qtleap

quality
translation
by deep
language
engineering
approaches

REPORT ON THE STATE OF THE ART OF NAMED ENTITY AND WORD SENSE DISAMBIGUATION

DELIVERABLE D5.1

VERSION 4.0 | 2015 JUNE 25

QTLeap

Machine translation is a computational procedure that seeks to provide the translation of utterances from one language into another language.

Research and development around this grand challenge is bringing this technology to a level of maturity that already supports useful practical solutions. It permits to get at least the gist of the utterances being translated, and even to get pretty good results for some language pairs in some focused discourse domains, helping to reduce costs and to improve productivity in international businesses.

There is nevertheless still a way to go for this technology to attain a level of maturity that permits the delivery of quality translation across the board.

The goal of the QTLeap project is to research on and deliver an articulated methodology for machine translation that explores deep language engineering approaches in view of breaking the way to translations of higher quality.

The deeper the processing of utterances the less language-specific differences remain between the representation of the meaning of a given utterance and the meaning representation of its translation. Further chances of success can thus be explored by machine translation systems that are based on deeper semantic engineering approaches.

Deep language processing has its stepping-stone in linguistically principled methods and generalizations. It has been evolving towards supporting realistic applications, namely by embedding more data based solutions, and by exploring new types of datasets recently developed, such as parallel DeepBanks.

This progress is further supported by recent advances in terms of lexical processing. These advances have been made possible by enhanced techniques for referential and conceptual ambiguity resolution, and supported also by new types of datasets recently developed as linked open data.

The project QTLeap explores novel ways for attaining machine translation of higher quality that are opened by a new generation of increasingly sophisticated semantic datasets and by recent advances in deep language processing.

www.qtleap.eu

Funded by

QTLeap is funded by the 7th Framework Programme of the European Commission.

Supported by

And supported by the participating institutions:

Faculty of Sciences, University of Lisbon

German Research Centre for Artificial Intelligence

Charles University in Prague

Bulgarian Academy of Sciences

Humboldt University of Berlin

University of Basque Country

University of Groningen

Higher Functions, Lda.

Revision History

version	date	author	organisation	description
1	2013 DEC 05	Nora Aranberri	UPV/EHU	First draft
2	2013 DEC 16	João Silva, António Branco, Martin Popel, Aljoscha Burchardt, Eneko Agirre, Nora Aranberri, Ander Barrena, Gorka Labaka, Kepa Sarasola	FCUL CUNI DFKI UPV/EHU	Second draft
3	2013 DEC 26	Gertjan van Noord Petya Osenova Eneko Agirre Nora Aranberri	UG IICT-BAS UPV/EHU	Implementation of Internal Review
4	2015 JUN 25	Eneko Agirre, Gorka Labaka, Iñaki Alegria, Mikel Artetxe	UPV/EHU	Implementation of recommendations in 1st year review. The rest of the deliverable was left as-is.

Statement of originality

This deliverable contains original unpublished work except where clearly indicated otherwise. Acknowledgement of previously published material and of the work of others has been made through appropriate citation, quotation or both.

REPORT ON THE STATE OF THE ART OF NAMED ENTITY AND WORD SENSE DISAMBIGUATION

DOCUMENT QTLEAP-2015-D5.1

EC FP7 PROJECT #610516

DELIVERABLE D5.1

completion

FINAL

status

SUBMITTED

dissemination level

PUBLIC

responsible

ENEKO AGIRRE (WP5 COORDINATOR)

reviewer

GERTJAN VAN NOORD

contributing partners

UPV/EHU, CUNI, DFKI, FCUL, IICT-BAS, UG

authors

**ENEKO AGIRRE, IÑAKI ALEGRIA, MIKEL ARTETXE , NORA ARANBERRI,
ANDER BARRENA, ANTÓNIO BRANCO, MARTIN POPEL,
ALJOSCHA BURCHARDT, GORKA LABAKA, PETYA OSENOVA,
KEPA SARASOLA, JOÃO SILVA,**

Table of Contents

Index of Tables

Index of Figures

Abbreviations

API	Application Programming Interface
CLEF	Cross-language Evaluation Forum
CDCR	Cross-document Coreference Resolution
CONLL	Conference on Natural Language Learning
IE	Information Extraction
IR	Information Retrieval
ESA	Explicit Semantic Analysis
KB	Knowledge Base
KBP	Knowledge Base Population
LDC	Linguistic Data Consortium
LOD	Linked Open Data
MFS	Most Frequent Sense
VSM	Vector Space Model
MUC	Message Understanding Conference
MT	Machine Translation
NEL	Named Entity Linking
NER	Named Entity Recognition
NERC	Named Entity Recognition and Classification
NED	Named Entity Disambiguation (conventionally referring to Entity Linking and/or Wikification)
NLP	Natural Language Processing
PoS	Part-of-Speech
RDF	Resource Description Framework
TAC	Text Analysis Conference
WSD	Word Sense Disambiguation

1 Introduction

This deliverable consists of an in-depth survey of the current state-of-the-art, data sources, tools and technology related to Named Entity Recognition and Classification (NERC), Named Entity Disambiguation (NED) and Word Sense Disambiguation (WSD) for the relevant working languages in QTLeap. It relies on similar deliverables presented in the OpeNER (ICT-296451 D3.21)[1] and NewsReader (ICT-316404 D4.1)[2] EU projects and has been adapted to suit the aims and needs of the QTLeap project and recent developments. The inclusion of content from those sources has been performed with permission of the respective authors.

Whereas word sense disambiguation tries to assign the correct sense of a word for a particular context, named entity disambiguation focuses on recognizing, classifying and linking every mention of a specific named entity in a text. For example, a named entity can be mentioned using a great variety of surface forms (Barack Obama, President Obama, Mr. Obama, B. Obama, etc.) and the same surface form can refer to a variety of named entities: for example, the form 'san juan' can be used to ambiguously refer to dozens of toponyms, persons, a saint, etc. (e.g., see http://en.wikipedia.org/wiki/San_Juan). Therefore, in order to provide an adequate and comprehensive account of named entities in text it is necessary to recognize the mention of a named entity, classify it as a type (e.g., person, location, etc.), link it or disambiguate it to a specific entity, and resolve every form of mentioning to the same entity in a text.

The use of WSD techniques in Machine Translation (MT) is an open research subject. Since the initial and disappointing work of Carpuat and Wu (2005), other ways to take WSD predictions into account have been proposed. Some of them achieved encouraging results in partial tasks such as word prediction, but WSD has not yet been integrated into a complete MT system (Apadaniaki et al. 2012). QTLeap will tackle this subject and measure the contribution lexical semantics can make in improving MT.

NERC-based techniques have been applied to statistical MT with success (Hálek et al. 2011, Li et al. 2013). The deep processing architecture of QTLeap brings a natural place where the success of those works can be further enhanced.

Finally, the interplay between NED and Linked Open Data (LOD) can open the avenue for further improvements. For example, the recognizing "Beethoven's 9th Symphony" in a source text as a named entity, can be used to link the string to the corresponding DBpedia entry, and then translate it using multilingual links into "Symphonie no 9 de Beethoven" (French) or "Bederatzigarren Sinfonia (Beethoven)" (Basque).

NERC, NED and WSD are three fields that have experienced great advancements in recent years. MT systems, in turn, have difficulty in dealing with these two tasks. And yet, no considerable effort has been made so far to port the improvements seen in NED and WSD into the MT field. The QTLeap project aims to contribute to the improvement of MT by integrating these two subtasks, among others, to the translation process.

To this end, QTLeap will advance the state of the art on multilingual and crosslingual NERC, NED and WSD. Parallel texts will allow the development of techniques that produce better quality annotations, leveraging the information available in one language with the information available in another language. In addition, we will advance the state of the art

[1] http://www.opener-project.eu/project/publications.html

[2] http://www.newsreader-project.eu/publications/deliverables/

on joint processing of the four problems, as they are highly related. This work will be developed mainly in Task 5.2 (Crosslingual named entity and word sense resolution) and Task 5.4 (Using semantic linking and resolving to improve MT). The objective of this deliverable is to describe the tasks of NERC, NED and WSD and present an overview of the resources - training data sets and tools - available today.

The rest of the document is structured as follows: Sections 2 to 4 present the tasks of NERC, NED and WSD, their aim, approaches, difficulties and state-of-the-art achievement. Section 5 presents available tools and datasets, and specifically targets resources for English, Basque, Bulgarian, Czech, Portuguese and Spanish, WP5 working languages in QTLeap. Section 6 looks into the role of lexical semantics in machine translation, specifically, in the tasks addressed in QTLeap. Some conclusions of this deliverable will be discussed in Section 7.

2 Named Entity Recognition and Classification

The development of NERC systems for a great variety of languages and domains has to overcome a dependency on large amount of manually-annotated data which is very expensive to produce and the adaptation to a variety of languages and domains. As we will see, current tools and services both in academia and industry are moving towards using large knowledge bases (KBs) such as Wikipedia and its Linked Data[3] derivations such as DBpedia[4], Freebase[5], or any other Linked Data sources.

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 like 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.

The impact of textual genre (journalistic, scientific, informal, etc.) and domain (gardening, sports, business, etc.) has been rather neglected in the NERC literature. Few studies are specifically devoted to diverse genres and domains. Researchers at the University of

[3] http://linkeddata.org /

[4] http://dbpedia.org

[5] http://www.freebase.com/

Sheffield designed a system for emails, scientific texts and religious texts (Maynard et al. 2001). Minkov et al. (Minkov, Wang, and Cohen 2005) created a system specifically designed for email documents. Perhaps unsurprisingly, these experiments demonstrated that although any domain can be reasonably supported, porting a system to a new domain or textual genre remains a major challenge. Another work tested some systems on both the MUC-6 collection composed of newswire texts, and on a proprietary corpus made of manual translations of phone conversations and technical emails (Poibeau and Kosseim 2001). They report a drop in performance for every system (some 20% to 40% of precision and recall). In this respect, the QTLeap project will build domain-specific machine translation systems, and therefore, the lexical semantic tools will be developed in line with this focus.

A good proportion of work in NERC research is devoted to the study of English but a proportion addresses language independence and multilingualism. With respect to the languages involved in QTLeap, Spanish is also quite well represented, whereas the remaining languages show scarce resources. Finally, there have been numerous studies for French (Petasis et al. 2001; Poibeau 2003) and Italian (Black, Rinaldi, and Mowatt 1998; Cucchiarelli and Velardi 2001).

Named Entities are usually thought of as consisting of one or more rigid designators (Kripke 1980). For instance, “the automotive company created by Henry Ford in 1903” can be referred to as Ford or Ford Motor Company. Rigid designators include proper names as well as certain natural kind terms like biological species and substances. There is a general agreement in the NERC community about the inclusion of temporal expressions and some numerical expressions such as amounts of money and other types of units. While some instances of these types are good examples of rigid designators (e.g., the year 2001 is the 2001st year of the Gregorian calendar) there are also many invalid ones (e.g., in June refers to the month of an undefined year – past June, this June, June 2020, etc.). It is arguable that the NE definition is loosened in such cases for practical reasons.

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 “entertainer” appear in the aforementioned work (Fleischman and Hovy 2002). In the ACE[6] 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[7]. 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), etc.

Other works do not limit the possible types to be extracted and are referred to as “open domain” NERC (Alfonseca and Manandhar 2002; Evans and Street 2004). For example, a named entity hierarchy has been defined which includes many fine grained subcategories,

[6] http://www.itl.nist.gov/iad/mig/tests/ace/

[7] http://www.timexportal.info/system:page-tags/tag/timex2

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, Santorini, and Marcinkiewicz 1993) by the BBN Pronoun Coreference 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 (many of these listed in Section 5.1.2) 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 expensive manually annotated data hinders the creation of NERC systems for most languages and domains. In fact, this has been a major impediment to adaptation of existing NERC systems to other domains, such as the scientific or the biomedical domain (M. 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 such 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 Muñoz 2006). However, it has been shown that this does not necessarily correspond to better results in NERC performance (Mikheev, Moens, and Grover 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 automatic low-cost annotation (in the form of so-called silver standard corpora which has been produced automatically) as opposed to supervised approaches highly dependent on large amounts of manually annotated data (gold-standard). A crucial role on the development of silver standard has been the rise to prominence of Wikipedia. Wikipedia provides a large source of world knowledge which can be potentially a source of silver-standard data for NE annotations (Richman and Schone 2008; Mika et al. 2008; Nothman, Curran, and Murphy 2008; Nothman et al. 2012).

Richman and Schone (2008) develop a system for six languages which is evaluated against the automatically-derived annotations of Wikipedia and on manually-annotated Spanish, French and Ukrainian newswire. Their evaluation uses Automatic Content Extraction entity types (LDC and others 2005), as well as MUC numerical and temporal annotations that are largely not derived from Wikipedia. Their results with a Spanish corpus built from over 50,000 Wikipedia articles are comparable to 20,000–40,000 words of gold-standard training data.

The use of Wikipedia infoboxes has also proved to be valuable (Mika et al. 2008). Using the summary as a list of key-value pairs their system tries to find instances of such values in the article's text labeling it with its corresponding key.

Other works (Nothman, Curran, and Murphy 2008; Nothman et al. 2012) produce silver-standard CONLL annotations from English Wikipedia, and show that Wikipedia training can perform better on manually-annotated news text than a gold-standard model trained on a different news source. Moreover, they show that a Wikipedia-trained model outperforms newswire models on a manually-annotated corpus of Wikipedia text (Balasuriya et al. 2009).

The use of large multilingual resources such as Wikipedia aims at overcoming the reliance on manually-annotated data for the development of NERC systems. Such systems are normally a core component in many natural language processing applications. Thus, many NERC modules listed in 5.1.2 are part of a more general NLP module.

3 Named Entity Disambiguation

As explained in Section 2, 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), e.g. OpenCalais service for English[8].

Once the named entities are recognised they can be identified with respect to an existing catalogue. Wikipedia has become the de facto standard as such a named entity catalogue. Entity Linking 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 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[9], CiceroLite from LCC[10]and, Zemanta[11], TAGME[12] or The Wiki Machine[13].

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 word sense disambiguation must be performed. For example, the Wikipedia disambiguation page lists many different articles that the term NH might refer to (a state within the United States of America, a Spanish-based hotel chain among fourteen possibilities[14]).

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

I stayed in **NH**[15] in **Brussels**[16] and **Zurich**[17] and I really liked them because of their modern and stylish design and big rooms.

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 1998). 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

[8] http://www.opencalais.com
[9] http://spotlight.dbpedia.org/demo/index.html
[10] http://demo.languagecomputer.com/cicerolite/
[11] http://www.zemanta.com
[12] http://tagme.di.unipi.it/
[13] http://thewikimachine.fbk.eu/html/index.html
[14] http://en.wikipedia.org/wiki/NH
[15] http://en.wikipedia.org/wiki/NH_Hoteles
[16] http://en.wikipedia.org/wiki/Brussels
[17] http://en.wikipedia.org/wiki/Zurich

and Pasca 2006). This means that, as opposed to Wikification, links were to be 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 will be used to refer to any of these three tasks indistinctly (Hachey et al. 2012).

Different approaches have been proposed for the NEL. A system typically searches for candidate entities and then disambiguates them, returning either the best candidate or nil. Thus, although both WSD and NEL address ambiguity and synonymy in natural language, there is an important difference. In NEL, it is not assumed that the KB is complete, as it is in WSD with respect to WordNet, which means that named entity mentions present in the text which do not have a reference in the KB must be marked as NIL. Moreover, the variation of named entity mentions is higher than that of lexical mentions in WSD, which makes the disambiguation process much noisier.

The rise to prominence of Wikipedia has allowed developing wide-coverage NEL systems. The most popular task, in which both datasets and NEL systems can be found, is the Knowledge Base Population (KBP) task at the NIST Text Analysis Conference (TAC). 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. The TAC 2012 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:

"Given a name (of a Person, Organization, or Geopolitical Entity) and a document containing that name, determine the KB node for the named entity, adding a new node for the entity if it is not already in the KB".[18]

The popularity of the KBP task has led to a huge number of NEL systems, although given that every participant was aiming at obtaining the highest accuracy, most of the systems differ in so many dimensions that it is rather unclear which aspects of the systems are actually necessary for good performance and which aspects are harming it (Hachey et al. 2012)

The first large set of manually annotated named entity linking data was prepared for the KBP 2009 edition (McNamee et al. 2010). In the KBP 2012 edition, the reference KB is derived from English Wikipedia, while source documents come from a variety of languages, including English, Chinese, and Spanish. Other NEL evaluation datasets have been compiled independently of the KBP TAC shared tasks. These, together with other resources based on Linked Data which can be used for NED will be listed and described in Section 5.2.1.

NED allows computing direct references to people, locations, organizations, etc. For example, in the financial domain NED can be used to link textual information about organizations to financial data, or in the tourist domain, NED can link information about hotels or destinations to particular opinions and/or facts about them.

In QTLeap, we will build NED systems that extract the appropriate semantic knowledge and properties concerning the named entities of interest for all the relevant working languages. 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).

[18] http://www.nist.gov/tac/2012/KBP/index.html

4 Word Sense Disambiguation

Word sense disambiguation stands for labeling 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 Veronis, 1998; Edmonds and Kilgarriff, 2002) and surveys (Navigli, et al. 2009); and books focusing on this issue (Ravin and Leacock, 2000; Stevenson, 2003; Agirre and Edmonds, 2006). Despite the work devoted to the task, 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; Zhong and Ng, 2010). Such a level of performance allows for improving tasks such as Machine Translation (Chan et al. 2007; Carpuat and Wu, 2007), syntactic parsing (Agirre et al. 2008), Information Extraction (Chai and Biermann, 1999), Information Retrieval (Stokoe et al. 2003; Liu et al. 2005b; Agirre et al. 2009c) and Cross-Linguistic Information Retrieval (Clough and Stevenson, 2004; Vossen et al. 2006).

WSD systems are usually classified as supervised or unsupervised. **Supervised methods** use machine learning on manually produced sense-annotated corpora. Supervised approaches (Màrquez et al. 2006), include probabilistic methods (such as Naïve Bayes or Maximum Entropy), similarity methods (such as Vector Space Models or K-Nearest Neighbors), methods based on discriminating rules (such as Decision Lists or Decision Trees) or margin-based methods (Support Vector Machines), etc.

Machine learning classifiers are undeniably 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. 1992). Due to this knowledge acquisition bottleneck problem, they are only available for words which occur in the training corpus. For most of the languages, no large-scale broad-vocabulary sense-annotated corpora exist. Note that acquiring such corpora is very expensive. For instance, (Ng, 1997) estimates that in order 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 16person-years per language, according to the experience of (Ng and Lee, 1996).

Beyond the scarcity of annotated corpora, 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 of bigrams, trigrams, and "bags of words" are too limited for modeling the contexts of the target words. Thus, some researchers have enriched the feature representation by including more sophisticated features such assyntactic dependencies (Chen and Palmer, 2009; Gaustad, 2004) or semantic classes (Izquierdo et al. 2010).

In order to overcome the scarcity of hand-annotated data, a number of research lines are being pursued. For instance, the use of 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 deLacalle, 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.

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 in performance when the training is carried out in one corpus and the testing in 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, but still a big drop in performance is observed.

Traditionally, **unsupervised approaches** are those not using manually sense-annotated data for training a supervised machine learning system. 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 completely "not supervised". Unsupervised methods can be grouped as knowledge-based methods, including graph-based methods, and corpus induction methods.

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 knowledge sources applied to Word Sense Disambiguation is summarized in (Agirre and Stevenson, 2005).

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; Laparra et al. 2010). 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 a given context. Graph-based methods have great advantages. Firstly, no training corpus 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 applied to a set of closely related words. For instance, using UKB[19] (Agirre and Soroa 2009), the KYOTO project developed knowledge-based WSD modules for English, Spanish, Basque, Italian, Dutch, Chinese and Japanese. Note also that this type of algorithms is also useful to compute semantic similarity of words and sentences (Agirre et al. 2010a).

Corpus-based induction 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.

[19] http://ixa2.si.ehu.es/ukb

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 disambiguates 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[20], despite the wide range of approaches investigated and the large effort devoted to tackle this problem, assigning the appropriate meaning to words in context has resisted all attempts to be fully successfully addressed.

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 ranges between 50% and 80% accuracy.

However, unsupervised methods and, in particular, graph-based methods present very appealing advantages. They are not dependent on a manually labeled corpus for training and obtain better results when applied to a set of closely related words than when applied to running text (Navigli and Velardi, 2005; Agirre et al. 2009b; Niemann and Gurevych, 2011).

When addressing WSD in specific domains, supervised methods perform worse compared to their performance in a general domain (Escudero et al. 2000; Martínez and Agirre, 2000). Following this direction (Agirre et al. 2009b; Agirre and Lopez de Lacalle, 2009) study the problem of domain WSD using different knowledge-based and machine learning techniques. They report that the best performing method which does not involve hand-annotating domain data is graph-based WSD.

5 Data Sources and Tools

This section lists the data sources and tools available for each of the lexical semantics tasks described in Sections 2-4, namely, NERC, NED and WSD, relevant to the QTLeap WP5. It will provide an overview of existing resources to be directly used or adapted for the specific domain and languages relevant to the QTLeap project.

5.1 Named Entity Recognition and Classification

In 5.1.1 the main existing data sources currently available for the development (both in industrial and academic settings) and evaluation of NERC systems is described. Generally, from the beginning of the MUC and CONLL shared tasks, these data sources have consisted of manually annotated data which serves as training sets of machine learning models for NERC classification. The performance of these systems is usually evaluated using the F-measure computed as the harmonic mean between *precision* and *recall.* As explained in 2.1 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).

[20] http://www.semeval.org

The tools and services for NERC described in Section 5.1.2 are mostly based on supervised machine learning approaches, although some systems make use of knowledge resources such as gazetteers.

5.1.1 NERC Data Sources

Table 1 lists the data sources, available for the 6 languages included in WP5 (English, Basque, Bulgarian, Czech, Portuguese and Spanish), in the form of annotated corpora for training and evaluating 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/...
- **Provision:** method and availability of the data. For example, API, WS, files, databases, etc.
- **Storage:** 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	Provision	Storage	Amount	Language	License	Website
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/
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/
BBN	Wall Street Journal 1998	Source files available at Linguistic Data Consortium	Plain text	1173766 annotated tokens for dev/train/test	English	US $1000.00 for non-members of the Linguistic Data Consortium, Private	Linguistic Data Consortium
Wikigold	Wikipedia 2008	Available at http://schwa.org/projects/resources/wiki/Wikiner	Plain text CONLL format	38007 tokens for testing	English	CC-BY 3.0 license, PU	http://schwa.org/projects/resources/wiki/Wikiner
WikiNER popular	Wikipedia pages	Available at http://schwa.org/projects/resources/wiki/Wikiner	Files in three annotation formats: CONLL, medium an fine	2322 popular Wikipedia pages	English	CC-BY 3.0 license, PU	http://schwa.org/projects/resources/wiki/Wikiner
WikiNER random	Wikipedia pages	Available at http://schwa.org/projects/resources/wiki/Wikiner	Files in three annotation formats: CONLL, medium an fine	4K random Wikipedia pages	English 2531 pages Spanish 203 pages Portuguese 202 pages	CC-BY 3.0 license, PU	http://schwa.org/projects/resources/wiki/Wikiner
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 all QTLeap languages except Basque	Free for research purposes. See license	http://langtech.jrc.it/JRC-Names.html

• Data Entity	Type of data	Provision	Storage	Amount	Language	License	Website
Ontonotes 4.0	Newswire and web text	Available at Linguistic Data Consortium	Treebank sentences with named-entity and coreference information	~1M words	English	Private	Linguistic Data Consortium
Ancora Corpus	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
Egunkaria 2000	Newswire articles made available by the Basque newspaper Euskaldunon Egunkaria, 2000	Contact UPV/EHU	XML	383 pieces of news, 62,187 words, 4,748 NEs	Basque	Contact UPV/EHU	N/A
CINTIL	Newspaper articles and fiction	Downloadable	XML	~1M tokens	Portuguese	ELDA	http://cintil.ul.pt/
HAREM	Various	Downloadable	XML	129 gold documents	Portuguese	Free	http://www.linguateca.pt/HAREM/
CNEC 1.0	News and books	http://hdl.handle.net/11858/00-097C-0000-0022-C73C-7	XML	5800 sentences	Czech	CC BY-NC-SA	http://ufal.mff.cuni.cz/cnec/
BulTreeBank	News media, Literature	free; available under license	Treebank sentences with NEs and sentence-internal coreference; in XML	15000 sentences, 21678 NEs	Bulgarian	free; available under license	www.bultreebank.org
Bulgarian Named Entity Lexicon	proper names from News Media	free; available under license	Lists of categorized NEs; in XML	26000 NEs	Bulgarian	free; available under license	www.bultreebank.org

Table 1: Data Sources for Named Entity Recognition and Classification.

5.1.1.1 CONLL 2002 datasets

The CONLL 2002 shared task 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. Among others, Spanish resources were made available. 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 of the Technical University of Catalonia (UPC) and the Center of Language and Computation (CLiC) of the University of Barcelona (UB).

5.1.1.2 CONLL 2003 datasets

The shared task of CoNLL-2003 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. 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 newswire articles from the Reuters Corpus. Due to copyright issues only the annotations were made available at CONLL and it is necessary to access the Reuters Corpus, which can be obtained from NIST for research purposes if one needs to build the complete datasets.

5.1.1.3 BBN Corpus

The BBN corpus (Weischedel and Brunstein 2005) supplements the one million word Penn Treebank corpus of Wall Street Journal texts (Marcus, Santorini, and Marcinkiewicz 1993). The corpus contains stand-off annotation of pronoun coreference, indicated by sentence and token numbers, as well as annotation of a variety of entity and numeric types. All annotation was performed by hand at the company BBN[21] using proprietary annotation tools.

The corpus contains two components:

- **Pronoun coreference**. Stand-off annotation of pronoun coreference of the WSJ corpus is provided in a single file. Pronouns and antecedents are indexed by sentence and token numbers.
- **Entity types**. The corpus includes annotation of 12 named entity types (Person, Facility, Organization, GPE, Location, Nationality, Product, Event, Work of Art, Law, Language, and Contact-Info), nine nominal entity types (Person, Facility, Organization, GPE, Product, Plant, Animal, Substance, Disease and Game), and seven numeric types (Date, Time, Percent, Money, Quantity, Ordinal and Cardinal). Several of these types are further divided into subtypes. Annotation for a total of 64 subtypes is provided.

5.1.1.4 Wikigold and WikiNER

WikiNER and Wikigold are resources developed at the University of Sidney and the company Capital Markets (Nothman et al. 2012). Wikigold (Balasuriya et al. 2009) is a gold-standard built from Wikipedia and consists of 39K annotated tokens. WikiNER is a much larger gold-standard which consists of two datasets:

- **Popular:** around 2k English Wikipedia pages classified using the type scheme shown above.
- **Random:** around 4k Wikipedia pages from 9 languages (Spanish and Portuguese are relevant for QTLeap).

These datasets are annotated with the following information separated by columns:

- **Type:** a hierarchical Named Entity type used to label Wikipedia pages.
- **CONLL:** a coarse-grained representation of the type based around CONLL NE types.
- **Medium:** a finer-grained representation of the type.
- **Fine:** a much finer-grained representation of the type.

[21] http://www.bbn.com/

5.1.1.5 JRC Names

JRC Names[22] 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 different scripts (Latin, Greek, Arabic, Cyrillic, Japanese, Chinese, etc.). JRC Names contains the most important names of the EMM name database[23], namely, 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, as well as about the same amount of variant spellings for these entities. Additionally, it contains a number of morphologically inflected variants of the names. The resource grows by about 230 new entities and an additional 430 new name variants per week.

5.1.1.6 OntoNotes 4.0

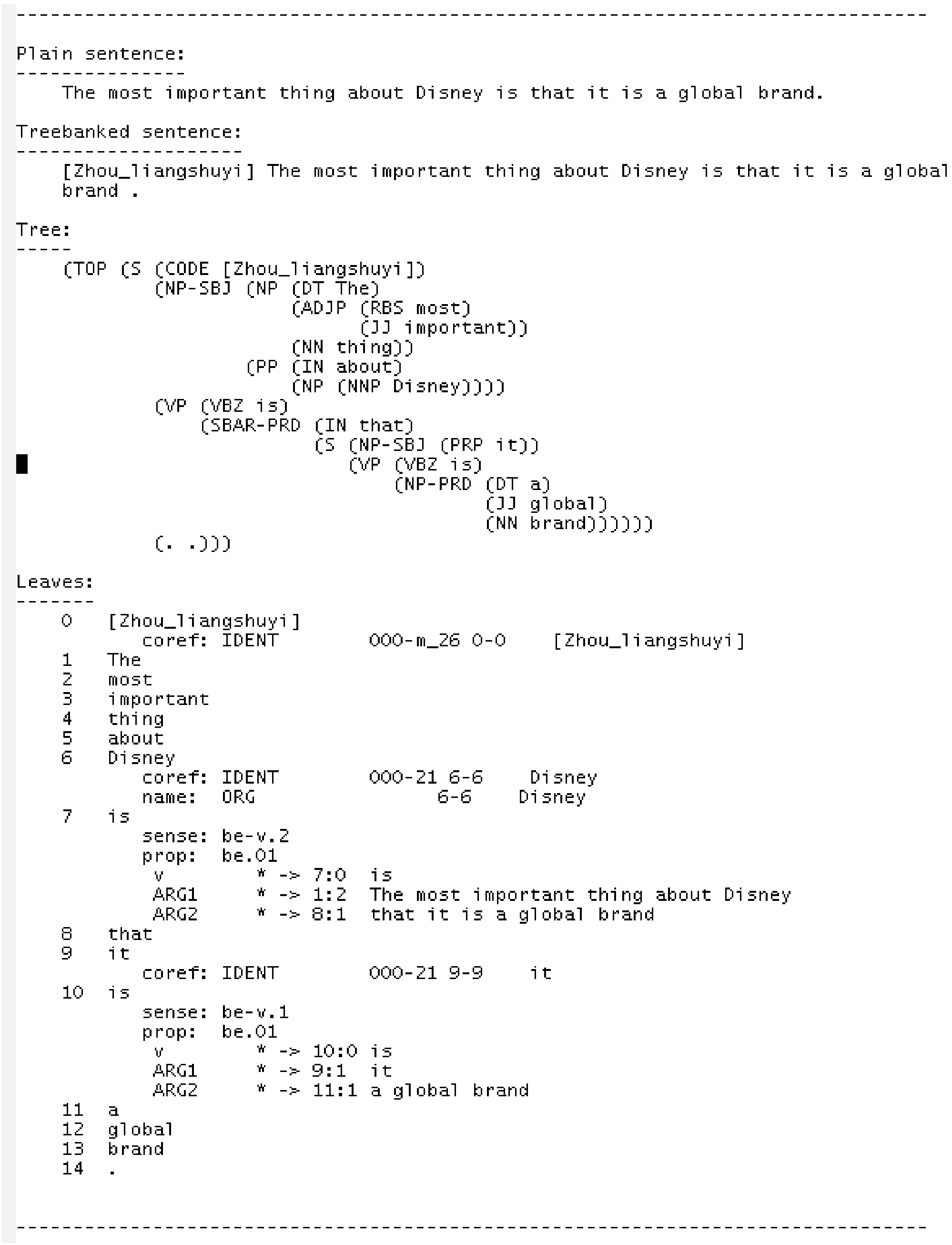
The OntoNotes project is a collaborative effort between BBN Technologies, Brandeis University, the University of Colorado, the University of Pennsylvania, and the University of Southern California's Information Sciences Institute to produce a rich semantic resource with annotation comprising various genres of text (news, conversational telephone speech, Web logs, Usenet newsgroups, broadcast, talk shows) in three languages (English, Chinese, and Arabic). The annotation includes syntax and predicate argument structure, and shallow semantics (word sense linked to an ontology and coreference) apart from named entity information (S. S. Pradhan et al. 2007). Figure 1 shows a sample of the annotation format in OntoNotes 4.0 for English.

5.1.1.7 Egunkaria 2000

The Egunkaria 2000 corpus is a collection of 383 newswire texts from 2000 built to train and test the Eihera+ NERC tool for Basque. It consists of 62,187 words and 4,748 manually reviewed named entity tags.

[22] http://langtech.jrc.it/JRC-Names.html

[23] http://emm.newsexplorer.eu



```
--------------------------------------------------------------------------------
Plain sentence:
---------------
    The most important thing about Disney is that it is a global brand.

Treebanked sentence:
--------------------
    [Zhou_liangshuyi] The most important thing about Disney is that it is a global
    brand .

Tree:
-----
    (TOP (S (CODE [Zhou_liangshuyi])
            (NP-SBJ (NP (DT The)
                        (ADJP (RBS most)
                              (JJ important))
                        (NN thing))
                    (PP (IN about)
                        (NP (NNP Disney))))
            (VP (VBZ is)
                (SBAR-PRD (IN that)
                          (S (NP-SBJ (PRP it))
                             (VP (VBZ is)
                                 (NP-PRD (DT a)
                                         (JJ global)
                                         (NN brand))))))
            (. .)))

Leaves:
-------
    0   [Zhou_liangshuyi]
           coref: IDENT        000-m_26 0-0    [Zhou_liangshuyi]
    1   The
    2   most
    3   important
    4   thing
    5   about
    6   Disney
           coref: IDENT        000-21 6-6    Disney
           name:  ORG                 6-6    Disney
    7   is
           sense: be-v.2
           prop:  be.01
            v        * -> 7:0  is
            ARG1     * -> 1:2  The most important thing about Disney
            ARG2     * -> 8:1  that it is a global brand
    8   that
    9   it
           coref: IDENT        000-21 9-9    it
    10  is
           sense: be-v.1
           prop:  be.01
            v        * -> 10:0 is
            ARG1     * -> 9:1  it
            ARG2     * -> 11:1 a global brand
    11  a
    12  global
    13  brand
    14  .

--------------------------------------------------------------------------------
```

Figure 1: OntoNotes Sample Annotation for English

5.1.1.8 Ancora

AnCora consists 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. Apart from named entities, the corpus provides annotation at various semantic levels including coreference.

5.1.1.9 CINTIL

CINTIL is a corpus of Portuguese with 1 Million annotated tokens, each one of which verified by human expert annotators. The annotation comprises information on part-of-speech, open classes lemma and inflection, multi-word expressions pertaining to the class of adverbs and to the closed POS classes, and multi-word proper names (for named entity recognition). Named entities are classified into Person, Location, Organization, etc. Over one third of the corpus is composed of transcribed spoken materials, with about half of that being the transcription of informal conversations. The remaining corpus is composed of written materials. The majority (58.73%) of this written corpus includes articles from newspapers and magazines, such as the Jornal Público, Diário de Notícias, Revista Visão, etc. The rest of the written corpus is mostly composed of fiction.

5.1.1.10 HAREM

HAREM is a joint evaluation task for NERC in Portuguese. A corpus of 129 documents from various genres, where entities have been manually annotated, is available.

5.1.1.11 CNEC 1.0

The Czech Named Entity Corpus 1.0 (CNEC 1.0) is the first publicly available corpus providing a large body of manually annotated named entities in Czech sentences (50,000 sentences), including a fine-grained classification (a two-level hierarchy of 80 NE types; nested NE annotated).

5.1.1.12 BulTreeBank

BulTreeBank is a syntactically annotated HPSG-based resource for Bulgarian, created under the BulTreeBank project (2002-2005) and funded by Volkswagen Stiftung Foundation, Germany. In 2006, for the CONLL contest, the treebank was converted into a dependency format. The Named Entities as well as the coreferences were manually annotated.

5.1.1.13 Bulgarian Named Entity Lexicon

The Bulgarian Named Entity Lexicon was compiled in two ways: 1. Data-driven (extracted from corpora) and 2. Dictionary-based (compiled from various existing proper name lists).

5.1.2 NERC Tools

Table 2 lists the services and available downloadable systems and tools to perform NERC for the 6 languages relevant to QTLeap. 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)
- **Responsible**: Name of the institution or company responsible of the service/system described. Responsibility is usually assigned according to the language of application.
- **Source availability**: Type of availability of the source code yes/no/partly
- **Provision**: Type of accessibility, namely, library, Web services, etc.
- **Programming Language**: Type of language used by the components: Java, C++, etc.
- **License**: 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	Source availability	Provision	Programming Language	License	Website URL
Open Calais	English, French, Spanish	No	Web service	Java, PHP, RDF	CC-SA	http://www.opencalais.com
BBN IdentiFinder Text Suite™	English, Spanish	No	Library or Service		Proprietary	http://bbn.com/technology/speech/identifinder
LingPipe	English	Yes	Library or Service	Java	Free for research, proprietary otherwise	http://alias-i.com/lingpipe
Stanford CoreNLP	English, German	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/
C&C Tools	English	Yes	Library	C++	Academic license, for research only	http://svn.ask.it.usyd.edu.au/trac/candc/
GATE	English, French, German	Yes	Library	Java	GNU GPLv2	http://gate.ac.uk
Kyoto NER	English, Dutch	Yes	Library	Java	Open Source	http://www.kyoto-project.eu
Eihera+	Basque	No				http://ixa2.si.ehu.es/demo/entitateak.jsp

System/Service	Languages	Source availability	Provision	Programming Language	License	Website URL
SProUT	Multilingual. Ready for English, Spanish, Czech				http://sprout.dfki.de/Licencing.html	http://sprout.dfki.de/
LX-NER	Portuguese	No	Local application, web service	Java, C	n/a	http://lxcenter.di.fc.ul.pt/services/en/LXServicesNer.html
Rembrandt	Portuguese	Yes	JAR	Java	GPL	http://xldb.di.fc.ul.pt/Rembrandt/
NameTag	Czech, English	Yes	Library	C++	GNU LGPL	https://redmine.ms.mff.cuni.cz/projects/nametag
CLaRK System	Bulgarian	No	jar	Java	free	http://www.bultreebank.org/clark/index.html
Bulgarian IT concept tagger	Bulgarian	No	jar	Java	free under license	http://www.bultreebank.org

Table 2: Systems and Services for Named Entity Recognition and Classification

5.1.2.1 OpenCalais

The OpenCalais Web Service automatically creates rich semantic metadata for the content you submit. Using machine learning and other methods, OpenCalais analyses your document and finds the entities within it. Calais goes beyond classic entity identification and provides some semantic processing returning facts and events from the text.

The web service is an API that accepts unstructured text (like 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 charge, although commercial and service support is available. It is available for use in commercial and non-commercial applications, the former at a cost.

OpenCalais defines entities as things like people, places, companies, geographies. Facts are relationships like John Doe is the CEO of Acme Corporation. Events are things that happened, like there was a natural disaster of type landslide in place Chula Vista. A number of Web applications using OpenCalais are listed in this URL: http://www.opencalais.com/showcase.

5.1.2.2 BBN IdentiFinder Text Suite™

The BBN IdentiFinder Text Suite™, is an industrial machine-learning tool to locate names of corporations, organizations, people, and places, including variations in names. It is a commercial application for which a service needs to be contracted. English and Spanish modules are available. No web service or public demo is available.

5.1.2.3 LingPipe

LingPipe is a software package from Alias-i and consists of several language processing modules: a statistical NERC, a heuristic sentence splitter, and a heuristic within-document coreference resolution system. LingPipe comes with an English language model. The types of named entities covered by LingPipe are locations, persons and organizations and offers pre-trained models for English. The LingPipe license allows its use for free for research purposes, but for commercial applications a rather costly licensing exists.

5.1.2.4 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 which can take raw English language text input and give, in a wide variety of output formats, the base forms of words, their parts of speech, named-entities, normalized dates, times, and numeric quantities, it marks up the structure of sentences in terms of phrases and word dependencies, and indicates which noun phrases refer to the same entities. Stanford CoreNLP is an integrated framework, which makes it very easy to apply a bunch of language analysis tools to a piece of text.

The Stanford CoreNLP code is written in Java and licensed under the GNU General Public License (v2 or later). Source is included. It requires at least 4GB to run. Although a German NERC module based on Stanford CoreNLP is also available (Faruqui and Padó 2010), the only language in QTLeap WP5 that benefits from this suite is English.

As listed in table 3, 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.

Type of model	Named Entities	Language
3 class	Location, Person, Organization	English
4 class	Location, Person, Organization, Misc	English, German
7 class	Time, Location, Organization, Person, Money, Percent, Date	English

Table 3: Stanford NERC models

5.1.2.5 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.

5.1.2.6 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, Màrquez, and Padró 2003). NERC is available in Freeling for English and Spanish. Freeling is licensed under the GPL. Each module requires about 2GB to run.

5.1.2.7 OpenNLP

OpenNLP is a general suite of NLP processing part of the Apache Software Foundation. The NERC module provides pre-trained models for English and Spanish (as well as Dutch) based on the CONLL datasets listed in Sections 5.1.1.1 and 5.1.1.2. It is developed in Java and distributed under the Apache license v.2.

5.1.2.8 C&C Tools

C&C tools is an NLP suite of processors developed in C++ and Prolog and build around Combinatory Categorial Grammar (CCG). This makes it suitable as a base for linguistic-based approaches to sentiment analysis that exploits compositionality (Simančík and Lee). C&C tools include a maximum-entropy-inspired NERC module for English based on CONLL 2003 data sources. C&C tools is released under an academic license for research purposes only.

5.1.2.9 GATE

Gate is a hugely comprehensive development environment and NLP suite for language processing developed at the University of Sheffield. It includes plug-ins for many applications including NLP tools such as LingPipe, TreeTagger, Stanford CoreNLP, OpenNLP, etc. Primarily, GATE supports NERC for English. GATE is released under the GPLv2 license.

5.1.2.10 KyotoNER

The Named-entity tagger developed as part of the Kyoto project[24] detects time points and places in KAF as named-entities. It applies named-entity disambiguation and represents the named-entities in a separate layer in KAF with GeoNames properties and WordNet mappings for locations.

5.1.2.11 Eihera+

Eihera+ is a NERC system for Basque. It classifies entities into three groups, namely, person, organization and location. It combines finite-state and machine learning technologies: recognition by rules, recognition by ML, classification by rules, classification by ML. It requires the use of Eustagger, a morphosyntactic tagger, as a previous step (Alegria et al. 2006).

5.1.2.12 SProUT

SProUT (Shallow Processing with Unification and Typed Feature Structures) is also a platform for development of multilingual shallow text processing and information extraction systems. It consists of several reusable Unicode-capable online linguistic processing components for basic linguistic operations ranging from tokenization to coreference matching. Since typed feature structures (TFS) are used as a uniform data structure for representing the input and output by each of these processing resources, they can be flexibly combined into a pipeline that produces several streams of linguistically annotated structures, which serve as an input for the shallow grammar interpreter, applied at the next stage. The grammar formalism in SProUT, called XTDL is a blend of very efficient finite-state techniques and unification-based formalisms which are known to guarantee transparency and expressiveness. Currently, the platform provides linguistic processing resources for several languages including among other English, German, French, Italian, Dutch, Spanish, Polish, Czech, Chinese, and Japanese.

5.1.2.13 LX-NER

LX-NER recognizes and classifies named expressions. It is composed by two sub-systems: a ruled-based NER for number-based expressions built upon handcrafted regular expressions; and a named-based component built upon stochastic procedures. Each sub-system achieves ~85% f-score.

5.1.2.14 Rembrandt

Rembrandt is a rule-based system that uses DBpedia as a knowledge base.

5.1.2.15 NameTag

A new named entity recognizer for the Czech language (Straková et al. 2013). The recognizer is based on Maximum Entropy Markov Model and a Viterbi algorithm decodes an optimal sequence labeling using probabilities estimated by a maximum entropy

[24] http://www.kyoto-project.eu/

classifier. The classification features utilize morphological analysis, two-stage prediction, word clustering and gazetteers.

5.1.2.16 CLaRK System

CLaRK system is an IICT-BAS in-house system for creation, annotation and maintenance of language resources. It supports core NLP processing steps via regular grammars, among which Named Entity Detection.

5.1.2.17 Bulgarian IT concept tagger

The Bulgarian IT concept tagger consists of cascaded regular grammars for identifying ontological concepts in IT domain. It is based on a domain ontology and related terminological lexicons.

5.2 Named Entity Disambiguation

This section describes the relevant data sources and tools for NED. The data sources are mainly either text corpora developed for NLP applications or Linked Data as part of the Linked Data[25] initiative. Most of the research on NED systems has been undertaken on text corpora, although some systems are already using Linked Data datasets such as DBpedia[26].

5.2.1 NED Data Sources

The data sources and systems described in this section will be those relevant to Wikification and Named Entity Linking. The term Named Entity Disambiguation will be used to refer to any of these two tasks indistinctly (Hachey et al. 2012).

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

As mentioned in Section 5.1.2, the first large datasets on NEL were created by the TAC for the KBP track. So far there have been 4 editions since 2009. Originally, the datasets were only available for English but the 2012 edition includes documents in Spanish. In addition to 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 reviewed in Sections 5.1.1.1 and 5.1.1.2 (Hoffart et al. 2011).

Other valuable datasets listed in table 4 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 be any type of named entity currently available in the Web. Well-known and large linked data resources in the NLP community are DBpedia, Freebase[27] and

[25] http://linkeddata.org/

[26] http://dbpedia.org

[27] http://www.freebase.com

Yago[28], 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.

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

Data Entity	Type of data	Provision	Storage	Amount	Language	License	Website
CSWA	Web pages	Text corpora	Annotated files	17200 annotated instances	English	Public	http://soumen.cse.iitb.ac.in/~soumen/doc/CSAW/
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	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
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/
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
TAGME	Wikipedia articles	Text Corpora	Files	~2 million fragments of Wikipedia articles	English, Italian	CC-BY-SA license	http://acube.di.unipi.it/tagme-dataset/

Data Entity	Type of data	Provision	Storage	Amount	Language	License	Website
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
Google Dictionary	Wikipedia Concepts	Dictionary	Text files	Mapping 175M of strings to related Wikipedia articles	English	Public	http://nlp.stanford.edu/pubs/crosswikis-data.tar.bz2
Dbpedia	Wikipedia articles	API, dump	Linked Data	~3.77 million named entities	Multilingual, including English, Spanish, German, Dutch, Italian, French	CC-BY-SA license	http://dbpedia.org
Freebase	Web pages	API, dump	Linked Data	~23 million of named entities	Multilingual	CC-BY 3.0 license, PU	http://www.freebase.com
YAGO2	Web pages, Wikipedia	API, dump	Linked Data	~10 million of named entities	Multilingual		http://www.mpi-inf.mpg.de/yago-naga/yago/
GeoNames	Web	Web services, dump, premium dump	Linked Data	~8 million geographic entities	Multilingual	CC-BY 3.0 license, PU	http://www.geonames.org/
LinkedGeo Data	Web	Web service, API, dump	Linked Data	~6 million location instances	Multilingual	CC-BY-SA license	http://linkedgeodata.org
Geo-Net-PT	?	Downloadable	OWL	?	Portuguese	CC-BY	http://www.linguateca.pt/GeoNetPT/
OKKAM	Entities	Web Service, Downloadable	Linked Data	7.5 millions	Multilingual	Apache v2.0	http://www.okkam.org

Table 4: Data Sources for Named Entity Disambiguation

5.2.1.1 CSWA

For the creation of this corpus researchers at IIIT Bombay built a ground truth collection using a browser-based annotation system. Documents for manual annotation were collected from the links within homepages of popular sites from a handful of domains including sports, entertainment, science and technology, and health. As with Cucerzan's dataset (Cucerzan 2007), the CSWA data also has high average ambiguity, although Cucerzan’s is higher because the spots are limited to common person and place names.

The authors collected a total of about 19,000 annotations by 6 volunteers. Unlike in previous work, volunteers were told to be as exhaustive as possible and tag all possible segments, even if to mark them as NA (not attached). The number of distinct Wikipedia entities that were linked to was about 3,800. About 40% of the spots was labeled as NA, highlighting the importance of back-offs. However, this also means that 60% of the spots were linked by the volunteers, which exceeds by far the token rate in other work (see KBP datasets, for example).

5.2.1.2 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 consisted of an ID that identified a document within a set of Reuters news articles, a mention string that occurred at least once within that document, and a node ID within the knowledge base. Each knowledge base node contained 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 were predicted to correspond to a named entity were included in the knowledge base. The annotators favoured mentions that were likely to be ambiguous, in order to provide a more challenging evaluation. If the entity referred to did not occur in the knowledge base, it was labeled NIL. A high percentage of queries in the 2009 test set did not map to any nodes in the knowledge base: the gold standard answer for 2,229 of the 3,904 queries was NIL.

In the 2010 challenge the same configuration as in the 2009 challenge was used with the same knowledge base. In this edition, however, a training set of 1,500 queries was provided, with a test set of 2,250 queries. In the 2010 training set, only 28.4% of the queries were NIL, compared to the 57.1% in the 2009 test data and the 54.6% in the 2010 test data. This mismatch between the training and test data showed the importance of the NIL queries and it is argued that it may have harmed performance for some systems. This is 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 a NIL. The most successful strategy to deal with this issue in the 2009 challenge was augmenting the knowledge base with extra articles from a recent Wikipedia dump. If a strong match against articles that did not have any corresponding node in the knowledge base was obtained, then NIL was return for these matches.

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

5.2.1.3 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.

5.2.1.4 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 was no corresponding page. Overall, 160 of the 500 mentions resolved to NIL (Fader et al. 2009).

5.2.1.5 Dredze 2010

In order to generate 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.

5.2.1.6 ACEtoWiki

ACEtoWIKI is the result of a joint effort between FBK[29] and CELCT[30]. The resource has been created by adding a manual annotation layer connecting the English ACE-2005 Corpus to Wikipedia. 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.*

5.2.1.7 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[31]. The CoNLL 2003 dataset is required to create the corpus.

5.2.1.8 TAGME Datasets

The TAGME Datasets is a collection of datasets that contain short text fragments drawn from the Wikipedia snapshot of Novembre 6, 2009. Fragments are composed by about 30 words, and they contain about 20 *non-stopwords* on average (Ferragina and Scaiella 2010). The authors gathered fragments of 2 types:

[29] http://www.fbk.eu/

[30] http://www.celct.it

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

- **WIKI-DISAMB30**, a list of 2M fragments each containing one ambiguous anchor. For each fragment two lines are deployed: the former contains the text (no *lower-case* was applied, we cleaned Wikipedia syntax by leveraging some heuristics), the latter contains the anchor (in *lower-case*) followed by the numeric ID of Wikipedia page which is pointed by the anchor. Anchor and ID are separated by a TABcharacter.
- **WIKI-ANNOT30**, a list of 186K fragments. The syntax is almost the same: the first line contains the text, the second one contains a list of annotated anchors found in the text, followed by numeric IDs of pages which are pointed by these anchors. A TABcharacter separates anchors and IDs in the list. Text and anchors are cleaned as for the previous dataset.

5.2.1.9 Illinois Wikifier Datasets

These datasets were created for the evaluation of the paper from which originated the Illinois Wikifier system (Ratinov et al. 2011) described in Section 5.2.2.5. They 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%, we manually corrected the annotations to obtain ground truth for our 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 cases the disambiguation task is trivial. The ACE-based corpus contains 257 mentions whereas the Wikipedia-based consists of 928 mentions.

5.2.1.10 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 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.

5.2.1.11 Google Wikipedia Concepts Dictionary

The Google Wikipedia Concepts dictionary is built by means of a mechanism for mapping between Wikipedia articles and a lower-level representation: free-form natural language strings in many languages.

The resource closely resembles a dictionary, with canonical English Wikipedia URLs on the one side, and relatively short natural language strings on the other. These strings come from several disparate sources, primarily: (i) English Wikipedia titles; (ii) anchor texts from English inter-Wikipedia links; (iii) anchor texts into the English Wikipedia from non-Wikipedia web-pages; and (iv) anchor texts from non-Wikipedia pages into non-English Wikipedia pages, for topics that have corresponding English Wikipedia articles. Unlike entries in traditional dictionaries, however, the strengths of associations between related pairs in their mappings can be quantified using basic statistics. They sorted the data using one particularly simple scoring function (a conditional probability) but all raw counts are also included.

5.2.1.12 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[32].

5.2.1.13 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[33].

5.2.1.14 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 spatial 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.

5.2.1.15 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 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.

[32] http://wiki.dbpedia.org/Datasets

[33] http://wiki.freebase.com/wiki/Freebase_data

5.2.1.16 LinkedGeoData

LinkedGeoData uses the comprehensive OpenStreetMap[34] 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.

5.2.1.17 Geo-Net-PT

Geo-Net-PT is an ontology of geographical information for locations in Portugal. Names are tagged with their translation into various languages (e.g. Lisboa also appears as Lisbon[EN], Lisbonne[FR] and Lissabon[DE]).

5.2.1.18 OKKAM

The overall goal of the OKKAM project[35] 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) which harvested entities (together with an automatically created profile) from some popular public data sources like Wikipedia/DBpedia, GeoNames, UNIProt, etc. They currently offer a repository of 7.5 million entities. There is a public demo of the ENS and the tools are available to download[36]. In particular, they offer the ENS Java Client which is distributed under Apache license v2.0.

5.2.2 NED Tools

Most of the currently available systems have been developed as a result of the popularity of the Wikification and KBP tasks introduced in Section 5.2. Furthermore, the rise of Linked Data datasets has 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). As in previous sections, table 5 lists the available systems and services for NED and thereafter some details of each system are provided.

[34] http://openstreetmap.org/
[35] http://www.okkam.org
[36] http://community.okkam.org/

System/Service	Languages	Sources availability	Provision	Programming Language	License	Website URL
Zemanta	English	NO	Browser add-on, API	Multiple	Free for non-commercial uses	http://www.zemanta.com
The Wiki Machine	English	Yes	Library			http://thewikimachine.fbk.eu
AlchemyAPI	English, Portuguese, Spanish	No	API	Multiple	Proprietary	http://www.alchemyapi.com
CiceroLite LCC	English	No	API, service		Proprietary	http://www.languagecomputer.com/
Illinois Wikifier	English	Yes	Jar, Library	Java	Public	http://cogcomp.cs.illinois.edu/page/software_view/Wikifier
DBpedia Spolight	Dutch, English, German, Portuguese, Spanish	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/
BasqueNED	Basque	No				

Table 5: Systems and Services for Named Entity Disambiguation

5.2.2.1 The Wiki Machine

The Wiki Machine is a Wikification system developed at the FBK in Trento, Italy. In addition to machine learning techniques, they use Linked Data to offer multilingual (QTLeap-relevant English and Portuguese) wikification via DBpedia and Freebase. They also offer a publicly available demo which compares their results with respect to AlchemyAPI, Zemanta and OpenCalais.

5.2.2.2 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.

5.2.2.3 AlchemyAPI

AlchemyAPI is a text mining platform providing a wide set of semantic analysis capabilities. AlchemyAPI enables customers to perform large-scale social media monitoring, target advertisements more effectively, track influencers and sentiment within the media, automate content aggregation and recommendation, etc. AlchemyAPI supports 8 languages, 3 of which are addressed in the QTLeap project. AlchemyAPI has 4 types of products regarding on how to access their service. A free service allows 1,000 API calls a day upon registration, and approved academic users may receive up to 30,000 API calls a day after contacting the company. They offer higher limits available to educational institutions and non-profit groups. The licensing terms are proprietary and it cannot be built to develop another commercial competitor system.

5.2.2.4 CiceroLite from LCC

LCC's CiceroLite family of entity extraction systems are provided for English, Arabic, and Chinese texts. CiceroLite includes more than 150 named entity types[37] and leverages resources such as Wikipedia and DBpedia for NED. CiceroLite is proprietary software.

5.2.2.5 Illinois Wikifier

The Illinois Wikifier system is developed at the Cognitive Computation Group at the University of Illinois at Urbana Champaign[38]. 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 not many approaches to NED have evaluated their results with the same datasets, the KBP participants being the general exception. A newer version of the tool exists (Cheng, et al. 2013).

[37] http://www.languagecomputer.com/EntityDocumentation

[38] http://cogcomp.cs.illinois.edu/

5.2.2.6 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 (Mendes et al. 2011; Daiber et al. 2013). 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, the machine learning professor or dbpedia:Michael_Jordan the basketball player).

DBpedia Spotlight can be used through their Web Application or Web Service endpoints. The Web Application is a user interface that allows entering 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 retrieval of data also in XML or JSON. DBpedia is released under the Apache License 2.0.

5.2.2.7 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.

5.2.2.8 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, described in Section 5.2.1.8, 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. A newer version of the tool exists (Cornolti, et al. 2013).

5.2.2.9 Basque NED

Basque NED is a NED system that combines of state-of-the-art methods for NED to work with Basque. It is based on the Basque Wikipedia. It uses MFS, VSM, ESA and UKB for linking ambiguous surface NE forms in a text with their corresponding Wikipedia entry in the Basque Wikipedia version (Fernández et al. 2011).

5.3 Word Sense Disambiguation

The following Section describes the relevant data sources and tools for WSD. Corpora annotated with word senses are the main resource for WSD. Also of great importance are the sets specifically annotated for shared tasks, mainly the SensEval/SemEval initiatives.

5.3.1 WSD Data Sources

5.3.1.1 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 this 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.).[39]

Based on SemCor, a bilingual corpus was created by (Bentivogli and Pianta, 2005): MultiSemCor is an English/Italian parallel corpus aligned at word level which provides, 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 word level. The original word sense tags from SemCor were then transferred to the aligned Italian words.

5.3.1.2 euSemCor

euSemCor is a Basque corpus of approximately 300,000 words annotated with word senses based on Basque WordNet senses or synsets (Pociello et al. 2010). It was created through a joint development of the Basque WordNet and a complementary corpus for Basque, the Basque SemCor. The words in the Basque WordNet were edited, EuSemCor was double-blind tagged with a referee, and further edited-tagged when required. EuSemCor consists of 1,355 tagged lemmas (59,968 occurrences, 47,8% of the total). The texts included in EuSemCor were chosen independently from the English SemCor.

5.3.1.3 OntoNotes

OntoNotes Release 4.0[40] (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 California 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

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

[40] http://www.ldc.upenn.edu/Catalog/catalogEntry.jsp?catalogId=LDC2011T03

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.

5.3.1.4Ancora

AnCora[41] (Taulé et al. 2008) consists 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
- Co-reference relations

The 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).

5.3.1.5 Meaning Bank

The Groningen Meaning Bank (GMB) consists of public domain English texts with corresponding syntactic and semantic representations (Basile, et al. 2012). The GMB is developed at the University of Groningen and the current (development) version of the GMB is accessible via the GMB Explorer.

The GMB supports deep semantics, opening the way to theoretically grounded, data-driven approaches to computational semantics. It integrates phenomena instead of covering single phenomena in isolation. This provides a better handle on explaining dependencies between various ambiguous linguistic phenomena, including word senses, thematic roles, quantifier scrope, tense and aspect, anaphora, presupposition, and rhetorical relations. In the GMB texts are annotated rather than isolated sentences, which provide a means to deal with ambiguities on the sentence level that require discourse context for resolving them.

5.3.1.6 Senseval/SemEval corpora

Since 1998, SensEval[42] and later on SemEval[43] organize an ongoing series of evaluations of computational semantic analysis systems. Along these years, multiple organizers have

[41] http://clic.ub.edu/corpus/en
[42] http://www.senseval.org/

provided a large number of multilingual datasets annotated at sense level (see table 6 for further details.)

Data Entity	Size	Language	License	Website
SemCor	234,000	English	GNU	http://www.cse.umt.edu/~rada/download.html#semcor
euSemCor	~300,000 1,355 words, 59,968 occurrences	Basque	Unknown	http://sisx04.si.ehu.es:8080/eusemcor/
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
Meaning Bank	1,020,367 tokens	English	Unknown	http://gmb.let.rug.nl/
MultiSemCor	92,420	English	CC-by-3.0	http://multisemcor.fbk.eu/index.php
SensEval2 English all-words WSD	5,000	English	Unknown	http://www.hipposmond.com/senseval2
SensEval3 Task 1 English all-words WSD	5,000	English	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	Unknown	http://xmlgroup.iit.cnr.it/SemEval2010/
SemEval2010 Task 03 Cross-lingual WSD	1,000	English, Spanish	Unknown	http://webs.hogent.be/~elef464/lt3_SemEval.html
SemEval2013 Task 10 Cross-lingual WSD	1,000	English, Spanish	Unknown	http://www.cs.york.ac.uk/semeval-2013/task10
SemEval2013 Task 12 Multilingual WSD	1,000	English, Spanish	Unknown	http://www.cs.york.ac.uk/semeval-2013/task12

Table 6: Data Sources for Word Sense Disambiguation

[43] http://en.wikipedia.org/wiki/SemEval

5.3.2 WSD Tools

5.3.2.1 SenseLearner

SenseLearner[44] (Mihalcea and Csomai, 2005) is a minimally supervised all-words WSD algorithm for English.

5.3.2.2 IMS

IMS (It Makes Sense)[45] (Zhong and Ng, 2010) is a supervised English all-words 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.

5.3.2.3 SuperSense Tagger

SuperSenseTagger[46] (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.

5.3.2.4 GWSD

GWSD[47] (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.

5.3.2.5 UKB

UKB[48] (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, KYOTO developed knowledge-based WSD modules for English, Spanish, Basque, Italian, Dutch, Chinese and Japanese. It has also been applied 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 is also useful to compute semantic similarity of words and sentences (Agirre et al. 2010a).

[44] http://www.cse.unt.edu/~rada/downloads.html#senselearner
[45] http://www.comp.nus.edu.sg/~nlp/software.html
[46] http://sourceforge.net/projects/supersensetag/
[47] http://www.cse.unt.edu/~rada/downloads.html#gwsd
[48] http://ixa2.si.ehu.es/ukb/

5.3.2.6 BabelNet API

BabelNet is both a multilingual encyclopedic dictionary, with lexicographic and encyclopedic coverage of terms, and a semantic network which connects concepts and named entities in a very large network of semantic relations, made up of more than 9 million entries, called Babel synsets (Navigli and Ponzetto, 2012a, 2012b). Each Babel synset represents a given meaning and contains all the synonyms which express that meaning in a range of different languages. Version 2.0 of the API was released October 2013.

Table 7 summarizes the WSD tools available.

System/Service	Languages	Source 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	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++	GPL v3	http://ixa2.si.ehu.es/ukb/
BabelNet API	Multilingual	?	Java	Creative Commons Attribution-Non Commercial-Share Alike 3.0 License.	http://babelnet.org/

Table 7: Systems and Services for Word Sense Disambiguation

6 Beyond the State of the Art: Lexical Semantics and Machine Translation in QTLeap

Semantic processing for lexical resolution is concerned with the resolution of referential ambiguity, more currently termed as named entity resolution and disambiguation, and the resolution of conceptual ambiguity, more currently termed as word sense disambiguation. The development and fast growth of LOD, ontologies and other semantic resources has now progressed to support an impactful application of lexical semantic processing that leverages a quantum leap in machine translation.

The first objective of WP5 in QTLeap is to advance uses of conceptual knowledge from LOD to develop new solutions that enhance parallel DeepBanks with referential and lexical ambiguity resolution and support deep processing for machine translation. To do so, the Consortium will work towards advancing the state-of-the-art on multilingual and cross-lingual named entity disambiguation and word sense disambiguation. Parallel texts will allow developing techniques that produce better quality annotations, leveraging the information available in one language with the information available in another. In addition, the joint processing of the different types of disambiguation, DeepBanks and the general translation process will be advanced.

The first step towards this aim will be the alignment of the available lexical resources for the project languages, namely, English, Basque, Bulgarian, Czech, Portuguese and Spanish to ontologies and instance data within selected **datasets from LOD**. The motivation for this is, for example, the work of Morris and Hirst (2004). They point out that most of the lexical relations necessary to determine the semantic content of lexical units are non-classical in contrast to the classical ones, i.e. hyponymy, meronymy, antonymy. The non-classical relations are specific to some classes of meanings, i.e. made-of, used-for, place-of-birth etc. The alignment will be done automatically by exploiting the existing alignments of wordnets, and by analysing related textual documents available on the web, such as the Wikipedia pages, with respect to the instances in the LOD datasets.

The use of **crosslingual links** in Wikipedia for NLP and MT has been studied (Filatova, 2009; Vasiljevs et al. 2012) with special emphasis on methods to gather comparable bilingual corpora (Otero and Lopez, 2010), to extract parallel sentences (Mohammadi and Aghaee, 2010), or as a source of lexical translations (Jones et al. 2008; Müller and Gurevych, 2009; Tyers and Pienaar, 2008; Arcan et al. 2014). For instance, Jones et al. (2008) used Wikipedia to augment a standard MT system with domain specific phrase dictionaries. Those dictionaries, automatically mined from the multilingual links between Wikipedia articles, were used to correct the output of a MT system. Experiments using our hybrid translation system with sample query logs demonstrate a large improvement in the accuracy of domain specific phrase detection and translation. In related work, Arcan et al. (2014) address the problems of automatic identification of bilingual terminology using Wikipedia as a lexical resource, and its integration into an SMT system. The correct translation equivalent of the disambiguated term identified in the monolingual text is obtained by taking advantage of the multilingual versions of Wikipedia. The small amount of high quality domain-specific terms is passed to the SMT system, which produced a significant relative translation improvement in BLEU.

QTLeap will focus on techniques similar to (Henrich et al. 2012) in order to prepare semi-automatically parallel and comparable corpora annotated with appropriate word senses. The parallel sentences will be extracted from the annotated corpora using techniques similar to (Jason et al. 2012). We will develop them further using the information from the selected LOD datasets. The actual annotation will be supported by the deep processing modules (NERC, NED, WSD) that will be used or implemented within the project. We will also explore techniques such as (Jones et al. 2012) that show that translation using deep representations as interlingua is feasible. Given the prevalence of English and its richer resources and corpora available, QTLeap will explore the use of English concepts and instances as the interlingual representation of choice. The project will explore several alternatives where expressions and their correlates within the fully fledged deep grammatical representations will be replaced by their interlingual conceptual representations from LOD and WordNet. We will also experiment with enriching the word-based representation with additional information, including the full probability vectors returned by the NER, NED and WSD components.

The use of **Named-Entity** related information is not novel. Hálek et al. (2011) propose a method to improve machine translation of named entities from English to Czech using Wikipedia. They first run a high recall named entity recognizer over the source text, and filter the recognized named entities that correspond to Wikipedia articles with some restrictions on their categories. They then identify the corresponding Czech Wikipedia article title for each of these named entities, as well as all of its inflected forms in Wikipedia, which will be the translation candidates. These candidates are then appropriately weighted and integrated into a Moses-based SMT system. Even though these leads to worse results when it comes to automatic metrics, a manual evaluation shows a slight improvement in translation quality. In more recent work, Li et al. (2013) develop a name-aware machine translation system for Chinese to English based on a hierarchical SMT system. They jointly recognize named entities in the parallel corpus and insert them as non-terminals in the phrase table. In the decoding phase, different techniques such as dictionary matching, statistical name transliteration and context information extraction based re-ranking are used to translate those named entities in the input text with less than 5 occurrences in the training corpus, which are dynamically added to the phrase table taking advantage of the non-terminals for named-entities introduced earlier. Their experimental results show a small improvement in terms of BLEU, and a more notable improvement for metrics that focus in named entity translation like the name-aware BLEU that they propose. Alternatively, work on name transliteration (Hermjakob et al. 2008) has been successfully used when translating from Arabic to English. Given a sequence of words to transliterate, their engine first identifies a list of English transliteration candidates from a large list of English words and phrases, and then choose the most similar one based on a rule-based scoring model. Instead of running a traditional named entity recognition system to identify the names to transliterate with the previous module, they build a task specific machine learning based system that tags those words and phrases in the input text that the transliteration module is likely to translate correctly. They then add these transliterations to the phrase table on the fly, with a special binary feature set to 1 whose weight is appropriately adjusted along with the rest in the tuning step during training. Their experimental results show an improvement in named entity translation accuracy as measured by the NEWA metric that they propose, and also a small gain in BLEU. The lessons learnt in this work will be considered for QTLeap, although the languages in QTLeap are not intensive in transliteration issues.

The use of **WSD techniques** in MT is an open research subject. Since the initial and disappointing work of Carpuat and Wu (2005), other ways to take WSD predictions into account have been proposed. Some of them achieved encouraging results in partial tasks such as word translation (Vickrey et al. 2005), but WSD has not yet been integrated into a complete MT system (Apadaniaki et al. 2012).

The task of word-sense disambiguation for MT is closely inter-linked with lexical selection (Cabezas and Resnik, 2005): if two senses of a word have the same translation, they do not need to be distinguished, and conversely, if a particular sense has two possible translations, the problem of selecting the correct one remains. Carpuat and Wu (2007) proposed to generalize the WSD system to perform fully phrasal multiword disambiguation, but this approach suffers from sparseness and computational problems.

More recently, Carpuat et al. (2013) have proposed SenseSpotting, systems that spot tokens that have new senses in new domain text, targeting the problem of domain adaptation for machine translation with an F-measure of 80% for new word types. The impact of MT has not been tested, however. Chan et al. (2007) introduced a way to modify the rule weights of a hierarchical translation system to reflect the WSD predictions. Specia et al. (2008) experimented with 10 highly ambiguous verbs by resorting to symbolic and probabilistic WSD systems: they considered standard n-best and expanded n-best re-ranking and demonstrated improvements in BLEU scores for both methods and WSD classifiers. Vintar et al. (2012) report that UKB-based WSD performs with a MT-relevant precision of 71% and that 21% of sense-related MT errors could be prevented by using unsupervised WSD. Yan and Kirchoff (2012) present significant improvements when experimenting with different approaches to unsupervised translation disambiguation within a SMT system for meeting-style speech.

Alternatives to explicit word sense disambiguation have also been explored in MT. Xiong and Zhang (2014) proposed a sense-based translation model to integrate word senses into statistical machine translation. They used word sense induction techniques to predict sense clusters and to annotate source words. Those senses were used in a sense-based translation model that enables the decoder to select appropriate translations for source words according to the inferred senses. The effectiveness of the proposed sense-based translation model was tested on a large-scale Chinese-to-English translation task, and results show that the proposed model substantially outperforms not only the baseline but also the previous reformulated word sense disambiguation. Although word sense induction was not planned in the project, this paper provides a very interesting and effective way to include lexical-semantic information in MT. In another strand of work, **distributional word representations** (also know as word embeddings or continuous word representations) have been explored. Mikolov et al. (2013) noted that dictionaries and phrase tables are the basis of modern statistical machine translation systems, and developed a method that can automate the process of generating and extending dictionaries and phrase tables. The method can translate missing word and phrase entries by learning language structures based on large monolingual data and mapping between languages from small bilingual data. It uses distributed representation of words and learns a linear mapping between vector spaces of languages. Despite its simplicity, the method was very effective. In later work, Zhao et al. (2015) applied related ideas for infrequent words and phrases, extracting translation rules for infrequent phrases based on phrases with similar continuous representations for which a translation is known. The approach of learning new translation rules improves a phrase-based baseline by up to 1.6 BLEU on Arabic-English translation.

QTLeap will develop further the usages of lexical semantics and conceptual knowledge, fostered with Linked Open Data, in the direction of supporting deep processing for machine translation. **WP5 in QTLeap** will explore the exploitation of the contribution of semantic linking and resolving to MT. This will be pursued through the undertaking of two types of experimental exercises. Firstly, we will perform off-line enrichment of language resources oriented to machine translation, i.e., parallel and comparable corpora gathered from the web and collections of specialized lexicons will be annotated (entities, multiword terms, categories, multilingual links, etc.) using a selection of the tools – and the corresponding datasets and LOD resources to train them – described in previous sections.

Secondly, experiments involving on-line gathering of multilingual information to improve translation will be performed. This is especially directed to the handling of unknown expressions by resorting to information sources of multilingual information whose content evolves very rapidly and is being constantly growing (Twitter, Wikipedia, news, etc.). For that purpose, resources and tools especially designed to collect and process comparable corpora from Twitter, Wikipedia, and news, will be used, as well as those that extract lexical information and name entities from them.

Crucially, the potential of these new datasets will be exploited for MT technology by creating new transduction algorithms that seek to anchor their key translation stage in deeper linguistic representations. These algorithms will use lexical-semantic information, integrate it in the translation framework, and use it to improve translation, including better lexical selection.

7 Conclusions

This deliverable provides a detailed survey about current availability of both datasets and tools to perform Named Entity Resolution (NERC and NED) and Word Sense Disambiguation for the languages relevant to the QTLeap project. This work is decisive in order to specify the requirements necessary to develop state-of-the-art lexical semantic tools for QTLeap. Crucially, the technology developed in this WP5 will contribute to the WP2 deep MT.

7.1 NERC

Of the six languages in the WP, and as it was expected, English is the language most represented both in terms of available data sources and systems. As shown in table 1, Spanish is also very well represented in terms of datasets for the development of NERC systems. This is partially due to their presence in CoNLL shared tasks where gold-standard datasets manually annotated for training and evaluation of NERC systems in those languages were created. Furthermore, there have been other manual annotation efforts including English and Spanish where the named entity annotations are part of a more general syntactic and semantic annotation, as shown by Ancora and OntoNotes corpora. Portuguese and Czech are also well supported with medium-sized annotated corpora such as CINTIL and HAREM, and CNEC 1.0 respectively. Bulgarian avails of a number of annotated corpora, BulTreeBank with 15,000 sentences and 21,678 NEs and a NE lexicon with 26,000 NEs. An annotated dataset for Basque is available, albeit limited (62,187 words with 4,748 NEs).

With respect to NERC systems, table 2 shows that there are several available for some of the languages: there is a wide range of tools available for English, as expected. Spanish is also well covered with a good number of options, e.g. Freeling, OpenNLP or BBN IdentiFinder Text Suite™. The remaining languages have at least one tool specifically built for their language, e.g. Eihera+ for Basque, LX-NER and Rembrandt for Portuguese, NameTag for Czech and CLaRK for Bulgarian (together with a IT-specific tagger). Additionally, multilingual systems such as SProUT are available that facilitate, to an extent, the development of new systems.

To conclude, sufficient tools and data for NERC in the general domain are available to implement baseline systems in all languages. Nonetheless, for the less-resourced languages such as Basque, some work will have to be done to enlarge the available datasets for improved tools.

7.2 NED

Most of the work on Named Entity Disambiguation has been done for English. Only with the increasing popularity of Linked Data data sources some tools have been developed which are multilingual, such as DBpedia Spotlight and Alchemy API. Two other systems, TAGME and The Wiki Machine perform Wikification for languages other than English, but unfortunately, none relevant to the project.

Out of the WP languages, DBpedia Spotlight currently supports English, Portuguese, Spanish, and in addition it also supports two other project languages like Dutch and German. The rest of the WP languages (Basque, Bulgarian, Czech) will need to be included into DBpedia Spotlight, but the task seems feasible given the availability of instructions for internationalization and DBpedia versions for those languages. In addition an in-house software exists for Basque, BasqueNED.

To the best of our knowledge, gold-standard publicly available annotations exist for English and Spanish, but not for the remaining languages of the WP. Note that a Basque annotated corpus exists.

7.3 WSD

Most of the WP languages lack large annotated datasets which could be used to train supervised WSD systems. The only exceptions are English and Spanish. Other languages like Basque do have more limited training datasets which would be useful for some target words (e.g. Basque with EuSemcor could be useful for 1,300 ambiguous nouns). Evaluation datasets are also limited. We are currently aware of datasets for English, Spanish and Basque, publicly available thanks to the SensEval and SemEval competitions.

As an alternative to supervised systems, knowledge-based systems, more specifically graph-based systems, provide good performance, using the information in the respective wordnets. Specifically UKB has been adapted to work with English, Spanish, and Basque. Given the widespread availability of wordnets for all the languages in the WP, UKB could be easily adapted for Bulgarian, Czech and Portuguese. In addition, BabelNet has released in October 2013 a multilingual resource which combines wordnets in several languages which could be also adapted to be used with UKB.

8 References

Agirre, Eneko and Aitor Soroa. 2009.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, Eneko and David Martíınez. 2000. 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, Eneko and David Martinez. 2001. 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.V´clav Matousek, Pavel Mautner, Roman Moucek, Karel a Tauser (eds.) Copyright Springer-Verlag. ISBN: 3-540-42557-8. ", 2001.

Agirre, Eneko and Oier Lopez de Lacalle. 2004. 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, Eneko and Oier Lopez de Lacalle. 2009. 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, Eneko, and Philip Edmonds, eds. 2006.*Word Sense Disambiguation: Algorithms and Applications*. 1st ed. Springer.

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

Agirre, Eneko, Montse Cuadros, German Rigau, and Aitor Soroa. 2010. 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, Eneko, Oier Lopez De Lacalle and Aitor Soroa. 2009a. "Knowledge-based wsd on specific domains: performing better than generic supervised wsd. In Proceedings of the 21st international jont conference on Artificial Intelligence, pages 1501-1506. Morgan Kaufmann Publishers Inc., 2009.

Agirre, Eneko, Oier Lopez deLacalle, Christiane Fellbaum, Shu-Kai Hsieh, Maurizio Tesconi, Monica Monachini, Piek Vossen, and Roxanne Segers. 2010. 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.

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

Alegria, Iñaki, Olatz Arregi, Nerea Ezeiza and Izaskun Fernandez. 2006. "Lessons from the Development of a Named Entity Recognizer." In Procesamiento del Lenguaje Natural, 36, 25-37. http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/2817/1316 Print edition ISSN:1135-5948, ISSN digital edition ISSN: 1989-7553.

Alfonseca, E., and S. Manandhar. 2002. "An Unsupervised Method for General Named Entity Recognition and Automated Concept Discovery." In *Proceedings of the 1st International Conference on General WordNet, Mysore, India*.

Apidianaki, Marianna, Guillaume Wisniewski, Artem Sokolov, Aurélien Max, and François Yvon, 2012, WSD for n-best reranking and local language modeling in SMT. SSST-6, Sixth Workshop on Syntax, Semantics and Structure in Statistical Translation, Jeju, Republic of Korea, 12 July 2012; pp.1-9.

Arcan, M. , C. Giuliano, M. Turchi and P. Buitelaar "Identification of Bilingual Terms from Monolingual Documents for Statistical Machine Translation", In Proceedings of the 4th International Workshop on Computational Terminology (Computerm). 2014.

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

Bagga, A., and B. Baldwin. 1998. "Entity-based Cross-document Coreferencing Using the Vector Space Model." In *Proceedings of the 17th International Conference on Computational linguistics-Volume 1*, 79–85. http://dl.acm.org/citation.cfm?id=980859.

Bagga, Amit, and Breck Baldwin. 1998. "Algorithms for Scoring Coreference Chains." In *Proceedings of the 1st International Conference on Language Resources and Evaluation (LREC1998)*, 563–566. Granada, Spain. ftp://ftp.cis.upenn.edu/pub/breck/scoring-paper.ps.gz.

Balasuriya, D., N. Ringland, J. Nothman, T. Murphy, and J. R. Curran. 2009. "Named Entity Recognition in Wikipedia." In *Proceedings of the 2009 Workshop on The People's Web Meets NLP: Collaboratively Constructed Semantic Resources*, 10–18. http://dl.acm.org/citation.cfm?id=1699767.

Basile, Valerio, Johan Bos, Kilian Evang, Noortje Venhuizen. 2012. "Developing a large semantically annotated corpus." In Proceedings of the Eight International conference on Language Resources and Evaluation (LREC 2012), pp 3196-3200, Istanbul, Turkey.

Bentivogli, Luisa 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.

Bick, E. 2004."A Named Entity Recognizer for Danish."In *Proc. of 4th International Conf. on Language Resources and Evaluation*, 305–308. http://www.lrec-conf.org/proceedings/lrec2004/pdf/99.pdf.

Black, W. J, F. Rinaldi, and D. Mowatt. 1998. "FACILE: Description of the NE System Used for MUC-7." In *Proceedings of the 7th Message Understanding Conference*. http://acl.ldc.upenn.edu/muc7/M98-0014.pdf.

Borrega, Oriol, Mariona Taulé, and MA Martí. What do we mean when we speak about named entities? In Proceedings of Corpus Linguistics, 2007.

Brockmann, Carsten and Mirella Lapata. 2003. 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.

Bunescu, R., and M. Pasca. 2006. "Using Encyclopedic Knowledge for Named Entity Disambiguation." In *Proceedings of EACL*, 6:9–16. http://acl.ldc.upenn.edu/E/E06/E06-1002.pdf.

Cabezas, Clara and Philip Resnik, ``Using WSD Techniques for Lexical Selection in Statistical Machine Translation'', Technical report CS-TR-4736/LAMP-TR-124/UMIACS-TR-2005-42, July 2005.

Carpuat, Marine and Dekai Wu, 2005, Word sense disambiguation vs. statistical machine translation, 2005, ACL-2005: 43rd Annual meeting of the Association for

Computational Linguistics, University of Michigan, Ann Arbor, 25-30 June 2005; pp. 387-394.

Carpuat, Marine and Dekai Wu, 2007, Improving statistical machine translation using word sense disambiguation, 2007, EMNLP-CoNLL-2007: Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, June 28-30, 2007, Prague, Czech Republic; pp. 61-72.

Carpuat, Marine, Hal Daume III, Katie Henry, Ann Irvine, Jagadeesh Jagalamudi and Rachel Rudinger. 2013. "SenseSpotting: Never let your parallel data tie you to an old domain". Association for Computational Linguistics. Sofia, Bulgaria: Aug 2013.

Carreras, X., I. Chao, L. Padró, and M. Padró. 2004. "Freeling: An Open-source Suite of Language Analyzers." In *Proceedings of the 4th LREC*.Vol. 4.

Carreras, X., L. Màrquez, and L. Padró. 2003. "A Simple Named Entity Extractor Using AdaBoost." In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003-Volume 4*, 152–155. http://dl.acm.org/citation.cfm?id=1119197.

Chai, Joyce Yue 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, Y. S., Ng, H. T., and Chiang, D., 2007, Word sense disambiguation improves statistical machine translation. In Annual Meeting-Association for Computational Linguistics (Vol. 45, No. 1, p. 33).

Chen, Jinying and Martha Palmer. 2009. 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.

Cheng, Xiao and D. Roth. 2013. "Relational Inference for Wikification." EMNLP 2013.

Chinchor, N. 1998."Overview of MUC-7."In *Proceedings of the Seventh Message Understanding Conference (MUC-7)*, 178–185.

Ciaramita, M., and Y. Altun. 2005. "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*.

Ciaramita, Massimiliano and Yasemin Altun. 2006. Broadcoverage 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, Paul and Mark Stevenson. Cross-language information retrieval using eurowordnet and word sense disambiguation. In Advances in information retrieval, pages 327–337. Springer, 2004.

Cornoti, Marco, Paolo Ferragina and Massimiliano Ciaramita. 2013. "A Framework for Benchmarking Entity-annotation Systems." In Web, 249-260.WWW '13. Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee, 2013. http://dl.acm.org/citation.cfm?id=2488388.2488411.

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

Cucchiarelli, A., and P. Velardi. 2001. "Unsupervised Named Entity Recognition Using

Syntactic and Semantic Contextual Evidence." *Computational Linguistics* 27 (1): 123–131.

Cucerzan, S. 2007. "Large-scale Named Entity Disambiguation Based on Wikipedia Data." In *Proceedings of EMNLP-CoNLL*, 2007:708–716. http://acl.ldc.upenn.edu/D/D07/D07-1074.pdf.

Daiber, Joachim, Max Jakob, Chris Hokamp and Pablo N. Mendes. 2013. "Improving Efficiency and Accuracy in Multilingual Entity Extraction." In Proceedings of the 9th International Conference on Semantic Systems (I-Semantics).

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

Dredze, M., P. McNamee, D. Rao, A. Gerber, and T. Finin. 2010. "Entity Disambiguation for Knowledge Base Population." In *Proceedings of the 23rd International Conference on Computational Linguistics*, 277–285. http://dl.acm.org/citation.cfm?id=1873813.

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

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

Escudero, Gerard, Lluís Màrquez, and German Rigau.2000. An empirical ıs a 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.

Etzioni, O., M. Cafarella, D. Downey, A. M Popescu, T. Shaked, S. Soderland, D. S Weld, and A. Yates. 2005. "Unsupervised Named-entity Extraction from the Web: An Experimental Study." *Artificial Intelligence* 165 (1): 91–134.

Fader, A., S. Soderland, O. Etzioni, and T. Center. 2009. "Scaling Wikipedia-based Named Entity Disambiguation to Arbitrary Web Text." In *WikiAI09 Workshop at IJCAI*.

Faruqui, M., and S. Padó. 2010. "Training and Evaluating a German Named Entity Recognizer with Semantic Generalization." *Semantic Approaches in Natural Language Processing*: 129.

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

Fernandez, Izaskun, Iñaki Alegria and Nerea Ezeiza. 2011. "Semantic Relatedness for Named Entity Disambiguation using a small Wikipedia." I. Habernal and V. Matouˇek (Eds.): TSD 2011, LNAI 6836, pp. 276–283. ISBN 978-3-642-23537-5.Springer Berlin / Heidelberg.https://ixa.si.ehu.es/Ixa/Argitalpenak/Artikuluak/1307532245/publikoak/SemanticRelatedness.pdf

Ferragina, P., and U. Scaiella. 2010. "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*, 1625–1628. http://dl.acm.org/citation.cfm?id=1871689.

Filatova, Elena, 2009, Directions for Exploiting Asymmetries in Multilingual Wikipedia. Proceedings of CLIAWS3, Third International Cross Lingual Information Access Workshop, in conjunction with NAACL-HLT 2009. Boulder. 2009.

Fleischman, M., and E. Hovy. 2002. "Fine Grained Classification of Named Entities." In *Proceedings of the 19th International Conference on Computational linguistics-Volume 1*, 1–7. http://dl.acm.org/citation.cfm?id=1072358.

Fox, Barbara, Dan Jurafsky, and Laura A. Michaelis. Cognition and Function in Language.

CSLI Publications, Stanford, CA., 1999.

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

Gaustad, Tanja. 2004. Linguistic Knowledge and Word Sense Disambiguation. PhD Thesis, Alfa-Informatica, University of Groningen, Groningen.

Grishman, R., and B. Sundheim. 1996. “Message Understanding Conference-6: A Brief History.” In *Proceedings of COLING*, 96:466–471. http://acl.ldc.upenn.edu/C/C96/C96-1079.pdf.

Hálek, Ondrej, Rudolf Rosa, Ales Tamchyna, and Ondrej Bojar. Named entities from Wikipedia for machine translation. Proceedings of ITAT. 2011

Hachey, Ben, Will Radford, Joel Nothman, Matthew Honnibal, and James R. Curran. 2012. “Evaluating Entity Linking with Wikipedia.” *Artificial Intelligence* (In press). doi:10.1016/j.artint.2012.04.005. http://www.sciencedirect.com/science/article/pii/S0004370212000446.

Henrich, Verena, Erhard Hinrichs, Tatiana Vodolazova. 2012. WebCAGe - A Web-Harvested Corpus Annotated with GermaNet Senses. EACL 2012.

Hermjakob, Ulf, Kevin Knight, and Hal Daumé III. "Name Translation in Statistical Machine Translation - Learning When to Transliterate." Proceedings of ACL. 2008.

Hoffart, J., M.A. Yosef, I. Bordino, H. Fürstenau, M. Pinkal, M. Spaniol, B. Taneva, S. Thater, and G. Weikum. 2011. “Robust Disambiguation of Named Entities in Text.” In *Proc. of EMNLP*, 27–31.

Hovy, Eduard, 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.

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

Izquierdo, Rubén, Armando Suárez, and German Rigau. 2009. 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, Rubén, Armando Suárez, and German Rigau. 2010. 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.

Jones, G. J., Fantino, F., Newman, E., and Zhang, Y. 2008. Domain-specific query translation for multilingual information access using machine translation augmented with dictionaries mined from Wikipedia. Proceedings of the 2nd International Workshop on “Cross Lingual Information Access” Addressing the Information Need of Multilingual Societies.

Jones, Bevan, Jacob Andreas, Daniel Bauer, Karl Moritz Hermann, Kevin Knight, 2012, Semantics-Based Machine Translation with Hyperedge Replacement Grammars.Proceedings of the Computational Linguistics Conference.Mumbai. 2012.

Kripke, S. 1980. *Naming and Necessity*.Harvard University Press.

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

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

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

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

Li, Haibo, Jing Zheng, Heng Ji, Qi Li, and Wen Wang. Name-aware Machine Translation. In Proceedings of the ACL, pages 604–614. 2013

Liu, Shuang, 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.

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

Marcus, Mitchell P., Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. "Building a Large Annotation Corpus of English: The Penn Treebank." *Computational Linguistics* 19 (2): 313–330.

Màrquez, Lluís, Gerard Escudero, German Rigau, and David Martínez. 2006. 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.

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

Martínez, David and Eneko Agirre. 2000. 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. ".

Maynard, D., V. Tablan, C. Ursu, H. Cunningham, and Y. Wilks. 2001. "Named Entity Recognition from Diverse Text Types." In *Recent Advances in Natural Language Processing 2001 Conference*, 257–274. https://gate.ac.uk/sale/ranlp2001/maynard-etal.pdf.

McNamee, P., H. T. Dang, H. Simpson, P. Schone, and S. M. Strassel. 2010. "An Evaluation of Technologies for Knowledge Base Population." In *Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC'10).European Language Resources Association (ELRA)*. http://hnk.ffzg.hr/bibl/lrec2010/pdf/634_Paper.pdf.

Mendes, Pablo, Max Jakob, Andres Garcia-Silva and Christian Bizer. 2011. "DBpedia Spotlight: Shedding Light on the Web of Documents." In Proceedings of the 7th International Conference on Semantic Systems (I-Semantics).

Mihalcea, R., and A. Csomai. 2007. "Wikify!: Linking Documents to Encyclopedic Knowledge." In *CIKM*, 7:233–242.

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

Mihalcea, Rada. 2005. 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, Rada. 2007. Using wikipedia for automatic word sense disambiguation. In Proceedings of the North American Chapter of the Association for Computational

Linguistics (NAACL), Rochester, 2007.

Mika, P., M. Ciaramita, H. Zaragoza, and J. Atserias. 2008. “Learning to Tag and Tagging to Learn: A Case Study on Wikipedia.” *IEEE Intelligent Systems* 23 (5): 26–33.

Mikolov, Tomas , Quoc V Le, and Ilya Sutskever. Exploiting similarities among languages for machine translation. arXiv preprint arXiv:1309.4168. 2013.

Mikheev, A., M. Moens, and C. Grover. 1999. “Named Entity Recognition Without Gazetteers.” In *Proceedings of the Ninth Conference on European Chapter of the Association for Computational Linguistics*, 1–8. http://dl.acm.org/citation.cfm?id=977037.

Miller, George A., 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, D., and I. H. Witten. 2008. “Learning to Link with Wikipedia.” In *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, 509–518. http://dl.acm.org/citation.cfm?id=1458150.

Minkov, E., R. C Wang, and W. W Cohen. 2005. “Extracting Personal Names from Email: Applying Named Entity Recognition to Informal Text.” In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, 443–450. http://dl.acm.org/citation.cfm?id=1220631.

Mohammadi, Mehdi, Nasser Ghasem Aghaee, 2010, Building Bilingual Parallel Corpora based on Wikipedia. Proceedings of the Second International Conference on Computer Engineering and Applications.IEEE. 2010.

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

Morris, Jane and Graeme Hirst. 2004. Non-Classical Lexical Semantic Relations. In: Proceedings of the HLT Workshop on Computational Lexical Semantics. Boston, Massachusetts, USA. pp 46-51.

Müller, Christof, Iryna Gurevych, 2009, Using Wikipedia and Wiktionary in Domain-Specific Information Retrieval. In Evaluating Systems for Multilingual and Multimodal Information Access, Springer. 2009.

Nadeau, D., and S. Sekine. 2007. “A Survey of Named Entity Recognition and Classification.” *Lingvisticae Investigationes* 30 (1): 3–26.

Navigli, R. and S. Ponzetto. 2012a. Multilingual WSD with Just a Few Lines of Code: the BabelNet API. Proc. of the 50th Annual Meeting of the Association for Computational Linguistics (ACL 2012), Jeju, Korea, July 9-11, 2012, pp. 67-72.

Navigli, R. and S. Ponzetto. 2012b. Joining Forces Pays Off: Multilingual Joint Word Sense Disambiguation. Proc. of the 2012 Conference on Empirical Methods in Natural Language Processing (EMNLP 2012), Jeju, Korea, July 12-14, 2012, pp. 1399-1410.

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

Navigli, Roberto and Paola Velardi. 2005. 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, Roberto. 2009. Word Sense Disambiguation: a survey. ACM Computing Surveys, 41(2):1–69, 2009.

Ng, H. T. 1997.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, H. T. and H. B. Lee. 1996. 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.

Niemann, Elisabeth 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.

Nothman, J., J. R. Curran, and T. Murphy. 2008. "Transforming Wikipedia into Named Entity Training Data." In *Proceedings of the Australian Language Technology Workshop*, 124–132. http://www.alta.asn.au/events/alta2008/proceedings/ALTA2008.pdf#page=132.

Nothman, J., N. Ringland, W. Radford, T. Murphy, and J. Curran. 2012. "Learning Multilingual Named Entity Recognition from Wikipedia." *Artificial Intelligence* (In press). doi:10.1016/j.artint.2012.03.006. http://www.sciencedirect.com/science/article/pii/S0004370212000276.

Otero, Pablo Gamallo, Isaac Gonzalez López, 2010, Wikipedia as Multilingual Source of Comparable Corpus.Proceedings of the Third Workshop on Building and Using Comparable Corpora, in conjunction with LREC 2010. Malta. 2010.

Pedersen, Ted. 2006. 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.

Petasis, G., F. Vichot, F. Wolinski, G. Paliouras, V. Karkaletsis, and C. D Spyropoulos. 2001. "Using Machine Learning to Maintain Rule-based Named-entity Recognition and Classification Systems." In *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*, 426–433. http://dl.acm.org/citation.cfm?id=1073067.

Philpot, Andrew, 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, Emanuele, 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.

Pociello, E., Agirre, E. and Aldezabal, I. 2010. "Methodology and construction of the Basque WordNet" In *Language Resources and Evaluation.Springer. ISSN 1574-020X. http://www.springerlink.com/content/40n24044u3k2k277/*

Poibeau, T. 2003. "The Multilingual Named Entity Recognition Framework."In *Proceedings of the Tenth Conference on European Chapter of the Association for Computational Linguistics-Volume 2*, 155–158. http://dl.acm.org/citation.cfm?id=1067772.

Poibeau, T., and L. Kosseim. 2001. "Proper Name Extraction from Non-journalistic Texts." *Language and Computers* 37 (1): 144–157.

Pradhan, Sameer S., Eduard Hovy, Mitch Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2007. "OntoNotes: A Unified Relational Semantic Representation." In *Proceedings of the International Conference on Semantic Computing*, 517–526.ICSC '07. Washington, DC, USA: IEEE Computer Society. doi:10.1109/ICSC.2007.67. http://dx.doi.org/10.1109/ICSC.2007.67.

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

Ratinov, L., and D. Roth. 2009. "Design Challenges and Misconceptions in Named Entity Recognition." In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning*, 147–155.

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

Richman, A.E., and P. Schone. 2008. "Mining Wiki Resources for Multilingual Named Entity Recognition." In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 1–9.

Rigau, German, Jordi Atserias, and Eneko Agirre. 1997. 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, Peter Mark. 1911. Roget's Thesaurus of English Words and Phrases... TY Crowell Company, 1911.

Sekine, S., and C. Nobata. 2004. "Definition, Dictionaries and Tagger for Extended Named Entity Hierarchy." In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, 1977–1980. https://nats-www.informatik.uni-hamburg.de/intern/proceedings/2004/LREC/pdf/65.pdf.

Simančík, F., and M. Lee."A CCG-based System for Valence Shifting for Sentiment Analysis." http://cicling-org.g-sidorov.org/2009/RCS-41/099-108.pdf.

Sinha, Ravi and Rada Mihalcea.2007. 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.

Specia, Lucia, Maria das Graças Volpe Nunes, and Mark Stevenson, 2006, Translation context sensitive WSD, 2006, EAMT-2006: 11th Annual Conference of the European Association for Machine Translation, June 19-20, 2006, Oslo, Norway. Proceedings; p.227-232.

Steinberger, Josef, Massimo Poesio, Mijail A. Kabadjov, and Karel Jeek. 2007. "Two Uses of Anaphora Resolution in Summarization." *Inf. Process. Manage.* 43 (6) (November): 1663–1680. doi:10.1016/j.ipm.2007.01.010.

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

Stokoe, Christopher, 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.

Straková Jana, Straka Milan, Hajič Jan: *A New State-of-The-Art Czech Named Entity Recognizer.* In: Lecture Notes in Computer Science, Vol. 8082, Text, Speech and Dialogue: 16th International Conference, TSD 2013. Proceedings, Copyright © Springer Verlag, Berlin / Heidelberg, ISBN 978-3-642-40584-6, ISSN 0302-9743, pp. 68-75, 2013

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

Tjong Kim Sang, E. F, and F. De Meulder. 2003. "Introduction to the CoNLL-2003 Shared Task: Language-independent Named Entity Recognition." In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003-Volume 4*, 142–147. http://dl.acm.org/citation.cfm?id=1119195.

Toral, A., and R. Munoz. 2006. "A Proposal to Automatically Build and Maintain Gazetteers for Named Entity Recognition by Using Wikipedia." In *NEW TEXT Wikis and Blogs and Other Dynamic Text Sources*, 56.http://acl.ldc.upenn.edu/eacl2006/ws12_newtext.pdf#page=64

Tyers, Francis M., Jacques A. Pienaar, 2008, Extracting bilingual word pairs from Wikipedia, 2008, Proceedings of the first workshop on Collaboration: interoperability between people in the creation of language resources for less-resourced languages, in conjunction with LREC 2008, Marrakech.

Vickrey, David, Luke Biewald, Marc Teyssier and Daphne Koller. 2005. Word-sense disambiguation for Machine Translation. In Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing, pages 771-778, Association for Computational Linguistics.

Vintar, Špela, Darja Fišer, and Aljoša Vrščaj: Were the clocks striking or surprising? Using WSD to improve MT performance, 2012, EACL Joint Workshop on Exploiting Synergies between Information Retrieval and Machine Translation (ESIRMT) and Hybrid Approaches to Machine Translation (HyTra): Proceedings of the workshop, 23-24 April 2012, Avignon, France; pp.87-92.

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

Weischedel, R., and A. Brunstein. 2005. "BBN Pronoun Coreference and Entity Type Corpus." *Linguistic Data Consortium, Philadelphia*.

Witten, I. H, Z. Bray, M. Mahoui, and W. J Teahan. 1999. "Using Language Models for Generic Entity Extraction." In *Proceedings of the ICML Workshop on Text Mining*. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.37.1678&rep=rep1&type=pdf.

Xiong, Deyi and Min Zhang. 2014. A Sense-Based Translation Model for Statistical Machine Translation. Proceedings of ACL. 2014

Yarowsky, David. 1992. 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.

Zhao, Kai , Hany Hassan and Michael Auli. Learning Translation Models from Monolingual Continuous Representations. Proceedings of NAACL. pages: 1527-1536. 2015.

Zhong, Zhi 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.