the University of Colorado, the University of Pennsylvania and the University of Southern Californias Information Sciences Institute. The goal of the project is to annotate a large corpus comprising various genres of text (news, conversational telephone speech, weblogs, usenet newsgroups, broadcast, talk shows) in three languages (English, Chinese, and Arabic) with structural information (syntax and predicate argument structure) and shallow semantics (word sense linked to an ontology and coreference). For English, OntoNotes contains 600k words of English newswire, 200k word of English broadcast news, 200k words of English broadcast conversation and 300k words of English web text. Its semantic representation includes word sense disambiguation for nouns and verbs, with each word sense connected to an ontology, and coreference. There are a total of 264,622 words in the combined corpora tagged with word sense information. These cover 1,338 noun and 2,011 verb types. A total of 6,147 WordNet word senses have been pooled and connected to the Omega Ontology [Philpot *et al.*, 2005]. The current goals call for annotation of over a million words of English.

7.1.3 Ancora

AnCora[99] [Taulé *et al.*, 2008] consist of a Catalan corpus (AnCora-CA) and a Spanish corpus (AnCora-ES), each of them of 500,000 words. The corpora are annotated at different levels:

- Lemma and Part of Speech
- Syntactic constituents and functions
- Argument structure and thematic roles
- Semantic classes of the verb
- Denotative type of deverbal nouns
- Nouns related to WordNet synsets
- Named Entities
- Coreference relations

[98] `http://www.ldc.upenn.edu/Catalog/catalogEntry.jsp?catalogId=LDC2011T03`
[99] `http://clic.ub.edu/corpus/en`

AnCora corpus is mainly based on journalist texts. For Spanish, the morphological and syntactic levels are already completed, while the semantic annotation covers 40% of the corpus (200,000 words). With respect to the semantic annotation, the corpora were annotated at different levels: 1) basic syntactic functions were tagged in a semiautomatic way with arguments and thematic roles taking into account the semantic class related to the verbal predicate [Taulé *et al.*, 2006]; 2) Spanish and Catalan WordNet synsets aligned to WN1.6 were manually assigned for all nouns in the corpora [Atserias *et al.*, 2004]; and 3) named entities were also manually annotated [Borrega *et al.*, 2007].

7.1.4 Senseval/SemEval corpora

Since 1998, SensEval[100] and later on SemEval[101] organize an ongoing series of evaluations of computational semantic analysis systems. Along these years, multiple organizers have provided a large number of multilingual datasets annotated at a sense level (see table 7 for further details.)

Data Entity	#Words	Language	License	Website
SemCor	234,000	English	GNU	http://www.cse.unt.edu/~rada/downloads.html#semcor
Semantically Annotated gloss corpus	454,439	English	Unknown	http://wordnet.princeton.edu/glosstag.shtml
OntoNotes	264,622	English	LDC	http://www.ldc.upenn.edu/Catalog/catalogEntry.jsp?catalogId=LDC2011T03
AnCora	<500,000	Spanish	Unknown	http://clic.ub.edu/corpus/en/ancora
MultiSemCor	92,420	English-Italian	CC-by-3.0	http://multisemcor.fbk.eu/index.php
SensEval2 English, Dutch all-words WSD	5,000	English-Dutch	Unknown	http://www.hipposmond.com/senseval2
SensEval3 Task 1 English all-words WSD	5,000	English	Unknown	http://www.senseval.org/senseval3
SensEval3 Task 2 Italian all-words WSD	5,000	Italian	Unknown	http://www.senseval.org/senseval3
SensEval3 Task 12 WSD of WordNet glosses	15,717	English	Unknown	http://www.senseval.org/senseval3
SemEval2007 Task 17 English LS, SRL, all-words WSD	5,000	English	Unknown	http://nlp.cs.swarthmore.edu/semeval/tasks/task17/description.shtml
SemEval2007 Task 09 Multilevel Semantic Annotation	Part of Ancora	Spanish	Unknown	http://www.lsi.upc.edu/~nlp/semeval/msacs.html
SemEval2010 Task 17 WSD-Domain	2,000	English-Dutch-Italian	Unknown	http://xmlgroup.iit.cnr.it/SemEval2010/
SemEval2010 Task 03 Cross-lingual WSD	1,000	English-Dutch-Italian-Spanish	Unknown	http://webs.hogent.be/~elef464/lt3_SemEval.html
SemEval2013 Task 10 Cross-lingual WSD	1,000	English-Dutch-Italian-Spanish	Unknown	http://www.cs.york.ac.uk/semeval-2013/task10
SemEval2013 Task 12 Multilingual WSD	1,000	English-Italian-Spanish	Unknown	http://www.cs.york.ac.uk/semeval-2013/task12
EVALITA WSD All-Word-Task	5,000	Italian	Unknown	http://www.evalita.it/2007/tasks/wsd
EVALITA SuperSense tagging	135,738	Italian	Unknown	http://www.evalita.it/2011/tasks/SST

Table 7: Data Sources for Word Sense Disambiguation

[100] `http://www.senseval.org/`
[101] `http://en.wikipedia.org/wiki/SemEval`

7.2 Tools

7.2.1 SenseLearner

SenseLearner[102] [Mihalcea and Csomai, 2005] is a minimally supervised all-words WSD algorithm for English.

7.2.2 IMS

IMS (It Makes Sense) [103] [Zhong and Ng, 2010] is a supervised English all-words word sense disambiguation (WSD) system. The flexible framework of IMS allows users to integrate different preprocessing tools, additional features, and different classifiers. By default, the system uses linear support vector machines as the classifier with multiple features. This implementation of IMS achieves state-of-the-art results on several SensEval and SemEval tasks.

7.2.3 SuperSenseTagger

SuperSenseTagger[104] [Ciaramita and Altun, 2006b] annotates English and Italian text with around 40 broad semantic categories (Wordnet lexicographic files or supersenses) for both nouns and verbs; i.e., it performs both sense disambiguation and named-entity recognition. The tagger implements a discriminatively-trained Hidden Markov Model.

7.2.4 GWSD

GWSD[105] [Sinha and Mihalcea, 2007b] is a system for unsupervised all-words graph-based word sense disambiguation. The algorithm annotates all the words in a text by exploiting similarities identified among word senses, and using centrality algorithms applied on the graphs encoding these sense dependencies.

7.2.5 UKB

UKB[106] [Agirre and Soroa, 2009] is a collection of programs for performing graph-based Word Sense Disambiguation and lexical similarity/relatedness using a pre-existing knowledge base. UKB applies the so-called Personalized PageRank on a Lexical Knowledge Base (LKB) to rank the vertices of the LKB and thus perform disambiguation. Moreover, the algorithm can be applied to any language having a wordnet or a large lexical knowledge base. For instance, using UKB[107], KYOTO developed knowledge-based WSD modules for English, Spanish, Basque, Italian, Dutch, Chinese and Japanese. It has also been applied

[102] http://www.cse.unt.edu/~rada/downloads.html#senselearner
[103] http://www.comp.nus.edu.sg/~nlp/software.html
[104] http://sourceforge.net/projects/supersensetag/
[105] http://www.cse.unt.edu/~rada/downloads.html#gwsd
[106] http://ixa2.si.ehu.es/ukb/
[107] http://ixa2.si.ehu.es/ukb

on WSD on specific domains [Agirre *et al.*, 2009a]. The algorithm can also be used to calculate lexical similarity/relatedness of words/sentences. This type of algorithms are also useful to compute semantic similarity of words and sentences [Agirre *et al.*, 2010a].

Table 8 summarizes the WSD tools available.

System/Service	Languages	Sources availability	Programming Language	License	URL
SenseLearner	English	Yes	Perl	GNU	`http://www.cse.unt.edu/~rada/downloads.html#senselearner`
IMS	English	Yes	Java	Unknown	`http://www.comp.nus.edu.sg/~nlp/software.html`
SuperSenseTagger	English-Italian	Yes	Java	Apache v2	`http://sourceforge.net/projects/supersensetag/`
GWSD	Multilingual	Yes	Perl	GNU	`http://www.cse.unt.edu/~rada/downloads.html#gwsd`
UKB	Multilingual	Yes	C++	Unknown	`http://ixa2.si.ehu.es/ukb/`

Table 8: Tools for Word Sense Disambiguation

8 Sentiment Analysis

Sentiment analysis and opinion mining is concerned with analysing opinions, sentiments, evaluations, attitudes, and emotions in text [Liu, 2012]. It is a useful natural language processing task for organisations who want to know how their brand or product is perceived by the public, and its popularity within and outside the research community has risen in the last decade. There are currently two dominant approaches to sentiment analysis: supervised machine learning using Naive Bayes, Support Vector Machines or Maximum Entropy classification and unsupervised methods or dictionary-based methods. [Chaovalit and Zhou, 2005] evaluated both techniques and found that supervised techniques slightly outperform unsupervised techniques (85% vs 77% accuracy). For a comprehensive overview of the state-of-the-art, the reader is referred to [Pang and Lee, 2008].

8.1 Data Sources

In Table 9 and 10, the resources marked up with sentiment information that are available to NewsReader are presented.

English	availability	authors	items	acquisition	evaluation

SenticNet 1.0 (2010)	`http://sentic.net`; only for research	[Cambria *et al.*, 2010]	5,700 items (i.e. words and combination of words), with polarity values ranging, from -1 (negative) to +1 (positive)	automatic	
SenticNet 2.0 (2012)	not yet available &only for research	[Cambria *et al.*, 2012]	14,000 items (i.e. words and combination of words), with polarity values, with affective labels like, Pleasantness, Attention, Sensitivity and Aptitude	automatic	extrinsically

Q-Wordnet 3.0 (2010)	available	[Agerri and García-Serrano, 2010]	- 16,000 items (i.e. synsets) - polarity categories (7402) - positive and 8108 negative)	automatic	intrinsically for smaller set of 5.000 items on MWOP : no accuracy, F-measure from 0.89 to 0.99% for positives and from 0.76 to 0.91 for negatives
Opinion Finder (2005) Lexicon aka Subjectivity Lexicon	`http://www.cs.pitt.edu/mpqa/subj_lexicon.html`; no restrictions	[Wiebe and Riloff, 2005]	- 8,221 items (i.e. words and - multi-word expressions (990)) - Labeled with reliability - (strong if they appear most often in subjective text vs. weak) and polarity and polarity (positive, negative, or neutral).	manually and augmented with entries learned from corpora	completely manually checked

General Inquirer (1966)	`http://www.wjh.havard.edu/~inquirer/`; for academic purposes	[Stone *et al.*, 1966]	- 1,915 positive items (i.e. words) - 2,291 negative items (i.e. words)	manually	manually checked
SentiWordNet (2006)	`http://sentiwordnet.isti.cnr.it/` freely available for research; restrictions for commercial use	[Esuli and Sebastiani, 2006]	- 35,000 items (i.e. synsets) - based on WordNet 2.0/3. - each synset has two polarity - values: one ranging from 0 to 1(positive) and one ranging from 0 to -1(negative)	automatic	evaluated in various classifation tasks and against MWOM
WordNet Affect (2004)	`http://wndomains.fbk.eu/wnaffect.html` freely available Creative Commons Attribution 3.0 Unported License	[Strapparava and Valitutti, 2004]	- 4,748 words organized in - 2,874 synsets - With affective labels like emotion, feeling, cognitive state, attitude, and behaviour	semi- automatic	the resource is started from a manually annotated list of 1903 words

OpinionLexicon (2005)	`http://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html` no restrictions	[Liu *et al.*, 2005a]	- 7,000 words (including - misspellings) from social media - Labeled with positive - (2,000) or negative (5,000) polarity	automatic	
Sentisense (2012)	`http://nlp.uned.es/~jcalbornoz/resources.html` available for research purposes	[de Albornoz *et al.*, 2012]	- 5,500 words organized in 2,200 synsets - Labeled with 10 emotional categories like love, fear, disgust etc.	semi- automatic	the process started from a list of 1200 manually annotated synsets
OpeNER general sentiment lexicon English	re-use of publicly available sentiment lexicon "Subjectivity-Clues", developed by [Wilson *et al.*, 2005]	OpeNER project	intensifiers, weakeners, polarity shifters, synonyms, near-synonyms, antonyms and hyponyms	semi-automatic	partly manually checked

UMass Amherst Linguistics Sentiment Corpora (also for Chinese, Japanese and German)	`http://semanticsarchive.net/Archive/jQOZGZiM/readme.html` free	[Constant *et al.*, 2009]	n-grams counts extracted from over 700,000 online product reviews in Chinese, English, German, and Japanese	Automatic from Amazon, Tripadvisor, Myprice
MPQA Opinion Corpus Product Debate Corpus	`http://www.cs.pitt.edu/mpqa/` free, GNU public license	[Wiebe *et al.*, 2005]	News articles from a wide variety of news sources manually annotated for opinions and other private states (i.e., beliefs, emotions, sentiments, speculations, etc.)	
Product Debate Corpus	`http://www.cs.pitt.edu/mpqa/` free, GNU public license	[Somasundaran and Wiebe, 2009]		

Political Debate Corpus	`http://www.cs.pitt.edu/mpqa/` free, GNU public license	[Somasundaran and Wiebe, 2010]			

Table 9: Generic sentiment lexicons for English

Dutch	availability	authors	items	acquisition	evaluation
Duoman Lexicon (2009)	Cornetto-based	[Jijkoun and Hofmann, 2009]	- 16,000 words (9,000 - negative/ 7,000 positive) - with polarity values ranging - from -1 to +1	automatic	top 6000 evaluated against gold standard (0.62% accuracy on positive polarity; 0.82% on negative polarity)
DutchPolarity Lexicon	Cornetto-based	[Maks and Vossen, 2011]	- 18,000 items (synset, word - sense and word version available) - labeled with polarity values - (positive and negative) and confidence values	automatic	evaluated against gold standard (accuracy 0.75%)

Dutch Adjective Lexicon (2012) book review domain	`https://www.clips.ua.ac.be/pages/pattern` open source with PDDL partly Cornetto based	[DeSmedt and Daelemans, 2012]	3,000 adjective words - (domain-dependent book reviews) - labeled with - polarity,subjectivity and intensity	automatic/ semi- automatic	1.100 manually annotated
OpeNER general sentiment lexicon Dutch	Cornetto based	OpeNER project	intensifiers, weakeners, polarity shifters, synonyms, near-synonyms, antonyms and hyponyms	semi-automatic	partly manually checked

Table 10: Generic sentiment lexicons for Dutch

For Spanish, we also have the TASS 2012[Villena-Román *et al.*, 2012], General Election Twitter Corpus[108], Movie Reviews[109] and SFU Reviews Corpus available[Brooke *et al.*, 2009].

The TASS corpus was compiled for the Tarea de Analysis de Sentimientos en la SEPLN (TASS) of 2012 by Daedalus[110]. The corpus contains 70,000 tweets, written in Spanish by 150 well-known personalities and celebrities of the world of politics, economy, communication, mass media and culture, between November 2011 and March 2012. Although the context of extraction has a Spain-focused bias, the diverse nationality of the authors, including people from Spain, Mexico, Colombia, Puerto Rico, USA and many other countries, give the corpus global coverage of the Spanish-speaking world. The user and TweetIDs are anonymised and each message is tagged with its global polarity, indicating whether the text expresses a positive, negative or neutral sentiment, or no sentiment at all. 5 levels have been defined: strong positive (P+), positive (P), neutral (NEU), negative

[108] http://www.lsi.us.es/ fermin/index.php/Datasets
[109] http://www.lsi.us.es/ fermin/index.php/Datasets
[110] `http://www.daedalus.es`

(N), strong negative (N+) and one additional no sentiment tag (NONE).

The General Election Twitter Corpus consists of 743 files of tweets conversations about the Spanish General election of 2011 in XML format.

Movie Reviews consists of 3,878 reviews of Spanish movies in XML and with part-of-speech tags and Dependency analysis.

SFU Reviews Corpus is a collection of 400 reviews on cars, hotels, washing machines, books, cell phones, music, computers, and movies. Each category contains 50 positive and 50 negative reviews, defined as positive or negative based on the number of stars given by the reviewer (1-2=negative; 4- 5=positive; 3-star review are not included). The reviews were collected from the website ciao.es. They are intended to be a Spanish parallel to the SFU Review Corpus (in English)[111].

8.2 Tools

In Table 11 available sentiment analysis tools are presented.

System	Languages	Responsible	Sources availability	How it is provided	Programming Language	License	URL
Opinion-Finder	English	University of Pittsburgh	Yes	Library	Python	Research Purposes	http://www.cs.pitt.edu/mpqa/opinionfinder 1_5.html
Sentiment Analysis-NLTK	English	NLTK Platform	Yes	Web service/Library	Python	Open Source	http://text-processi com/docs/sentiment.html

[111] http://www.sfu.ca/~mtaboada/research/SFU_Review_Corpus.html

TM Text Mining Package	English	Vienna University of Economics and Business	Yes	Library	R	GPL	http://tm.r-forge.r-project.org/index.html https://r-forge.r-project.org/R/?group_id=1048
Sentiment140	English, Spanish	Stanford University	No	API/Web service	Java	Non commercial (free version limited)	http://help.sentiment140.com/api
AlchemyAPI	English, French, German, Italian, Portuguese, Russian, Spanish and Swedish	AlchemyAPI	Yes	API/Web service	Android, JAva, Perl, Ruby, Python, PHP, C, C++, C#	Non commercial (free version limited)	http://www.alchemyapi.com
DUOMAN	Dutch	University of Amsterdam/TST Centrale	Not yet	Not yet known	Not yet known	Not yet known	Not yet known
LingPipe	English	Atlas-i	Yes	Library of Service	Java	Free for research, proprietary otherwise	http://alias-i.com/lingpipe
GATE	English	University of Sheffield	Yes	Library	Java	GNU GPLv2	http://gate.ac.uk/sentiment

NaturalOpinions	English, Spanish	Bitext.com	No	API to JSON, XML and CSV formats	C++	Commercial	http://www.bitext.com
Olery	Dutch, English, Italian, German	Olery	No	API/Web service	Ruby, Python, C++, Java	Commercial	http://www.olery.com
OpeNER Polarity Tagger	Dutch, English, German	VU University Amsterdam	Yes	API/Web service	Python	-	http://ic.vupr.nl:9081/vu_polarity_tagger_basic
OpeNER Opinion Detector	Dutch, English	VU University Amsterdam	Yes	API/Web service	Python	-	http://ic.vupr.nl:9081/vu_opinion_detector_basic_en_nl

Table 11: Sentiment Analysis Tools

9 Semantic Role Labeling

Semantic Role Labeling (SRL) is a task involving recognition of semantic arguments of predicates on top of their syntactic constituents [Baker *et al.*, 1998a; Gildea and Jurafsky, 2002a]. Usual semantic roles include Agent, Patient, Instrument or Location. Such quite general and widely-recognized labels are usual in building corpora and other linguistic resources [Dorr, 1997; Alonso *et al.*, 2005]. Furthermore, advantages and/or disadvantages of a more fine-grained lexical role specification, such as buyer, seller, killer, victim or time period [Fillmore *et al.*, 2003; García-Miguel and Albertuz, 2005] deserve to be closely analyzed when working on domains. In the last decade many lexical databases have included Semantic Roles as a feature of predicates (i.e. FrameNet [Ruppenhofer *et al.*, 2002], among others). Also, several corpora have been labelled with Semantic Roles (i.e. PropBank [Palmer *et al.*, 2005b], among others). From a linguistic point of view, SRs are situated in the syntax-semantics interface; empirically, argument identification is closely related to syntax and argument classification is more related to semantics.

SRL is a crucial task for establishing Who does What, Where, When and Why. A technology which has proved to be key for applications such as Information Extraction, Question Answering, Summarization and probably every NLP task involving any level of semantic interpretation ([Carreras and Màrquez, 2005a; Zapirain *et al.*, 2008; Lluís and Màrquez, 2008].

Semantic parsing is considerably more complex than Semantic Role Labeling (SRL). In fact, there are not many semantic interpretation systems for unrestricted domains. For English, the three most advanced Semantic Parsers are those of Shalmaneser [Erk and Pado, 2006], Lingo/LKB [Copestake, 2002], and Boxer [Bos, 2008]. Moreover, it does not seem simple to adapt these systems to other languages.

9.1 Data Sources

PropBank ([Palmer *et al.*, 2005a]) is the most widely used corpus for training SRL systems, probably because it contains running text from the Penn Treebank corpus with annotations on all verbal predicates. However, a serious criticisms to the PropBank corpus refers to the role set used in this corpus, which consists of a set of numbered core arguments, whose semantic translation is verb-dependent ([Zapirain *et al.*, 2008]). In this section we describe the most known role repositories traditionally used for training SRL systems.

9.1.1 PropBank and Nombank

The PropBank and NomBank corpus are the result of adding a semantic layer to the syntactic structures of Penn Treebank II ([Palmer *et al.*, 2005a]). Specifically, they provide information about predicate-argument structures to all verbal and nominal predicates of the Wall Street Journal section of the treebank. The role set is theory–neutral and consists of a set of numbered core arguments (Arg0, Arg1, ...). Each verb has a *frameset* listing its allowed role labels and mapping each numbered role to an English-language description of its semantics.

Different senses for a polysemous verb have different framesets, but the argument labels are semantically consistent in all syntactic alternations of the same verb–sense. For instance in "Kevin broke [the window]$_{Arg1}$" and in "[The door]$_{Arg1}$ broke into a million pieces", for the verb *broke.01*, both Arg1 arguments have the same semantic meaning, that is "broken entity". Nevertheless, argument labels are not necessarily consistent across different verbs (or verb senses). For instance, the same Arg2 label is used to identify the Destination argument of a proposition governed by the verb *send* and the Beneficiary argument of the verb *compose*. This fact might compromise generalization of systems trained on PropBank, which might be focusing too much on verb–specific knowledge. It is worth noting that the two most frequent arguments, Arg0 and Arg1, are intended to indicate the general roles of Agent and Theme and are usually consistent across different verbs. However, this correspondence is not total. According to the study by ([Yi *et al.*, 2007]), Arg0 corresponds to Agent 85.4% of the time, but also to Experiencer (7.2%), Theme (2.1%), and Cause (1.9%). Similarly, Arg1 corresponds to Theme in 47.0% of the occurrences but also to

Topic (23.0%), Patient (10.8%), and Product (2.9%), among others. Contrary to core arguments, adjuncts (Temporal and Location markers, etc.) are annotated with a closed set of general and verb-independent labels.

9.1.2 VerbNet

VerbNet ([Kipper *et al.*, 2000]) is a computational verb lexicon in which verbs are organized hierarchically into classes depending on their syntactic/semantic linking behavior. The classes are based on Levin's verb classes ([Levin, 1993]) and each contains a list of member verbs and a correspondence between the shared syntactic frames and the semantic information, such as thematic roles and selectional constraints. There are 23 thematic roles (Agent, Patient, Theme, Experiencer, Source, Beneficiary, Instrument, etc.) which, unlike the PropBank numbered arguments, are considered as general verb-independent roles.

This level of abstraction makes them, in principle, better suited (compared to PropBank numbered arguments) for being directly exploited by general NLP applications. But, VerbNet by itself is not an appropriate resource to train SRL systems. As opposed to PropBank, the number of tagged examples is far more limited in VerbNet. Fortunately, in the last years a twofold effort has been made in order to generate a large corpus fully annotated with thematic roles. Firstly, the SemLink[112] resource ([Loper *et al.*, 2007]) established a mapping between PropBank framesets and VerbNet thematic roles. Secondly, the SemLink mapping was applied to a representative portion of the PropBank corpus and manually disambiguated ([Loper *et al.*, 2007]). The resulting corpus is currently available for the research community and makes possible comparative studies between role sets.

9.1.3 FrameNet

FrameNet is a lexical database of English that it is based on a theory of meaning called Frame Semantics, deriving from the work of Charles J. Fillmore and colleagues [Fillmore, 1976; Fillmore, 1977; Fillmore, 1982; Fillmore, 1985; Fillmore and Baker, 2001; Fillmore and Baker, 2009]. In FrameNet word meanings or Lexical Units are connected with particular Semantic Frames, which are basically descriptions of events and their participants or Frame Elements.

FrameNet annotations derive from two sources. In pursuing the goal of recording the range of semantic and syntactic combinatory possibilities (valences) of each word in each of its senses, they normally concentrate on a particular target LU and extract sentences from the different texts of a corpus containing that LU. Then they annotate a selection of the extracted sentences in respect to the target LU. In another kind of work that represents a much smaller percentage of our overall annotations, they annotate running text. Full-text annotation differs from sentence annotation mostly in that the sentences are chosen for them, so to speak, by the author of the text. The annotation of running text is technically possible thanks to the annotation layering technique: FN lexicographers can one by one declare each word in a sentence a target, select a frame relative to which the new

[112]http://verbs.colorado.edu/semlink/

target is to be annotated, get a new set of annotation layers (frame element, grammatical function, phrase type) and appropriate frame element tags, and then annotate the relevant constituents.

9.2 Tools

Since Gildea and Jurafsky's initial work "Automatic Labeling of Semantic Roles" ([Gildea and Jurafsky, 2002b]) on FrameNet-based SRL, many researchers have devoted their efforts on this exciting and relatively new task. Several evaluation exercises on SRL were conducted by the "shared tasks" of CoNLL-2004 ([Carreras and Màrquez, 2004]), CoNLL-2005 ([Carreras and Màrquez, 2005a]), CoNLL-2008 ([Surdeanu *et al.*, 2008]) and CoNLL-2009 ([Hajič *et al.*, 2009]) conferences, bringing to scene a comparative analysis of competitive systems trained on the PropBank corpus. From there, PropBank became the most widely used corpus for training SRL systems, leaving VerbNet and FrameNet based tasks ([Pradhan *et al.*, 2007a] and [Litkowski, 2004], respectively) in a more modest position.

9.2.1 Mate-Tools

The Mate tools[113] provide a pipeline of modules that carry out lemmatization, part-of-speech tagging, dependency parsing, and PropBank based semantic role labeling of a sentence. The system's two main components draw on improved versions of a state-of-the-art dependency parser and semantic role labeler ([Björkelund *et al.*, 2009b]) developed independently by the authors. The tools are language independent, provide a very high accuracy and are fast. The dependency parser had the top score for German and English dependency parsing in the CoNLL shared task 2009.

9.2.2 SwiRL

SwiRL[114] is a PropBank based Semantic Role Labeling (SRL) system for English constructed on top of full syntactic analysis of text. The syntactic analysis is performed using Eugene Charniak's parser. SwiRL trains one classifier for each argument label using a rich set of syntactic and semantic features. The classifiers are learned using one-vs-all AdaBoost classifiers. SwiRL is a free (GPL) SRL system.

9.2.3 SENNA

SENNA[115] is a software package that is distributed under a non-commercial license, which outputs a host of Natural Language Processing (NLP) predictions: part-of-speech (POS) tags, chunking (CHK), name entity recognition (NER), semantic role labeling (SRL) and syntactic parsing (PSG). It is fast and uses a simple architecture, self-contained because it

[113] `http://code.google.com/p/mate-tools/`
[114] `http://surdeanu.info/mihai/swirl/`
[115] `http://ml.nec-labs.com/senna/`

does not rely on the output of existing NLP system, and accurate because it offers state-of-the-art or near state-of-the-art performance. Written in ANSI C, with about 3,500 lines of code, SENNA requires about 200MB of RAM and should run on any IEEE floating point computer.

9.2.4 SEMAFOR

SEMAFOR [116] –Semantic Analysis of Frame Representations– is a tool for automatic analysis of the frame-semantic structure of English text. This tool attempts to find which words in text evoke which semantic frames, and to find and label each frame's arguments. It takes as input a file with English sentences, one per line, and performs the following steps:

- Preprocessing: The sentences are lemmatized, part-of-speech tagged, and syntactically parsed (optionally using a syntactic parsing running in server mode.)
- Target identification: Frame-evoking words and phrases (”targets”) are heuristically identified in each sentence.
- Frame identification: a log-linear model, trained on FrameNet 1.5 data with full-text frame annotations, produces for each target a probability distribution over frames in the FrameNet lexicon (optionally constrained by a semi-supervised filter). The target is then labeled with the highest-scoring frame.
- Argument identification: A second log-linear model, trained on the same data, considers every role of each labeled frame instance and identifies a span of words in the sentence–or NULL–as filling that role. A subsequent step ensures that none of a frame's overt arguments overlap using beam search; an alternate strategy using Alternating Directions Dual Decomposition uses two other constraints used in FrameNet for argument identification.
- Output: An XML file is produced containing the text of the input sentences, augmented with the frame-semantic information (target-frame and argument-role pairings) predicted by the system. See the papers listed below (”Further Reading”) for algorithmic details and experimental evaluation of the components of this system.

9.2.5 Shalmaneser

Shalmaneser ([Erk and Pado, 2006])is a supervised learning toolbox for shallow semantic parsing, i.e. the automatic assignment of semantic classes and roles to text. The system was developed for Frame Semantics; thus they use Frame Semantics terminology and call the classes frames and the roles frame elements. However, the architecture is reasonably general, and with a certain amount of adaption, Shalmaneser should be usable for other

[116] http://www.ark.cs.cmu.edu/SEMAFOR/

paradigms (e.g., PropBank roles) as well. Shalmaneser caters both for end users, and for researchers.

For end users, they provide a simple end user mode which can simply apply the pre-trained classifiers for English (FrameNet annotation / Collins parser) and German (SALSA Frame annotation / Sleepy parser). For researchers interested in investigating shallow semantic parsing, our system is extensively configurable and extendable.

9.3 Implicit Semantic Role Labeling

Traditionally, Semantic Role Labeling (SRL) systems have focused in searching the fillers of those explicit roles appearing within sentence boundaries [Gildea and Jurafsky, 2000; Gildea and Jurafsky, 2002b; Carreras and Màrquez, 2005b; Surdeanu *et al.*, 2008; Hajič *et al.*, 2009]. These systems limited their search-space to the elements that share a syntactical relation with the predicate. However, when the participants of a predicate are implicit this approach obtains incomplete predicative structures with null arguments. The following example includes the gold-standard annotations for a traditional SRL process:

(1) [*arg0* The network] had been expected to have [*np* **losses**] [*arg1* of as much as $20 million] [*arg3* on baseball this year]. It isn't clear how much those [*np* **losses**] may widen because of the short Series.

The previous analysis includes annotations for the nominal predicate **loss** based on the NomBank structure [Meyers *et al.*, 2004]. In this case the annotator identifies, in the first sentence, the arguments *arg0*, the entity losing something, *arg1*, the thing lost, and *arg3*, the source of that loss. However, in the second sentence there is another instance of the same predicate, **loss**, but in this case no argument has been associated with it. Traditional SRL systems facing this type of examples are not able to fill the arguments of a predicate because their fillers are not in the same sentence of the predicate. Moreover, these systems also let unfilled arguments occurring in the same sentence, like in the following example:

(2) Quest Medical Inc said it adopted [*arg1* a shareholders' rights] [*np* **plan**] in which rights to purchase shares of common stock will be distributed as a dividend to shareholders of record as of Oct 23.

For the predicate **plan** in the previous sentence, a traditional SRL process only returns the filler for the argument *arg1*, the theme of the plan.

However, in both examples, a reader could easily infer the missing arguments from the surrounding context of the predicate, and determine that in (1) both instances of the predicate share the same arguments and in (2) the missing argument corresponds to the subject of the verb that dominates the predicate, *Quest Medical Inc*. Obviously, this additional annotations could contribute positively to its semantic analysis. In fact, [Gerber and Chai, 2010] pointed out that implicit arguments can increase the coverage of argument structures in NomBank by 71%.

The first attempt for the automatic annotation of implicit semantic roles was proposed by [Palmer *et al.*, 1986]. This work applied selectional restrictions together with coreference chains, in a very specific domain. In a similar approach, [Whittemore *et al.*, 1991] also attempted to solve implicit arguments using some manually described semantic constraints for each thematic role they tried to cover. Another early approach was presented by [Tetreault, 2002]. Studying another specific domain, they obtained some probabilistic relations between some roles. These early works agree that the problem is, in fact, a special case of anaphora or coreference resolution.

Recently, the task has been taken up again around two different proposals. On the one hand, [Ruppenhofer *et al.*, 2010] presented a task in SemEval-2010 that included an implicit argument identification challenge based on FrameNet [Baker *et al.*, 1998b]. The corpus for this task consisted in some novel chapters. They covered a wide variety of nominal and verbal predicates, each one having only a small number of instances. Only two systems were presented for this subtask obtaining quite poor results (F1 below 0,02). VENSES++ [Tonelli and Delmonte, 2010] applied a rule based anaphora resolution procedure and semantic similarity between candidates and thematic roles using WordNet [Fellbaum, 1998]. The system was tuned in [Tonelli and Delmonte, 2011] improving slightly its performance. SEMAFOR [Chen *et al.*, 2010b] is a supervised system that extended an existing semantic role labeler to enlarge the search window to other sentences, replacing the features defined for regular arguments with two new semantic features. Although this system obtained the best performance in the task, data sparseness strongly affected the results. Besides the two systems presented to the task, some other systems have used the same dataset and evaluation metrics. [Ruppenhofer *et al.*, 2011], [Laparra and Rigau, 2012], [Gorinski *et al.*, 2013] and [Laparra and Rigau, 2013c] explore alternative linguistic and semantic strategies. These works obtained significant gains over previous approaches. [Silberer and Frank, 2012] adapted an entity-based coreference resolution model to extend automatically the training corpus. Exploiting this additional data, their system was able to improve previous results. Following this approach [Moor *et al.*, 2013] present a corpus of predicate-specific annotations for verbs in the FrameNet paradigm that are aligned with PropBank and VerbNet.

On the other hand, [Gerber and Chai, 2010; Gerber and Chai, 2012] studied the implicit argument resolution on NomBank. They uses a set of syntactic, semantic and coreferential features to train a logistic regression classifier. Unlike the dataset from SemEval-2010 [Ruppenhofer *et al.*, 2010], in this work the authors focused on a small set of ten predicates. But for those predicates, they annotated a large amount of instances in the documents from the Wall Street Journal that were already annotated for PropBank [Palmer *et al.*, 2005b] and NomBank. This allowed them to avoid the sparseness problems and generalize properly from the training set. The results of this system were far better than those obtained by the systems that faced the SemEval-2010 dataset. This works represent the deepest study so far of the features that characterizes the implicit arguments [117]. However, many of the most important features are lexically dependent on the predicate and cannot been generalized. Thus, specific annotations are required for each new predicate to be analyzed.

[117] [Gerber and Chai, 2012] includes a set of 81 different features.

Finally, the most recent approach to this problem is the ImpAr[118] algorithm presented in [Laparra and Rigau, 2013b]. In that work, the authors studied the coherence of the predicate and argument realization in discourse. In particular, the followed a similar approach to the one proposed by [Dahl *et al.*, 1987] who filled the arguments of anaphoric mentions of nominal predicates using previous mentions of the same predicate. [Laparra and Rigau, 2013b] present an extension of this idea assuming that in a coherent document the different ocurrences of a predicate, including both verbal and nominal forms, tend to be mentions of the same event, and thus, they share the same argument fillers. Following this approach, ImpAr, a deterministic algorithm, was developed that obtains competitive results with respect to supervised methods, moreover, ImpAr can be potentially applied to any predicate without training data.

10 Recognising and Interpreting Time

Recognising and interpreting temporal expressions is a vital task to information extraction as it allows us to ground extracted information in time. Recognition (or detection) of temporal expressions is concerned with identifying phrases in text that express a date or time, and possibly a temporal relationship. Interpreting temporal expressions is concerned with normalising temporal expressions in text to a common format and disambiguate them in cases of underspecified temporal expressions (such as 'yesterday' which can only be grounded with respect to the date of the utterance). Some tools only perform one of the two subtasks, others attempt to recognise and interpret temporal expressions within one system. In the domain of recognising temporal expressions, machine learning methods dominate, whereas for the full task of recognising and interpreting temporal expressions,rule-based methods dominate [Negri and Marseglia, 2004].

10.1 Resources

Several temporal corpora have been created over the years, most of which adhere to some version of the TimeML temporal annotation standard. TimeBank started as an illustration and proof of concept of the TimeML specifications. TimeBank 1.1 was created in the early days of TimeML and follows the 1.1 version of the specifications. The more recent TimeBank 1.2 and the AQUAINT corpus were compiled following the 1.2.1 specifications. The TempEval1 corpus was created for the SemEval-2007 workshop[119] at the ACL 2007 Conference[120] in Prague, Czech Republic. It contains the same documents as TimeBank 1.2 but uses a simplified set of temporal relations, grouped in three separate tasks. The Tempeval2 corpus is a multilingual corpus created for the Semeval-2010 workshop[121] in Up-

[118] `http://adimen.si.ehu.es/web/ImpAr`
[119] `http://nlp.cs.swarthmore.edu/semeval/`
[120] `http://ufal.mff.cuni.cz/acl2007/`
[121] `http://stel.ub.edu/semeval2010-coref/`

psala, Sweden. TempEval3 was created for the SemEval-2013 workshop[122] in conjunction with *SEM 2013 Conference[123] in Atlanta, GA, USA.

An overview of the most important temporal corpora is given in Table 12.

Name	Description	Annotation	URL
MUC-6	The corpus from the 6th Message Understanding Conference, available at LDC under the catalogue number LDC2003T13.	MUC-6 TIMEX	`http://www.ldc.upenn.edu/Catalog/catalogEntry.jsp?catalogId=LDC2003T13`
MUC-7	The corpus from the 7th Message Understanding Conference, available at LDC under the catalogue number LDC2001T02.	MUC-7 TIMEX	`http://www.ldc.upenn.edu/Catalog/catalogEntry.jsp?catalogId=LDC2001T02`
ACE-2004	This is the corpus used at the Automatic Content Extraction (ACE) evaluations in 2004, available at LDC under the catalogue number LDC2005T07.	TIMEX2 2003 v.1.3 (April 2004)	`http://www.ldc.upenn.edu/Catalog/catalogEntry.jsp?catalogId=LDC2005T07`
ACE-2005 Dev	This is the corpus used at the Automatic Content Extraction (ACE) evaluations in 2005, available at LDC under the catalogue number LDC2006T06.	TIMEX2 (April 2005)	`http://www.ldc.upenn.edu/Catalog/catalogEntry.jsp?catalogId=LDC2006T06`

[122] `http://www.cs.york.ac.uk/semeval-2013/`
[123] `http://clic2.cimec.unitn.it/starsem2013/`

ACE-2007 Dev	This was the development corpus, consisting of selected domains in Arabic and Spanish only, used at the Automatic Content Extraction (ACE) evaluations in 2007. Corpora does not seem available anymore.	TIMEX2 (April 2005)	
TimeBank 1.1	The TimeBank corpus in the 1.1 version, used to be available from the mitre website.	TIMEX3 (TimeML 1.1)	`http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2006T08`
TimeBank 1.2	The TimeBank corpus in the 1.2 version, available at LDC under the catalogue number LDC2006T08.	TIMEX3 (TimeML 1.2.1)	`http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2006T08`
AQUAINT TimeML Corpus	The AQUAINT TimeBank contains 73 news report documents. It is very similar in content to, and uses the same specifications as, TimeBank 1.2.	TIMEX3 (TimeML 1.2.1)	`http://www.timeml.org/site/timebank/aquaint-timeml/aquaint_timeml_1.0.tar.gz`
WikiWars	A corpus of English Wikipedia articles about wars.	TIMEX2 (Sep 2005)	`http://www.timexportal.info/wikiwars`

ModeS TimeBank 1.0	This is a corpus of Modern Spanish (17th and 18th centuries) annotated with temporal and event information expressed in TimeML markups and annotated with spatial information following the SpatialML scheme.	TIMEX3 (TimeML)	`http://www.ldc.upenn.edu/Catalog/catalogEntry.jsp?catalogId=LDC2012T01`
TempEval1	Temporal relation task Semeval 2007	TIMEX3	`http://www.timeml.org/site/timebank/tempeval/tempeval_training.tar.gz`
TempEval2	Semeval 2010 Languages: Chinese, English, French, Italian, Korean and Spanish	TIMEX3	download: `http://www.timeml.org/site/timebank/tempeval/tempeval2-data.zip`
TempEval3	Semeval 2013, Languages: English, Spanish	TIMEX3	`http://www.cs.york.ac.uk/semeval-2013/task1/index.php?id=data`

Table 12: Resources for Temporal Information Extraction

10.2 Tools

In Table 13, the current-state-of-the-art tools for temporal information extraction are presented.

Name	Creator	Language	Type	Software available	License	URL
HeidelTime	[Strötgen and Gertz, 2013]	English, German, Dutch, Vietnamese, Arabic, Spanish and Italian	rule-based	yes	GPLv3	`https://code.google.com/p/heideltime/` `http://heideltime.ifi.uni-heidelberg.de/heideltime/`
ManTime	[Filannino *et al.*, 2013]	English	CRF+rule-based normaliser	demo + download	GPLv2	`http://www.cs.man.ac.uk/~filannim/projects/tempeval-3/`
SUTime	[Chang and Manning, 2013]	English	rule-based	demo+download	GPLv2	`http://nlp.stanford.edu/software/sutime.shtml`
ClearTK	[Bethard, 2013]	English	SVM	download	BSD-3 clause	`https://code.google.com/p/cleartk/`
Timextag	[Ahn *et al.*, 2007]	English	SVM	download	LGPL	`http://ilps.science.uva.nl/resources/timextag`

Timen	[Llorens *et al.*, 2012]	English	rule-based	download	AGPL / Apache	http://www.timen.org/
TipSem	[Llorens *et al.*, 2010]	English, Spanish	CRF	download	educational/ research purposes; TreeTagger & Freeling license cond.	http://www.timexportal.info/tipsem
Tarsqi	http://www.timeml.org/site/tarsqi/index.html	English	rule-based	download	CC BY-NC-SA 3.0 US	http://www.timeml.org/site/tarsqi/toolkit/index.html
TextPro	FBK	English, Italian	SVM	available to the project	Free for research, propetary otherwise	http://textpro.fbk.eu/

Table 13: Tools for Temporal Information Extraction

11 Factuality Module for Events

To distinguish between factual information and speculative information, the NWR pipeline requires a factuality module. This module is to classify whether an article, utterance or extracted event happened, or has not happened (yet). Determining the factuality score of an utterance in text is a task that has not yet received much attention in the research community, hence resources and tools are scarce.

11.1 Resources

The main resource for factuality detection is FactBank[Saurí and Pustejovsky, 2009b][124]. FactBank is a resource containing annotations that indicate whether an event mention describes actual situations in the world, situations that have not happened, or situations of uncertain interpretation. FactBank was built on top of TimeBank (see Section 10), as tense and other temporal markers play a vital role in determining factuality.

11.2 Tools

The few automatic factuality detection methods that are known to us are still in experimental state. [Saurí and Pustejovsky, 2012] describe De Facto, an algorithm that determines the factuality of an event based on the source of the utterance, factuality markers (such as modality markers), and context values constructed from the surrounding syntax. [van den Hoven *et al.*, 2010] describe a machine learning based approach that is aimed to detect in news articles whether a strike that is discussed took place or did not using linguistic features. Neither tools are available for download.

12 Event Detection and Classification

12.1 Event types

Events have been studied in linguistics for a long time [Tenny and Pustejovsky, 2000]. Nevertheless, the detection and classification of events is mostly not considered as a separate task in NLP. Most research on *event detection* refers to the detection of significant or relevant signals within a stream of data (both textual such as twitter, and non-textual such as sensor-based). As for the analysis of the text itself, most tools and approaches simply assume that all verbs represent events. Some other tools also consider nominalizations and abstract nouns but this requires some type of resource to distinguish nouns that can denote events from nouns that do not. The main goal of these tools is to extract more detailed information in addition to the main predicate such as: semantic roles, even-participant relations, event-relations, semantic parsing.

[124] `http://www.ldc.upenn.edu/Catalog/catalogEntry.jsp?catalogId=LDC2009T23`

An important framework for the definition of events in NLP is TimeML [Pustejovsky *et al.*, 2010]. TimeML is an international standard (ISO 24617-1:2009, ISO-TimeML) for annotating the event and temporal structure of a text. It provides a standard definition of *event* that has been used in the annotation the TimeBank corpus [Pustejovsky *et al.*, 2003] and for the annotation of news events in the 2010 TempEval competition [Verhagen *et al.*, 2010a]. TimeML ([Pustejovsky *et al.*, 2010], page 2-3) defines events in the following way:

> *Events* are considered as a cover term for situations that happen or occur. Events can be punctual (1-2) or last for a period of time (3-4). We also consider as events those predicates describing states or circumstances in which something obtains or holds true (5):
>
> 1. 1. Ferdinand Magellan, a Portuguese explorer, first [event reached] the islands in search of spices.
> 2. 2. A fresh flow of lava, gas and debris [event erupted] there Saturday.
> 3. 3. 11,024 people, including local Aeta aborigines, were [event evacuated] to 18 disaster relief centers.
> 4. 4. "We are [event expecting] a major eruption," he said in a telephone interview early today.
> 5. 5. Israel has been scrambling to buy more masks abroad, after a [event shortage] of several hundred thousand gas masks. [Roser Saurí and Pustejovsky, 2005]
>
> TimeML events may be expressed by tensed (erupted) and untensed (expecting) verbs, nominalizations (invasion), predicative clauses (is the president), adjectives (dormant) or prepositional phrases (on board).

The TimeML guidelines also consider direct speech, negated, hypothetical and modal events and even light verbs and aspectuals as events that need to be marked.

What expressions and words in text are considered events is also determined by the available semantic resources. For English, a large variety of resources is available that indirectly define what words and expressions qualify as events. These resources can be divided into annotated text corpora and lexical/ontological resources. The most well-known text corpora with event annotations are TimeBank [Pustejovsky *et al.*, 2006], FactBank [Saurí and Pustejovsky, 2009a], PropBank [C. *et al.*, 2010] and NomBank [Meyers *et al.*, 2004]. Whatever is labeled as an event in these corpora implicitly defines what an event is and what does not count as an event. In the case of PropBank, each verb in the Penn Treebank tree denotes an event, whereas NomBank specifies all nominals in Penn Treebank that reflect a predicate-argument structure. Not all nouns are 'markable' according to the NomBank guidelines. Three conditions are given for markable nouns:

1. a Noun Phrase (NP) must contain at least one (unincorporated) argument of the head noun.

2. The head of the NP must be of a propositional type (representing an event, state, etc.) and the NP must contain at least one proposition-modifying adjunct.

3. The head of the NP takes an argument that matches the arguments of verbal predicates in clauses.

The TimeBank corpora (TimeBank 1.2 (183 news articles), the AQUAINT corpus (73 news reports), and TempEval1 and 2 (multilingual)) are annotated according to the TimeML guidelines. These include both verbal, nominal and adjectival constructs.

Even though the actual annotation in these corpora provides a resource of classified events, most tools also rely on generic lexical and ontological resources that define expressions as events independently of an annotated corpus. Again, the most elaborate classifications of predicates are available for English: VerbNet [Kipper *et al.*, 2006], WordNet [Fellbaum, 1998], and FrameNet [Baker *et al.*, 1998b].

VerbNet provides 274 semantic classes, originally based on [Levin, 1993], for 3,769 English verbs. These classes have some further structuring in sub-classes but overall the typing is shallow. FrameNet provides over a 1,000 frames for nearly 12K lexical units, most of which belong to verbs. FrameNet does not provide an overarching model for events. In order to find the event denoting lexical units, a separate classification need to be made on top of the 1,000 frames. WordNet is the largest semantic resource available for English and also for many other languages. However, WordNet does not have a well-defined hierarchical structure either. In order to find all the events, a range of hypernym synsets need to be selected manually (e.g. the nominal synset S: (n) event: something that happens at a given place and time) with the assumption that they govern all event-denoting predicates in the language and no more than that. SemLink, [Kipper K. and M., 2009], provides mappings across Verbnet, Wordnet, FrameNet and Propbank. This provides a more complete typology of events but still it is a loose framework.

Some more structured top-down definitions of events are provided by larger ontologies, especially when linked to for example WordNet, most notably: SUMO [Niles and Pease, 2001] and DOLCE [C. *et al.*, 2002]. SUMO has a single top class Process that subsumes all event concepts. DOLCE has a top class endurant that subsumes all statives and all dynamic events. SUMO has been mapped to WordNet, [Niles and A., 2003], and likewise, all predicates that are somehow related to the Process class (directly or through subclass or hypernym relations) can be considered to be able to refer to events. The KYOTO project[125] resulted in an extension of the DOLCE and to a comprehensive mapping of the classes to WordNet [Laparra E. and P., 2012]. Likewise, it is possible to find all event denoting predicates through the DOLCE endurant class. In general, any semantic typing that has been assigned to the English WordNet can be transferred to any other language with a wordnet linked to the English WordNet. The above typologies of events are thus in principle reusable for other languages in NewsReader.

Finally, a completely different way of defining events is provided by the International

[125] `www.kyoto-project.eu`

Press Telecommunications Council.[126] They defined a thesaurus that classifies news events in terms of 1,405 topics spread over 3 levels. These topics are loose thematic groupings but they are oriented towards events in reality and as such useful as a classification of events. Unfortunately, IPTC classes have not been mapped to any language resources so far.

In addition to determining which expression refers to an event in text, we also may need to decide on the specific type of event. Most of the above resources do provide more specific subtypes. These subdivisions follow different insights and approaches in linguistic and semantic theories and are often created for different purposes. There is no uniform and standardized system for typing events. However for NewsReader there is one distinction that is more important. News reporting on events that took place in the world of today or that may take place there are 3 broad categories of events that need to be distibhuished:

1. Speech acts and mental events that indicate the provenance of the information that is expressed and their private state or opinion towards the information.
2. Grammatical constructions, mostly using verbs, that do not represent separate events in reality but properties of events or relations between events expressed in their adjuncts.
3. Events describing the world around us about which the news articles report.

These distinctions are partially found also in FactBank [Saurí and Pustejovsky, 2009a]. FactBank is a corpus annotated with information concerning the factuality of events. It identifies the most common linguistic devices to express the factuality of events, for which [Saurí and Pustejovsky, 2009a] introduce the notion of event-selecting predicates (ESPs). ESPs are defined as "predicates (either verbal, nominal, or adjectival) that select for an argument denoting an event of some sort" ([Saurí and Pustejovsky, 2009a]:234). Saurí and Pustejovsky distinguish two types of ESPs: source introducing predicates (SIPs) and non-source introducing predicates (NSIPs). SIPs correspond to our type 1 event and introduce the agents of speech acts, holders of opinions, experiencers of psychological reactions etc. as an additional source relative to which the factuality of the embedded event is assessed. In other words: these sources are committing to the factuality of the event. The NSIPs do not introduce a source and correspond to our type 2 event. Examples of these are auxiliaries expressing tense (*be, have, will*) and modal properties (*do, do not, can, will*) and expressions for aspectual properties (*start, continue, stop*). Within the events describing the world (type 3), any further differentiation can be adopted as far as needed by the applications that will use the data. This depends on what actually occurs in the data collections used and what groupings are most appropriate.

12.2 Tools

As explained before, there are not many tools that only do event detection and classification. On the one hand, there are text classification tools that determine the overall topic

[126] `http://www.iptc.org/site/Home/`

of the event, and on the other hand, there are many tools that do a deeper analysis of the event-argument structure of expressions and detect the event as a subtask.

The Evita (*Events InText Analyzer*), [Roser Saurí and Pustejovsky, 2005], is an event recognition system developed under the ARDA-funded TARSQI research framework. TARSQI is devoted to the more complex task of parsing text to TimeML specifications. Within TARSQI's framework, Evita's role is locating and tagging all event-referring expressions in the input text that can be temporally ordered. Evita combines linguistic- and statistically-based techniques to better address all subtasks of event recognition. For example, the module devoted to recognizing temporal information that is expressed through the morphology of certain event expressions (such as tense and aspect) uses grammatical information, whereas disambiguating nouns that can have both eventive and non-eventive interpretations is carried out by a statistical module. The functionality of Evita breaks down into two parts: event identification and analysis of the event-based grammatical features that are relevant for temporal reasoning purposes. Both tasks rely on a preprocessing step which performs part-of-speech tagging and chunking, and on a module for clustering together chunks that refer to the same event.

Another approach that performs direct event detection was developed during the KYOTO project. In KYOTO, a sequence of modules was developed in which expressions in text ultimately are typed any ontology linked to WordNet [Vossen P. and A., 2013]. What classifies as an event is the result of the decisions made by the POS tagging, the WSD (scoring each synset) and the ontological mapping. This was demonstrated for various ontologies linked to wordnet, among which the extension to DOLCE developed in KYOTO.

SEMAFOR, [Chen *et al.*, 2010a], was developed for frame-semantic parsing, assigning FramNet frames and frame elements to text. It treat the task as a structure prediction problem. It finds words that evoke FrameNet frames, selects frames for them, and locates the arguments for each frame. The system uses two feature-based, discriminative probabilistic (log-linear) models, one with latent variables to permit disambiguation of new predicate words. They use a probabilistic framework that cleanly integrates the FrameNet lexicon and available training data. The training data comes from the SemEval'07 task.[127] For comparison, the MATE tool [Björkelund *et al.*, 2009b], that assigns Propbank annotations to text through a pipeline of basic processing (involving lemmatization, part-of-speech tagging, dependency parsing), assumes that the predicates are already identified and only assigns the argument structure for each predicate.

Other systems consider event-detection and classification within a more narrow perspective of a specific task. For example, [Bethard and Martin, 2006] describe a system for detecting events in a question-answer system. They determine which expressions are events and what their type is based on TimeBank using the subclasses OCCURRENCE, PERCEPTION, REPORTING, ASPECTUAL, STATE, I_STATE, I_ACTION, and MODAL. They view event identification as a classification task using a word-chunking paradigm implemented using SVM and a wide range of features. The training data was derived from

[127] `http://framenet.icsi.berkeley.edu/semeval/FSSE.html`

TimeBank.

13 Event Coreference

The task of event-coreference is only partly comparable to coreference for entities 5 . Whereas entities may be found in external databases and otherwise are more stable and fixed, events are seldom listed in resources and have less clear boundaries. Events are usually not referred to by names and often also other expressions play a role in defining the events than just the main verb or noun phrase. Likewise, the variation in referring to the same event is much bigger and the process is more complex.

In recent years, event-coreference received more and more attention, e.g. [Bejan and Harabagiu, 2010], [Chen *et al.*, 2011] and [Lee *et al.*, 2012]. Bejan and Harabagiu use nonparametric Bayesian models, employing a combination of lexical, class and WordNet features (WordNet synonyms and super-senses) as well as predicate – argument structures. On the ACE (2005, restricted set of event types) data set, they achieved the highest results of 83.8% B3 F-score (B3 [Bagga and Baldwin, 1998b]) / 76.7% CEAF F [Luo, 2005b]. On their newly created EventCorefBank (articles on 43 different topics from the GoogleNews archive) they reached ca. 90% B3 and 86.5% CEAF F-score. Chen et al propose a framework for resolution of co-reference between event actions and their objects. They employ support vector machine with tree kernels and spectral graph partitioning. They use a combination of lexical, PoS, semantic and syntactic features (amongst others an argument matching feature to account for different syntactic structures and a semantic type feature with types such as person, location etc). Within-document-coreference is solved between descriptions of events and objects with 46.91% B3 F-score on the OntoNotes 2.0 corpus, annotated with coreference between all event mentions (not using any predefined concept types as in the ACE corpus).

These approaches do not explicitly account for partial coreference of events, where some of the event components are related through hyponymy or part-of relationship. Bejan and Harabagiu noted in their paper that not accounting for partial coreference is the reason for one of the common errors in their output. The approach of Chen et al accounts for synonymy relations between mentions but also neither for meronymy nor hyponymy relations.

Soft matching has been successfully used for entity coreference coreference resolution. Semantic similarity measures based on WordNet taxonomy as well as semantic relatedness (Wikipedia based) were used as features in a machine learning approach to entity coreference by [Ponzetto and Strube, 2006a]. Some semantic features based on synset relationships in WordNet are used by [Ng and Cardie, 2002b] and [Ng, 2005], while [Harabagiu *et al.*, 2001] use hyponymy, meronymy and other semantic relations from WordNet for NP coreference. They employ WordNet to distinguish between individuals and groups amongst entities of the semantic category PERSON.

Lee et. al [Lee *et al.*, 2012] merge entities and event clusters by means of linear regression, using semantic role dependencies as features. Event coreference is boosted depending on the number of shared arguments. Partial coreference is incorporated into this study by using distributional similarity as one of the features for cluster comparison. This approach achieved 62.7% MUC [Vilain *et al.*, 1995b] / 67.7% B3 / 33.9% (entity based) CEAF / 71.7% BLANC F-score on the extended version of the ECB corpus. Lee et. al. employ here the idea of modeling coreference resolution of events and entities jointly in an explicit way, while other approaches tend to use entities for event coreference in an indirect way for instance [Bejan and Harabagiu, 2008] and [Bejan and Harabagiu, 2010] by using semantic roles as features for their SVM multi-class classifiers. [Bejan and Harabagiu, 2010] account for synonymy amongst heads of semantic roles within the task of event coreference. And Chen and Ji ChenJi+'09 check for verbal argument compatibility and whether there are conflicts in the value of arguments with Time-Within and Place roles. Chen and Ji results indicate that features referring to event arguments only slightly (ca. plus 1% MUC, B3 and ECM F-score) improve event coreference, but possibly due to incorrect argument labeling.

A theory-oriented discussion about the nature of full identity, near-identity and non-identity and a continuum approach to entity coreference is presented in [Recasens *et al.*, 2011]. A discussion of full and quasi identity of events, pointing out the significance of partial coreference for coreference resolution, is held in [Hovy *et al.*, 2013]. Full identity and partial coreference as well as event membership and subevent relations between events in text were the focus of a study which resulted in creation of gold standard annotation of two corpora – the Intelligence Community (IC) Corpus, annotated with within-document violent event coreference, membership and subevent relations, and the Biography (Bio) Corpus, annotated with inter-textual full and quasi event coreference.

Using semantic shifts in NLP applications is not a new idea: [Mulkar-Mehta, 2011] investigated granularity shifts and granularity structures in natural language text. They focused on modeling part-whole relations between entities and events and causal relations between coarse and fine granularities. Finally, [Howald and Abramson, 2012] use granularity types as features for prediction of rhetorical relations. Their results show that inclusion of granularity types significantly improves the performance of prediction of rhetorical relations amongst clauses. In our work, we use shifts in granularity but also in abstraction for the purpose of event coreference resolution. Likewise, [Cybulska and Vossen, 2013] combine granularity with similarity to model fine and coarse-grained matches across event descriptions that are likely to happen across different documents and sources. In their approach, event co-reference is based on action matches, participant overlap and time and location matches. Matches take hypnymic relations and granularity shifts into account.

13.1 Data Sources

Data Entity	Type of data	How it is provided	Language
Intelligence Community (IC) Corpus	Newswire	annotated with within-document violent event coreference, membership and subevent relations	English
Biography (Bio) Corpus	Biographies	annotated with intertextual full and quasi event coreference	English
EventCorefBank	Articles on 43 different topics from the GoogleNews archive		English

Table 14: Resources for Event Coreference

13.2 Tools

To our knowledge there are no off-the-shelf tools for event-coreference. This is partly due to the fact that the technology is still in its early stage and involves complex mixtures of technology and pre-processing.

14 Event Relations

The identification of event relations is the task of identifying the relation holding between two given events in context. This process takes in input the events detected and classified as described in Section 12 and delivers in output the types of pairwise relations holding between them. The main relations that will be considered in NewsReader are coreferential (see Section 13), temporal and causal ones. The two latter relations are the main focus of the current section.

14.1 Data Sources

14.1.1 Temporal relations

The most relevant resources for encoding temporal relations between events have been all created in the last year within the **TimeBank** project, following ISO-TimeML specification [Pustejovsky *et al.*, 2003]. For a complete list of such resources, see Section 10. In this framework, event relations are usually provided together with other additional information on event types and temporal expressions. For the TempEval evaluation campaigns [Verhagen *et al.*, 2007; Verhagen *et al.*, 2010b], TimeML-like annotations have been provided also for other languages such as Spanish [Saurı and Barcelona, 2010] and Italian [Caselli, 2010].

Temporal relations in TimeML are marked via TLINKs. Each event (or time) is assigned a unique identifier, and these identifiers are used by TLINK annotations to assign one of the following temporal relations: BEFORE, AFTER, INCLUDES, IS_INCLUDED, DURING, DURING_INV, SIMULTANEOUS, IAFTER, IBEFORE, BEGINS, BEGUN_BY, ENDS or ENDED_BY. Given the complexity of this temporal framework, TempEval competitions tried to simplify the annotation scheme, annotating only temporal relations in certain syntactic constructions (e.g. the main events in adjacent sentences) and adopting a simpler relation set: BEFORE, AFTER, OVERLAP, BEFORE-OR-OVERLAP, OVERLAP-OR-AFTER and VAGUE. However, during the last TempEval campaign ended in April 2013 [UzZaman *et al.*, 2013] the full set of TimeML temporal relations has been used instead of the coarse-grained version of previous editions.

14.1.2 Causal relations

Compared to temporal relations, less resources have been annotated with causal information, probably due to the lack of agreement on a standard annotation scheme for causal phenomena. In fact, causality can be expressed in several ways, for instance through causal signals such as "because", or by specific verbs of causation. It can also be left implicit, so that the reader can infer a causal relation between events from the discourse context. Several datasets are available, each of them capturing few specific aspects of causality. We list them below.

PropBank [Kingsbury and Palmer, 2002b]: Causal relations have been annotated in the form of predicate-argument relations, and tagged as ARGM-CAU. In this resource, relations are annotated between a verbal and a nominal event, where the latter is a syntactic dependent of the former. See for instance the following example: "The highway was [closed $_{Pred}$] [because of the snow$_{Argm-Cau}$]."

SemEval 2007 Task4 [Girju *et al.*, 2007]: Causal relations have been annotated, among other relations, between pairs of nominals in text. The training and test data include 210 manually tagged pairs[128]. It is to note that inter-annotator agreement on causal relations was the highest one among the 7 relations proposed in the task, being 86.1%.

SemEval 2012 Task7 [Gordon *et al.*, 2012]: The COPA (Choice Of Plausible Alternatives) data set created for this competition consists of 1,000 questions, each composed of a premise and two alternatives, where the task is to select the alternative that more plausibly has a causal relation with the premise. The data are available at: `http://people.ict.usc.edu/~gordon/copa.html`.

[128] `http://nlp.cs.swarthmore.edu/semeval/tasks/task04/description.shtml`

Corpus of Temporal-Causal Structure[Bethard *et al.*, 2008]: The corpus includes 1,000 event pairs annotated with temporal and causal relations (in parallel). All events are connected by "and". The corpus is available at: `http://verbs.colorado.edu/~bethard/treebank-verb-conj-anns.xml`. The event pairs have been annotated with the goal to investigate the overlap between causal and temporal relations when a highly ambiguous connective like "and" is used.

Penn Discourse TreeBank (PDTB) [The PDTB Research Group, 2008]. In PDTB, relations are not annotated between specific event pairs but between two text spans called *Arguments*. However if we extract the main predicate from such spans, we may straightforwardly derive relations between events. Causal relations in PDTB are identified when the situations described in `Argument1` and `Argument2` are causally influenced, but they are not in a conditional relation. Directionality is specified at the level of subtype with two different labels: "reason" ((‖`Arg2`‖<‖`Arg1`‖[129]) and "result" (‖`Arg1`‖<‖`Arg2`‖) specifying which situation is the cause and which the effect. The typical connective for the first relation subtype is indeed *because*. On the contrary, for the latter (i.e. "result") , typical connectives are *so that*, *therefore*, *as a result*.

If there is no causal influence between `Arg1` and `Arg2`, but `Arg2` provides rather a justification for the claim expressed in `Arg1`, another type of cause has been introduced, called "Pragmatic Cause". We report an example below:

> *Mrs Yeargin is lying* [because] **they found students in an advanced class a year earlier who said she gave them similar help**.

In PDTB, annotated relations are both implicit and explicit (e.g. marked by a causal connective).

TimeBank: Although TimeML does not foresee a specific link for causal constructions, the annotation guidelines provide instructions on how to annotate some of them through TLINKs. Specifically, when two events e_1 and e_2 are connected through a causative predicate e_c, e_1 and e_c are connected through an 'Identity' TLINK, while e_1 and e_2 are connected through a 'Before' TLINK. A set of causative predicates is listed including *cause, stem from, lead to, breed, engender, hatch, induce, occasion, produce, bring about, produce, secure.*

14.2 Tools

To our knowledge, no system has been made available that identifies relations between events. However, for some specific types of relations, some applications have been produced. This process has been boosted by the **TempEval** campaigns for the evaluation of temporal processing systems. In the last edition [UzZaman *et al.*, 2013], 5 participants took part to the subtasks related to the identification of temporal relations, namely *i)* identification of pairs of entities connected by a TLINK and relation classification, and *ii)* Classification of

[129]The symbol < used in the PDTB categories means "causes".

the temporal relation, given the gold entities and the pairs involved in a relation. In the first task, the best performing system (*ClearTK-2*) achieved F1 36.26, while in the second task the first-ranked system (*UTTime-1*) scored F1 56.45. All TimeML relations were included, which made the task much more difficult than in the past evaluation campaign editions. All participants used partially or fully machine learning-based systems, trained on TimeBank and AQUAINT. The task participants report also that using temporal inference typically increased systems recall. Morphosyntactic and lexical-semantic information was used by all systems, although semantic features were proved to be less effective than morphosyntactic ones, because the low performance of semantic parsers may affected the quality of the features.

Largely inspired by TimeML annotation, two systems for temporal processing have been developed within the Terence European Project[130], one for English and one for Italian. The systems annotate temporal and causal relations between events, as well as temporal expressions, signals and participants. A demo can be accessed at `http://ariadne.cs.kuleuven.be/TERENCEStoryService/`. The English version is largely based on the technology presented in [Kolomiyets *et al.*, 2012], while the Italian version is rule-based and relies on morphosyntactic and semantic information information provided by the TextPro NLP suite [Pianta *et al.*, 2008].

15 Structured Data RDF

15.1 Tools

Several tools are available to convert structured data (e.g., databases, spreadsheets) from an application-specific format into RDF for use with RDF tools and integration with other data. An up-to-date list is maintained on the W3C web site[131]. Next, we recap some of the most prominent approaches, especially in view of the typology of structured data that may be exploited in the project.

15.1.1 Databases-to-RDF

Triplify Triplify[132] is a tool that, by defining some relational database queries, enables to retrieve information from a database-driven web application, and to convert the results of these queries into RDF, JSON and Linked Data.

RDBToOnto RDBToOnto[133] allows to automatically generate fine-tuned OWL ontologies from relational databases. It allows to produce structured ontologies with deeper hierarchies by exploiting both the database schema and the stored data. It can be used

[130] `http://www.terenceproject.eu/`
[131] `http://www.w3.org/wiki/ConverterToRdf`
[132] `http://triplify.org/Overview`
[133] `http://sourceforge.net/projects/rdbtoonto/`

in conjunction with Triplify to generate highly accurate RelationalDB-to-RDF mapping rules.

Virtuoso Sponger The Virtuoso Sponger[134] is the Linked Data middleware component of the Virtuoso Triple Store. It generates Linked Data from a variety of data sources (including database-driven web application, e.g., CrunchBase), and supports a wide variety of data representation and serialization formats. Content from external data sources can be easily retrieved through the Virtuoso's SPARQL Query Processor.

15.1.2 XML-to-RDF

Krextor: The KWARC RDF Extractor Krextor[135] is an extensible XSLT-based framework for extracting RDF from XML. The translation is based on templates (a number of templates for some input formats is already provided) that maps the input schema of the XML file to RDF statements. The extracted RDF graph will in most cases be an outline of the semantic structure of an XML document, abstracting from the concrete syntax.

XML2RDF mapping The XML2RDF mapping[136], part of the ReDeFer project, allows to map XML content (XML instances) to RDF (RDF statements), enriching it with semantics. The semantics have to be explicited by mapping the XSD of the XML file to OWL (using the XSD2OWL tool). The XML2RDF mapping can be tested on-line in the ReDeFer project web page.

15.1.3 Spreadsheet-to-RDF

RDF Refine RDF Refine[137] support, by means of a graphical interface, exporting data of Google Refine projects as interlinked RDF data, so that they can be queried through SPARQL endpoint or stored in RDF repositories. The export functionality allows to define the intended structure of the RDF data by drawing a template graph. The exporter iterates through the project rows, evaluates expressions in the template graph and produces an equivalent RDF subgraph per row. The final result is the merge of all the subgraphs.

XLWrap XLWrap[138] is a spreadsheet-to-RDF wrapper which is capable of transforming spreadsheets to arbitrary RDF graphs based on a mapping specification. It supports Microsoft Excel and OpenDocument spreadsheets such as comma- (and tab-) separated value (CSV) files. It works both with files on a local filesystem, or available at some url.

[134] `http://virtuoso.openlinksw.com/dataspace/doc/dav/wiki/Main/VirtSponger`
[135] `http://kwarc.info/projects/krextor/`
[136] `http://rhizomik.net/html/redefer/xml2rdf/`
[137] `http://refine.deri.ie/docs`
[138] `http://xlwrap.sourceforge.net`

16 Conclusions

This deliverable provides a detailed survey about current availability of resources and tools to perform event detection for English, Dutch, Italian and Spanish. Event Detection (WP04) addresses the development of text processing modules that detect mentions of events, participants, their roles and the time and place expressions. Thus, text-processing requires basic and generic NLP steps, such as tokenization, lemmatization, part-of-speech tagging, parsing, word sense disambiguation, named entity and semantic role recognition for all the languages in NewsReader. Furthermore, named entities are as much as possible linked to possible Wikipedia pages as external sources (Wikification) and entity identifiers.

The semantic interpretation of the text is directed towards the detection of event mentions and those named entities that play a role in these events, including time and location expressions. This implies covering all expressions (verbal, nominal and lexical) and meanings that can refer to events, their participating named entities, time and place expressions but also resolving any coreference relations for these named entities and explicit (causal) relations between different event mentions. Processing events also implies the detection of expressions of factuality of event mentions and the authority of the source of each event mention. Now, we summarize the current state-of-the art with respect each task.

- **Named Entity Recognition and Classification** tools recognize information units such as names, including person, organization and location names, and numeric expressions including time, date, money and percent expressions. Nowadays, there are good tools and data for NERC in the news texts on the languages covered by this project.

- **Coreference resolution** is the task of linking noun phrases to the entities that they refer to. Most of the coreference systems have been developed for English. But, some systems are available for Dutch, Italian and Spanish too.

- Most of the work in **Named Entity Disambiguation** has been done in English. However, there are some multilingual tools such as DBpedia Spotlight. Moreover, the Wiki Machine performs Wikification in both English and Italian.

- **Word Sense Disambiguation** stands for labelling every word in a text with its appropriate meaning or sense depending on its context. Lately, graph-based WSD systems are gaining growing attention. These methods are language independent since only requires a local wordnet connected to the Princeton WordNet. For instance, using UKB it is possible to implement WSD modules for English, Dutch, Italian and Spanish.

- **Sentiment analysis** and **Opinion Mining** is concerned with analysing opinions, sentiments, evaluations, attitudes, and emotions in text. Current resources and tools allow the appropriate analysis of sentiments and opinions for the languages of the project.

- **Semantic Role Labeling** is a task involving recognition of semantic arguments of predicates on top of their syntactic constituents. Usual semantic roles include Agent, Patient, Instrument or Location. PropBank is the most used corpus for training SRL systems, but, depending on the language to deal with, different resources such as VerbNet and FrameNet provide a complementary prepective for the task. All these resources and tools are going to be considered within NewsReader.

- Recognising and interpreting **Temporal Expressions** is a vital task to information extraction as it allows us to ground extracted information in time. Most of the corpora follow the TimeML specification. HeidelTime is one of the few multilingual tools that could be used for all the languages of the project. However, FBK developed TextPro to deal with English and Italian. We will study which is the best option to deal for Dutch and Spanish.

- The NewsReader project requires of a module to classify whether an article, utterance or extracted event happened, or has not happened (yet). Determining the **Factuality** score of an utterance in text is a task that has not yet received much attention in the research community. Hence, resources and tools are scarce. As a consequence, the project will possibly create and implement its own resources.

- The **Detection and Classification of Events** is mostly not considered as a separate task in NLP. Most research on event detection refers to the detection of significant or relevant signals within a stream of data.

- The task of **Event-Coreference** is only partly comparable to coreference for entities. Whereas entities may be found in external databases and otherwise are more stable and fixed, events are seldom listed in resources and have less clear boundaries. Events are usually not referred to by names and often also other expressions play a role in defining the events than just the main verb or noun phrase. Likewise, the variation in referring to the same event is much bigger and the process is more complex. To our knowledge there are no off-the-shelf tools for event-coreference. Thus, a new tool to deal with even-coreference would be implemented within the project.

- The identification of **Event Relations** is the task of identifying the relation holding between two given events in context. This process takes in input the events detected and delivers in output the types of pairwise relations holding between them. The main relations that will be considered in NewsReader are coreferential, temporal and causal ones. To our knowledge, no system has been made available that identifies relations between events. However, for some specific types of relations, some applications have been produced.

This survey has helped in the specification of the requirements necessary to the first prototype to be delivered in month 9 of the project, deliverable D4.2.1 (Event detection, version 1). Deliverable D4.1 could be updated if new technologyis detected.

References

[Agerri and García-Serrano, 2010] R. Agerri and A. García-Serrano. Q-wordnet: extracting polarity from wordnet senses. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, 2010.

[Aggarwal and Zhai, 2012] Charu C Aggarwal and ChengXiang Zhai. A survey of text clustering algorithms. In *Mining Text Data*, pages 77–128. Springer, 2012.

[Agirre and Edmonds, 2006] Eneko Agirre and Philip Glenny Edmonds. *Word sense disambiguation: Algorithms and applications*, volume 33. Springer Science+ Business Media, 2006.

[Agirre and Lopez de Lacalle, 2004] Eneko Agirre and Oier Lopez de Lacalle. Publicly available topic signatures for all wordnet nominal senses. In *Proceedings of the 4rd International Conference on Language Resources and Evaluations (LREC). Lisbon, Portugal, pp. 1123-1126. ISBN: 2 - 9517408 - 1 - 6*, 2004.

[Agirre and Lopez deLacalle, 2009] Eneko Agirre and Oier Lopez deLacalle. Supervised domain adaptation for wsd. In *Proceedings of The 12th Conference of the European Chapter for Computational Linguistics (EACL09), pp 42-50. ISBN 978-1-932432-16-9"*, 2009.

[Agirre and Martínez, 2000] Eneko Agirre and David Martínez. Exploring automatic word sense disambiguation with decision lists and the web. In *Proceedings of the COLING workshop on Semantic Annotation and Intelligent Annotation*, Luxembourg, 2000.

[Agirre and Martinez, 2001] Eneko Agirre and David Martinez. Knowledge sources for word sense disambiguation. In *Proceedings of the Fourth International Conference TSD 2001, Plzen (Pilsen), Czech Republic. Published in the Springer Verlag Lecture Notes in Computer Science series. Václav Matousek, Pavel Mautner, Roman Moucek, Karel Tauser (eds.) Copyright Springer-Verlag. ISBN: 3-540-42557-8. "*, 2001.

[Agirre and Soroa, 2009] Eneko Agirre and Aitor Soroa. Personalizing pagerank for word sense disambiguation. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 33–41. Association for Computational Linguistics, 2009.

[Agirre and Stevenson, 2005] E. Agirre and M. Stevenson. Knowledge sources for word sense disambiguation. In *Word Sense Disambiguation: Algorithms, Applications and Trends. Kluwer*, 2005.

[Agirre *et al.*, 2008] Eneko Agirre, Timothy Baldwin, and David Martinez. Improving parsing and pp attachment performance with sense information. *Proceedings of ACL-08: HLT*, pages 317–325, 2008.

[Agirre *et al.*, 2009a] Eneko Agirre, Oier Lopez De Lacalle, Aitor Soroa, and Informatika Fakultatea. Knowledge-based wsd on specific domains: performing better than generic supervised wsd. In *Proceedings of the 21st international jont conference on Artifical intelligence*, pages 1501–1506. Morgan Kaufmann Publishers Inc., 2009.

[Agirre *et al.*, 2009b] Eneko Agirre, Oier Lopez deLacalle, and Aitor Soroa. Knowledge-based wsd on specific domains: Performing better than generic supervised wsd. In *Proceedigns of IJCAI. pp. 1501-1506. ISBN 978-1-57735-429-1."*, 2009.

[Agirre *et al.*, 2009c] Eneko Agirre, Arantxa Otegi, and Hugo Zaragoza. Using semantic relatedness and word sense disambiguation for (cl)ir. In *Working Notes of the Cross-Lingual Evaluation Forum, Corfu, Greece"*, 2009.

[Agirre *et al.*, 2010a] Eneko Agirre, Montse Cuadros, German Rigau, and Aitor Soroa. Exploring knowledge bases for similarity. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC10). European Language Resources Association (ELRA). ISBN*, pages 2–9517408, 2010.

[Agirre *et al.*, 2010b] Eneko Agirre, Oier Lopez deLacalle, Christiane Fellbaum, Shu-Kai Hsieh, Maurizio Tesconi, Monica Monachini, Piek Vossen, and Roxanne Segers. Semeval-2010 task 17: All-words word sense disambiguation on a specific domain. In *Proceedings of the 5th International Workshop on Semantic Evaluation. 75–80. Uppsala, Sweden."*, 2010.

[Ahn *et al.*, 2007] David Ahn, Joris van Rantwijk, and Maarten de Rijke. A cascased machine learning approach for interpreting temporal expressions. In *Proceedings of HLT-NAACL 2007*, 2007.

[Aikhenvald, 2004] Alexandra Y Aikhenvald. *Evidentiality*. Oxford University Press, 2004.

[Alfonseca and Manandhar, 2002] Enrique Alfonseca and Suresh Manandhar. An unsupervised method for general named entity recognition and automated concept discovery. In *Proceedings of the 1st International Conference on General WordNet*, 2002.

[Alonso *et al.*, 2005] Laura Alonso, Joanan Capilla, Irene Castellón, Ana Fernández-Montraveta, and Glòria Vázquez. The SENSEM project: syntactico-semantic annotation of sentences in spanish. In *Recent Advances in Natural Language Processing (RANLP'05)*, 2005.

[Artiles *et al.*, 2007] Javier Artiles, Julio Gonzalo, and Satoshi Sekine. The semeval-2007 weps evaluation: establishing a benchmark for the web people search task. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, SemEval '07, pages 64–69, Stroudsburg, PA, USA, 2007. Association for Computational Linguistics.

[Artiles *et al.*, 2009] Javier Artiles, Julio Gonzalo, and Satoshi Sekine. WePS 2 Evaluation Campaign: overview of the Web People Search Clustering Task. In *2nd Web People Search Evaluation Workshop (WePS 2009), 18th WWW Conference*, 2009.

[Artiles *et al.*, 2010] Javier Artiles, A. Borthwick, J. Gonzalo, S. Sekine, and E. Amigo. Weps-3 evaluation campaign: Overview of the web people search clustering and attribute extraction tasks. In *Proceedings of the Conference on Multilingual and Multimodal Information Access Evaluation*, 2010.

[Atserias *et al.*, 2004] Jordi Atserias, Luis Villarejo, and German Rigau. Spanish wordnet 1.6: Porting the spanish wordnet across princeton versions. In *LREC*, 2004.

[Bagga and Baldwin, 1998a] Amit Bagga and Breck Baldwin. Algorithms for scoring coreference chains. In *Proceedings of the 1st International Conference on Language Resources and Evaluation (LREC1998)*, pages 563–566, Granada, Spain, 1998.

[Bagga and Baldwin, 1998b] Amit Bagga and Breck Baldwin. Algorithms for scoring coreference chains. In *Proceedings of LREC*, 1998.

[Bagga and Baldwin, 1998c] Amit Bagga and Breck Baldwin. Entity-based cross-document coreferencing using the vector space model. In *Proceedings of the 17th international conference on Computational linguistics - Volume 1*, COLING '98, pages 79–85, Stroudsburg, PA, USA, 1998. Association for Computational Linguistics.

[Baker *et al.*, 1998a] Collin F Baker, Charles J Fillmore, and John B Lowe. The berkeley framenet project. In *Proceedings of the 17th international conference on Computational linguistics-Volume 1*, pages 86–90. Association for Computational Linguistics, 1998.

[Baker *et al.*, 1998b] Collin F. Baker, Charles J. Fillmore, and John B. Lowe. The berkeley framenet project. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*, ACL '98, pages 86–90, Montreal, Quebec, Canada, 1998.

[Bejan and Harabagiu, 2008] Cosmin Adrian Bejan and Sanda Harabagiu. A linguistic resource for discovering event structures and resolving event coreference. In *Proceedings of LREC*, 2008.

[Bejan and Harabagiu, 2010] Cosmin Adrian Bejan and Sanda Harabagiu. Unsupervised event coreference resolution with rich linguistic features. In *the 48th Annual Meeting of the Association for Computational Linguistics, Uppsala, Sweden*, 2010.

[Bell, 1991] Allan Bell. *The language of news media*. Blackwell Oxford, 1991.

[Bell, 1998] Allan Bell. The discourse structure of news stories. *Approaches to media discourse*, pages 64–104, 1998.

[Bengtson and Roth, 2008] Eric Bengtson and Dan Roth. Understanding the value of features for coreference resolution. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, pages 294–303, Honolulu, Hawaii, 2008.

[Bentivogli and Pianta, 2005] Luisa Bentivogli and Emanuele Pianta. Exploiting parallel texts in the creation of multilingual semantically annotated resources: the multisemcor corpus. *Natural Language Engineering*, 11(3):247–262, 2005.

[Bethard and Martin, 2006] Steven Bethard and James H. Martin. Identification of event mentions and their semantic class. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, EMNLP '06, pages 146–154, Stroudsburg, PA, USA, 2006. Association for Computational Linguistics.

[Bethard *et al.*, 2008] Steven Bethard, William Corvey, Sara Klingenstein, and James H. Martin. Building a corpus of temporal-causal structure. In European Language Resources Association (ELRA), editor, *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, may 2008.

[Bethard, 2013] Steven Bethard. Cleartk-timeml: A minimalist approach to tempeval 2013. In *Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 10–14, 2013.

[Bick, 2004] Eckhard Bick. A named entity recognizer for danish. In *LREC*, 2004.

[Björkelund *et al.*, 2009a] Anders Björkelund, Love Hafdell, and Pierre Nugues. Multilingual semantic role labeling. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning: Shared Task*, pages 43–48. Association for Computational Linguistics, 2009.

[Björkelund *et al.*, 2009b] Anders Björkelund, Love Hafdell, and Pierre Nugues. Multilingual semantic role labeling. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning: Shared Task*, CoNLL '09, pages 43–48, Boulder, Colorado, USA, 2009.

[Black *et al.*, 1998] William Black, Fabio Rinaldi, and David Mowatt. Facile: Description of the ne system used for muc-7. In *Proceedings of the 7th Message Understanding Conference*, 1998.

[Bontcheva *et al.*, 2002] Kalina Bontcheva, Marin Dimitrov, Diana Maynard, Valentin Tablan, and Hamish Cunningham. Shallow methods for named entity coreference resolution. In *Proceedings of the Traitement Automatique des Langues Naturelles*, TALN '02, pages 79–85, Nancy, France, June 24–27 2002.

[Borrega *et al.*, 2007] Oriol Borrega, Mariona Taulé, and MA Martı. What do we mean when we speak about named entities? In *Proceedings of Corpus Linguistics*, 2007.

[Bos, 2008] Johan Bos. Wide-coverage semantic analysis with boxer. In *Proceedings of the 2008 Conference on Semantics in Text Processing*, pages 277–286. Association for Computational Linguistics, 2008.

[Bosma *et al.*, 2009] Wauter Bosma, Piek Vossen, Aitor Soroa, German Rigau, Maurizio Tesconi, Andrea Marchetti, Monica Monachini, and Carlo Aliprandi. Kaf: a generic semantic annotation format. In *Proceedings of the GL2009 Workshop on Semantic Annotation*, 2009.

[Brockmann and Lapata, 2003] Carsten Brockmann and Mirella Lapata. Evaluating and combining approaches to selectional preference acquisition. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics"*, pages 27–34, Budapest, 2003.

[Brooke *et al.*, 2009] J. Brooke, M. Tofiloski, and M. Taboada. Cross-linguistic sentiment analysis: From english to spanish. In *Proceedings of RANLP 2009, Recent Advances in Natural Language Processing*, Borovets, Bulgaria, 2009.

[Bryl *et al.*, 2010] Volha Bryl, Claudio Giuliano, Luciano Serafini, and Kateryna Tymoshenko. Supporting natural language processing with background knowledge: coreference resolution case. In *The Semantic Web–ISWC 2010*, pages 80–95. Springer, 2010.

[Bunescu and Pasca, 2006] Razvan Bunescu and Marius Pasca. Using encyclopedic knowledge for named entity disambiguation. In *Proceesings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL-06)*, pages 9–16, Trento, Italy, 2006.

[C. *et al.*, 2002] Masolo C., Borgo S., Gangemi A., Guarino N., Oltramari A., and Schneider L. Wonderweb deliverable d17. the wonderweb library of foundational ontologies and the dolce ontology. Technical report, ISTC-CNR, 2002.

[C. *et al.*, 2010] Bonial C., Babko-Malaya O., Choi J., Hwang J., and Palmer M. Propbank annotation guidelines. Technical report, http://www.ldc.upenn.edu/Catalog/docs/LDC2011T03/propbank/english-propbank.pdf, 2010.

[Cambria *et al.*, 2010] Erik Cambria, Robert Speer, Catherine Havasi, and Amir Hussain. Senticnet: a publicly available semantic resource for opinon mining. In *Commonsense Knowledge: Papers from the AAAI Fall Symposium*, pages 14–18, 2010.

[Cambria *et al.*, 2012] Erik Cambria, Catherine Havasi, and Amir Hussain. Senticnet 2: A semantic and affective resource for opinion mining and sentiment analysis. In *Proceedings of Twenty-Fifth International FLAIRS Conference*, pages 202–207, 2012.

[Carpuat and Wu, 2007] Marine Carpuat and Dekai Wu. Improving statistical machine translation using word sense disambiguation. In *Proceedings of Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL 2007)*, Prague, 2007.

[Carreras and Màrquez, 2004] X. Carreras and L. Màrquez. Introduction to the CoNLL-2004 Shared Task: Semantic Role Labeling. In *Proceedings of the Eigth Conference on Computational Natural Language Learning (CoNLL-2004)*, pages 89–97, Boston, MA, USA, 2004.

[Carreras and Màrquez, 2005a] Xavier Carreras and Lluís Màrquez. Introduction to the conll-2005 shared task: Semantic role labeling. In *Proceedings of the Ninth Conference on Computational Natural Language Learning*, pages 152–164. Association for Computational Linguistics, 2005.

[Carreras and Màrquez, 2005b] Xavier Carreras and Lluís Màrquez. Introduction to the conll-2005 shared task: Semantic role labeling. In *Proceedings of the 9th Conference on Computational Natural Language Learning*, CoNLL '05, pages 152–164, Ann Arbor, Michigan, USA, 2005.

[Carreras *et al.*, 2003] Xavier Carreras, Lluís Màrquez, and Lluís Padró. A simple named entity extractor using adaboost. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003 - Volume 4*, CONLL '03, pages 152–155, Stroudsburg, PA, USA, 2003. Association for Computational Linguistics.

[Carreras *et al.*, 2004] Xavier Carreras, Isaac Chao, Lluís Padró, and Muntsa Padró. Freeling: An open-source suite of language analyzers. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC'04)*, 2004.

[Caselli, 2010] T Caselli. It-timeml: Timeml annotation scheme for italian version 1.3. Technical report, Technical Report, 2010.

[Cavnar and Trenkle, 1994] William B Cavnar and John M Trenkle. N-gram-based text categorization. *Ann Arbor MI*, 48113(2):161–175, 1994.

[Chai and Biermann, 1999] Joyce Yue Chai and Alan W. Biermann. The use of word sense disambiguation in an information extraction system. In *Proceedings of the sixteenth national conference on Artificial intelligence and the eleventh Innovative applications of artificial intelligence conference innovative applications of artificial intelligence*, AAAI '99/IAAI '99, pages 850–855, Menlo Park, CA, USA, 1999. American Association for Artificial Intelligence.

[Chan *et al.*, 2007] Yee Seng Chan, Hwee Tou Ng, and David Chiang. Word sense disambiguation improves statistical machine translation. In *Proceedings of the 45th Annual Meeting-Association for Computational Linguistics*, 2007.

[Chang and Manning, 2013] A. Chang and C. D. Manning. Sutime: Evaluation in tempeval-3. In *Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 78–82, 2013.

[Chaovalit and Zhou, 2005] P. Chaovalit and Lina Zhou. Movie review mining: A comparison between supervised and unsupervised classification approaches. In *Proceedings of the Hawaii International Conference on System Sciences (HICSS)*, 2005.

[Chapman *et al.*, 2001] Wendy W Chapman, Will Bridewell, Paul Hanbury, Gregory F Cooper, and Bruce G Buchanan. A simple algorithm for identifying negated findings and diseases in discharge summaries. *Journal of biomedical informatics*, 34(5):301–310, 2001.

[Charniak, 2000] Eugene Charniak. A maximum-entropy-inspired parser. In *Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference*, NAACL 2000, pages 132–139, Seattle, Washington, 2000.

[Chen and Palmer, 2009] Jinying Chen and Martha Palmer. Improving english verb sense disambiguation performance with linguistically motivated features and clear sense distinction boundaries, springer netherland: Semeval2007. *Language Resources and Evaluation*, 43:181—208, 2009.

[Chen *et al.*, 2010a] Desai Chen, Nathan Schneider, Dipanjan Das, and Noah A Smith. Semafor: Frame argument resolution with log-linear models. In *Proceedings of the 5th international workshop on semantic evaluation*, pages 264–267. Association for Computational Linguistics, 2010.

[Chen *et al.*, 2010b] Desai Chen, Nathan Schneider, Dipanjan Das, and Noah A. Smith. Semafor: Frame argument resolution with log-linear models. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, SemEval '10, pages 264–267, Los Angeles, California, USA, 2010.

[Chen *et al.*, 2011] Bin Chen, Jian Su, Sinno Jialin Pan, and Chew Lim Tan. A unified event coreference resolution by integrating multiple resolvers. In *Proceedings of the 5th International Joint Conference on Natural Language Processing, Chiang Mai, Thailand*, 2011.

[Chinchor, 1998] Nancy Chinchor. Overview of muc-7. In *Proceedings of the Seventh Message Understanding Conference*, MUC-7, pages 178–185, 1998.

[Ciaramita and Altun, 2005] M. Ciaramita and Y. Altun. Named-entity recognition in novel domains with external lexical knowledge. In *Proceedings of the NIPS Workshop on Advances in Structured Learning for Text and Speech Processing*, 2005.

[Ciaramita and Altun, 2006a] Massimiliano Ciaramita and Yasemin Altun. Broad-coverage sense disambiguation and information extraction with a supersense sequence tagger. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, EMNLP '06, pages 594–602, Sydney, Australia, 2006.

[Ciaramita and Altun, 2006b] Massimiliano Ciaramita and Yasemin Altun. Broad-coverage sense disambiguation and information extraction with a supersense sequence tagger. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 594–602. Association for Computational Linguistics, 2006.

[Clough and Stevenson, 2004] Paul Clough and Mark Stevenson. Cross-language information retrieval using eurowordnet and word sense disambiguation. In *Advances in information retrieval*, pages 327–337. Springer, 2004.

[Constant *et al.*, 2009] Noah Constant, Christopher Davis, Christopher Potts, and Florian Schwarz. The pragmatics of expressive content: Evidence from large corpora. *Sprache und Datenverarbeitung*, pages 5–21, 2009.

[Copestake, 2002] Ann Copestake. *Implementing typed feature structure grammars*, volume 110. CSLI publications Stanford, 2002.

[Cuadros and Rigau, 2008] M. Cuadros and G. Rigau. KnowNet: Building a Large Net of Knowledge from the Web. In *Proceedings of COLING*, 2008.

[Cucchiarelli and Velardi, 2001] Alessandro Cucchiarelli and Paola Velardi. Unsupervised named entity recognition using syntactic and semantic contextual evidence. *Comput. Linguist.*, 27(1):123–131, March 2001.

[Cucerzan, 2007] Silviu Cucerzan. Large-scale named entity disambiguation based on Wikipedia data. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 708–716, Prague, Czech Republic, June 2007. Association for Computational Linguistics.

[Cybulska and Vossen, 2013] Agata Cybulska and Piek Vossen. Semantic relations between events and their time, locations and participants for event coreference resolution. In *Proceedings of RANLP*, 2013.

[Dahl *et al.*, 1987] Deborah A. Dahl, Martha S. Palmer, and Rebecca J. Passonneau. Nominalizations in pundit. In *In Proceedings of the 25th Annual Meeting of the Association for Computational Linguistics*, ACL '87, pages 131–139, Stanford, California, USA, 1987.

[de Albornoz *et al.*, 2012] Jorge Carrillo de Albornoz, Laura Plaza, and Pablo Gervás. Sentisense: An easily scalable concept-based affective lexicon for sentiment analysis. In *Proceedings of LREC 2012*, 2012.

[DeSmedt and Daelemans, 2012] Bart DeSmedt and Walter Daelemans. "vreselijk mooi" (terribly beautiful): a subjectivity lexicon for dutch adjectives. In *Proceedings of LREC 2012*, 2012.

[Diab *et al.*, 2009] T. M. Diab, L. Levin, T. Mitamura, O. Rambow, V. Prabhakaram, and W. Guo. Committed belief annotation and tagging. In *Proceedings of the ACL-IJCNLP 2009*, pages 68–73. Citeseer, 2009.

[Dorr, 1997] Bonnie J. Dorr. Large-scale dictionary construction for foreign language tutoring and interlingual machine translation. *MACHINE TRANSLATION*, 12(4):271–322, 1997.

[Dredze *et al.*, 2010] Mark Dredze, Paul McNamee, Delip Rao, Adam Gerber, and Tim Finin. Entity disambiguation for knowledge base population. In *Proceedings of the 23rd International Conference on Computational Linguistics*, COLING '10, pages 277–285, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[Edmonds and Cotton, 2001] P. Edmonds and S. Cotton. Senseval-2: Overiew. In *Proceedings of Senseval-2; Second International Workshop on Evaluating Word Sense Disambiguation Systems*, pages 1–5, Toulouse, France, 2001.

[Edmonds and Kilgarriff, 2002] Philip Edmonds and Adam Kilgarriff. Introduction to the special issue on evaluating word sense disambiguation systems. *Nat. Lang. Eng.*, 8(4):279–291, 2002.

[Erk and Pado, 2006] Katrin Erk and Sebastian Pado. Shalmaneser - a flexible toolbox for semantic role assignment. In *Proceedings of LREC 2006*, Genoa, Italy, 2006.

[Escudero *et al.*, 2000] Gerard Escudero, Lluís Màrquez, and German Rigau. An empirical study of the domain dependence of supervised word sense disambiguation systems. In *Proceedings of Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC'00)*, Hong Kong, China., 2000.

[Esuli and Sebastiani, 2006] A. Esuli and F. Sebastiani. Sentiwordnet: A publicly available lexical resource for opinion mining. In *Proceedings of LREC-2006*, 2006.

[Etzioni *et al.*, 2005] Oren Etzioni, Michael Cafarella, Doug Downey, Ana-Maria Popescu, Tal Shaked, Stephen Soderland, Daniel S. Weld, and Alexander Yates. Unsupervised named-entity extraction from the web: an experimental study. *Artif. Intell.*, 165(1):91–134, June 2005.

[Evans and Street, 2004] Richard Evans and Stafford Street. A framework for named entity recognition in the open domain. *Recent advances in natural language processing III: selected papers from RANLP 2003*, 260:267, 2004.

[Fader *et al.*, 2009] Anthony Fader, Stephen Soderland, Oren Etzioni, and Turing Center. Scaling wikipedia-based named entity disambiguation to arbitrary web text. In *Proceedings of the IJCAI Workshop on User-contributed Knowledge and Artificial Intelligence: An Evolving Synergy, Pasadena, CA, USA*, pages 21–26. Citeseer, 2009.

[Fellbaum, 1998] Christiane Fellbaum. *WordNet: an electronic lexical database*. MIT Press, 1998.

[Ferragina and Scaiella, 2010] Paolo Ferragina and Ugo Scaiella. Tagme: on-the-fly annotation of short text fragments (by wikipedia entities). In *Proceedings of the 19th ACM international conference on Information and knowledge management*, CIKM '10, pages 1625–1628, New York, NY, USA, 2010. ACM.

[Ferro *et al.*, 2007] Lisa Ferro, Laurie Gerber, Inderjeet Mani, Beth Sundheim, and George Wilson. Tides 2005 standard for the annotation of temporal expressions (2005). *MITRE Corporation*, 2007.

[Filannino *et al.*, 2013] M. Filannino, G. Brown, and G. Nenadic. ManTIME: Temporal expression identification and normalization in the tempeval-3 challenge. In *Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 53–57, 2013.

[Fillmore and Baker, 2001] C.J. Fillmore and C.F. Baker. Frame semantics for text understanding. Proceedings of NAACL 2001, WordNet and Other Lexical Resources Workshop, Pittsburgh, Pennsylvania, June 2001.

[Fillmore and Baker, 2009] C. J. Fillmore and C. F. Baker. A frames approach to semantic analysis. In The Oxford Handbook of Linguistic Analysis, B. Heine and H. Narrog, eds., pp. 313-340, Oxford University Press, 2009.

[Fillmore *et al.*, 2003] C. Fillmore, C. Johnson, and M. Petruck. Background to FrameNet. *International Journal of Lexicography*, 16(3):235–250, 2003.

[Fillmore, 1976] C.J. Fillmore. Frame semantics and the nature of language. Annals of the New York Academy of Sciences: Conference on the Origin and Development of Language and Speech, Vol. 280, pp. 20-32, 1976.

[Fillmore, 1977] C.J. Fillmore. Scenes-and-frames semantics, linguistic structures processing. Zampolli, Antonio (Ed.): Fundamental Studies in Computer Science, No. 59, North Holland Publishing, pp. 55-88, 1977.

[Fillmore, 1982] C.J. Fillmore. Frame semantics. Linguistics in the Morning Calm, Seoul, South Korea, Hanshin Publishing Co., pp. 111-137, 1982.

[Fillmore, 1985] C. J. Fillmore. Frames and the semantics of understanding. Quaderni di Semantica, Vol. 6, No. 2, pp. 222-254, December 1985.

[Fleischman and Hovy, 2002] Michael Fleischman and Eduard Hovy. Fine grained classification of named entities. In *Proceedings of the 19th international conference on Computational linguistics - Volume 1*, COLING '02, pages 1–7, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics.

[Fox *et al.*, 1999] Barbara Fox, Dan Jurafsky, and Laura A. Michaelis. *Cognition and Function in Language*. CSLI Publications, Stanford, CA., 1999.

[Gale *et al.*, 1992a] W. A. Gale, Kenneth W. Church, and David Yarowsky. Estimating upper and lower bounds on the performance of word sense disambiguation. In *Proceedings of the 30th Annual Meeting of the Association for Computational Linguistics*. ACL, 1992.

[Gale *et al.*, 1992b] William Gale, Kenneth Church, and David Yarowsky. A method for disambiguating word senses in a large corpus. *Computers and the Humanities*, 26:415–439, 1992. 10.1007/BF00136984.

[García-Miguel and Albertuz, 2005] José M García-Miguel and Francisco J Albertuz. Verbs, semantic classes and semantic roles in the adesse project. In *Proceedings of the Interdisciplinary Workshop on the Identification and Representation of Verb Features and Verb Classes*, 2005.

[Gerber and Chai, 2010] Matthew Gerber and Joyce Y. Chai. Beyond nombank: a study of implicit arguments for nominal predicates. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, pages 1583–1592, Uppsala, Sweden, 2010.

[Gerber and Chai, 2012] Matthew Gerber and Joyce Chai. Semantic role labeling of implicit arguments for nominal predicates. *Computational Linguistics*, 38(4):755–798, December 2012.

[Gildea and Jurafsky, 2000] Daniel Gildea and Daniel Jurafsky. Automatic labeling of semantic roles. In *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, ACL '00, pages 512–520, Hong Kong, 2000.

[Gildea and Jurafsky, 2002a] Daniel Gildea and Daniel Jurafsky. Automatic labeling of semantic roles. *Computational Linguistics*, 28(3):245–288, 2002.

[Gildea and Jurafsky, 2002b] Daniel Gildea and Daniel Jurafsky. Automatic labeling of semantic roles. *Computational Linguistics*, 28(3):245–288, September 2002.

[Girju *et al.*, 2007] Roxana Girju, Preslav Nakov, Vivi Nastase, Stan Szpakowicz, Peter Turney, and Deniz Yuret. Semeval-2007 task 04: Classification of semantic relations between nominals. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, pages 13–18. Association for Computational Linguistics, 2007.

[Goldin and Chapman, 2003] I Goldin and Wendy W Chapman. Learning to detect negation with "not" in medical texts. In *Proc Workshop on Text Analysis and Search for Bioinformatics, ACM SIGIR*, 2003.

[Gordon *et al.*, 2012] Andrew S Gordon, Zornitsa Kozareva, and Melissa Roemmele. Semeval-2012 task 7: Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, pages 394–398. Association for Computational Linguistics, 2012.

[Gorinski *et al.*, 2013] Philip Gorinski, Josef Ruppenhofer, and Caroline Sporleder. Towards weakly supervised resolution of null instantiations. In *Proceedings of the 10th International Conference on Computational Semantics*, IWCS '13, pages 119–130, Potsdam, Germany, 2013.

[Grishman and Sundheim, 1996] Ralph Grishman and Beth Sundheim. Message understanding conference-6: a brief history. In *Proceedings of the 16th conference on Computational linguistics - Volume 1*, COLING '96, pages 466–471, Copenhagen, Denmark, 1996.

[Grosz *et al.*, 1995] Barbara J. Grosz, Aravind K. Joshi, and Scott Weinstein. Centering: A Framework for Modeling the Local Coherence of Discourse. *Computational Linguistics*, 21:203–225, 1995.

[Hachey *et al.*, 2013] Ben Hachey, Will Radford, Joel Nothman, Matthew Honnibal, and James R. Curran. Evaluating entity linking with wikipedia. *Artificial Intelligence*, 194(0):130–150, 2013.

[Haghighi and Klein, 2009] Aria Haghighi and Dan Klein. Simple coreference resolution with rich syntactic and semantic features. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3 - Volume 3*, EMNLP '09, pages 1152–1161, Singapore, 2009.

[Hajič *et al.*, 2009] Jan Hajič, Massimiliano Ciaramita, Richard Johansson, Daisuke Kawahara, Maria Antònia Martí, Lluís Màrquez, Adam Meyers, Joakim Nivre, Sebastian Padó, Jan Štěpánek, Pavel Straňák, Mihai Surdeanu, Nianwen Xue, and Yi Zhang. The CoNLL-2009 shared task: Syntactic and semantic dependencies in multiple languages. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning: Shared Task*, CoNLL '09, pages 1–18, Boulder, Colorado, USA, 2009.

[Harabagiu *et al.*, 2001] Sanda M. Harabagiu, Razvan C. Bunescu, and Steven J. Maiorano. Text and knowledge mining for coreference resolution. In *Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies*, 2001.

[Hobbs, 1977] Jerry R. Hobbs. Pronoun resolution. *Intelligence/sigart Bulletin*, pages 28–28, 1977.

[Hoffart *et al.*, 2011] Johannes Hoffart, Mohamed Amir Yosef, Ilaria Bordino, Hagen Fürstenau, Manfred Pinkal, Marc Spaniol, Bilyana Taneva, Stefan Thater, and Gerhard Weikum. Robust disambiguation of named entities in text. In *Proceedings of the*

Conference on Empirical Methods in Natural Language Processing, EMNLP '11, pages 782–792, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

[Hovy *et al.*, 2006] Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. Ontonotes: the 90% solution. In *Proceedings of the human language technology conference of the NAACL, Companion Volume: Short Papers*, pages 57–60. Association for Computational Linguistics, 2006.

[Hovy *et al.*, 2013] Eduard Hovy, Teruko Mitamura, Felisa Verdejo, Jun Araki, and Andrew Philpot. Events are not simple: Identity, non-identity, and quasi-identity. In *Proceedings of the NAACL HLT 2013*, 2013.

[Howald and Abramson, 2012] Blake Stephen Howald and Martha Abramson. The use of granularity in rhetorical relation prediction. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, pages 44–48. Association for Computational Linguistics, 2012.

[Huang and Lowe, 2007] Yang Huang and Henry J Lowe. A novel hybrid approach to automated negation detection in clinical radiology reports. *Journal of the American Medical Informatics Association*, 14(3):304–311, 2007.

[Hyland, 1998] Ken Hyland. *Hedging in Scientific Resarch Articles*, volume 54. John Benjamins Publishing Company, 1998.

[Ide and Véronis, 1998] Nancy Ide and Jean Véronis. Word sense disambiguation: The state of the art. *Computational Linguistics*, 24:1–40, 1998.

[Izquierdo *et al.*, 2009] Rubén Izquierdo, Armando Suárez, and German Rigau. An empirical study on class-based word sense disambiguation. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 389–397. Association for Computational Linguistics, 2009.

[Izquierdo *et al.*, 2010] Rubén Izquierdo, Armando Suárez, and German Rigau. Gplsi-ixa: Using semantic classes to acquire monosemous training examples from domain texts. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 402–406, Uppsala, Sweden, July 2010. Association for Computational Linguistics.

[Jiang and Conrath, 1997] Jay J Jiang and David W Conrath. Semantic similarity based on corpus statistics and lexical taxonomy. *arXiv preprint cmp-lg/9709008*, 1997.

[Jijkoun and Hofmann, 2009] V. Jijkoun and K. Hofmann. Generating a non- english subjectivity lexicon: Relations that matter. In *Proceedings of EACL-2009*, 2009.

[Kamp and Reyle, 1993] Hans Kamp and Uwe Reyle. *From discourse to logic: Introduction to modeltheoretic semantics of natural language, formal logic and discourse representation theory.* Kluwer Academic Pub, 1993.

[Kingsbury and Palmer, 2002a] Paul Kingsbury and Martha Palmer. From treebank to propbank. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC-2002)*, pages 1989–1993. Citeseer, 2002.

[Kingsbury and Palmer, 2002b] Paul Kingsbury and Martha Palmer. From treebank to propbank. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC-2002)*, pages 1989–1993. Citeseer, 2002.

[Kipper *et al.*, 2000] K. Kipper, H. T. Dang, and M. Palmer. Class Based Construction of a Verb Lexicon. In *Proceedings of the 17th National Conference on Artificial Intelligence (AAAI-2000)*, Austin, TX, USA, 2000.

[Kipper *et al.*, 2006] Karin Kipper, Anna Korhonen, Neville Ryant, and Martha Palmer. Extending verbnet with novel verb classes. In *Proceedings of LREC*, 2006.

[Kipper K. and M., 2009] Ryant N. Kipper K., Korhonen A. and Palmer M. Semlink: Linking propbank, verbnet and framenet. In *Proceedings of the Generative Lexicon Conference, Pisa*, 2009.

[Kolomiyets *et al.*, 2012] Oleksandr Kolomiyets, Steven Bethard, and Marie-Francine Moens. Extracting narrative timelines as temporal dependency structures. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, Stroudsburg, PA, 2012.

[Kučera and Francis, 1967] Henry Kučera and Winthrop Nelson Francis. *Computational analysis of present-day American English*. Dartmouth Publishing Group, 1967.

[Laparra and Rigau, 2012] Egoitz Laparra and German Rigau. Exploiting explicit annotations and semantic types for implicit argument resolution. In *6th IEEE International Conference on Semantic Computing*, ICSC '12, pages 75–78, Palermo, Italy, 2012.

[Laparra and Rigau, 2013a] E. Laparra and G. Rigau. Impar: A deterministic algorithm for implicit semantic role labelling. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL 2013)*, pages 33–41. Association for Computational Linguistics, 2013.

[Laparra and Rigau, 2013b] Egoitz Laparra and German Rigau. Impar: A deterministic algorithm for implicit semantic role labelling. In *The 51st Annual Meeting of the Association for Computational Linguistics*, ACL '13, Sofia, Bulgaria, 2013.

[Laparra and Rigau, 2013c] Egoitz Laparra and German Rigau. Sources of evidence for implicit argument resolution. In *Proceedings of the 10th International Conference on Computational Semantics*, IWCS '13, pages 155–166, Potsdam, Germany, 2013.

[Laparra *et al.*, 2010] Egoitz Laparra, German Rigau, and Montse Cuadros. Exploring the integration of wordnet and framenet. In *Proceedings of the 5th Global WordNet Conference (GWC 2010), Mumbai, India*, 2010.

[Laparra E. and P., 2012] Rigau G. Laparra E. and Vossen P. Mapping wordnet to the kyoto ontology. In *Proceedings of the 8th international conference on Language Resources and Evaluation, N. Calzolari, K. Choukri, T. Declerck, M. Ugur Dogan, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis (Eds.), Publ. European Language Resources Association (ELRA), ISBN 978-2-9517408-7-7*, pages 2584–2589, 2012.

[Lappin and Leass, 1994] Shalom Lappin and Herbert J. Leass. An algorithm for pronominal anaphora resolution. *Computational Linguistics*, 20(4):535–561, December 1994.

[Leacock *et al.*, 1998] C. Leacock, M. Chodorow, and G. A. Miller. Using corpus statistics and wordnet relations for sense identification. *Computational Linguistics*, 24(1):147–166, 1998.

[Lee *et al.*, 2011] Heeyoung Lee, Yves Peirsman, Angel Chang, Nathanael Chambers, Mihai Surdeanu, and Dan Jurafsky. Stanford's multi-pass sieve coreference resolution system at the conll-2011 shared task. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*, pages 28–34. Association for Computational Linguistics, 2011.

[Lee *et al.*, 2012] Heeyoung Lee, Marta Recasens, Angel Chang, Mihai Sur-deanu, and Dan Jurafsky. Joint entity and event coreference resolution across documents. In *Proceedings of the 2012 Conference on Empirical Methods in Natural Language Processing and Natural Language Learning*, 2012.

[Lenat, 1995] Douglas B. Lenat. Cyc: A large-scale investment in knowledge infrastructure. *Communications of the ACM*, 38(11):33–38, 1995.

[Lesk, 1986] M. Lesk. Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from a ice cream cone. In *Proceedings of SIGDOC'86*, 1986.

[Levin, 1993] B. Levin. *English Verb Classes and Alternations: A Preliminary Investigation.* The University of Chicago Press, Chicago, 1993.

[Lin, 1998] Dekang Lin. An information-theoretic definition of similarity. In *Proceedings of the 15th international conference on Machine Learning*, volume 1, pages 296–304. San Francisco, 1998.

[Litkowski, 2004] K. C. Litkowski. Senseval-3 task: Automatic Labeling of Semantic Roles. In *Proceedings of the 3rd International Workshop on the Evaluation of Systems for the Semantic Analysis of Text (Senseval-3)*, pages 9–12, Barcelona, Spain, 2004.

[Liu *et al.*, 2005a] Bing Liu, Minqing Hu, and Junsheng Cheng. Opinion observer:analyzing and comparing opinions on the web. In *Proceedings of WWW-2005*, Chiba, Japan, 2005.

[Liu *et al.*, 2005b] Shuang Liu, Clement Yu, and Weiyi Meng. Word sense disambiguation in queries. In *Proceedings of the 14th ACM international conference on Information and knowledge management*, pages 525–532. ACM, 2005.

[Liu, 2012] Bing Liu. *Sentiment Analysis and Opinion Mining.* Morgan Claypool, 2012.

[Llorens *et al.*, 2010] Hector Llorens, Estela Saquete, and Borja Navarro. Tipsem (english and spanish): Evaluating crfs and semantic roles in tempeval-2. In *Proceedings of the 5th International Workshop on Semantic Evaluation (SEMEVAL-2010)*, pages 284–291, 2010.

[Llorens *et al.*, 2012] Hector Llorens, Leon Derczynski, Robert Gaizauskas, and Estela Saquete. Timen: An open temporal expression normalisation resource. In *Proceedings of LREC 2012*, 2012.

[Lluís and Màrquez, 2008] Xavier Lluís and Lluís Màrquez. A joint model for parsing syntactic and semantic dependencies. In *Proceedings of CoNLL-2008 Shared Task*, 2008.

[Loper *et al.*, 2007] Edward Loper, Szu-Ting Yi, and Martha Palmer. Combining Lexical Resources: Mapping between PropBank and VerbNet. In *Proceedings of the 7th International Workshop on Computational Linguistics*, Tilburg, the Netherlands, 2007.

[Luo, 2005a] Xiaoqiang Luo. On coreference resolution performance metrics. In *Proceedings of the Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT-EMNLP2005)*, pages 25–32, Vancouver, B.C., Canada, 2005.

[Luo, 2005b] Xiaoqiang Luo. On coreference resolution performance metrics. In *Proceedings of the Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, 2005.

[Maks and Vossen, 2011] Isa Maks and Piek Vossen. Different approaches to automatic polarity annotation at synset level. In *Proceedings of WOLER*, 2011.

[Mann and Yarowsky, 2003] Gideon S. Mann and David Yarowsky. Unsupervised personal name disambiguation. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003 - Volume 4*, CONLL '03, pages 33–40, Stroudsburg, PA, USA, 2003. Association for Computational Linguistics.

[Manning and Schütze, 1998] Christopher D. Manning and Hinrich Schütze. *Foundations of Statistical Natural Language Processing.* MIT Press, 1998.

[Marcus *et al.*, 1993] Mitchell P. Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. Building a large annotated corpus of english: the penn treebank. *Comput. Linguist.*, 19(2):313–330, June 1993.

[Màrquez *et al.*, 2006] Lluís Màrquez, Gerard Escudero, German Rigau, and David Martínez. *Word Sense Disambiguation. Algorithms and Applications*, volume 33 of *Text, Speech and Language Technology Series*, chapter Supervised Corpus-based Methods for Word Sense Disambiguation, pages 167–216. Springer, 2006.

[Martínez and Agirre, 2000] David Martínez and Eneko Agirre. One sense per collocation and genre/topic variations. In *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora. Hong Kong, 2000. "*, 2000.

[Martinez, 2004] David Martinez. *Supervised Word Sense Disambiguation: Facing Current Challenges*. PhD thesis, Informatika Fakultatea, UPV-EHU, 2004.

[Maynard *et al.*, 2001] Diana Maynard, Valentin Tablan, Cristian Ursu, Hamish Cunningham, and Yorick Wilks. Named entity recognition from diverse text types. In *In Recent Advances in Natural Language Processing 2001 Conference, Tzigov Chark*, pages 257–274, 2001.

[McCallum, 1996] Andrew Kachites McCallum. Bow: A toolkit for statistical language modeling, text retrieval, classification and clustering. http://www.cs.cmu.edu/ mccallum/bow, 1996.

[McCallum, 2002] Andrew Kachites McCallum. Mallet: A machine learning for language toolkit. http://mallet.cs.umass.edu, 2002.

[Mendes *et al.*, 2011] Pablo N Mendes, Max Jakob, Andrés García-Silva, and Christian Bizer. Dbpedia spotlight: shedding light on the web of documents. In *Proceedings of the 7th International Conference on Semantic Systems*, pages 1–8. ACM, 2011.

[Meyers *et al.*, 2004] A. Meyers, R. Reeves, C. Macleod, R. Szekely, V. Zielinska, B. Young, and R. Grishman. The nombank project: An interim report. In *In Proceedings of the NAACL/HLT Workshop on Frontiers in Corpus Annotation*, HLT-NAACL '04, pages 24–31, Boston, Massachusetts, USA, 2004.

[Mihalcea and Csomai, 2005] Rada Mihalcea and Andras Csomai. Senselearner: Word sense disambiguation for all words in unrestricted text. In *Proceedings of the ACL 2005 on Interactive poster and demonstration sessions*, pages 53–56. Association for Computational Linguistics, 2005.

[Mihalcea and Csomai, 2007] Rada Mihalcea and Andras Csomai. Wikify!: linking documents to encyclopedic knowledge. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, CIKM '07, pages 233–242, New York, NY, USA, 2007. ACM.

[Mihalcea and Moldovan, 1999] Rada Mihalcea and Dan Moldovan. An automatic method for generating sense tagged corpora. In *Proceedings of the 16th National Conference on Artificial Intelligence*. AAAI Press, 1999.

[Mihalcea, 2005] Rada Mihalcea. Unsupervised large-vocabulary word sense disambiguation with graph-based algorithms for sequence data labeling. In *Proceedings of the Joint Conference on Human Language Technology / Empirical Methods in Natural Language Processing (HLT/EMNLP)*, Vancouver, 2005.

[Mihalcea, 2007] Rada Mihalcea. Using wikipedia for automatic word sense disambiguation. In *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL)*, Rochester, 2007.

[Mika *et al.*, 2008] Peter Mika, Massimiliano Ciaramita, Hugo Zaragoza, and Jordi Atserias. Learning to tag and tagging to learn: A case study on wikipedia. *IEEE Intelligent Systems*, 23(5):26–33, September 2008.

[Mikheev *et al.*, 1999] Andrei Mikheev, Marc Moens, and Claire Grover. Named entity recognition without gazetteers. In *Proceedings of the ninth conference on European chapter of the Association for Computational Linguistics*, EACL '99, pages 1–8, Stroudsburg, PA, USA, 1999. Association for Computational Linguistics.

[Miller *et al.*, 1993] George A Miller, Claudia Leacock, Randee Tengi, and Ross T Bunker. A semantic concordance. In *Proceedings of the workshop on Human Language Technology*, pages 303–308. Association for Computational Linguistics, 1993.

[Milne and Witten, 2008] David Milne and Ian H. Witten. Learning to link with wikipedia. In *Proceedings of the 17th ACM conference on Information and knowledge management*, CIKM '08, pages 509–518, New York, NY, USA, 2008. ACM.

[Mitkov *et al.*, 2002] Ruslan Mitkov, Richard Evans, and Constantin Orasan. A new, fully automatic version of mitkov's knowledge-poor pronoun resolution method. In *Proceedings of the Third International Conference on Computational Linguistics and Intelligent Text Processing*, CICLing '02, pages 168–186, London, UK, UK, 2002.

[Montoyo *et al.*, 2005] A. Montoyo, A. Suárez, G. Rigau, and M. Palomar. Combining knowledge- and corpus-based word-sense-disambiguation methods. *Journal of Artificial Intelligence Research*, 23:299–330, 2005.

[Moor *et al.*, 2013] Tatjana Moor, Michael Roth, and Anette Frank. Predicate-specific annotations for implicit role binding: Corpus annotation, data analysis and evaluation experiments. In *Proceedings of the 10th International Conference on Computational Semantics*, IWCS '13, pages 369–375, Potsdam, Germany, 2013.

[Morante and Daelemans, 2009] Roser Morante and Walter Daelemans. Learning the scope of hedge cues in biomedical texts. In *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing*, pages 28–36. Association for Computational Linguistics, 2009.

[Mulkar-Mehta, 2011] Rutu Mulkar-Mehta. *Granular Causality for Learning by Reading*. PhD thesis, University of Southern California, 2011.

[Mutalik *et al.*, 2001] Pradeep G Mutalik, Aniruddha Deshpande, and Prakash M Nadkarni. Use of general-purpose negation detection to augment concept indexing of medical documents a quantitative study using the umls. *Journal of the American Medical Informatics Association*, 8(6):598–609, 2001.

[Nadeau and Sekine, 2007] David Nadeau and Satoshi Sekine. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26, 2007.

[Navigli and Lapata, 2007] Roberto Navigli and Mirella Lapata. Graph connectivity measures for unsupervised word sense disambiguation. In *Proceedings of the* 20^{th} *International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1683–1688, Hyderabad, India, 2007.

[Navigli and Velardi, 2005] Roberto Navigli and Paola Velardi. Structural semantic interconnections: a knowledge-based approach to word sense disambiguation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 27(7):1063–1074, 2005.

[Navigli, 2009] Roberto Navigli. Word Sense Disambiguation: a survey. *ACM Computing Surveys*, 41(2):1–69, 2009.

[Negri and Marseglia, 2004] M. Negri and L. Marseglia. Recognition and normalization of time expressions: ITC-irst at TERN 2004. Technical report, ITC-irst, Trento, 2004.

[Ng and Cardie, 2002a] Vincent Ng and Claire Cardie. Identifying anaphoric and non-anaphoric noun phrases to improve coreference resolution. In *Proceedings of the 19th international conference on Computational linguistics - Volume 1*, COLING '02, pages 1–7, Taipei, Taiwan, 2002.

[Ng and Cardie, 2002b] Vincent Ng and Claire Cardie. Improving machine learning approaches to coreference resolution. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 2002.

[Ng and Lee, 1996] H. T. Ng and H. B. Lee. Integrating multiple knowledge sources to disambiguate word sense: An exemplar-based approach. In *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics*. ACL, 1996.

[Ng, 1997] H. T. Ng. Getting serious about word sense disambiguation. In *Proceedings of the ACL SIGLEX Workshop on Tagging Text with Lexical Semantics: Why, What, and How?*, pages 1–7, Washington, D.C., USA., 1997.

[Ng, 2005] Vincent Ng. Machine learning for coreference resolution: From local classification to global ranking. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, 2005.

[Niemann and Gurevych, 2011] Elisabeth Niemann and Iryna Gurevych. The people's web meets linguistic knowledge: Automatic sense alignment of wikipedia and wordnet. In *Proceedings of the International Conference on Computational Semantics (IWCS)*, pages 205–214, Jan 2011.

[Niles and A., 2003] I. Niles and Pease A. A. linking lexicons and ontologies: Mapping wordnet to the suggested upper merged ontology. In *Proceedings of the 2003 International Conference on Information and Knowledge Engineering*, 2003.

[Niles and Pease, 2001] I. Niles and A. Pease. Towards a standard upper ontology. In *Proceedings of the 2nd Inter. Conf. on Formal Ontology in Information Systems*, 2001.

[Nothman *et al.*, 2008] Joel Nothman, James R. Curran, and Tara Murphy. Transforming Wikipedia into named entity training data. In *Proceedings of the Australasian Language Technology Workshop*, Hobart, Australia, 2008.

[Nothman *et al.*, 2012] Joel Nothman, Nicky Ringland, Will Radford, Tara Murphy, and James R. Curran. Learning multilingual named entity recognition from Wikipedia. *Artificial Intelligence*, 2012. (in press).

[Orăsan *et al.*, 2003] Constantin Orăsan, Ruslan Mitkov, and Laura Hasler. Cast: a computer-aided summarisation tool. In *Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics - Volume 2*, EACL '03, pages 135–138, Budapest, Hungary, 2003.

[Øvrelid *et al.*, 2010] Lilja Øvrelid, Erik Velldal, and Stephan Oepen. Syntactic scope resolution in uncertainty analysis. In *Proceedings of the 23rd international conference on computational linguistics*, pages 1379–1387. Association for Computational Linguistics, 2010.

[Palmer *et al.*, 1986] Martha S. Palmer, Deborah A. Dahl, Rebecca J. Schiffman, Lynette Hirschman, Marcia Linebarger, and John Dowding. Recovering implicit information. In *Proceedings of the 24th annual meeting on Association for Computational Linguistics*, ACL '86, pages 10–19, New York, New York, USA, 1986.

[Palmer *et al.*, 2005a] M. Palmer, D. Gildea, and P. Kingsbury. The Proposition Bank: An Annotated Corpus of Semantic Roles. *Computational Linguistics*, 31(1):71–105, 2005.

[Palmer *et al.*, 2005b] Martha Palmer, Daniel Gildea, and Paul Kingsbury. The proposition bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–106, March 2005.

[Pang and Lee, 2008] Bo Pang and Lilian Lee. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2), 2008.

[Pedersen, 2006] Ted Pedersen. Unsupervised corpus based methods for wsd. In Agirre E. and Edmonds P., editors, *Word Sense Disambiguation*, volume 33 of *Text, Speech and Language Technology Series*, pages 133–166. Springer, 2006.

[Philpot *et al.*, 2005] Andrew Philpot, Eduard Hovy, and Patrick Pantel. The omega ontology. In *Proceedings of the ONTOLEX Workshop at the International Conference on Natural Language Processing (IJCNLP)*, 2005.

[Pianta *et al.*, 2002] Emanuele Pianta, Luisa Bentivogli, and Christian Girardi. Developing an aligned multilingual database. In *Proc. 1st International Conference on Global WordNet*, 2002.

[Pianta *et al.*, 2008] Emanuele Pianta, Christian Girardi, and Roberto Zanoli. The TextPro tool suite. In *Proceedings of LREC, 6th edition of the Language Resources and Evaluation Conference*, Marrakech (Morocco), 2008.

[Poesio and Magnini, 2009] Massimo Poesio and Bernardo Magnini. Content extraction meets the social web in the livememories project. In *Workshop on Advanced Technologies for Digital Libraries 2009 AT4DL 2009*, page 42, 2009.

[Ponzetto and Strube, 2006a] Simone Paolo Ponzetto and Michael Strube. Exploiting semantic role labeling, wordnet and wikipedia for coreference resolution. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL, pages 192-199*, 2006.

[Ponzetto and Strube, 2006b] Simone Paolo Ponzetto and Michael Strube. Semantic role labeling for coreference resolution. In *Proceedings of the Eleventh Conference of the European Chapter of the Association for Computational Linguistics: Posters & Demonstrations*, EACL '06, pages 143–146, Trento, Italy, 2006.

[Pradhan *et al.*, 2007a] S. Pradhan, E. Loper, D. Dligach, and M. Palmer. SemEval-2007 Task 17: English Lexical Sample, SRL and All Words. In *Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007)*, pages 87–92, Prague, Czech Republic, 2007.

[Pradhan *et al.*, 2007b] Sameer S. Pradhan, Eduard Hovy, Mitch Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. Ontonotes: A unified relational semantic representation. In *Proceedings of the International Conference on Semantic Computing*, ICSC '07, pages 517–526, Irvine, California, USA, 2007.

[Pradhan *et al.*, 2011] Sameer Pradhan, Lance Ramshaw, Mitchell Marcus, Martha Palmer, Ralph Weischedel, and Nianwen Xue. Conll-2011 shared task: modeling unrestricted coreference in ontonotes. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*, CONLL Shared Task '11, pages 1–27, Portland, Oregon, 2011.

[Procter, 1987] P. Procter, editor. *Longman Dictionary of Common English.* Longman Group, Harlow, Essex, England, 1987.

[Pustejovsky *et al.*, 2003] James Pustejovsky, José Castano, Robert Ingria, Roser Sauri, Robert Gaizauskas, Andrea Setzer, Graham Katz, and Dragomir Radev. Timeml: Robust specification of event and temporal expressions in text. *New directions in question answering*, 3:28–34, 2003.

[Pustejovsky *et al.*, 2006] James Pustejovsky, Jessica Littman, Roser Saurí, and Marc Verhagen. Timebank 1.2 documentation. Technical report, Brandeis University, April 2006.

[Pustejovsky *et al.*, 2010] James Pustejovsky, Kiyong Lee, Harry Bunt, and Laurent Romary. Iso-timeml: An international standard for semantic annotation. In *CHAIR), NCC, CHOUKRI, K., MAEGAARD, B., MARIANI, J., ODIJK, J., PIPERIDIS, S., ROSNER, M. et TAPIAS, D., éditeurs: Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC10), Valletta, Malta. European Language Resources Association (ELRA)*, 2010.

[Qiu *et al.*, 2004] Long Qiu, Min-Yen Kan, and Tat-Seng Chua. A public reference implementation of the rap anaphora resolution algorithm. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation*, LREC '04, Lisbon, Portugal, 2004.

[Raghunathan *et al.*, 2010] Karthik Raghunathan, Heeyoung Lee, Sudarshan Rangarajan, Nathanael Chambers, Mihai Surdeanu, Dan Jurafsky, and Christopher Manning. A multi-pass sieve for coreference resolution. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP '10, pages 492–501, Cambridge, Massachusetts, USA, 2010.

[Rahman and Ng, 2009] Altaf Rahman and Vincent Ng. Supervised models for coreference resolution. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2 - Volume 2*, EMNLP '09, pages 968–977, Singapore, 2009.

[Ratinov and Roth, 2009] Lev Ratinov and Dan Roth. Design challenges and misconceptions in named entity recognition. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning*, CoNLL '09, pages 147–155, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics.

[Ratinov *et al.*, 2011] Lev Ratinov, Dan Roth, Doug Downey, and Mike Anderson. Local and global algorithms for disambiguation to wikipedia. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 1375–1384, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

[Ravin and Leacock, 2000] Y. Ravin and C. Leacock. *Polysemy: Theoretical and Approaches.* Oxford University Press, 2000.

[Recasens and Hovy, 2011] M. Recasens and E. Hovy. Blanc: Implementing the rand index for coreference evaluation. *Nat. Lang. Eng.*, 17(4):485–510, October 2011.

[Recasens and Martí, 2010] Marta Recasens and M. Antònia Martí. Ancora-co: Coreferentially annotated corpora for spanish and catalan. *Lang. Resour. Eval.*, 44(4):315–345, December 2010.

[Recasens *et al.*, 2010] Marta Recasens, Lluís Màrquez, Emili Sapena, M. Antònia Martí, Mariona Taulé, Véronique Hoste, Massimo Poesio, and Yannick Versley. Semeval-2010 task 1: Coreference resolution in multiple languages. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, SemEval '10, pages 1–8, Los Angeles, California, USA, 2010.

[Recasens *et al.*, 2011] Marta Recasens, Eduard Hovy, and M. Antònia Martí. Identity, non-identity, and near-identity: Addressing the complexity of coreference. *Lingua*, 121(6):1138–1152, 2011.

[Richman and Schone, 2008] Alexander E. Richman and Patrick Schone. Mining wiki resources for multilingual named entity recognition. In *Proceedings of the ACL*, pages 1–9. Association for Computational Linguistics, 2008.

[Rigau *et al.*, 1997] German Rigau, Jordi Atserias, and Eneko Agirre. Combining unsupervised lexical knowledge methods for word sense disambiguation. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics. Joint ACL/EACL*, pages 48–55, Madrid, Spain, July 1997.

[Roget, 1911] Peter Mark Roget. *Roget's Thesaurus of English Words and Phrases...* TY Crowell Company, 1911.

[Rokach *et al.*, 2008] Lior Rokach, Roni Romano, and Oded Maimon. Negation recognition in medical narrative reports. *Information Retrieval*, 11(6):499–538, 2008.

[Roser Saurí and Pustejovsky, 2005] Marc Verhagen Roser Saurí, Robert Knippen and James Pustejovsky. Evita: A robust event recognizer for qa systems. In *Proceedings of the Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 700–707, 2005.

[Rubin *et al.*, 2006] Victoria L Rubin, Elizabeth D Liddy, and Noriko Kando. Certainty identification in texts: Categorization model and manual tagging results. In *Computing attitude and affect in text: Theory and applications*, pages 61–76. Springer, 2006.

[Ruppenhofer *et al.*, 2002] Josef Ruppenhofer, Collin F. Baker, and Charles J. Fillmore. The FrameNet database and software tools. In Anna Braasch and Claus Povlsen, editors, *Proceedings of the European Association for Lexicography*, pages 371–377, Copenhagen, 2002. EURALEX, EURALEX.

[Ruppenhofer *et al.*, 2010] Josef Ruppenhofer, Caroline Sporleder, Roser Morante, Collin Baker, and Martha Palmer. Semeval-2010 task 10: Linking events and their participants in discourse. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, SemEval '10, pages 45–50, Los Angeles, California, USA, 2010.

[Ruppenhofer *et al.*, 2011] Josef Ruppenhofer, Philip Gorinski, and Caroline Sporleder. In search of missing arguments: A linguistic approach. In *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011*, RANLP '11, pages 331–338, Hissar, Bulgaria, 2011.

[Sapena *et al.*, 2011] Emili Sapena, Lluís Padró, and Jordi Turmo. Relaxcor participation in conll shared task on coreference resolution. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*, CONLL Shared Task '11, pages 35–39, Portland, Oregon, 2011.

[Saurı and Barcelona, 2010] Roser Saurı and Catalonia Barcelona. Tempeval 2 spanish data release. Technical report, Tech. rep. BM 2010-05, Barcelona Media Technical Report, 2010.

[Saurí and Pustejovsky, 2009a] Roser Saurí and James Pustejovsky. Factbank: A corpus annotated with event factuality. *Language Resources and Evaluation*, 43(3):227–268, 2009.

[Saurí and Pustejovsky, 2009b] Roser Saurí and James Pustejovsky. FactBank: a corpus annotated with event factuality. *Language Resources and Evaluation*, 43:227–268, 2009.

[Saurí and Pustejovsky, 2012] Roser Saurí and James Pustejovsky. Are you sure that this happened? assessing the factuality degree of events in text. *Computational Linguistics*, 38(2):261–299, 2012.

[Sebastiani, 2002] Fabrizio Sebastiani. Machine learning in automated text categorization. *ACM computing surveys (CSUR)*, 34(1):1–47, 2002.

[Sekine and Nobata, 2004] Satoshi Sekine and Chikashi Nobata. Definition, dictionaries and tagger for extended named entity hierarchy. In *Proceedings of LREC*, 2004.

[Silberer and Frank, 2012] Carina Silberer and Anette Frank. Casting implicit role linking as an anaphora resolution task. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics*, *SEM '12, pages 1–10, Montréal, Canada, 2012.

[Singh *et al.*, 2010] Sameer Singh, Dustin Hillard, and Chris Leggetter. Minimally-supervised extraction of entities from text advertisements. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association* *for Computational Linguistics*, pages 73–81. Association for Computational Linguistics, 2010.

[Sinha and Mihalcea, 2007a] Ravi Sinha and Rada Mihalcea. Unsupervised graph-based word sense disambiguation using measures of word semantic similarity. In *Proceedings of the IEEE International Conference on Semantic Computing (ICSC 2007), Irvine, CA*, 2007.

[Sinha and Mihalcea, 2007b] Ravi Sinha and Rada Mihalcea. Unsupervised graph-basedword sense disambiguation using measures of word semantic similarity. In *Semantic Computing, 2007. ICSC 2007. International Conference on*, pages 363–369. IEEE, 2007.

[Somasundaran and Wiebe, 2009] Swapna Somasundaran and Janyce Wiebe. Recognizing stances in online debates. In *Joint conference of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing*, Singapore, August 2-7 2009.

[Somasundaran and Wiebe, 2010] Swapna Somasundaran and Janyce Wiebe. Recognizing stances in ideological on-line debates. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, pages 116–124, Los Angeles, California, USA, 2010.

[Steinberger and Pouliquen, 2007] Ralf Steinberger and Bruno Pouliquen. Cross-lingual named entity recognition. *Lingvisticæ Investigationes*, 30(1):135–162, 2007.

[Steinberger *et al.*, 2007] Josef Steinberger, Massimo Poesio, Mijail A. Kabadjov, and Karel Jeek. Two uses of anaphora resolution in summarization. *Inf. Process. Manage.*, 43(6):1663–1680, November 2007.

[Steinberger *et al.*, 2012] Ralf Steinberger, Mohamed Ebrahim, and Marco Turchi. Jrc eurovoc indexer jex - a freely available multi-label categorisation tool. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may 2012. European Language Resources Association (ELRA).

[Stevenson, 2003] Mark Stevenson. *Word Sense Disambiguation: The Case for Combinations of Knowledge Sources*. CSLI Publications, Stanford, CA., 2003.

[Stokoe *et al.*, 2003] Christopher Stokoe, Michael P Oakes, and John Tait. Word sense disambiguation in information retrieval revisited. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 159–166. ACM, 2003.

[Stone *et al.*, 1966] P. J. Stone, D. C. Dunphy, M.S. Smith, and D. M. Ogilvie. *The General Inquirer: A Computer Approach to Content Analysis.* MIT Press, 1966.

[Stoyanov *et al.*, 2010] Veselin Stoyanov, Claire Cardie, Nathan Gilbert, Ellen Riloff, David Buttler, and David Hysom. Coreference resolution with reconcile. In *Proceedings of the ACL 2010 Conference Short Papers*, ACLShort '10, pages 156–161, Uppsala, Sweden, 2010.

[Strapparava and Valitutti, 2004] Carlo Strapparava and Alessandro Valitutti. Wordnet-affect: an affective extension of wordnet. In *Proceedings of LREC 2004*, 2004.

[Strassel *et al.*, 2008] Stephanie Strassel, Mark A. Przybocki, Kay Peterson, Zhiyi Song, and Kazuaki Maeda. Linguistic resources and evaluation techniques for evaluation of cross-document automatic content extraction. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation*, LREC'08, Marrakech, Morocco, 2008.

[Strötgen and Gertz, 2013] J. Strötgen and M. Gertz. Multilingual and cross-domain temporal tagging. *Language Resources and Evaluation*, pages 269–298, 2013.

[Surdeanu *et al.*, 2008] Mihai Surdeanu, Richard Johansson, Adam Meyers, Lluís Màrquez, and Joakim Nivre. The CoNLL-2008 shared task on joint parsing of syntactic and semantic dependencies. In *Proceedings of the Twelfth Conference on Natural Language Learning*, CoNLL '08, pages 159–177, Manchester, United Kingdom, 2008.

[Taulé *et al.*, 2006] M Taulé, J Castellvı, and MA Martı. Semantic classes in cess-lex: Semantic annotation of cess-ece. In *Proceedings of the Fifth Workshop on Treebanks and Linguistic Theories (TLT-2006)*, 2006.

[Taulé *et al.*, 2008] Mariona Taulé, Maria Antònia Martí, and Marta Recasens. Ancora: Multilevel annotated corpora for catalan and spanish. In *LREC*, 2008.

[Tenny and Pustejovsky, 2000] C. Tenny and J. Pustejovsky. History of events in linguistic theory. In C. Tenny and J. Pustejovsky, editors, *Events as Grammatical Objects*. Cambridge University Press, 2000.

[Tetreault, 2002] Joel R. Tetreault. Implicit role reference. In *International Symposium on Reference Resolution for Natural Language Processing*, pages 109–115, Alicante, Spain, 2002.

[The PDTB Research Group, 2008] The PDTB Research Group. The PDTB 2.0. Annotation Manual. Technical Report IRCS-08-01, Institute for Research in Cognitive Science, University of Pennsylvania, 2008.

[Tjong Kim Sang and De Meulder, 2003] Erik F. Tjong Kim Sang and Fien De Meulder. 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*, CONLL '03, pages 142–147, Stroudsburg, PA, USA, 2003. Association for Computational Linguistics.

[Tonelli and Delmonte, 2010] Sara Tonelli and Rodolfo Delmonte. Venses++: Adapting a deep semantic processing system to the identification of null instantiations. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, SemEval '10, pages 296–299, Los Angeles, California, USA, 2010.

[Tonelli and Delmonte, 2011] Sara Tonelli and Rodolfo Delmonte. Desperately seeking implicit arguments in text. In *Proceedings of the ACL 2011 Workshop on Relational Models of Semantics*, RELMS '11, pages 54–62, Portland, Oregon, USA, 2011.

[Toral and Munoz, 2006] Antonio Toral and Rafael Munoz. A proposal to automatically build and maintain gazetteers for named entity recognition by using wikipedia. *NEW TEXT Wikis and blogs and other dynamic text sources*, page 56, 2006.

[UzZaman *et al.*, 2013] Naushad UzZaman, Hector Llorens, Leon Derczynski, Marc Verhagen, James Allen, and James Pustejovsky. Semeval-2013 task 1: Tempeval-3: Evaluating events, time expressions, and temporal relations. In *Proceedings of the 7th International Workshop on Semantic Evaluation (SemEval 2013)*, 2013.

[van den Hoven *et al.*, 2010] Martha van den Hoven, Antal van den Bosch, and Kalliopi Zervanou. Beyond reported history: Strikes that never happened. In *Proceedings of the First International AMICUS Workshop on Automated Motif Discovery in Cultural Heritage and Scientific Communication Texts*, pages 20–28, Vienna, Austria, 2010.

[Verhagen *et al.*, 2007] Marc Verhagen, Robert Gaizauskas, Frank Schilder, Mark Hepple, Graham Katz, and James Pustejovsky. Semeval-2007 task 15: Tempeval temporal relation identification. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 75–80, Prague, Czech Republic, June 2007. Association for Computational Linguistics.

[Verhagen *et al.*, 2010a] Marc Verhagen, Roser Sauri, Tommaso Caselli, and James Pustejovsky. Semeval-2010 task 13: Tempeval-2. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 57–62. Association for Computational Linguistics, 2010.

[Verhagen *et al.*, 2010b] Marc Verhagen, Roser Saurrí, Tommaso Caselli, and James Pustejovsky. Semeval-2010 task 13: Tempeval-2. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 57–62, Uppsala, Sweden, July 2010. Association for Computational Linguistics.

[Versley *et al.*, 2008] Yannick Versley, Simone Paolo Ponzetto, Massimo Poesio, Vladimir Eidelman, Alan Jern, Jason Smith, Xiaofeng Yang, and Alessandro Moschitti. Bart: a modular toolkit for coreference resolution. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Demo Session*, HLT-Demonstrations '08, pages 9–12, Columbus, Ohio, 2008.

[Vieira and Poesio, 2000] Renata Vieira and Massimo Poesio. An empirically ased system for processing definite descriptions. *Comput. Linguist.*, 26(4):539–593, December 2000.

[Vilain *et al.*, 1995a] Marc Vilain, John Burger, John Aberdeen, Dennis Connolly, and Lynette Hirschman. A model-theoretic coreference scoring scheme. In *Proceedings of the 6th conference on Message understanding*, MUC6 '95, pages 45–52, Stroudsburg, PA, USA, 1995.

[Vilain *et al.*, 1995b] Marc Vilain, John Burger, John Aberdeen, Connolly Dennis, and Lynette Hirschman. A model theoretic coreference scoring scheme. In *Proceedings of MUC-6*, 1995.

[Villena-Román *et al.*, 2012] Julio Villena-Román, Sara Lana-Serrano, Eugenio Martínez-Cámara, and José Carlos González-Cristóbal. TASS-workshop on sentiment analysis at SEPLN. *Procesamiento del Lenguaje Natural*, 50:37–44, November 2012.

[Vossen *et al.*, 2006] Piek Vossen, German Rigau, Iñaki Alegria, Eneko Agirre, David Farwell, Manuel Fuentes, et al. Meaningful results for information retrieval in the meaning project. In *Proc. of the 3rd Global Wordnet Conference*, pages 22–26, 2006.

[Vossen P. and A., 2013] Rigau G. Vossen P., Agirre E. and Soroa A. Kyoto: a knowledge-rich approach to the interoperable mining of events from text. In Oltramari A., Lu-Qin, Vossen P., Qin L., and Hovy E., editors, *New Trends of Research in Ontologies and Lexical Resources*, ISBN 978-3-642-31781-1. Springer Verlag, Heidelberg, 2013.

[Weischedel and Brunstein, 2005] Ralph Weischedel and Ada Brunstein. Bbn pronoun coreference and entity type corpus, 2005.

[Whittemore *et al.*, 1991] Greg Whittemore, Melissa Macpherson, and Greg Carlson. Event-building through role-filling and anaphora resolution. In *Proceedings of the 29th annual meeting on Association for Computational Linguistics*, ACL '91, pages 17–24, Berkeley, California, USA, 1991.

[Wiebe and Riloff, 2005] J. Wiebe and E. Riloff. Creating subjective and objective sentence classifiers from unannotated texts. In *Proceedings of CICLing-2005*, 2005.

[Wiebe *et al.*, 2004] Janyce Wiebe, Theresa Wilson, Rebecca Bruce, Matthew Bell, and Melanie Martin. Learning subjective language. *Computational linguistics*, 30(3):277–308, 2004.

[Wiebe *et al.*, 2005] Janyce Wiebe, Theresa Wilson, and Claire Cardie. Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, 39(2-3):165–210, 2005.

[Wilson *et al.*, 2005] Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of HLT-EMNLP-2005*, 2005.

[Witten *et al.*, 1999] Ian H. Witten, Zane Bray, Malika Mahoui, and W. J. Teahan. Using language models for generic entity extraction. In *In International Conference on Machine Learning Workshop on Text Mining*, 1999.

[Yang and Liu, 1999] Yiming Yang and Xin Liu. A re-examination of text categorization methods. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 42–49. ACM, 1999.

[Yarowsky, 1992] David Yarowsky. Word-sense disambiguations using statistical models of roget's categories trained on large corpora. In *Proceedings of the 14th International Conference on Computational Linguistics, COLING*, Nantes, France, 1992.

[Yi *et al.*, 2007] Szu-Ting Yi, Edward Loper, and Martha Palmer. Can Semantic Roles Generalize Across Genres? In *Proceedings of the Human Language Technology Conferences/North American Chapter of the Association for Computational Linguistics (HLT/NAACL-2007)*, Rochester, NY, USA, 2007.

[Zapirain *et al.*, 2008] Benat Zapirain, Eneko Agirre, and Lluís Màrquez. Robustness and generalization of role sets: Propbank vs. verbnet. In *Proceedings of ACL-08: HLT*, pages 550–558, 2008.

[Zhong and Ng, 2010] Zhi Zhong and Hwee Tou Ng. It makes sense: A wide-coverage word sense disambiguation system for free text. In *Proceedings of the ACL 2010 System Demonstrations*, pages 78–83. Association for Computational Linguistics, 2010.

[Özgür and Radev, 2009] A. Özgür and D. Radev. Detecting speculations and their scopes in scientific text. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3-Volume 3*, pages 1398–1407. Association for Computational Linguistics, 2009.