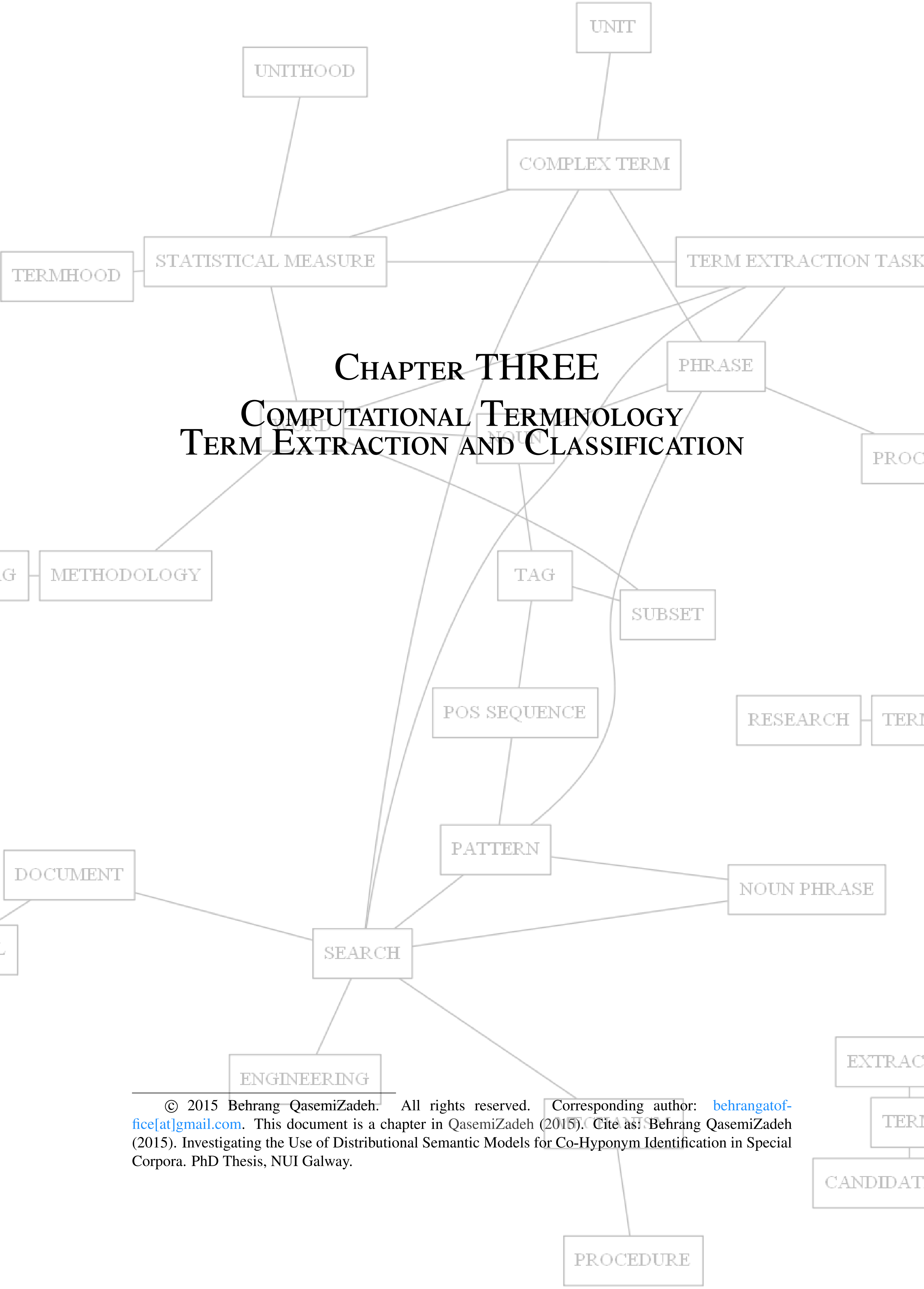
# Chapter THREE

# Computational Terminology
# Term Extraction and Classification



Corresponding author: behrangatoffice[at]gmail.com. This document is a chapter in QasemiZadeh (2015). Cite as: Behrang QasemiZadeh (2015). Investigating the Use of Distributional Semantic Models for Co-Hyponym Identification in Special Corpora. PhD Thesis, NUI Galway.

This page is intentionally left blank.

# Contents

This page is intentionally left blank.

# List of Figures

This page is intentionally left blank.

# List of Tables

This page is intentionally left blank.

# Chapter 3

# Computational Terminology: Term Extraction and Classification

Systematic terminology collection, management, and maintenance are significant tasks in any application that deals with knowledge. These processes are the subjects of study in terminology and subsequently computational terminology. Apart from established applications in knowledge management systems, recent endeavours such as information retrieval, machine translation, ontology learning and semantic search have stimulated research in terminology mining. With a focus on term extraction, this chapter provides an overview of the basic definitions and tasks in computational terminology.

Section 3.1 provides an overview of terminology mining methods. Sections 3.2 describes the common employed mechanism in these methods. Section 3.3, and 3.4 details the processes of candidate term extraction and scoring, respectively. Section 3.5 touches the subject of term organisation. Section 3.6 briefly discusses the use of machine learning techniques in terminology mining. Finally, the chapter concludes with a brief discussion on the evaluation in Section 3.7.

## 3.1 Introduction to Computational Terminology

Computational terminology embraces a set of algorithms that extract terms from *special corpora* and arrange them in domain-specific knowledge structures such as a vocabulary, thesaurus or ontology. As defined by Sinclair (1996), special corpora contain sublanguage material. Hence, according to this definition, computational terminology is concerned with the automatic analysis of *languages for special purposes*, for example, in order to facilitate interoperability when communicating specialised knowledge.

Computational terminology inherits its complexities from difficulties in the interpretation of meaning in language. In terminology, these complexities are often summarised by the question what counts as a *term*? The Oxford Dictionary defines a term as:

> a word or phrase used to describe a thing or to express a concept, specially in a particular kind of language or branch of study (Term[Def. 1], 2014).

According to the International Organisation for Standardisation (ISO), a term is

> a verbal designation of a general concept in a specific subject field (ISO 1087-1, 2000).

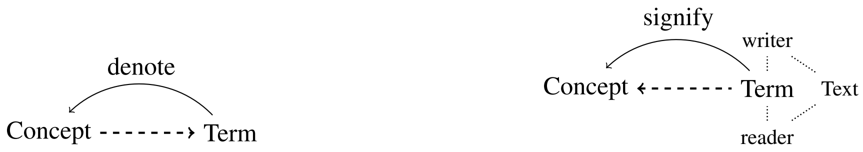
As stated by Cabré (2010), linguistically, terms are *lexical units* and carry a special *meaning* in particular *contexts*. A lexical unit is often considered as a *lexical form*—a single token, part of a word, a word or a combination of these—that is paired with a single meaning and serves as the basic element of a language's vocabulary. Similarly, as suggested by L'Homme (2014), terms are the denomination of items of knowledge—that is, concepts.

According to their lexical forms, terms are usually classified as *simple* or *complex*. Simple terms consist of one token; complex terms are composed of more than one token or word. For instance, 'lexicography' and 'multilingual terminology management' are, respectively, examples of a simple and a complex term in the domain of computational linguistics. The extracted lexical units constitute a *terminological resource*, also known as *terminology*: a specialised vocabulary of knowledge in a domain. Terms and their use are studied in a relatively young discipline, which is also called *terminology* (Cabré, 2003; Kageura, 1999):

> the field of activity concerned with the collection, description, processing and presentation of terms (Sager, 1990).

While terminology can be approached from several perspectives—for example, as a branch of philosophy, sociology, or cognitive science—terminology is dominantly considered a linguistic and cognitive activity. Modern terminology is therefore pursued within a linguistic framework and as the study of specialised languages—that is, languages for special purposes (Faber and Rodríguez, 2012).

In terminology, the meanings of terms and the process of concept denomination are studied within the framework of a *theory of terminology*. As stated in Cabré (2003), a theory of terminology elaborates the fundamental problem of interpretation of meaning



(a) The General Theory of Terminology

(b) Modern Terminology

Figure 3.1: Association of meaning in the GTT compared to recent theories of terminology: the GTT starts with concepts. Terms are only labels and denote concepts existing a priori. In recent theories of terminology such as the CTT, however, terms are treated like other linguistic units. They signify concepts in a communicative context. In the figures above, the dashed lines indicate the direction in which the meaning of a term is elaborated according to these theories. The indicated communicative context (the dotted triangle in Figure b) can be extended in a number of ways, for example, by considering the application of terms.

into a set of questions in which the definition of a terminological unit—and its characteristics—is often the nucleus, for example:[1]

- What are the basic units of terminological knowledge?
- How are they defined and acquired?
- Where are they observed?
- How are they recognised and what are their characteristics?

The general theory of terminology (GTT) by Wüster (1974, as cited in Campo (2013, chap. 2)) is widely recognised as the first theory of terminology. The GTT, which is also known as traditional terminology, puts concepts first; terms are merely unambiguous labels for concepts that exist a priori (Faber and L'Homme, 2014) (Figure 3.1a). Put simply, in the GTT, knowledge is gained independently of the language, and thus the usage of terms. As implied by the given definition in ISO 1087-1(2000), the GTT has been one of the major adopted theories amongst terminologists.[2] The sequel to the GTT can also be found in early computational terminology research (e.g., see Ananiadou, 1994). Consequently, the GTT regards terms and concepts as having mono-referential relationships (Figure 3.2a). The objective behind the GTT, understandably, is to eliminate ambiguity in natural language in order to improve clarity in technical communication.

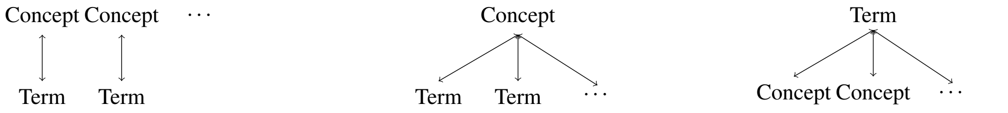
In an authoritative institutional organisation[3] that promotes or enforces standards, terms can be *made* and shared in a top-down manner; hence, the meaning of terms can be interpreted by the mechanism described in the GTT.[4] However, in practice and in many organisations, new terms are introduced in a bottom-up *synthesis* process. A lexical form (which may or may not be newly invented) in contexts that bear a concept (which may

[1] For a comprehensive list of questions and possible answers, see Cabré (2003).

[2] Accordingly, Felber (1982) defines terminology as 'the combined action of groups of subject specialists (terminology commissions) of specialised organisations'.

[3] Here, the *organisation* can be a scientific discipline, a technical domain, a company, etc., that requires a specialised language for effective communication.

[4] it is, perhaps, best demonstrated in the applications of controlled natural languages.



(a) One-to-One Relation (b) Synonymous Relation (c) Polysemous Relation

Figure 3.2: Relationships between terms and the concepts they signify: Figure 3.2a illustrates a mono-referential, unambiguous relationship between terms and concepts. Figure 3.2b shows an ambiguity that may arise when several terms denote the same concept in a synonymous relation. Figure 3.2c illustrates an ambiguous term-concept relation, a polysemous relationship where a term may denote several concepts.

or may not be newly invented) is used frequently inasmuch as it becomes a term[1] in the organisation. In practice, therefore, terms can be ambiguous: a term can refer to several concepts—similar to polysemy–homonymy in lexical semantics (Figure 3.2c); or, contrariwise, a particular concept can be denoted by several terms (Figure 3.2b). Heid and Gojun (2012) suggest that the rapid evolution of organisations as well as multi-players that are involved in an uncoordinated way, specifically in multidisciplinary domains, reinforces this situation and thus contributes to term ambiguity.

In contrast to the GTT, recent theories of terminology—for example, the communicative theory of terminology (CTT) by Cabré (1999, chap. 3) and the lexical-semantic approach that is promoted by Faber and L'Homme (2014)—acknowledge the situation stated above and take an empiricist approach to terminology in the sense that the meanings of terms, and as a result the elements of domain knowledge, are not preconceived. Simply put, in modern theories of terminology, knowledge is a posteriori that is dependent upon the language. Hence, terms are understood differently with regards to the communicative context, for example, by the text surrounding them, the application they are used in and so on.

Putting this discussion into the structuralist framework of distributional semantics, terms are *linguistic units* that signify concepts by syntagmatic and paradigmatic relations that they hold in a specialised communicative discourse (Figure 3.1b).[2]

The importance of a theory of terminology lies in the fact that it outlines practical issues that must be addressed in terminology. According to the adopted theory of terminology, computational terminology tasks are formulated differently and are thus approached from alternative perspectives. Consequently, the perspective presented by a theory of terminology establishes boundaries for the definition and classification of the tasks that are currently addressed in computational terminology. However, as indicated by Cabré (2003) in her *theory of doors*[3], the mere fact of the existence of these issues is not affected by the way they are formulated. Research in computational terminology addresses these

[1] That is, a norm.

[2] It becomes evident that the main difference between the GTT and modern terminology theories is the interpretation of the process of pairing concepts and lexical units—that is, as suggested in Chapater 1, the result of the GTT's rationalist vs. the CTT's empiricist approach to comprehend the process of gaining knowledge and communicating meanings.

[3] In the theory of doors, Cabré (2003) elaborates on her position as follows:

practical issues. Inevitably, although computational terminology is often associated with the task of automatic term recognition, it goes beyond that and embraces a number of research tasks.

In computational terminology, the task of automatic term recognition (ATR) has been at the centre of discussion as an essential component of modern information systems. In ATR, the input is a large collection of documents, that is, a special corpus, and the output is a terminological resource. In ATR, the meaning of the generated terms is interpreted in a wide spectrum of concepts in the domain that is being investigated and represented by the input domain-specific corpus. Since ATR facilitates the automatic construction of terminological resources, it is a significant processing resource in knowledge engineering tasks for a multitude of applications such as information retrieval and machine translation.

As articulated by Kageura and Umino (1996), ATR deals with the computation of measures known as *unithood* and *termhood*. It is believed that the majority of terms in a domain are complex terms. Unithood indicates the degree to which a sequence of tokens can be combined to form a complex term. It is, thus, a measure of the *syntagmatic* relation between the constituents of complex terms: a lexical association measure to identify collocations. In the absence of explicit linguistic criteria to distinguish complex terms from other natural language text phrases, a unithood measure construes the problem of complex term identification as the identification of *stable* lexical units (Sager, 1990).[1]

Termhood, on the other hand, 'is the degree that a linguistic unit is related to ··· some domain-specific concepts' (Kageura and Umino, 1996). It characterises a *paradigmatic* relation between lexical units—either simple or complex terms—and the communicative context that verbalises domain-concepts. Termhood, thus, conveys the measurement of meaning. In the absence of a formal answer to the question 'what are domain-specific concepts?'—for instance, see the discussions in Laurence and Margolis (1999); Fodor and Lepore (2012)—devising a termhood measure for distinguishing terms and non-terms is a difficult and often conflictual task.

Computational terminology, however, embraces a set of techniques other than ATR, which also aim to extract stable lexical units. In ATR, the communicative context is a domain-specific corpus. Therefore, ATR should not be confused with tasks such as keyword extraction and entity recognition that bear a close resemblance to it. These tasks are similar to ATR in the sense that they extract stable lexical units from natural language text. However, they are different from ATR, because the meaning of the extracted lexical units—thus the termhood measure—is interpreted in a context other than a special corpus (Figure 3.3). For example, an automatic keyphrase extraction algorithm pulls out lexical units from a single document that best describe the content of this document. Both unithood and termhood must be also measured in automatic keyphrase extraction. However, the criterion for their definition and the information available for their computation are

> This theory is suitably represented by the image of a house; let us assume a house with several entrance doors. We can enter any one of its rooms through a different door, but the choice of the door conditions the way to the inside of the house. The internal arrangement of rooms is not altered, what does change is the way one chooses to get there.

[1] See Evert (2004) on applications of lexical association measures for the identification of lexical units.

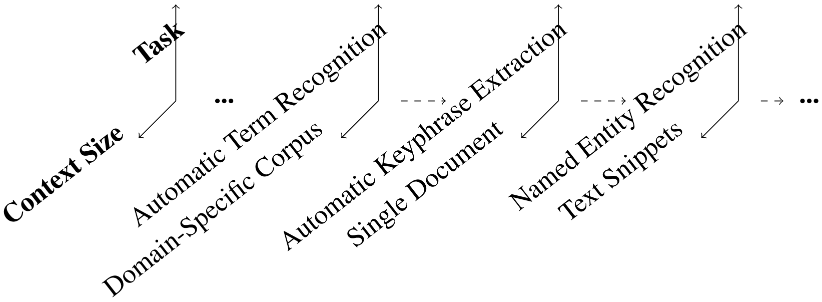


Figure 3.3: Lexical unit extraction tasks and the granularity in which they interpret the of the meanings of a lexical item. Although all the tasks listed in this figure extract lexical items that denote salient domain concept, the scope and the granularity in which they interpret the meanings of lexical units is different. At the highest level of granularity, automatic term recognition tasks investigate the meanings of lexical units across the set of documents that constitute a domain-specific corpus. At the least level of granularity, entity recognition tasks decide about the meanings of lexical units in a given snippet of text. The diagram can be extended by adding new dimensions that take into consideration characteristics of the communicative context other than the size of the input text. This diagram can form a basis to suggest taxonomies of tasks that extract lexical units from text.

different from ATR.

Categorisation of term extraction tasks can be extended by considering characteristics of communicative contexts other than the size of the input text. Cabré et al. (2007) classify term extraction tasks as *intermediary* and *terminal* with respect to the end-users' interaction with the extracted terminological resources. An intermediary application constructs a terminological resource—for example, a domain-specific ontology—that will be exploited as a component of other information systems; for example, to address problems such as information extraction and retrieval. Hence, in an intermediary application, end-users do not interact directly with the constructed terminological resource. However, in terminal applications, a terminological resource is constructed to be accessed and used directly by a particular user.

Besides the communicative context, the term extraction techniques are often classified by the linguistic characteristics of the extracted terms. For instance, Yangarber et al. (2002) distinguish tasks that address the extraction of *proper* names from those that focus on the extraction of *generalised* names. Accordingly, Yangarber et al. (2002) relate named entity recognition tasks to the former category of term extraction methods since their output is limited to the names of people, organisations, locations, and so on. For the latter category, they enumerate methods that extract mentions of concepts such as the name of biological agents based on the rationale that these terms are not proper names. Similarly, one may place keyphrase extraction methods in this category.

Tasks that are addressed in computational terminology can be further distinguished by the direction in which they bridge the gap between terminological resources and text. Recent developments of ontological resources have stimulated a research strand that targets the reverse of intermediary term extraction tasks. The goal of these applications is to fill

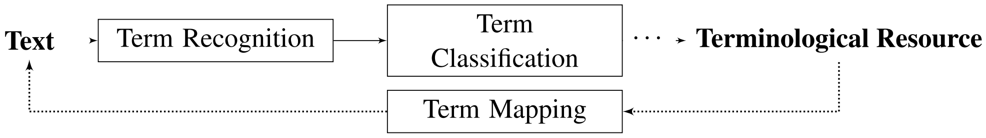


Figure 3.4: Significant processes in computational terminology. Whereas term extraction and classification techniques distil a terminological resource from text, a set of techniques in computational terminology try to bridge the gap between terminological resources—such as domain ontologies—to natural language text.

the gap between an available knowledge base—for example, an ontology—and natural language text. In these tasks, given a particular concept in a knowledge base (e.g., a class and its instances in an ontology), a method—which is called *term mapping* by Krauthammer and Nenadic (2004)—decides if this concept or its instances have been mentioned in a given text snippet. Entity linking, which has been promoted by the series of Text Analysis Conferences,[1] is another term that characterises these research efforts (see also Rao et al., 2013).

In contrast to term mapping techniques, there are methods that organise constituent terms of a terminological resource into a variety of classes. Given a terminological resource, in these methods, the usage of terms in a corpus is assessed to decide their membership in concept classes. If the classes are known prior to the assignment task, then the task is known as term classification (e.g., see Nigel et al., 1999). Otherwise, if the classes are *unknown*, the task is called term clustering (e.g., see Dupuch et al., 2014). As described in Chapter 5, from a linguistic point of view, these methods address the identification of *hypernym/hyponym* relationships between the entries of a terminological resource. Krauthammer and Nenadic suggest that these three tasks—that is, term recognition, term classification, and term mapping—are essential to form a closed loop between terminology and natural language text, for the facilitation of automatic construction and maintenance of terminological resources (Figure 3.4).

A more elaborate taxonomy of techniques in computational terminology can be obtained by discerning elements and characteristics of the communicative context other than what is discussed here. As implied in the discussions, besides the methods that are named above, the outlook of ‘terms as units of language’—as named by L’Homme (2014)—underlines the requirements for addressing a number of challenges such as *term variation* and *acquisition of semantic relations* for systematic management of terminological resources. Each of these problems is an active research topic in computational terminology, beyond the scope of this thesis.

In the remaining sections, the common mechanism of term extraction methods is discussed in Sections 3.2. The involved processes, that is, candidate term extraction and the scoring procedure are explained in Sections 3.3, and 3.4, respectively. In Section 3.5, organising terminologies is discussed briefly. The use of machine learning methods and a number of term classification techniques are explained in Section 3.6. Section 3.7 concludes this chapter by explaining the evaluation of theses methods.

[1] See http://www.nist.gov/tac/about/.

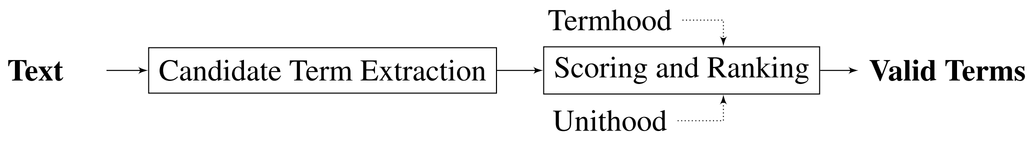


Figure 3.5: Prevalent architecture of terminology mining methods.

## 3.2 Prevalent Mechanism in Term Extraction Tasks

As suggested in Nakagawa (2001a), the algorithms for term recognition are usually in the form of a two-step procedure: candidate term extraction followed by a term scoring and ranking process (Figure 3.5).

Candidate term extraction deals with the term formation and the extraction of candidate terms. The latter is not a trivial task since usually there are no clear differences between a term and general words and phrases in the language at the text surface level. In particular domains such as molecular biology, a share of new terms—for example, the name of new genes—are single-token simple terms. These terms are usually formed and invented using a set of common predefined morphological patterns. The identification of these patterns, for example, as suggested in Ananiadou (1994) and in Zweigenbaum and Grabar (1999), can be helpful in the process of candidate term extraction. However, this kind of term formation is not employed in a large number of domains. Therefore, solutions such as morphological pattern analysis may not always be useful for identifying simple terms. Furthermore, as suggested by Nakagawa (2001a), multitudes of terms are complex terms in the form of uninterrupted collocations. Similar to other types of multi-word expressions, distinguishing these complex terms from phrasal structures in the language has remained a research challenge. Several methods for the extraction of candidate terms are suggested, which will be reviewed in the next section.[1]

Although several Categorisations of the scoring and ranking methods can be given from a methodological point of view (e.g., statistics-based, machine learning-based, rule-based, etc.) or by the kind of information that is exploited for weighting (e.g., linguistic-based, statistical-based, hybrid), as stated earlier, all these techniques rely on the text and take a corpus-based distributional approach to score and rank terms. The usage of candidate terms in a communicative context (e.g., domain-corpus) is formulated in a machine-tractable format—for example, in the form of a contingency table or a vector space model. To compute a score for each candidate term, the collected data is then assessed using statistical measures, similarity metrics, language models or a set of rules. The scoring methodology is determined by the metric employed for scoring candidate terms (e.g., only termhood, only unithood, or a combination of both) as well as the ob-

[1] As can be inferred, this processing pattern is very similar to the extraction of multi-word expressions. However, aside from the difference in scope of research, one notable difference between the research in multi-word expressions and terminology extraction is the scoring procedure in these areas. In term extraction, both unithood and termhood are employed to weight terms, whereas multi-word expression research leans towards unithood measurement (see Baldwin and Kim, 2010, for an overview of research in multi-word expressions).

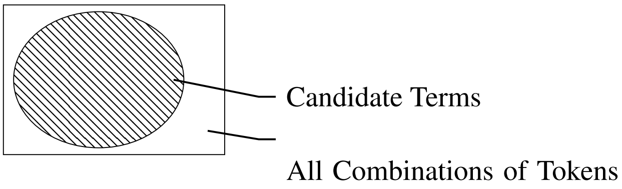


Figure 3.6: Output of the candidate term extraction process: a subset of all combinations of tokens in input text corpus.

jective of the task in hand, which often decides the type of paradigmatic relation that the termhood measure characterises.

This two-step term extraction procedure can be followed by a number of additional processes. For instance, following the two-step procedure, a term selection process may discard a number of extracted terms that have a score below a particular threshold. The strategy for designing this kind of post-processing technique is determined by the intended application for the extracted terms and therefore is not considered as a core process in a term extraction task. Similarly, depending on the employed methodology, a number of pre-processings—for example, part-of-speech tagging, syntactic analysis, etc.—might be required prior to the two-step term extraction procedure.

## 3.3 Candidate Term Extraction

The first step in most term extraction tasks is to extract candidate terms from text. As suggested earlier, candidate term extraction is a non-trivial task. Terms' boundaries cannot be distinguished easily from other words and phrases in the text surface. Whereas earlier research in term extraction suggested that terms show particular morphological or syntactic behaviours, recent research in terminology indicates that terms show a similar linguistic behaviour as general words and phrases in a language. From a radical perspective, in a given text, any combination of tokens and words can be a term. Consequently, choosing candidate terms can be seen as the problem of finding a subset of tokens' sequences (which are likely to be terms) in an exponentially large search space, thus resulting in an *NP-hard problem*. Luckily, a number of linguistic observations suggest particular criteria for the terms' linguistic behaviours—for example, the frequency and the length of terms—which are utilised to define a set of heuristics to limit this search space (Figure 3.6).

In a limited number of domains, knowledge workers may have a guideline for introducing new terms, particularly simple terms. For example, in molecular biology the names of genes are often a combination of letters and numbers. Similar regulations can be found in automotive engine technologies. As suggested earlier, these observations, coupled with the traditional terminology's outlook, led to a number of research methods that assume term formation is a planned, conscious, and well-structured process (Ananiadou, 1994). Hence, in order to extract candidate terms, these methods pay extra attention to the internal morphosyntactic structure of terms and often ignore the context in which they appear (Accordingly, Maynard and Ananiadou (2001) classify these techniques as *intrinsic* approaches). In these methods, a terminological resource is often available prior to the extraction task and it is employed to identify new candidate terms.

While the above-mentioned morphosynatic-based methods have been employed in a few domains, they are not applicable in a large number of sublanguages; for example, creation of new terms may not follow particular morphosyntactic patterns and a terminological resource may not be available prior to the extraction task. Besides, a simple search in a terminological resource shows that the majority of terms are multi-word complex terms. The extraction of these terms introduces additional complexity to the process of candidate term extraction.

Hence, apart from the aforementioned morphosyntactic-based methods that focus on the terms' internal structure, several other techniques have been introduced to address the problem of candidate term extraction. Five major methods can be identified for the extraction of candidate terms:

- the *n*-gram-based techniques;
- linguistic filtering using part-of-speech tag sequence patterns;
- linguistic filtering using syntactic relation patterns;
- techniques that rely on the presence of particular markers in text;
- contrastive approaches.

A combination of these techniques can also be employed to improve the results (e.g., see Aubin and Hamon, 2006). In the following section, each of these methods are described.

### 3.3.1 The *N*-Gram-Based Methods

In the context of candidate extraction, an *n-gram* is a contiguous sequence of *n* tokens from text. In *n*-gram-based methods, the *n*-gram is usually bound to a text window of a particular size (often, $1 \leq n \leq 6$). The most common size for *n* is two in which two-word collocations (bigrams) are considered as the potential candidate terms. In order to reduce the number of undesirable sequences of tokens and restrict the size of the set of the extracted candidate terms, a number of heuristics are employed to filter the extracted *n*-grams. For instance, *n*-grams that contain *stop words*—such as articles, particular prepositions, auxiliary verbs, etc.—are discarded. A major advantage of the *n*-gram-based techniques is that they can be employed in the absence of linguistic analysis tools. Hence, they allow the terminology extraction task to be carried out with purely statistical approaches. Therefore, *n*-gram-based techniques are desirable when dealing with the under-resourced languages where the linguistic analysis tools are usually not available (e.g., see Pinnis et al., 2012).

Compared to other techniques of candidate term extraction, the use of *n*-gram-based methods often results in lower precision. The *n*-gram-based methods generate a large set of candidate terms of which the number of correct terms compared to incorrect terms is expectedly very low. For example, in the context of a keyphrase extraction application, Hulth (2003) investigates the performance of a few candidate term extraction methods including an *n*-gram-based technique. In her methodology, the extracted candidate terms using different techniques are classified as valid or invalid keyphrase using the same supervised machine learning technique. Subsequently, she compares the keywords assigned by the classifier with a list of the author's provided keywords in order to estimate the performance of the candidate term extraction techniques. In these experiments, the employed *n*-gram-based method shows one of the worst performances. Similar results can be found for an automatic term extraction task in Zadeh and Handschuh (2014).

### 3.3.2 Part-of-Speech-Based Methods

Linguistic filters in the form of part-of-speech (PoS) tag sequence patterns have been widely employed for the extraction of candidate terms (Justeson and Katz, 1995). These methods are often affiliated by the linguistic approaches to term recognition. In this category of techniques, patterns of particular PoS tag sequences are employed to extract candidate terms. These patterns are often represented by regular expressions. The use of these patterns yields to the assumption that the construct of terms is restricted to grammatical structures of particular PoS sequences. For example, by observing the target domain's terms, Justeson and Katz (1995) only consider candidate terms that are composed of a combination of nouns ($W_N$), adjectives ($W_A$) and prepositions ($W_P$) and satisfy the following PoS pattern:

$$((W_A|W_N)^+|(W_A|W_N)^*(W_N W_P)^?(W_A|W_N)^*)W_N$$

Bourigault's (1992) LEXTER is another system that employs PoS-based linguistic filtering for the extraction of candidate terms. However, instead of defining desirable PoS patterns, LEXTER employs *negative* knowledge about the form of terminological units, by identifying patterns that do not meet the requirements for forming candidate terms. In the proposed approach, similar to noun phrase chunking, punctuations and particular PoS tags such as verbs and conjunctions—which Bourigault calls frontier markers—are used for determining the boundaries of sequences of tokens that can form candidate terms.[1] A recent example of this methodology can be found in Meyers et al. (2014).

Park et al.'s (2002) GlossEx is another example of a term extractor system that employs PoS tag sequence patterns to extract words and phrases in order to construct domain-specific glossaries. The automatic extraction of candidate terms in GlossEx is limited to the $P_{\text{Noun Phrase}}$ structure that is defined by the following regular expressions:

$$P_{\text{Noun Phrase}} = W_{\text{DT}}^?(W_{\text{VBG}}|W_{\text{VBN}})^? P_{\text{Modifier}}^*(W_{\text{NN}}|W_{\text{NP}}|W_{\text{NPS}}),$$

in which $P_{\text{Modifier}}$ is defined as:

$$P_{\text{Modifier}} = ((W_{\text{JJ}}(W_{\text{CC}}W_{\text{JJ}})^*)|(W_{\text{NN}}|W_{\text{NP}}|W_{\text{NPS}})^?).$$

In these patterns, $W_X$ denotes a word of the particular PoS category $X$ in which $X$ is a PoS tag from the inventory of the tags employed in the Penn Treebank Project. Table 3.1 shows the Penn Treebank PoS tags and their corresponding definitions (Taylor et al., 2003).

[1] The idea behind the method is best described in Abney (1992).

| | | | |
|---|---|---|---|
| CC | Coordinating conj. | RB | Adverb |
| CD | Cardinal number | RBR | Adverb, comparative |
| DT | Determiner | RBS | Adverb, superlative |
| EX | Existential there | RP | Particle |
| FW | Foreign word | SYM | Symbol |
| IN | Preposition | TO | infinitival to |
| JJ | Adjective | UH | Interjection |
| JJR | Adjective, comparative | VB | Verb, base form |
| JJS | Adjective, superlative | VBD | Verb, past tense |
| LS | List item marker | VBG | Verb, gerund/present participle |
| MD | Modal | VBN | Verb, past participle |
| NN | Noun, singular or mass | VBP | Verb, non-3rd ps. sg. Present |
| NNS | Noun, plural | VBZ | Verb, 3rd ps. sg. present |
| NNP | Proper noun, singular | WDT | Wh-determiner |
| NNPS | Proper noun, plural | WP | Wh-pronoun |
| PDT | Predeterminer | WP$ | Possessive wh-pronoun |
| POS | Possessive ending | WRB | Wh-adverb |
| PRP | Personal pronoun | LRB | Left bracket character |
| PP$ | Possessive pronoun | RRB | Right bracket character |

Table 3.1: The list of part-of-speech tags employed in the Penn Treebank Project: *ps.* and sg. denote *person* and *singular*, respectively.

In contrast to the above-mentioned methods that define PoS sequence patterns—thus candidate terms—of arbitrary length, a number of research restrain the length of candidate terms. For instance, Daille (1995) limits the length of their employed patterns to four words, whereas Frantzi (1997) employs patterns that are only two words long. Empirical evidences show that the length of terms is often limited to a few words/tokens. For instance, Maynard (2000) states that in most applications the length of term is usually up to 4 words and it is extremely rare for a term to exceed 8 words in length. Hence, limiting the length of candidate terms may enhance the accuracy of the candidate term extraction process without necessarily decreasing its recall.

Using PoS-based filters implies the need for autoamtic PoS tagging prior to the process of candidate term extraction. Ittoo et al. (2010) highlight problems that can arise due to the presence of noise in the output of this automatic PoS tagging process, particularly when dealing with irregular texts with subtle language patterns and malformed sentences. For instance, in the reported experiment by Ittoo et al., authors noticed that many nouns in their evaluation corpus are tagged incorrectly as progressive-verbs, and therefore resulting in misleading and inaccurate detection of candidate terms. To make the employed PoS patterns tolerant to these errors and solve the problem, Ittoo et al. refer to the actual output of their employed PoS tagger and define patterns that encompass progressive-verbs:

$$(W_{\text{VBG}}^{?})(W_{A}^{*})(W_{N}^{+})$$

where $W_{\text{VBG}}$, $W_A$, and $W_N$ respectively denote progressive verb, adjectives, and nouns.

Dorji et al. (2011) use PoS patterns for the automatic extraction of candidate terms that are used as index terms in a document classification task. By observing appropriate terms

| | |
|---|---|
| Daille (1995) | $(AN\|NN)$ |
| Frantzi (1997) | $(N\|A) + N$ |
| Nakagawa (2001b) | $N^?$<br>$A(N\|A)^*N$<br>$NP_{\text{OF}}N$<br>$F$ |
| Zervanou (2010) | $(A\|N\|V_{\text{BG}}\|V_{\text{BN}})^+N$<br>$(A\|N\|V_{\text{BG}}\|V_{\text{BN}})C(A\|N\|V_{\text{BG}}\|V_{\text{BN}})N$<br>$(A\|N\|V_{\text{BG}}\|V_{\text{BN}})^+NCN$<br>$NP(A\|N\|V_{\text{BG}}\|V_{\text{BN}})^*N$<br>$NP(A\|N\|V_{\text{BG}}\|V_{\text{BN}})^*NCN$<br>$NCNP(A\|N\|V_{\text{BG}}\|V_{\text{BN}})^*N$<br>$(A\|N\|V_{\text{BG}}\|V_{\text{BN}})C(A\|N\|V_{\text{BG}}\|V_{\text{BN}})N$<br>$(A\|N\|V_{\text{BG}}\|V_{\text{BN}})^+NCN$ |
| Bonin et al. (2010a) | $N^+(P^+(N\|A)^+\|N\|A)$ |

Table 3.2: Proposed PoS sequence patterns for Candidate Term Extraction. $A$ denotes adjectives; $N$ denotes nouns; $C$ denotes conjunctions; $P$ denotes prepositions; $P_{\text{OF}}$ denotes the preposition *of*; $F$ denotes foreign words; $V_{\text{BG}}$ denotes verbs in gerund form; and, $V_{\text{BN}}$ denotes verbs in the past participle form.

in their application, Dorji et al. have adopted PoS sequence patterns with various lengths of two to ten words. However, instead of specifying the complete PoS sequence patterns, they define seven core patterns of lengths two to four words. These sequences of PoS tags can in turn be followed by an arbitrary number of nouns to form patterns of maximum length ten words. Similarly, Eck et al. (2010) only consider a subset of noun phrases that do not contain any preposition. The use of PoS sequence patterns is not limited to what is reported here and has been widely employed in term extraction tasks (e.g., see Anick et al., 2014; Zervanou, 2010; Hsu, 2010; Bonin et al., 2010a; Barrón-Cedeño et al., 2009).

Apart from algorithmic variances, the coverage of patterns is the major difference between techniques that employ PoS-based patterns for candidate term extraction. The higher coverage of patterns yield a higher recall, but usually at the expense of lower precision. Preference for precision requires a strict filter which permits a limited sequence of words as candidate terms, whereas preference for recall demands a filter with relaxed restrictions on the permitted sequences of words (Frantzi et al., 2000a). In addition, Eck et al. (2010) emphasise that the choice of an appropriate PoS pattern depends on the common structures that are employed by the sublanguage of the corpus. The definition of patterns using PoS sequences, thus, is an open question and no best universal pattern can be found. The reported experiment by Hulth (2003) states that considering PoS tags can result in a dramatic improvement of precision. Moreover, in her evaluation, the highest recall has been reported for the candidate term extraction based on a set of PoS tag patterns (surprisingly even in comparison to the $n-$gram technique). Table 3.2 shows additional examples of the employed PoS sequence patterns in research literature.

### 3.3.3 Syntactic-Based Methods

Research literature reports the use of linguistic filters that employ syntactic relations for the extraction of candidate terms. The first category of these methods employs syntactic patterns for the identification of term variations rather than the extraction of candidate terms. For example, Jacquemin and Tzoukermann (1999) report the use of a transformational unification-based syntactic parser together with morphosyntactic analysis for the identification of term variants in a controlled vocabulary environment. If a dictionary of terms is available prior to the extraction task, this method can be used for generating candidate terms.

The second category of syntactic-based methods use shallow parsing for the extraction of candidate terms. Instead of the extraction of collocations with specific PoS patterns, noun phrase chunks are extracted as candidate terms (e.g., see Evans and Zhai, 1996; Nakagawa, 2001a; Fan and Chang, 2008)[1]. In the reported results by Hulth's (2003), this technique gives the highest precision amongst PoS-based and $n$-gram techniques. However, in an experiment that I have reported in Zadeh and Handschuh (2014), whereas noun phrase chunking outperforms an $n$-gram-based technique, it underperforms a PoS-based method.

The third category of syntactic-based filters considers the role of compounding in term formation and employs syntactic relations according to the head-modifier principle (e.g., see Jakubíček et al., 2014; Hippisley et al., 2005). By observing the role of compounding in term formation, Hippisley et al. (2005) apply the head-modifier principle in compounding word formation for the extraction of complex candidate terms. According to the head-modifier principle, in a syntactic construct, one of the constituents acts as the head. The head has a strong association to the core semantics of the construct, and it is modified by the other dependent constituents. In the proposed method in Hippisley et al. (2005), candidate terms are extracted by identifying particular syntactic relations to the left and the right side of the head. The major advantage of these techniques is that the head-modifier principle can additionally be used for deconstructing complex terms. Therefore, the proposed approach by Hippisley et al. is more popular within the context of machine translation applications for multilingual term extraction.

A detailed description of a head-modifier-based technique for candidate term extraction can be found in Wong (2009). Using dependency relations, the proposed method starts with a search for the heads in a sentence. Using the acquired head-modifier information from the dependency parse, the head is then extended to both left and right direction to identify maximal-length noun phrases. In the proposed method, the head-driven filter restricts the PoS tags of modifiers to nouns (except possessive nouns), adjectives, and foreign words. This process is followed by the use of a statistical measure in order to attach terms that appear immediately after one another, or terms that are separated by a preposition or coordinating conjunction.

[1] Perhaps, a number of methods that are listed in Section 3.3.2 can also be added under this category.

The use of syntactic relations for the extraction of candidate terms is not limited to the above-listed methodologies. For example, Seretan et al. (2004) describe a sophisticated technique for the extraction of multi-word complex terms. In the proposed method, a set of pairs of words that are connected directly through a syntactic relationship are first extracted. Instead of the sequence of tokens in the input corpus, the extracted pairs of words are searched for extracting candidate terms. The set of extracted pairs of words is then utilised for the extraction of compound words, idioms and collocations from French and English parallel corpora.

### 3.3.4 Methods Based on Particular Structures in Text

An alternative approach to candidate term extraction exploits specific properties of the input text. A growing numbers of research exploits the presence of mark-ups in input text to extract candidate terms. For instance, Brunzel (2008) uses the HTML mark-ups in order to extract candidate terms and Hartmann et al. (2011) and Toral and Munoz (2006) exploit the semi-structured representation of text in Wikipedia's articles in order to form a set of candidate terms. The use of these techniques therefore is limited to domains in which text is annotated by mark-ups.

In the same way, in particular domains, candidate terms can be extracted with the help of specific lexical patterns or the presence of mark-ups in input text. For instance, in biotechnology, Rindflesch et al. (2000) describe a method for the extraction of candidate terms that employs a list of general *binding words*. In the proposed application domain, the presence of *binding words* in a noun phrases qualifies it as a candidate term. Similar method for the extraction of disease risk factors for metabolic syndrome in biomedical text is reported by Fiszman et al. (2007). Fiszman et al. (2007) suggest the use of indicative words including specific lists of verbs and nouns. Similar methods are proposed in Hazen et al. (2011) for the extraction of terms related to imaging observations in radiology and in Gooch and Roudsari (2011) for the extraction of clinical terms.

### 3.3.5 Contrastive Approaches

*Contrastive approaches* exploit a reference corpus of general language to identify simple and complex candidate terms from input text (Drouin, 2004, 2003). To form the hypothesis space of likely candidate terms, these methods rely on one of the techniques listed in the previous sections, for example, an *n*-gram-based method. Candidate terms are extracted from both the target special corpus and a general language corpus (e.g., the British National Corpus[1] when processing English text) or a special corpus in knowledge domain other that the target special corpus. The extracted candidate terms and their frequencies in these two corpora are exploited to distil a set of likely candidate terms in the given special corpus. Similar methods can be found in Basili et al. (2001).

[1] See `http://www.natcorp.ox.ac.uk/`.

### 3.3.6 A Summary of Methods

To summarise, methods that employ linguistic information such as PoS tags and syntactic relations demand more resources than methods that rely only on the text surface structure. Methods that employ PoS-based sequence patterns require a PoS tagger with an acceptable performance. Similarly, syntactic-based methods demand a form of chunking or a syntactic parsing prior to the extraction task. These methods have been reported to deliver high precision; however, their required resources may not be available for all languages or domains. On the other hand, the *n*-gram-based techniques do not require such resources and are language-independent. However, these methods are reported to have a low precision, which can diminish the performance of the subsequent ranking process. The application of techniques such as the use of text structure, or using lexical indicators may not be applicable to all domains. Lastly, as suggested in Bonin et al. (2010b), the use of contrastive techniques can enhance the results.

In real-world applications, in order to improve the results, a combination of the above-listed methods are employed. For instance, Aubin and Hamon (2006) consider a combination of PoS sequence patterns, head-modifier relationship as well as a contrastive technique to extract a list of candidate terms. In another example, Hulth (2003) reports the highest F-Score in her experiments when candidate term extraction is carried out using a combination of *n*-gram techniques and PoS tag sequence patterns.

## 3.4 Methods for Scoring Candidate Terms

In automatic term recognition tasks, the scoring and ranking process follows the extraction of candidate terms. It is assumed that the set of extracted candidate terms contains both *valid* and *invalid* terms. Put simply, a candidate term is valid if it denotes a concepts from the knowledge domain that is represented by the input special corpus to the term extractor.[1] Hence, the main goal of term scoring process is to distinguish valid terms from invalid terms. This goal is often achieved by a ranking and filtering mechanism. The scoring process assign a score to each candidate term, ideally according to the significance of the concepts that they represent in the target knowledge domain. After this process, candidate terms with a score below a certain threshold are usually discarded and the rest are ranked and accepted as valid terms for further processes (Figure 3.7).

Traditionally and from a methodological perspective, terminology extraction approaches are often classified as *linguistically-motivated*, *statistically-oriented*, and *hybrid* methods ( e.g., see Kageura and Umino, 1996, description on the topic). In this classification, often the candidate term extraction and scoring procedure are not heeded independently from each other. Hence, linguistically-motivated methods often encompass techniques that employ linguistic filtering for the extraction of candidate terms (although recent methods also use linguistic information as an attribute in statistical models).[2] In this

[1] As discussed earlier in Section 3.1, there is no straightforward definition of valid terms.

[2] The use of linguistically-motivated approaches can be traced in information retrieval tasks for the problem of index term extraction (e.g., see Baxendale, 1958).

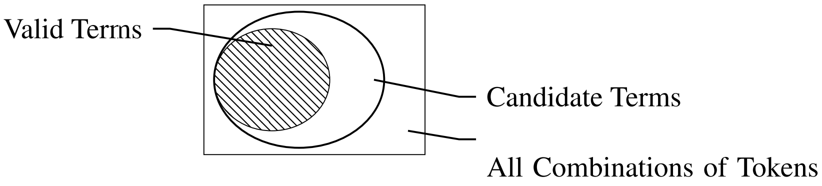


Figure 3.7: It is assumed that the output of the candidate term extraction process—that is, a subset of all combinations of tokens in input special corpus—contains both valid and invalid terms. Hence, a scoring and ranking process is employed to distinguish valid terms—that is, a subset of candidate terms.

classification, the statistical methods employ a mathematical model such as probabilities to perform the extraction task and ignore linguistic structure of terms and their context. As expected, the methods in this category often use *n*-gram-based methods for the extraction of candidate terms. The third category of methods in this classification, known as hybrid methods, offers solutions that combine both linguistic information and statistical measures. In fact, since the majority of the methods for terminology extraction rely on the text and adopt a corpus-based approach, they use a kind of statistical information derived from the corpus at some stage in the process. Hence, corpus-based methods are classified as statistically-oriented or hybrid technique.

Alternatively, as suggested earlier, the procedure of term extraction can be analysed and classified from a functional perspective: (a) methods that deal with the identification of atomic meaning-bearing lexical units and (b) methods that indicate the desirability of the extracted candidate terms as a unit of meaning in a terminology database. As suggested by Kageura and Umino (1996), in the former group, the focus is on the unithood measurement, thus the extraction of candidate terms that form stable lexical units. However, the focus of the former methods is on the termhood measurement, that is, scoring atomic lexical units by their significance in the target knowledge domain.

In the framework of distributional semantics, the computation of unithood is perceived as the identification of *syntagmatic* relationships between words that constitute a complex term. These relationships are often in the form of collocations. Therefore, the first category of methods deals with lexical association measures. A general account of these methods can be found in Evert (2004); Hoang et al. (2009); and, Pecina (2010). Similarly, in the framework of distributional semantics, the computation of termhood implies the identification of paradigmatic relations. These paradigmatic relations characterise the relevance of the meaning of terms to significant concepts in the knowledge domain and with respect to the communicative context, that is (in its simplest form), the special corpus.[1]

In corpus-based distributional approaches, the text and the statistics that are induced from its analysis are the major source of information to characterise these paradigmatic relations. As detailed in the next few sections, the statistical information about the usage of terms can be modelled and presented in a variety of ways, for example, as simple as

[1] In fact, the communicative context goes beyond the special corpus. It is a complex system consisting of several elements such as the knowledge the users, the intended application, and so on.

computing *tf-idf* of terms to sophisticated learning algorithms. To characterise termhood, techniques other than corpus-based approaches are also feasible. For example, Maynard (2000) draws attention to the incorporation of knowledge-bases and their internal structure for the development of terminology extraction systems. The study of these methods, however, remains out of the scope of this thesis.

As described in the preamble of this section, statistical measures employed in terminology extraction can be classified into two categories: measures that address unithood and those that address termhood. However, drawing such a clear line is sometimes not possible (Kageura and Umino, 1996). According to Kageura and Umino (1996), statistical measures in terminology extraction are employed by relying on the following hypotheses:

- a lexical unit that appears frequently in a special corpus is likely to be a term of the domain knowledge that the special corpus represents;
- a lexical unit that appears only in one special corpus is likely to be a term of the domain knowledge that the special corpus represents;
- a lexical unit that appears more frequently in a special corpus than in a general language corpus is likely to be a term in the domain knowledge that is represented by the special corpus.

As discussed earlier, unithood is only defined for complex terms. The examples of statistical measures that have been used to measure unithood are numerous: Pearson's chi-square test and Log-likelihood, mutual information (e.g., as employed in Church and Hanks, 1990); coefficients for sequential data such as the Ochiai and Kulczynski coefficient suggested by Daille (1995); customised measures such as paradigmatic modifiability by Wermter and Hahn (2005); mutual expectation as suggested in Dias and Kaalep (2003), and so on.

Likewise, a long list of statistical measures have been employed to characterise termhood: inverse document frequency (*idf*) suggested in Jones (1972); term frequency–inverse document frequency (*tf-idf*) as used in Salton (1992) and its modifications such as Feiyu et al.'s (2002) *kfidf*; Frantzi and Ananiadou's (1996) C-*value* and NC-*value*; and, the statistical barrier measure proposed in Nakagawa (2001a) are a few examples.

### 3.4.1 Unithood Measures

Pearson's chi-square test ($\chi^2$ test) is an intuitive statistical measure that can be used for characterising both unithood and termhood. $\chi^2$ is measured by the comparison of the observed and expected frequencies under *the null hypothesis of independence*:

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}, \tag{3.1}$$

where $f_o$ is the observed frequency and $f_e$ is the expected frequency (see Manning and Schütze, 1999, for further explanations). If $f_o$ and $f_e$ are derived from the observed frequencies in the collocations of constituent words in complex terms—for example, as suggested in Dunning (1993)—then the computed $\chi^2$ value can be interpreted as a measure of unithood. However, if $f_o$ and $f_e$ are derived from the observed occurrences of terms in documents—for example, as suggested in Kilgarriff (1996)—then the computed $\chi^2$ value can be interpreted as a measure of the terms's association to documents, hence termhood (see also Rayson et al., 2004). It is important to note that the chi-squared measure is meaningful only when the collected frequencies are greater than 5.

Log-likelihood ratio test (LL) is another statistical measure that has been used for characterising unithood. According to Dunning (1993), LL shows one of the best performances, particularly when frequencies are collected from small corpora. As described in Daille (1995) and Korkontzelos et al. (2008), LL can be seen as a refinement of the $\chi^2$ test. Instead of relying on the assumption of a normal distribution of words in collocations, LL compares the observed frequency counts in a sub-corpus with the counts that would be expected in a reference corpus to measure the likelihood of co-occurrence. For bigrams $w_i w_j$, LL can be computed as follows:

$$\mathrm{LL} = \log_2 \frac{P_s(w_i, w_j)}{P(w_i, w_j)}, \tag{3.2}$$

where $P(w_i, w_j)$ is the probability of observing $w_i$ and $w_j$ as a bigram in the reference corpus, and $P_s(w_i, w_j)$ is the probability of their occurance as bigram in the subset $s$ of the corpus (i.e., the target domain). Similar to the interpretation of $\chi^2$ test, a high LL means that observed and expected values diverge significantly, and thus indicates that $w_i$, and $w_j$ do not co-occur by chance. In contrast, a LL value close to 0 indicates that $w_i$, and $w_j$ do co-occur by chance. LL ratio is highest when $w_i$, and $w_j$ only appear as bigrams next to each other. However, as mentioned in Korkontzelos et al. (2008), the LL ratio is also high for rare bigrams. Hence, the LL ratio of noisy bigrams such as typographical errors is also high, which consequentially may negatively affect the performance.

Similar to LL and $\chi^2$, point wise mutual information (PMI) can also be used to measure the unithood of complex candidate terms in a corpus. PMI, however, estimates the expected probabilities using the products of the probabilities of the constituent words of complex terms. For terms that consist of two words $w_i$ and $w_j$, PMI is defined as:

$$PMI = \log_2 \frac{P(w_i, w_j)}{P(w_i)P(w_j)}, \tag{3.3}$$

where it is assumed that $w_i$ and $w_j$ appear independently. A high PMI value shows a strong association between the constituent words of the candidate terms. Hence, candidate terms that have high PMI value are assumed to be valid complex terms. In contrast to LL, PMI gives a low score to the rare candidate terms. The Dice measure, Z-score, and rank aggregation as suggested in Dinu et al. (2014) are other methods that can be used to evaluate the unithood of complex terms. As stated earlier, any method of sequential data modelling can be used to estimate unithood. Moreover, the use of statistical information other than words occurrence information is also feasible. For example, Tsvetkov and Wintner (2014) construct of a Bayasian network by integrating diverse statistical information to extract multi-word expressions.[1]

[1] As suggested by Evert (2009) and Kilgarriff (2005), in this context, the assumption of independence is not reasonable and thus can decrease the performance of the method.

### 3.4.2 Termhood Measures

The tf-idf measure, a term weighting score often used in information retrieval, is perhaps one of the most applied statistical measures for characterising termhood. In automatic term recognition tasks, tf-idf is usually used as a baseline for the comparison of termhood measures (Zhang et al., 2008). The tf-idf score is the product of two statistics: inverse document frequency and term frequency. Inverse document frequency $idf(t_i)$ measures the general importance of a term $t_i$ in a collection of documents $D$ by counting the number of documents that contain $t_i$, usually but not necessarily in a logarithmic scale:

$$idf(t_i) = \log \frac{|D|}{\left|\left\{d_j \in D : t_i \in d_j\right\}\right|}, \tag{3.4}$$

where $|D|$ denotes the cardinality of $D$, and the denominator indicates the number of documents that contain $t_i$. Subsequently, tf-idf for the term $t_i$ over $D$ is give by:

$$tf\text{-}idf(t_i) = tf(t_i) \times idf(t_i), \tag{3.5}$$

where $tf(t_i)$ can be the frequency of the term $t_i$ in the corpus. This definition of the tf-idf score is employed by assuming that important terms occur in particular documents frequently whereas they are relatively rare in the input corpus (i.e., they occur in a small number of documents). This assumption can be refined; hence, alternative definitions of $tf(t_i)$ and $idf(t_i)$ may be used.

Another approach to estimate a termhood score is that of *corpus comparison*—or, contrastive methods as explained earlier for candidate term extraction. In these methods, a corpus is compared against a general language corpus. It is often assumed that the distribution of valid terms and invalid terms varies in corpora of different types (Knoth et al., 2009). One implicit way to implement this logic is the use of statistical hypothesis testing, for example, as described earlier for Equation 3.1 and 3.2 and as employed in Kilgarriff (2001) and Rayson and Garside (2000). Alternatively, a category of contrastive approaches define statistical measures that explicitly exploit the observed frequencies in different corpora (e.g., see Drouin, 2004; Ittoo and Bouma, 2013). Liu and Kit (2008) suggest that these approaches are more desirable than techniques that only utilise a special corpus since they employ intrinsic statistical characteristics of valid terms in different corpora. Ahmad et al.'s (1999) *Weirdness* score is a classic example of this category of techniques that can be used to assign a termhood measure to a candidate terms $t$ in a special corpus:

$$Weirdness(t) = \frac{f_s(t)/_{n_s}}{f_g(t)/_{n_g}}, \tag{3.6}$$

where $f_s(t)$ and $f_g(t)$ are the frequency of $t$ in the special and a general corpus, respectively; similarly, $n_s$ and $n_g$ are the total frequency of terms in the respective corpora.

### 3.4.3 Hybrid Measures and a Little More of the Context

Amongst the statistical methods for termhood and unithood measurement, Frantzi and Ananiadou's (1996) c-*value* measure has attracted much attention. In contrast to statistics measures introduced previously, the C-*value* score can be seen as a *hybrid termhood-unithood* measure hence its definition considers statistical information that concerns both unithood and termhood of terms. For each candidate term $t$, the C-*value* score of $t$, is calculated using four criteria (Frantzi et al., 2000b): the frequency of $t$ in the corpus; the frequency of $t$ when it appears nested in other terms longer than $t$; the number of those longer terms; and the number of the constituent words of $t$ shown by $|t|$. The C-*value* of $t$ is given by

$$\textsc{c}\text{-}\mathit{value}(t) = \begin{cases} \log_2 |t| f(t) & \text{if } t \notin \text{nested} \\ \log_2 |t| \left(f(t) - \frac{1}{|T_t|} \sum_{b \in T_t} f(b)\right) & \text{otherwise} \end{cases}, \tag{3.7}$$

where $T_t$ denotes the set of all the terms that contain $t$ and are longer than $t$, and $f(s)$ denotes the frequency of an arbitary term $s$ in the corpus. The greater the C-*value*($t$), the more likely $t$ is a valid term.

Following the C-*value* score, Frantzi et al. (2000b) introduce the NC-*value* score. The NC-*value* score is perhaps one of the first widely employed scores that implements the idea of *terms in context* by Pearson (1998). The NC-*value* score improves the C-*value* score by considering the frequency of words surrounding the terms. Frantzi et al. (2000b) hypothesise that valid term appears with a *closed* set of neighbour words. Accordingly, the occurrence of these words around a candidate term is a *positive clue* that can be used in determining the termhood of the candidate term. This idea is implemented with the help of a function called *context weighting factor*. First, a set of as context words—which consists of nouns, adjectives, and verbs that appear in the vicinity of candidate terms—is extracted. Each word in this set is assigned to a context weight:

$$\mathit{weight}(w) = \frac{t(w)}{n}, \tag{3.8}$$

where $t(w)$ is the number of terms that $w$ co-occur with, and $n$ is the total number of candidate terms considered. The *weight*($w$) is then considered to indicate the important of $w$ as a context word. Subsequently, the NC-*value* for the term $t$ is computed by

$$\textsc{nc}\text{-}\mathit{value}(t) = 0.8\textsc{c}\text{-}\mathit{value}(t) + 0.2 \sum_{b \in C_t} f_t(b)\mathit{weight}(b), \tag{3.9}$$

where $C_t$ is the set of distinct context words that co-occur with term $t$, and $f_t(b)$ is the frequency of the co-occurrences of the word $b$ and the term $t$.

Following the NC-*value*, Maynard and Ananiadou (2000) introduce the SNC-*value* score by incorporating further information about the context in which candidate terms appear. To compute SNC-*value*, Maynard and Ananiadou suggest the use of three kinds of contextual information: *syntactic*, *terminological*, and *semantic* information. The syntactic information, as its name suggests, is mostly concerned with the distance between a candidate term and its context words. The terminological information suggests the use of co-occurrence counts of candidate terms and previously known terms (context terms). Finally, semantic information takes similarities of context terms into consideration by computing distances between them in a pre-constructed taxonomy of the context terms—similar to WordNet-based methods such as Wu and Palmer (1994).[1]

By incorporating contextual information in their implementations (e.g., as implied by the last few techniques in this section), statistical techniques can go beyond the simple classic intuitions that are listed in the beginning of this section. Incorporating the contextual information in these models not only enhances the performance of methods that assign unithood and termhood scores to candidate terms, but also enables the design of methods that can model the semantics of terms. Hence, during the past decade, the terminology extraction methods have leaned further towards the implementation of the idea of terms in context, often in the form of *supervised* machine learning techniques. Perhaps, this is partly due to the availability of the language resources that are required for implementing this type of methods.

## 3.5 Organising Terminologies

Modern approaches to terminology encourage perspectives of terminology management similar to the way that lexical items are handled in general language. As discussed in the beginning of this chapter, whereas traditional terminology considers terms as *labels for concept*—untouched by context and detached from linguistic characteristics and interpretations—it has become evident that terms, like other lexical units in general language, are subject to linguistic norms. As suggested by Faber and L'Homme (2014), this latter perspective is perhaps best characterised by the term *lexical-semantic approaches to terminology*, in which conceptual modelling and knowledge representation is one of the major concerns (see also Buitelaar et al., 2009, for a similar discussion in the context of *ontology engineering*).

In order to organise lexical resources, lexical-semantic frameworks identify and employ a set of semantic relations such as synonymy and hyponymy between words. The well-known example of such a general lexical resource is WordNet (Miller, 1995). In WordNet, lexical units are grouped into *synsets*. Each synset contains a set of synonymous words—that is, words that have a similar meaning. Subsequently, these synsets are organised into a hierarchy of lexical concepts by defining a hyponym relationship between them—that is, in simple terms, a *type-of* or is-a relationship. Lexical items can be grouped by mechanisms other than synsets (e.g., see Pustejovsky et al., 1993) and organised by a variety of relationships other than synonym and hyponym relationships between lexical units (e.g., see Khoo and Na, 2006, for a survey on semantic relations).

Driven by demands in information system, in modern terminology, a similar principle is suggested for organising terminological resources. Manual encoding of semantic relationships between terms, however, is a time-consuming and tedious task. Moreover, terminological resources are required to be updated frequently; new terms are often introduced and they must be identified and organised in a terminological resource. More challenges are imminent when other properties of terms, such as their life cycle,[2] is con-

[1] I would like to draw your attention to the paradigm change in the series of research by Ananiadou in terminology extraction: from a rationalist approach similar to the GTT in Ananiadou (1994) to empiricist *term in context* techniques in Maynard and Ananiadou (2000).

[2] As discussed in the introduction of the thesis, too.

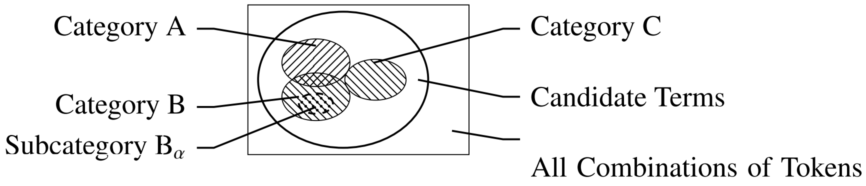


Figure 3.8: A Venn diagram that illustrates organisation of terms with respect to their concept categories. The dashed area shows valid terms. The set of valid terms enfolds several categories of terms, and each characterise a major concept in knowledge-domain. Hence, the identification of terms can be seen as the identification of a number of categories of terms. As discussed earlier, a term may belong to more than one category of concepts. Similarly, a category of concepts may include several subcategories. Entity recognition and term classification tasks are meant to identify particular categories of terms—that is, a subset of valid terms.

sidered (see L'Homme, 2014). Hence, a body of research in terminology mining has paid attention to the automatic organisation of terminological resources and the identification of semantic relationships between terms.

Amongst conceivable semantic relationships between terms, the detection of synonym relationships for the identification of term variations, and hyponym relationships for characterising an organisation of terms in a 'conceptual structure' have been at the centre of attention. The study of research literatures that address the identification of semantic relationships goes beyond the scope of this thesis. However, to provide a complementary view on the term classification task investigated in the later chapters, I briefly review research literature that aim for the identification of *type-of* relationships between terms (see also L'Homme and Bernier-Colborne, 2012; L'Homme, 2014, for an elaboration of the use of semantic relationships in terminological resources).

Methods that address automatic organisation of terminological resources by identifying a *type-of* relationship between terms are all similar in the sense that they assume terms can be organised in several categories to form a taxonomy.[1] Each category (taxon) characterises a group of terms from *similar* concepts in the domain of study (see Figure 3.8). For example, in computational linguistics, the terms *lexicon* and *multilingual corpus* can be categorised under the concept category of *language resources*, while *parsing* and *speech recognition* can be categorised under the concept of *methods and technologies*. Scoring techniques discussed in the earlier sections target distinguishing invalid candidate terms from valid terms and thus result in terminological resources that have a flat organisation (as opposed to the structure of taxonomies). To organise terms in an structure, therefore, an additional classification process is employed.

These classification methods can be distinguished with respect to several factors. For example, Weeds et al. suggest that these methods can be grouped by the type of information that they employ. Similar to what is suggested earlier in Section 3.4, Weeds et al. (2005) identify methods that rely on *internal* information (i.e., the lexical properties of

[1] How these categories are defined and observed is a controversial matter (e.g., see Kilgarriff, 1997) that goes beyond the scope of this thesis.

the words that constitute terms) or *external* information (i.e., statistical, contextual, or ontological information about terms). As discussed earlier, except early works that rely on internal information, recent methods usually adopt a distributional approach towards modelling the semantics of terms, hence they often rely on external information or a combination of both external and internal information.[1]

From a methodological perspective, Weeds et al. (2005) suggest that the majority of these classification methods employ machine learning techniques in the form of a *supervised* classification problem. However, other types of methodologies are also possible. For example, Fukuda et al.'s (1998) PROPER system—a bio-entity tagger—employs a rule-based method. The use of rule-based methods, however, is hindered by their requirements for hand-crafted rules. I extend this study by distinguishing the way that the task of organising terminologies and the classification method are formulated.

If a prior knowledge of the concept categories is not available, automatic organisation of terminologies can be carried out using a method of clustering. These clustering methods are *unsupervised* since no manual effort is required prior to the classification (clustering) task. These methods suggest an organisation of terms by automatic identification of a number of concept categories. Recent examples can be found in Bertels and Speelman (2014); Dupuch et al. (2014, 2012). Terms are first grouped by a measure of similarity—usually, with the help of a distributional approach. Depending on the application context, the obtained clusters of terms can be labelled, which may introduce further complications to the process. One of the main applications of these methods is *ontology learning*, where these clustering techniques can be used as an assistive tool in the process of *ontology engineering*.

Concept categories, however, are typically known prior to the extraction of terms (or, at least, a partial knowledge of them exists). In these scenarios, a typical task is to find terms that belong to particular concept categories. The most established example of this kind of task is the identification of terms that correspond to instances of concepts that are of interest to biologists, namely bio-entity recognition (Nigel et al., 1999). These tasks rely heavily on manually annotated corpora: each mention of a term and its category-concept is annotated in a special corpus. The manual annotations are then employed to develop an entity tagger in a supervised fashion and, often, in the form of a sequence classifier—for example, using a machine learning technique such as the conditional random field method, etc. As reported previously, provided that enough training data is available, it is possible to attain a reasonable performance in these recognition tasks (e.g., see Kim et al., 2004).

In an alternative use case, the targeted concept categories—similar to entity recognition tasks—are known. However, no manual annotation is available for the training and development of a term/entity tagger. The lack of language resources is a familiar problem if a terminological resource with a taxonomic structure must be constructed for a new domain and only using text (i.e., from scratch). This is a task with many real-world

[1] See Chapter 2 of this thesis for an introduction to the distributional methods. Maynard et al. (2008) articulate the basic idea behind these methods through an example: as a person's social life can provide valuable insight into their personality, so we can gather much information about a term by analysing the company that it keeps.

applications (e.g., see Chakraborty et al., 2014; Anick et al., 2014), which can also be employed to address *ontology population* (e.g., see Tanev and Magnini, 2008; Maynard et al., 2008; Andersson et al., 2014). Lastly, a restored interest in these methods is signalled by the trending task of *cold-start knowledge base population* (see Ellis et al., 2012; Mayfield et al., 2014). As previously stated, one of the common challenge that these methods address is the lack of sufficient language resources for the development of classifiers.

Similar to terminology extraction and in contrast to entity recognition task, in these methods the communicative context is often the special corpus. Hence, these methods do not deal with individual term mentions. However, in contrast to terminology recognition techniques (which extracts terms from diverse concept categories in a specific domain knowledge) and similar to entity recognition, the objective of these methods is to extract a subset of terms from a similar category of concepts in a specific domain knowledge. From a lexical-semantic perspective, given a term in a special corpus, these methods can be used to discover the major *senses* of the term in the corpus. Therefore, the outcome can also be beneficial in ontology-based information systems, in which terms are often used as labels to access concepts. Similarly, these methods can be used for the knowledge base population using the so-called distant supervision technique (e.g., see Dredze et al., 2010). As suggested in the introduction chapter, this thesis investigates the development of a term classification method from this category.

Disregard of the methodology for extracting the term and its concept category, these methods assume terms have non-compositional semantics. The targeted hyponymy/hypernymy relationships are then modelled as a paradigmatic relationship. The same approach is often applied to synonymy identification and addressing the problem of term variation.

## 3.6 Machine Learning in Terminology Mining

Machine learning techniques have been widely used for extracting terms and constructing organised terminological resources. The extraction of candidate terms—particularly, complex candidate terms—is expectedly the first juncture that learning methods are utilised. In these applications, though implicitly, a learning method is employed to estimate lexical associations and thus unithood. The simplest example is the use of learning techniques for chunking and extracting nominal phrases. More sophisticated examples of this kind can be found in the context of multiword expression extraction in which the extraction of candidate multi-word lexical units often goes beyond extracting nominal collocations (e.g., see Pecina, 2008).[1]

Apart from the use of machine learning techniques for bracketing and candidate term extraction, in the research literature that investigates terminology mining, they are employed in two additional broad applications.

In the first category, a learning technique is employed to combine various scores from different sources of information in order to enhance the computed scores for the extracted

[1] Hence, although important in many natural language processing applications, not all the applied methods for extracting multiword expressing are relevant to terminology mining.

candidate terms. Usually, several types of unithood and termhood measures are merged to synthesise a new score. A classic example in this category is Vivaldi et al. (2001) in which a term scoring process is enhanced by combining multiple scores using a *boosting algorithm*. A more recent research in this line is presented by Hamon et al. (2014). Hamon et al. suggest a parametrised c-*value* scoring technique in which the introduced parameters are learned through an optimisation process based on the principles of *Genetic algorithm*.

In the next category, as suggested in the previous section, learning methods are often employed to organise a terminology by identifying co-hyponym relationships, or comparably, to extract terms that belong to a particular category of concepts (see Figure 3.8). In most applications, as discussed, the learning techniques are often used in the form of a supervised classifier. Based on the reasoning shown in Figure 3.3 and apart from the discussion in the previous section, machine learning-based methods that are employed in terminology mining can be also grouped by the type of communicative context that they model.

In the first group, a snippet of text that contains a mention of a candidate term is assumed to be a sufficient representative of the communicative context. In these applications, the identification of candidate terms and their corresponding concept categories are done simultaneously. In the second group, however, the communicative context is the special corpus. In these methods, the extraction of candidate terms and their Categorisation are usually, but not necessarily, performed in a two-step procedure. The first group, understandably, consists of machine learning-based *entity recognisers*, which aim for the identification of entity mentions in text. The second subcategory, however, encompasses methods that are commonly known as *term classification* methods.

The first group of learning-based methods—that is, entity recognition—is situated at the convergence point of the automatic term extraction and the classic named entity recognition (NER) tasks. The goal of NER is to recognise and classify *proper nouns* and numerical values into particular classes of entities such as location, organisation, time, and date (see Mohit, 2014; Nadeau and Sekine, 2007, for a survey on NER). However, as suggested by Yangarber et al. (2002), these recognition tasks can be generalised to other types of nominal compounds other than proper nouns. Therefore, techniques that have been previously applied to NER, have been widely adopted for the recognition of terms, inasmuch as some research does not differentiate between NER and other term classification methods (e.g., see Spasić and Ananiadou, 2004). The best examples of these tasks can be found in molecular biology domain and the task of *bio-entity recognition*. A bio-entity recogniser aims to identify mentions of a particular class of biological instances in text snippets (e.g., see Kim et al., 2004).

Various learning algorithms and a diverse set of features have been proposed to address the task of bio-entity recognition. For instance, Yamamoto et al. (2003) propose a system that employs a support vector machine to identify protein names from sentences in a set of abstracts from scientific publications—that is, from Kim et al.'s (2003) GENIA corpus. The proposed method relies on several kinds of features: morphological characteristics of candidate terms, the surface form as well as the lemma of the set of words that co-occur with candidate terms in the training set, part-of-speech tags and syntactic information, and features extracted from available dictionaries in the domain. Many more examples of this kind can be found in biomedical text mining research.

The application of entity recognisers is not limited to the identification of biological instances. Kovačević et al. (2012) suggest a method to identify *methodology mentions* in scientific publications and classify them into four categories: *tasks*, *methods*, *resources*, and *implementations*. The term recognition and classification are merged and formalised as a sequence tagging problem using conditional random fields—a classifier per concept category. In the proposed method, sentences that describe a methodology are identified. The identified sentences are then passed to each of the trained classifiers in order to extract text segments that correspond to the methodology mentions. Similarly, QasemiZadeh et al. (2012) employ *support vector machines* to extract technical terms.

In the second group—that is, term classification—the process of mapping terms to concept categories is often modelled as an ad-hoc process. A classic example of this kind of method is Nigel et al. (1999), in which *decision trees* are employed to classify terms extracted from abstracts in the domain of molecular biology. Similarly, Spasić and Ananiadou (2004) propose another two-step approach for the classification of biomedical terms. In the proposed approach, terms are first extracted using dictionary look-ups and C-*value* and NC-*value* scoring techniques. The extracted terms are then classified by help of *verb selectional patterns* and using a nearest neighbour and genetic algorithm. Likewise, Afzal et al. (2008) propose a two-step method; however, for the identification of terms that signal *bioinformatics services and tools* and using a different set of features and learning technique. A similar method and application can be found in Houngbo and Mercer (2012).

Although in the above-mentioned examples a term classification process follows a term recognition process to select a subset of valid terms, as suggested by Maynard and Ananiadou (2001), the recognition and classification process can be merged. In this way, the scoring process in the term recognition system is replaced by the scoring mechanism that is implemented by the classifier; hence, candidate terms can be directly assessed and classified by the term classifier system(e.g., see Foo and Merkel, 2010; Judea et al., 2014). The type of information that is employed during the classification is what makes these methodologies different from the entity recognisers. These methods are also different due to the type of the output that they generate. The entity recognisers mark the boundaries of terms mentioned in a given sentence or text snippet, whereas the term classifiers are often used to organise terms in a knowledge structure such as ontologies and thesaurus. As a result, term classification methods have been widely employed for learning, populating and extending domain ontologies (e.g., see Wong et al., 2012).

Lastly, a large number of methods proposed for automatic thesaurus construction are comparable to term classification tasks (e.g., see Navigli and Ponzetto, 2012). Whereas automatic thesaurus construction deals with the processing of concept hierarchies in general domain language, terminology classification methods deal with special corpora and sublanguages.

## 3.7 Evaluation Techniques

Evaluation of the majority of natural language processing systems has posed itself as a research challenge. Several factors can be named as a barrier to an objective evaluation of these systems (see Jones and Galliers, 1995, for a full depiction of these problems):

- disagreements on the basic concepts' definitions—for example, what is *semantics*?
- complexity of the tasks—for example, how to model a communication system? how to model users' background and psychological state? how to measure these factors and study their influences on the performance of a system?
- a large number of interdependent variables that play a role in the performance of a system;
- qualitative nature of the evaluation in a number of applications;
- multi-stage, intermediate, or different representations of the output;
- irreproducible evaluation situations and hence outputs;
- and, the absence of a common baseline on which to establish evaluations.

The most widely adopted framework for the evaluation of natural language processing tasks, including terminology mining methods, is the evaluation approach promoted in the series of message understanding conferences (MUC) for the assessment of information extraction systems. The MUC-style evaluation framework emphasises quantitative evaluations. This evaluation style accommodates a systematic reproducible assessment of the participating methods, which is methodologically clear and understandable. In this framework, the evaluation is carried out by comparing system-generated responses and hand-coded expected outputs (manual annotations) , which is expressed by a quantitative scoring measure. Figure 3.9 illustrates the evaluation's elements and procedure in this framework.

In an MUC-style evaluation, the most important building blocks are the manually annotated reference corpus[1] and the scoring measure. In the past decades, a number of research initiatives[2] and evaluation campaigns[3] have resulted in the development of a number reference corpora and datasets that are successfully employed for the development and evaluation of language processing techniques. Creating corpora for benchmarking terminology extraction techniques has been addressed in several research efforts, too.

The GENIA corpus is a well-known example of such reference datasets in bio-text mining: a corpus of 2000 abstracts from scientific publications in biological literature that is accompanied by the annotations of 100,000 terms organised in a well-defined ontology (Kim et al., 2003). The Colorado Richly Annotated Full Text Corpus (CRAFT) is another example of a bio-text mining dataset, which consists of 97 articles from the PubMed Central Open Access subset annotated with biomedical concepts such as *mouse*

[1] As evident, the development of the methods.

[2] For example, the Expert Advisory Group on Language Engineering Standard (The EAGLES Evaluation Working Group, 1996).

[3] For example, the series of automatic content extraction evaluation (see http://www.itl.nist.gov/iad/mig/tests/ace/), text analysis conference (http://www.nist.gov/tac/), as well as the series of workshops on semantic evaluation (http://aclanthology.info/venues/semeval).

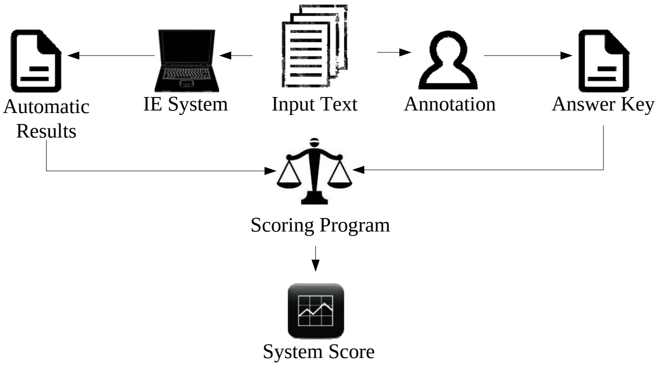


Figure 3.9: MUC-style evaluation for information extraction tasks (Lehnert et al., 1994).

*genes* (Bada et al., 2012). In a more recent effort, Bernier-Colborne and Drouin (2014) report on creating a corpus for the evaluation of term extraction in the domain of automotive engineering. Similarly, Zadeh and Handschuh (2014) introduce the ACL RD-TEC, a dataset of manually annotated terms in the domain of computational linguistics.

In quantitative evaluations, *precision* and *recall* are the two most widely-used scoring measures. Precision shows the ratio of the correct automatically generated results against all the information generated by the system. The correct automatically generated results are often those that *match* the answer keys provided through the manual annotation. Recall, however, measures the ratio of correct automatically generated information against all the available information in the reference corpus expected to be generated/extracted by the system. A combination of these measures such as $F$-score is used for scoring the systems. For an automatic term recognition (ATR) system, precision is the proportion of correct terms in the overall list of extracted candidate terms:

$$Precision = \frac{number\ of\ extracted\ valid\ terms}{number\ of\ candidate\ terms}. \tag{3.10}$$

Recall, on the other hand, is the proportion of extracted terms to the complete set of terms in the corpus:

$$Recall = \frac{number\ of\ extracted\ valid\ terms}{number\ of\ all\ valid\ terms\ in\ the\ corpus}. \tag{3.11}$$

And, usually but not necessarily, the $F$-score is given by

$$F_{measure} = \frac{2 * Recall * Precision}{Recall + Precision}. \tag{3.12}$$

The use of precision and recall is limited to the availability of manual annotations. In many real-world applications, manual annotations for all the system generated results are not available. For example, manual annotations are not available for all the candidate terms generated by an ATR system. In this case, precision thus cannot be computed; similarly, the complete set of expected information is not available. For example, the complete set of valid terms in a corpus, which must be extracted by an ATR system, is unknown; hence, recall cannot be computed. Moreover, in a number of use-cases, other quantitative aspects of the generated results are required—for example, the number of valid information items discovered by the system but not annotated/presented in the reference dataset (e.g., see the evaluation in Roark and Charniak, 1998).[1] In these situations, figures of merit other than precision and recall are employed.

In terminology extraction, one popular measure that often replaces precision and recall is *precision at n* (i.e., $P_{@n}$). Given a sorted list of $m$ candidate terms, precision at $n$, $n \leq m$, measures the precision (i.e., the number of valid terms $|v|$) in the list of top $n$ candidate term that are sorted by the scores assigned by an ATR system:

$$P_{@n} = \frac{|v|}{n}. \tag{3.13}$$

For example, $P_{@n}$ for $n = 10$ is the number of valid terms in the list of top 10 candidate terms sorted by their ATR-computed scores. It becomes evident that if a single number is used to summarise the performance, then the value of $n$ and $m$ can have a large impact on the computed performances. Hence, $P_{@n}$ is often replaced by an averaged precision.

Amongst techniques for obtaining an average of precision, *non-interpolated average precision* for $k$ valid terms ($NAP_k$) is often used to report the performance of methods as a single number (e.g., see Zhang et al., 2008; Fahmi, 2009, chap. 4). As suggested by Schone and Jurafsky (2001), $NAP_k$ is given by

$$NAP_k = \frac{1}{k} \sum_{i=1}^{k} P^i, \tag{3.14}$$

where $k$ is the number of valid terms that are targeted to be seen in the list of sorted candidate terms, and $P^i$ is the observed precision for pulling out $i$ valid terms. That is, $P^i = \frac{i}{|H_i|}$, in which $i$ is the number of valid terms, and $|H_i|$ is the number of candidate terms that are required to be checked in order to find this $i$ valid terms. Compared to $P_{@k}$, $NAP_k$ signify the distribution of valid terms in the extracted sorted lists of candidate terms. Depending on the evaluation context, one of these measures is usually used to show a method's performance.

### 3.7.1 Some Evaluation Caveats and Questions

Even with the availability of language resources, MUC-style quantitative evaluation framework cannot always replace qualitative assessments. For instance, depending on the design principles adopted for the development of reference corpora, quantitative evaluations may not provide proper perspective on the scalability and portability of the systems participating in an evaluation. In addition, as suggested by Lehnert et al. (1994), this quantitative assessment cannot be used to assess the time and effort that is required to develop these systems. Therefore, in a number of occasions, qualitative assessments may still be required for a comprehensive evaluation.[2]

A number of critics draw attention to the way the output of a system matches the provided answer keys in the manual annotations. For example, Esuli and Sebastiani

[1] One controversy here is that while the answer keys cannot be used, how to decide whether an information item is valid.

[2] This can also be discussed in the context of black box vs. glass box evaluation frameworks.

(2010) suggest that the evaluation of an extraction method can be enhanced by permitting the notion of *true negative*, incorporating a measure that is sensitive to the *degree* of overlap between the correct expected answers and the outputs of the extraction system, and allowing for multiple correct output. Other researchers go further and question the basis in which some of the measures such as precision and recall are employed in evaluation scenarios. For instance, Cowie and Wilks (2000) suggest that precision and recall are designed for information retrieval tasks; hence, they are not appropriate for the evaluation of a number of information extraction tasks. For example, in a multi-slot template filling task, counting correct results can produce some paradoxical outcomes and attention should be paid to the details of how performance scores are calculated.

Lavelli et al. (2008) address the evaluation of machine learning-based information extraction systems and the assessment of the ability of these algorithms to *learn*. Besides the factors discussed above, the authors argue that establishing an evaluation methodology and the availability of gold standard corpora do not guarantee a reliable comparison between different approaches and algorithms. Lavelli et al. suggest that considering the influential variables in the overall performance of such systems, for example, the number of *features* and setting of algorithm-specific parameters, is beneficial for a meaningful comparison of learning methods.

To avoid a number of barriers to an objective evaluation of information extraction systems, apart from the *intrinsic* MUC-style evaluations, *extrinsic* or indirect evaluation has been suggested. Extrinsic evaluations measure the quality of the output of a method by assessing the performance of a third system that employs the generated output. For example, a common method of extrinsic evaluation for an information extraction system is to utilise its output in a document classification problem and assess the extraction task by studying the precision and recall of the classification task (Yangarber et al., 2000).

As suggested earlier in this section, as with other information extraction tasks, the evaluation of terminology mining methods is often carried out by comparing the output of a term extractor against a gold standard dataset, manually checking the output of the method with the help of a terminologist/a domain-expert, or an extrinsic evaluation such as the one suggested in Kit et al. (2008).

A number of concerns in the evaluation of terminology mining methods is similar to those that are listed for other information extraction systems. For instance, the evaluation of perfect and imperfect recognition has been one of the concerns in the evaluation of ATR systems (e.g., see Lauriston, 1995). Maynard et al. (2008) suggest that in modern applications, for example, ontology learning, performance metrics such as precision and recall are not sufficient since they provide a binary decision of correctness—that is, a term is either right or wrong and nothing in between. Therefore, they suggest the use of matching techniques that acknowledge partial correctness such as using edit distance as employed in the *balanced distance metric* by Maynard (2005) and the *SOLD* measure by Spasic and Ananiadou (2005).

However, the complexity of the evaluation of terminology mining methods goes beyond the common problems such as partial matching. As is rightly argued by Vivaldi and Rodríguez (2007), in short, the evaluation of these methods inherits its complexity from the definition of terms. In order to have an overall evaluation of terminologies, Vivaldi and Rodríguez suggest that three dimensions of terms' characteristics, namely, unithood, termhood, and their specialised usage, must first be assessed and then combined. This multi-faceted characteristic of terms often makes it hard to find an objective judgement when preparing reference corpora, annotating terms, and preparing an evaluation framework.

Lastly, assuming that all the terms in a corpus are annotated with high confidence, do all these terms have the same importance in domain-knowledge? Is it ever possible to introduce a measure to quantify their importance objectively? These are all questions that still must be addressed in an ideal evaluation framework of terminology mining.

## 3.8 Summary

In this chapter, terminology extraction methods are reviewed, in the application context of this thesis in which the use of distributional models will be investigated. The discussion started with the definition of the term *term* to highlight the complexity of terminology mining methods; the wide-range of task that it embraces; and, the wide spectrum of problems that it encounters.

In Section 3.2, the general two-step mechanism of a typical terminology mining method is discussed. In Section 3.3, a review of candidate term extraction techniques was provided, followed by a study of term scoring methods in Section 3.4. Organising terminologies was discussed briefly in Section 3.5. This discussion was followed by an introduction to term classification techniques often used to form co-hyponym groups in Section 3.6. Finally, this chapter concluded with a brief study of the common practices for the evaluation of terminology mining methods.

The presented study in this chapter set the background for the proposed co-hyponym term extraction method in Chapter 5. However, it is worth mentioning that it only scratches the surface of the vast amount of ongoing research in computational terminology.

# Reference List

Abney, S. (1992). Parsing by chunks. In *Principle-Based Parsing*, Studies in Linguistics and Philosophy, pages 257–278. Kluwer Academic Publishers. 77

Afzal, H., Stevens, R., and Nenadic, G. (2008). Towards semantic annotation of bioinformatics services: Building a controlled vocabulary. In Salakoski, T., Schuhmann, D. R., and Pyysalo, S., editors, *Proceedings of the Third International Symposium on Semantic Mining in Biomedicine (SMBM 2008)*, pages 5–12, Turku, Finland. Turku Centre for Computer Science (TUCS). 93

Ahmad, K., Gillam, L., and Tostevin, L. (1999). University of Surrey participation in TREC 8: Weirdness indexing for logical document extrapolation and retrieval (WILDER). In Voorhees, E. M. and Harman, D. K., editors, *NIST Special Publication 500-246: The Eighth Text REtrieval Conference (TREC-8)*, pages 717–724. Department of Commerce, National Institute of Standards and Technology. 86

Ananiadou, S. (1994). A methodology for automatic term recognition. In *COLING 94: The 15th Conference on Computational Linguistics: Proceedings*, volume 2, pages 1034–1038, Kyoto, Japan. Association for Computational Linguistics. 69, 74, 75, 88

Andersson, L., Lupu, M., Palotti, J. R. M., Piroi, F., Hanbury, A., and Rauber, A. (2014). Insight to hyponymy lexical relation extraction in the patent genre versus other text genres. In Jung, H., Mandl, T., Womser-Hacker, C., and Xu, S., editors, *Proceedings of the First International Workshop on Patent Mining and its Applications (IPaMin 2014)*, volume 1292, Hildesheim, Germany. CEUR Workshop Proceedings. 91

Anick, P., Verhagen, M., and Pustejovsky, J. (2014). Extracting aspects and polarity from patents. In Meyers, A., He, Y., and Grishman, R., editors, *Proceedings of the COLING Workshop on Synchronic and Diachronic Approaches to Analyzing Technical Language (SADAATL 2014)*. Association for Computational Linguistics and Dublin City University. 79, 91

Aubin, S. and Hamon, T. (2006). Improving term extraction with terminological resources. In Salakoski, T., Ginter, F., Pyysalo, S., and Pahikkala, T., editors, *Advances in Natural Language Processing*, volume 4139 of *Lecture Notes in Computer Science*, pages 380–387. Springer Berlin Heidelberg. 76, 82

Bada, M., Eckert, M., Evans, D., Garcia, K., Shipley, K., Sitnikov, D., Baumgartner, W., Cohen, K., Verspoor, K., Blake, J., and Hunter, L. (2012). Concept annotation in the CRAFT corpus. *BMC Bioinformatics*, 13(1):161. 95

Baldwin, T. and Kim, S. N. (2010). Multiword expressions. In Indurkhya, N. and Damerau, F. J., editors, *Handbook of Natural Language Processing*, pages 267–292. CRC Press, Taylor and Francis Group, second edition. ISBN 978-1420085921. 74

Barrón-Cedeño, A., Sierra, G., and Ananiadou, P. D. S. (2009). An improved automatic term recognition method for Spanish. In Gelbukh, A., editor, *Computational Linguistics and Intelligent Text Processing*, volume 5449 of *Lecture Notes in Computer Science*, pages 125–136. Springer Berlin Heidelberg. 79

Basili, R., Moschitti, A., Pazienza, M. T., and Zanzotto, F. M. (2001). A contrastive approach to term extraction. In *TIA 2001: Terminologie et Intelligence Artificielle*, pages 119–128, Nancy, France. INIST-CNRS. 81

Baxendale, P. B. (1958). Machine-made index for technical literature – an experiment. *IBM Journal of Research and Development*, 2(4):354–361. 82

Bernier-Colborne, G. and Drouin, P. (2014). Creating a test corpus for term extractors through term annotation. *Terminology*, 20(1):50–73. 95

Bertels, A. and Speelman, D. (2014). Clustering for semantic purposess: Exploration of semantic similarity in a technical corpus. *Terminology: International Journal of Theoretical and Applied Issues in Specialized Communication*, 20(2):279–303. 90

Bonin, F., Dell'Orletta, F., Montemagni, S., and Venturi, G. (2010a). A contrastive approach to multi-word extraction from domain-specific corpora. In Calzolari, N., Choukri, K., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., Rosner, M., and Tapias, D., editors, *Proceedings of the Seventh conference on International Language Resources and Evaluation*, pages 3222–3229, Valletta, Malta. European Language Resources Association. 79

Bonin, F., Dell'Orletta, F., Venturi, G., and Montemagni, S. (2010b). Contrastive filtering of domain-specific multi-word terms from different types of corpora. In *Proceedings of the 2010 Workshop on Multiword Expressions: from Theory to Applications (MWE 2010)*, pages 77–80, Beijing, China. Coling 2010 Organizing Committee. 82

Bourigault, D. (1992). Surface grammatical analysis for the extraction of terminological noun phrases. In *COLING 1992 Volume 3: The 15th International Conference on Computational Linguistics*, pages 977–981, Nantes, France. International Committee on Computational Linguistics. 77

Brunzel, M. (2008). The XTREEM methods for ontology learning from Web documents. In Buitelaar, P. and Cimiano, P., editors, *Ontology Learning and Population: Bridging the Gap between Text and Knowledge*, volume 167 of *Frontiers in Artificial*

*Intelligence and Applications*, pages 3–26. IOS Press, Amsterdam, The Netherlands. 81

Buitelaar, P., Cimiano, P., Haase, P., and Sintek, M. (2009). Towards linguistically grounded ontologies. In Aroyo, L., Traverso, P., Ciravegna, F., Cimiano, P., Heath, T., Hyvonen, E., Mizoguchi, R., Oren, E., Sabou, M., and Simperl, E., editors, *The Semantic Web: Research and Applications*, volume 5554 of *Lecture Notes in Computer Science*, pages 111–125. Springer Berlin Heidelberg. 88

Cabré, M. T. (1999). *Terminology: Theory, Methods and Applications*. John Benjamins. 70

Cabré, M. T. (2003). Theories of terminology their description, prescription and explanation. *Terminology*, 9(2):163–199. 68, 69, 70

Cabré, M. T. (2010). Terminology and translation. In dins Gambier, Y. and Van Doorslaer, L., editors, *Handbook of translation studies*, volume 1, pages 356–365. John Benjamins Publishing Company. 68

Cabré, M. T., Condamines, A., and Ibekwe-SanJuan, F. (2007). Introduction: Application-driven terminology engineering. In *Application-Driven Terminology Engineering*, volume vii, pages 1–19. John Benjamins. 72

Campo, Á. (2013). *The reception of Eugen Wüster's work and the development of terminology*. PhD thesis, Université de Montréal. 69

Chakraborty, S., Subramanian, L., and Nyarko, Y. (2014). Extraction of (key,value) pairs from unstructured ads. In *AAAI Fall Symposium Serie*, pages 10–17, Arlington, Virginia. AAAI Press. 91

Church, K. W. and Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22–29. 84

Cowie, J. and Wilks, Y. (2000). Information extraction. In Dale, R., Moisl, H., and Somers, H., editors, *Handbook of Natural Language Processing*. New York: Marcel Dekker. 97

Daille, B. (1995). Combined approach for terminology extraction: Lexical statistics and linguistic filtering. In Wilson, A. and McEnery, T., editors, *UCREL Technical Papers*. Lancaster University. 78, 79, 84, 85

Dias, G. and Kaalep, H.-J. (2003). Automatic extraction of multiword units for Estonian: Phrasal verbs. In Metslang, H. and Rannut, M., editors, *Languages in Development*, volume 41 of *Linguistics Edition*, pages 81–90. LINCOM. 84

Dinu, A., Dinu, L., and Sorodoc, I. (2014). Aggregation methods for efficient collocation detection. In Calzolari, N., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Ninth*

*International Conference on Language Resources and Evaluation*, pages 4041–4045, Reykjavik, Iceland. European Language Resources Association. 85

Dorji, T. C., sayed Atlam, E., Yata, S., Fuketa, M., Morita, K., and ichi Aoe, J. (2011). Extraction, selection and ranking of field association (FA) terms from domain-specific corpora for building a comprehensive FA terms dictionary. *Knowledge and Information Systems*, 27(1):141–161. 78, 79

Dredze, M., McNamee, P., Rao, D., Gerber, A., and Finin, T. (2010). Entity disambiguation for knowledge base population. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 277–285, Beijing, China. Coling 2010 Organizing Committee. 91

Drouin, P. (2003). Term extraction using non-technical corpora as a point of leverage. *Terminology*, 9(1):99–115. 81

Drouin, P. (2004). Detection of domain specific terminology using corpora comparison. In Lino, M. T., Xavier, M. F., Ferreira, F., Costa, R., Silva, R., Pereira, C., Carvalho, F., Lopes, M., Catarino, M., and Barros, S., editors, *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, pages 79–82, Lisbon, Portugal. European Language Resources Association. 81, 86

Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1):61–74. 84, 85

Dupuch, M., Dupuch, L., Hamon, T., and Grabar, N. (2014). Exploitation of semantic methods to cluster pharmacovigilance terms. *Journal of Biomedical Semantics*, 5(18). 73, 90

Dupuch, M., Hamo, T., Dupuch, L., and Grabar, N. (2012). Semantic distance and terminology structuring methods for the detection of semantically close terms. In *Proceedings of the 2012 Workshop on Biomedical Natural Language Processing (BioNLP 2012)*, pages 20–28, Montreal, Canda. Association for Computational Linguistics. 90

Eck, N. J. V., Waltman, L., Noyons, E. C., and Buter, R. K. (2010). Automatic term identification for bibliometric mapping. *Scientometrics*, 82:581–596. 79

Ellis, J., Li, X., Griffitt, K., Strassel, S. M., and Wright, J. (2012). Linguistic resources for 2012 knowledge base population evaluations. In *Proceedings of the Fifth Text Analysis Conference (TAC 2012)*, Maryland, USA. National Institute of Standards and Technology. 91

Esuli, A. and Sebastiani, F. (2010). Evaluating information extraction. In Agosti, M., Ferro, N., Peters, C., Rijke, M., and Smeaton, A., editors, *Multilingual and Multimodal Information Access Evaluation*, volume 6360 of *Lecture Notes in Computer Science*, pages 100–111. Springer Berlin Heidelberg. 96

Evans, D. A. and Zhai, C. (1996). Noun-phrase analysis in unrestricted text for information retrieval. In *34th annual meeting on Association for Computational Linguistics: Proceedings of the Conference*, pages 17–24, California, USA. Association for Computational Linguistics. 80

Evert, S. (2004). *The Statistics of Word Cooccurrences Word Pairs and Collocations*. PhD thesis, Institut für maschinelle Sprachverarbeitung, Universität Stuttgart. 71, 83

Evert, S. (2009). Corpora and collocations. In Lüdeling, A. and Kytö, M., editors, *Corpus Linguistics: An International Handbook*, volume 2 of *Handbooks of Linguistics and Communication Science*. Mouton de Gruyter. 85

Faber, P. and L'Homme, M.-C. (2014). Lexical semantic approaches to terminology: An introduction. *Terminology*, 20(2):143–150. 69, 70, 88

Faber, P. and Rodríguez, C. I. L. (2012). Terminology and specialized language. In Faber, P., editor, *A Cognitive Linguistics View of Terminology and Specialized Language*, volume 20 of *Applications of Cognitive Linguistics*, pages 9–33. Walter de Gruyter. 68

Fahmi, I. (2009). *Automatic term and relation extraction for medical question answering system*. PhD thesis, University Library Groningen. 96

Fan, T.-K. and Chang, C.-H. (2008). Exploring evolutionary technical trends from academic research papers. In Kise, K. and Sako, H., editors, *DAS 2008: Proceedings of the 8th IAPR International Workshop on Document Analysis Systems*, pages 574–581, Nara, Japan. IEEE Computer Society. 80

Feiyu, X., Kurz, D., Piskorski, J., and Schmeier, S. (2002). A domain adaptive approach to automatic acquisition of domain relevant terms and their relations with bootstrapping. In *Proceedings of the 3rd International Conference on Language Resources an Evaluation*, number 224–230, Las Palmas, Canary Islands, Spain. European Language Resources Association. 84

Felber, H. (1982). Computerized terminology in Termnet: The role of terminological data banks. In *Term banks for tomorrow's world: Translating and the Computer 4*, pages 8–20, London, UK. Aslib. Conference Proceedings. 69

Fiszman, M., Rosemblat, G., Ahlers, C. B., and Rindflesch, T. C. (2007). Identifying risk factors for metabolic syndrome in biomedical text. In *AMIA Annual Symposium Proceedings*, volume 2007, page 249. American Medical Informatics Association. 81

Fodor, J. and Lepore, E. (2012). What sort of science is semantics? In Peter, G. and KrauSSe, R.-M., editors, *Selbstbeobachtung der modernen Gesellschaft und die neuen Grenzen des Sozialen*, pages 217–226. Springer Fachmedien Wiesbaden. 71

Foo, J. and Merkel, M. (2010). Using machine learning to perform automatic term recognition. In *Proceedings of the LREC 2010 Workshop on Methods for Automatic Acquisition of Language Resources and their Evaluation Methods*, pages 49–54. 93

Frantzi, K. T. (1997). Incorporating context information for the extraction of terms. In *Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics*, EACL '97, pages 501–503, Stroudsburg, PA, USA. Association for Computational Linguistics. 78, 79

Frantzi, K. T. and Ananiadou, S. (1996). Extracting nested collocations. In *COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics*, pages 41–46, Copenhagen, Denmark. Association for Computational Linguistics. 84, 86

Frantzi, K. T., Ananiadou, S., and Mima, H. (2000a). Automatic recognition of multiword terms: The C-value/NC-value method. *International Journal on Digital Libraries*, 3(2):115–130. 79

Frantzi, K. T., Ananiadou, S., and Tsujii, J. (2000b). The C-value/NC-value method of automatic recognition for multi-word terms. In *Research and Advanced Technology for Digital Libraries*, volume 3 of *Lecture Notes in Computer Science*, pages 585–604. Springer Berlin Heidelberg. 87

Fukuda, K., Tsunoda, T., Tamura, A., and Takagi, T. (1998). Toward information extraction: Identifying protein names from biological papers. In *Pacific Symposium on Biocomputing*, volume 3, pages 707–718. 90

Gooch, P. and Roudsari, A. V. (2011). Automated recognition and post-coordination of complex clinical terms. In Borycki, E. M., Bartle-Clar, J. A., Househ, M. S., Kuziemsky, C. E., and Schraa, E. G., editors, *International Perspectives in Health Informatics*, Studies in Health Technology and Informatics, pages 8–12. IOS Press. 81

Hamon, T., Engström, C., and Silvestrov, S. (2014). Term ranking adaptation to the domain: Genetic algorithm-based optimisation of the C-value. In Przepiórkowski, A. and Ogrodniczuk, M., editors, *Advances in Natural Language Processing*, volume 8686 of *Lecture Notes in Computer Science*, pages 71–83. Springer International Publishing. 92

Hartmann, S., Szarvas, G., and Gurevych, I. (2011). Mining multiword terms from Wikipedia. In Pazienza, M. T. and Stellato, A., editors, *Semi-Automatic Ontology Development: Processes and Resources*, pages 226–258. IGI Global, Hershey, PA, USA. 81

Hazen, R., Esbroeck, A., Mongkolwat, P., and Channin, D. (2011). Automatic extraction of concepts to extend RadLex. *Journal of Digital Imaging*, 24:165–169. 81

Heid, U. and Gojun, A. (2012). Term candidate extraction for terminography and CAT: An overview of TTC. In Fjeld, R. V. and Torjusen, J. M., editors, *Proceedings of the*

*15th Euralex International Congress*, pages 585–594, University of Oslo, Norway. 70

Hippisley, A., Cheng, D., and Ahmad, K. (2005). The head-modifier principle and multilingual term extraction. *Natural Language Engineering*, 11(2):129–157. 80

Hoang, H. H., Kim, S. N., and Kan, M.-Y. (2009). A re-examination of lexical association measures. In *MWE 2009: Proceedings of the 2009 Workshop on Multiword Expressions: Identification, Interpretation, Disambiguation and Applications*, pages 31–39, Suntec, Singapore. The Association for Computational Linguistics and The Asian Federation of Natural Language Processing. 83

Houngbo, H. and Mercer, E. R. (2012). Method mention extraction from scientific research papers. In Kay, M. and Boite, C., editors, *Proceedings of COLING 2012: Technical Papers*, pages 1211–1222, Mumbai, India. The COLING 2012 Organizing Committee. 93

Hsu, L.-F. (2010). Mining on terms extraction from Web news. In Pan, J.-S., Chen, S.-M., and Nguyen, N., editors, *Computational Collective Intelligence. Technologies and Applications*, volume 6421 of *Lecture Notes in Computer Science*, pages 188–194. Springer Berlin Heidelberg, Kaohsiung, Taiwan. 79

Hulth, A. (2003). Improved automatic keyword extraction given more linguistic knowledge. In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*, pages 216–223, Sapporo, Japan. Association for Computational Linguistics. 76, 79, 80, 82

International Organization for Standardization (2000). ISO 1087-1:2000(en) terminology — vocabulary — part 1: Theory and application.

Ittoo, A. and Bouma, G. (2013). Term extraction from sparse, ungrammatical domain-specific documents. *Expert Systems with Applications*, 40(7):2530–2540. 86

Ittoo, A., Maruster, L., Wortmann, H., and Bouma, G. (2010). Textractor: A framework for extracting relevant domain concepts from irregular corporate textual datasets. In Abramowicz, W. and Tolksdorf, R., editors, *Business Information Systems*, volume 47 of *Lecture Notes in Business Information Processing*, pages 71–82. Springer. 78

Jacquemin, C. and Tzoukermann, E. (1999). NLP for term variant extraction: Synergy between morphology, lexicon, and syntax. In Strzalkowski, T., editor, *Natural Language Information Retrieval*, volume 7 of *Text, Speech and Language Technology*, pages 25–74. Springer Netherlands. 80

Jakubíček, M., Kilgarriff, A., Kovář, V., Rychlý, P., and Suchomel, V. (2014). Finding terms in corpora for many languages with the Sketch Engine. In *Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association* *for Computational Linguistics*, pages 53–56, Gothenburg, Sweden. Association for Computational Linguistics. 80

Jones, K. S. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28(1):11–21. 84

Jones, K. S. and Galliers, J. R. (1995). *Evaluating Natural Language Processing Systems: An Analysis and Review*, volume 1083 of *Lecture Notes in Artificial Intelligence*. Springer-Verlag Berlin Heidelberg, Secaucus, NJ, USA, 1 edition. 94

Judea, A., Schütze, H., and Bruegmann, S. (2014). Unsupervised training set generation for automatic acquisition of technical terminology in patents. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 290–300, Dublin, Ireland. Dublin City University and Association for Computational Linguistics. 93

Justeson, J. S. and Katz, S. M. (1995). Technical terminology: Some linguistic properties and an algorithm for identification in text. *Natural Language Engineering*, 1(1):9–27. 77

Kageura, K. (1999). On the study of dynamics of terminology: A proposal of a theoretical framework. *Research Bulletin of the NACSIS*, 11:1–10. 68

Kageura, K. and Umino, B. (1996). Methods of automatic term recognition: A review. *Terminology*, 3(2):259–289. 71, 82, 83, 84

Khoo, C. S. and Na, J.-C. (2006). Semantic relations in information science. *Annual Review of Information Science and Technology*, 40(1):157–228. 88

Kilgarriff, A. (1996). Which words are particularly characteristic of a text? A survey of statistical approaches. Technical Report ITRI-96-08, University of Brighton, Brighton, UK. 85

Kilgarriff, A. (1997). “I don’t believe in word senses”. *Computers and the Humanities*, 31(2):91–113. 89

Kilgarriff, A. (2001). Comparing corpora. *International journal of Corpus Linguistics*, 6(1):97–133. 86

Kilgarriff, A. (2005). Language is never, ever, ever, random. *Corpus Linguistics and Linguistic Theory*, 1(2):1613–7027. 85

Kim, J.-D., Ohta, T., Tateisi, Y., and Tsujii, J. (2003). GENIA corpus—a semantically annotated corpus for bio-textmining. *Bioinformatics*, 19(Suppl 1):i180–i182. 92, 94

Kim, J.-D., Ohta, T., Tsuruoka, Y., Tateisi, Y., and Collier, N. (2004). Introduction to the bio-entity recognition task at JNLPBA. In Collier, N., Ruch, P., and Nazarenko, A., editors, *JNLPBA: Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and Its Applications*, pages 70–75, Geneva, Switzerland. Association for Computational Linguistics. 90, 92

Kit, Chunyu, and Liu, X. (2008). Measuring mono-word termhood by rank difference via corpus comparison. *Terminology*, 14(2):204–229. 97

Knoth, P., Schmidt, M., Smrz, P., and Zdrahal, Z. (2009). Towards a framework for comparing automatic term recognition methods. In *Conference Znalosti 2009*, Brno, Czech Republic. 86

Korkontzelos, I., Klapaftis, I. P., and Manandhar, S. (2008). Reviewing and evaluating automatic term recognition techniques. In Nordström, B. and Ranta, A., editors, *Advances in Natural Language Processing*, volume 5221 of *Lecture Notes in Computer Science*, pages 248–259. Springer Berlin Heidelberg. 85

Kovačević, A., Konjović, Z., Milosavljević, B., and Nenadic, G. (2012). Mining methodologies from NLP publications: A case study in automatic terminology recognition. *Computer Speech and Language*, 26(2):105–126. 93

Krauthammer, M. and Nenadic, G. (2004). Term identification in the biomedical literature. *Journal of Biomedical Informatics*, 37(6):512–526. 73

Laurence, S. and Margolis, E. (1999). Concepts and cognitive science. In Laurence, S. and Margolis, E., editors, *Concepts: Core Readings*, pages 3–81. MIT Press Cambridge, MA. 71

Lauriston, A. (1995). Criteria for measuring term recognition. In *Proceedings of the Seventh Conference on European Chapter of the Association for Computational Linguistics*, pages 17–22, Dublin, Ireland. Association for Computational Linguistics. 97

Lavelli, A., Califf, M., Ciravegna, F., Freitag, D., Giuliano, C., Kushmerick, N., Romano, L., and Ireson, N. (2008). Evaluation of machine learning-based information extraction algorithms: Criticisms and recommendations. *Language Resources and Evaluation*, 42(4):361–393. 97

Lehnert, W., Cardie, C., Fisher, D., McCarthy, J., Riloff, E., and Soderland, S. (1994). Evaluating an information extraction system. *Journal of Integrated Computer-Aided Engineering*, 1(6). 95, 96

L'Homme, M.-C. (2014). Terminologies and taxonomies. In Taylor, J. R., editor, *The Oxford Handbook of the Word*. Oxford University Press. 68, 73, 89

L'Homme, M.-C. and Bernier-Colborne, G. (2012). Terms as labels for concepts, terms as lexical units: A comparative analysis in ontologies and specialized dictionaries. *Applied Ontology*, 7(4):387–400. 89

Liu, X. and Kit, C. (2008). An improved corpus comparison approach to domain specific term recognition. In *Proceedings of the 22nd Pacific Asia Conference on Language, Information and Computation: PACLIC 22*, pages 253–261, Cebu City, Philippines. De La Salle University. 86

Manning, C. D. and Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. The MIT Press. 84

Mayfield, J., McNamee, P., Harmon, C., Finin, T., and Lawrie, D. (2014). KELVIN: Extracting knowledge from large text collections. In *Natural Language Access to Big Data: Papers from the 2014 AAAI Fall Symposium*, pages 34–41, Arlington, Virginia. AAAI Press. 91

Maynard, D. (2005). Benchmarking ontology-based annotation tools for the Semantic Web. In Cox, S., editor, *Proceedings of the UK e-Science All Hands Conference*, Nottingham, UK. Engineering and Physical Sciences Research Council. 97

Maynard, D. and Ananiadou, S. (2000). Identifying terms by their family and friends. In *Proceedings of the 18th International Conference on Computational Linguistics*, pages 530–536, Saarbrucken, Germany. Association for Computational Linguistics. 87, 88

Maynard, D. and Ananiadou, S. (2001). Term extraction using a similarity-based approach. In Bourigault, D., Jacquemin, C., and L'Homme, M.-C., editors, *Recent Advances in Computational Terminology*, volume xviii, pages 261–278. John Benjamins. 76, 93

Maynard, D., Li, Y., and Peters, W. (2008). NLP techniques for term extraction and ontology population. In Buitelaar, P. and Cimiano, P., editors, *Ontology Learning and Population: Bridging the Gap between Text and Knowledge*, volume 167 of *Frontiers in Artificial Intelligence and Applications*, pages 3–26. IOS Press, Amsterdam, The Netherlands. 90, 91, 97

Maynard, D. G. (2000). *Term recognition using combined knowledge sources*. PhD thesis, Manchester Metropolitan University. 78, 84

Meyers, A., Glass, Z., Grieve-Smith, A., He, Y., Liao, S., and Grishman, R. (2014). Jargon-term extraction by chunking. In Meyers, A., He, Y., and Grishman, R., editors, *Proceedings of the COLING Workshop on Synchronic and Diachronic Approaches to Analyzing Technical Language (SADAATL 2014)*, pages 11–20. Association for Computational Linguistics and Dublin City University. 77

Miller, G. A. (1995). WordNet: A lexical database for English. *Communications of the ACM*, 38(11):39–41. 88

Mohit, B. (2014). Named entity recognition. In Zitouni, I., editor, *Natural Language Processing of Semitic Languages*, Theory and Applications of Natural Language Processing, pages 221–245. Springer. 92

Nadeau, D. and Sekine, S. (2007). A survey of named entity recognition and classification. *Special issue of Linguisticae Investigationes*, 30(1):3–26. 92

Nakagawa, H. (2001a). Automatic term recognition based on statistics of compound nouns. *Terminology*, 6(2):195–210. 74, 80, 84

Nakagawa, H. (2001b). Disambiguation of single noun translations extracted from bilingual comparable corpora. *Terminology*, 7(1):63–83. 79

Navigli, R. and Ponzetto, S. P. (2012). BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250. 93

Nigel, C. N., Collier, N., and Tsujii, J. (1999). Automatic term identification and classification in biology texts. In *Proceedings of the 5th Natural Language Pacific Rim Symposium (NLPRS'99)*, pages 369–374, Beijing, China. 73, 90, 93

Park, Y., Byrd, R. J., and Boguraev, B. K. (2002). Automatic glossary extraction: Beyond terminology identification. In Tseng, S.-C., editor, *COLING 2002: Proceedings of the 19th International Conference on Computational Linguistics*, Taipei, Taiwan. Association for Computational Linguistics and Chinese Language Processing. 77

Pearson, J. (1998). *Terms in Context*, volume 1 of *Studies in Corpus Linguistics*. John Benjamins, Amsterdam, The Netherlands. 87

Pecina, P. (2008). A machine learning approach to multiword expression extraction. In *Proceedings of the LREC Workshop: Towards a Shared Task for Multiword Expressions (MWE 2008)*, pages 54–57, Marrakech, Morocco. European Language Resources Association. 91

Pecina, P. (2010). Lexical association measures and collocation extraction. *Language Resources and Evaluation*, 44(1-2):137–158. 83

Pinnis, M., Ljubešić, N., Ştefănescu, D., Skadiņa, I., Tadić, M., and Gornostay, T. (2012). Term extraction, tagging, and mapping tools for under-resourced languages. In de CeaMari Carmen Suarez-Figueroa Raul Garcia-Castro Elena Montiel-Ponsoda, G. A., editor, *Proceedings of the 10th Conference on Terminology and Knowledge Engineering: New frontiers in the constructive symbiosis of terminology and knowledge engineering*, pages 193–208, Spain, Madrid. 76

Pustejovsky, J., Anick, P., and Bergler, S. (1993). Lexical semantic techniques for corpus analysis. *Computational Linguistics*, 19(2):331–358. 88

QasemiZadeh, B. (2015). *Investigating the Use of Distributional Semantic Models for Co-Hyponym Identification in Special Corpora*. PhD thesis, National University of Ireland, Galway. i

QasemiZadeh, B., Buitelaar, P., Chen, T. Q., and Bordea, G. (2012). Semi-supervised technical term tagging with minimal user feedback. In Calzolari, N., Choukri, K., Declerck, T., Doğan, M. U., Maegaard, B., Mariani, J., Odijk, J., and Piperidis, S., editors, *Proceedings of the Eighth International Conference on Language Resources* *and Evaluation (LREC-2012)*, pages 617–621, Istanbul, Turkey. European Language Resources Association (ELRA). 93

Rao, D., McNamee, P., and Dredze, M. (2013). Entity linking: Finding extracted entities in a knowledge base. In Poibeau, T., Saggion, H., Piskorski, J., and Yangarber, R., editors, *Multi-source, Multilingual Information Extraction and Summarization*, Theory and Applications of Natural Language Processing, pages 93–115. Springer Berlin Heidelberg. 73

Rayson, P., Berridge, D., and Francis, B. (2004). Extending the Cochran rule for the comparison of word frequencies between corpora. In G., P., C., F., and A., D., editors, *Proceedings of the 7th International Conference on Statistical Analysis of Textual Data (JADT 2004)*, volume II, pages 926–936, Louvain-la-Neuve, Belgium. Presses universitaires de Louvain. 85

Rayson, P. and Garside, R. (2000). Comparing corpora using frequency profiling. In *WCC '00: Proceedings of the Workshop on Comparing Corpora*, volume 9, pages 1–6, Hong Kong. Association for Computational Linguistics. 86

Rindflesch, T. C., Rajan, J. V., and Hunter, L. (2000). Extracting molecular binding relationships from biomedical text. In *6th Conference on Applied Natural Language Processing: Proceedings of the Conference (ANLP-2000)*, pages 188–195, Seattle, Washington. Association for Computational Linguistics. 81

Roark, B. and Charniak, E. (1998). Noun-phrase co-occurence statistics for semi-automatic semantic lexicon construction. In *COLING-ACL'98: 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics: Proceedings of the Conference*, volume 2, pages 1110–1116, Montréal, Quebec, Canada. Morgan Kaufmann Publishers. 96

Sager, J. C. (1990). *A practical course in terminology processing*. John Benjamins Publishing. 68, 71

Salton, G. (1992). The state of retrieval system evaluation. *Information processing & management*, 28(4):441–449. 84

Schone, P. and Jurafsky, D. (2001). Is knowledge-free induction of multiword unit dictionary headwords a solved problem? In *Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing*, Pittsburgh, PA USA. Association for Computational Linguistics. 96

Seretan, V., Nerima, L., and Wehrli, E. (2004). A tool for multi-word collocation extraction and visualization in multilingual corpora. In Williams, G. and Vessier, S., editors, *Proceedings of the Eleventh EURALEX International Congress, EURALEX 2004*, volume II, pages 755–766, Lorient, France. Universite de Bretagne. 81

Sinclair, J. (1996). Preliminary recommendations on corpus typology. Technical Report EAG–TCWG–CTYP/P, Expert Advisory Group on Language Engineering Standards (EAGLES). 68

Spasić, I. and Ananiadou, S. (2004). Using automatically learnt verb selectional preferences for classification of biomedical terms. *Journal of Biomedical Informatics*, 37(6):483–497. 92, 93

Spasic, I. and Ananiadou, S. (2005). A flexible measure of contextual similarity for biomedical terms. In Altman, R. B., Dunker, A. K., Hunter, L., Jung, T. A., and Klein, T. E., editors, *Pacific Symposium on Biocomputing 2005*, pages 197–208, Hawaii, USA. World Scientific. 97

Tanev, H. and Magnini, B. (2008). Weakly supervised approaches for ontology population. In Buitelaar, P. and Cimiano, P., editors, *Ontology Learning and Population: Bridging the Gap Between Text and Knowledge*, volume 167 of *Frontiers in Artificial Intelligence and Applications*, pages 129–143. IOS Press, Amsterdam, The Netherlands. 91

Taylor, A., Marcus, M., and Santorini, B. (2003). The Penn treebank: An overview. In Abeillé, A., editor, *Treebanks*, volume 20 of *Text, Speech and Language Technology*, pages 5–22. Springer Netherlands. 77

Term (2014). In *Oxford Dictionary of English*. Oxford University Press. Retrieved June 20, 2015, from from http://www.oxforddictionaries.com/definition/english/term. 68

The EAGLES Evaluation Working Group (1996). Evaluation of natural language processing systems. FINAL REPORT EAGLES DOCUMENT EAG-EWG-PR.2, Expert Advisory Group on Language Engineering. 94

Toral, A. and Munoz, R. (2006). A proposal to automatically build and maintain gazetteers for named entity recognition by using Wikipedia. In *Proceedings of the ACL 2006 Workshop on New Text Wikis and Blogs and Other Dynamic Text Sources*, Trento, Italy. Association for Computational Linguistics. 81

Tsvetkov, Y. and Wintner, S. (2014). Identification of multiword expressions by combining multiple linguistic information sources. *Computational Linguistics*, 40(2):449–468. 85

Vivaldi, J., Màrquez, L., and Rodríguez, H. (2001). Improving term extraction by system combination using boosting. In De Raedt, L. and Flach, P., editors, *Machine Learning: ECML 2001*, volume 2167 of *Lecture Notes in Computer Science*, pages 515–526, London, UK. Springer Berlin Heidelberg. 92

Vivaldi, J. and Rodríguez, H. (2007). Evaluation of terms and term extraction systems: A practical approach. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication*, 13:225–248. 97

Weeds, J., Dowdall, J., Schneider, G., Keller, B., and Weir, D. (2005). Using distributional similarity to organise BioMedical terminology. *Terminology*, 11(1):3–4. 89, 90

Wermter, J. and Hahn, U. (2005). Finding new terminology in very large corpora. In *Proceedings of the 3rd International Conference on Knowledge Capture*, Alberta, Canada. ACM Press. 84

Wong, W. (2009). Determination of unithood and termhood for term recognition. In Song, M. and Wu, Y.-F. B., editors, *Handbook of Research on Text and Web Mining Technologies*, chapter 30, pages 500–529. IGI Global. 80

Wong, W., Liu, W., and Bennamoun, M. (2012). Ontology learning from text: A look back and into the future. *ACM Computing Surveys*, 44(4):20:1–20:36. 93

Wu, Z. and Palmer, M. (1994). Verbs semantics and lexical selection. In *32nd Annual Meeting of the Association for Computational Linguistics: Proceedings of the Conference*, pages 133–138, New Mexico, USA. Association for Computational Linguistics. 88

Wüster, E. (1974). Die allgemeine Terminologielehre–ein Grenzgebiet zwischen Sprachwissenschaft, Logik, Ontologie, Informatik und den Sachwissenschaften. *Linguistics*, 12(119):61–106. 69

Yamamoto, K., Kudo, T., Konagaya, A., and Matsumoto, Y. (2003). Protein name tagging for biomedical annotation in text. In Ananiadou, S. and Tsujii, J., editors, *Proceedings of the ACL 2003 Workshop on Natural Language Processing in Biomedicine*, pages 65–72, Sapporo, Japan. Association for Computational Linguistics. 92

Yangarber, R., Grishman, R., Tapanainen, P., and Huttunen, S. (2000). Automatic acquisition of domain knowledge for information extraction. In *The 18th Conference on Computational Linguistics: Proceedings of the Conference*, volume 2, pages 940–946, Saarbrucken, Germany. Association for Computational Linguistics. 97

Yangarber, R., Lin, W., and Grishman, R. (2002). Unsupervised learning of generalized names. In *COLING 2002: The 19th International Conference on Computational Linguistics*, Taipei, Taiwan. Association for Computational Linguistics. 72, 92

Zadeh, B. Q. and Handschuh, S. (2014). The ACL RD-TEC: A dataset for benchmarking terminology extraction and classification in computational linguistics. In Drouin, P., Grabar, N., Hamon, T., and Kageura, K., editors, *COLING 2014: Computerm 2014: 4th International Workshop on Computational Terminology: Proceedings of the Workshop*, pages 52–63, Dublin, Ireland. Association for Computational Linguistics and Dublin City University. 77, 80, 95

Zervanou, K. (2010). UvT: The UvT term extraction system in the keyphrase extraction task. In *SemEval 2010: 5th International Workshop on Semantic Evaluation: Proceedings of the Workshop*, pages 194–197, Uppsala, Sweden. Association for Computational Linguistics. 79

Zhang, Z., Iria, J., Brewster, C., and Ciravegna, F. (2008). A comparative evaluation of term recognition algorithms. In Calzolari, N., Choukri, K., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., and Tapias, D., editors, *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, pages 2108–2113, Marrakech, Morocco. European Language Resources Association. 86, 96

Zweigenbaum, P. and Grabar, N. (1999). Automatic acquisition of morphological knowledge for medical language processing. In Horn, W., Shahar, Y., Lindberg, G., Andreassen, S., and Wyatt, J., editors, *Artificial Intelligence in Medicine*, volume 1620 of *Lecture Notes in Computer Science*, pages 416–420. Springer Berlin Heidelberg. 74

This page is intentionally left blank.