Automatic Term Extraction

Kris Heylen and Dirk De Hertog

KU Leuven (University of Leuven) Research group QLVL

Abstract

This chapter focuses on computational approaches to the automatic extraction of terms from domain specific corpora. The different subtasks of Automatic Term Extraction are presented in detail, including corpus compilation, unithood, termhood and variant detection, and system evaluation.

1. Introduction

The general aim of Term Extraction (TE) is to identify the core vocabulary of a specialised domain. Traditional Manual Term Extraction (MTE) is carried out by a terminologist who lists potential Term Candidates (TC) and then consults with a domain expert to arrive at a final list of validated terms. However, in a rapidly changing world with an ever growing technical vocabulary, the manual maintenance, or in the case of new technological fields, the manual exploration, indexation and description of a domain's core vocabulary is a labour-intensive enterprise. Automatic Term Extraction (ATE) is meant first and foremost as a computerised aid to alleviate this time-consuming task. For now, ATE concentrates on automating the preliminary identification of Term Candidates. In the long run, ATE might replace MTE completely.

ATE is also known as Terminology Extraction, Terminology Mining, Term Recognition, Glossary Extraction, Term Identification and Term Acquisition. It is based on the computerised analysis of text corpora. ATE offers some advantages to MTE. Firstly, ATE involves a computerised agent, which bases itself objectively on corpus evidence. Because an expression's terminological status is often a matter of degree and open to individual variation, ATE can circumvent an expert's subjectivity that potentially influences the TE-process. Secondly, ATE saves the expert the effort of manually investigating the full text and serves as a first filter to preselect TC's, a task well suited for an automatic agent. Despite these advantages of ATE, it must be noted that terms are inherently *semantically* defined, as referring to a domain specific *concept*, and the full automatic modelling of semantics is still out of reach for computers. The final confirmation of an expression's term status therefore still has to be done manually by domain specialists.

ATE is a well-established research domain within Natural Language Processing and Information Retrieval since the early 1990s (see Cabré Castellví, Estopà, and Vivaldi 2001 for a review of early systems). It consists of a number of modular subtasks that are typically carried out consecutively. The following subtasks can be distinguished:

1. *Corpus Collection* - the compilation of a representative domain specific corpus. If contrastive approaches to term extraction are used, also a general language corpus is required. Depending on the requirements of methods used further on in the ATE-process, the corpora undergo pre-processing such as lemmatisation, part-of-speech tagging, chunking or full syntactic parsing;

2. *Detection of Unithood (UH)* - The identification of linguistic elements that constitute a Multiword Unit (MWU) and refer to one conceptual unit;

3. *Detection of Termhood (TH)* - A method that ranks, or classifies, the extracted units in terms of the likelihood that they constitute a valid term for the domain at hand;

4. *Detection of Term Variants (TV)* - The identification of different linguistic realisations of the same domain specific concept;

5. *Evaluation and Validation* - A procedure to assess the quality of the automatic term extraction relative to manual extraction by a domain expert.

Term Extraction is usually not a goal in itself. Its output, the Term Candidate (TC) list, is the input for other tasks in Terminology Management. The exact interpretation and importance given to each of the modules above, largely depends on the intended further use of the TC list. In general, three practical applications can be identified (Thurmair 2003):

1. *Terminography* - The TC list is used as input for compiling a terminological dictionary or an electronic term database of a domain's specialised vocabulary. This type of terminology work is the focus of this handbook and relies on academically underpinned, concept-based criteria for termhood. As such, validation and the detection of term variants referring to the same concept are important subtasks;

2. *Translation Support* - The TC list functions as an ad-hoc glossary for a (manual or automatic) translation project and aims to identify unknown words whose translation needs looking up, or helps to maintain translation consistency throughout the project. What counts as a term is quite flexible and often determined opportunistically. Whereas the detection of multi-word units, which need to be translated consistently, is quite important in this application, termhood detection and validation only play a minor role;

3. *Information Retrieval (IR)* - The TC-list is the basis for indexing a document collection, so that users can query or browse the collection for domain-specific topics. The relevance of TCs is defined based on the users' search needs and the validation is often external and application-based. The compilation of document collections is an important aspect of IR and if the IR is concept-based, variant detection is an important subtask as well.

From these potential applications, it becomes clear that ATE is strongly related to some neighbouring disciplines: in a translation setting, it is closely related to Term Alignment, also called Bilingual Term Extraction. Alignment in general is the task of matching equivalent sentences, words and phrases in parallel corpora. Term alignment focuses on the pairing of domain specific terms in two or more languages. In a terminography or IR setting, ATE is often the first step in Ontology Construction. Ontology Construction then, is the discipline that identifies the relationships that hold between terms in a certain domain. Those relations, like synonymy, type-of relations or part-whole relations, are used to construct a relational network that offers the user an accessible overview of the domain's terminology.
In the remainder of this chapter, we concentrate on ATE proper, leaving Bilingual Term Extraction and Ontology Construction aside. More specifically, we focus on the theoretical and methodological foundations of each of the 5 subtasks identified above.

2. Corpus Collection

Any ATE method has to be based on a text corpus that is *representative* of the specialised domain whose terminology is to be charted. In some ATE applications, the specialised domain is quite restricted and the relevant texts to be analysed form a finite and well defined set. For example, when a company or organisation wants an inventory of its in-house terminology, the text corpus corresponds naturally to the document collection that the company or organisation provides to the terminologist. However, when a project aims to analyse the terminology of a domain at large, like "Marine Biology" or "Aeronautics", corpus compilation necessarily involves the *sampling* of texts from that domain and both design and practical issues come into play. For contrastive approaches (Section 3.5), the acquisition of a

representative corpus of general language use is equally important. For a detailed discussion of corpus design issues in both the compilation of general and specialised corpora, we refer to the previous chapter and to the overviews in Biber (1993) and McEnery, Xiao and Tono (2006, 13-21). Rizzo (2010) offers a practical guide to specialised corpus compilation.

In the remainder of this section, we only briefly discuss some recent online corpus compilation approaches that are explicitly aimed at ATE and that follow an incremental procedure to collect large corpora of specialised language use with relative ease and speed, be it at the expense of rigorous design and text quality control.

Baroni and Bernardini's (2004) BootCat system[1] starts from a small set of manually selected *seed terms* that are highly representative of the intended specialised domain. The seed terms may also come from a preliminary ATE analysis on a (small) domain specific corpus. In a first phase, random combinations of seed terms are submitted as a query to a general purpose search engine like Google in order to retrieve domain specific URLs. The URLs' web pages are downloaded and their content is checked against the seed term list to ensure they indeed belong to the intended domain. If so, they are added to the incrementally compiled corpus. The newly added texts are submitted to ATE and the initial term list is extended with additional terms. This extended term list is the input to a second phase of URL retrieval. The procedure is repeated until the corpus is large enough, or until no new URLs and/or terms can be retrieved. In de Groc 2011, this approach is extended with a web crawling phase, in which the retrieved URLs are used as seeds to recursively traverse linked web pages that are also checked for domain specificity and added to the corpus. These online collected corpora are then the input for the next steps in the ATE process that are described below.

[1] Available for download at http://bootcat.sslmit.unibo.it/. Also available as part of The Sketch Engine (http://www.sketchengine.co.uk).

3. Unithood

3.1 Introduction

Unithood is defined as “the degree of strength or stability of syntagmatic combinations and collocations” (Kageura and Umino 1996). Historically, the detection of UH was the first (sub)task to be covered by ATE when it established itself as a discipline in the late 1980s and early 1990s. There are several reasons for this clear focal point.

First of all, multiword units, mostly in the form of noun phrases, are argued to be highly prevalent in technical domains. They are therefore considered to be the most important target for ATE. Nakagawa and Mori (1998, 2002) claim that 85% of the TC targets are identified as technical noun phrases consisting of 2 or more words.

Secondly, the theoretical terminological ideal that a term has a one on one relationship with the concept it represents, serves as an immediate steppingstone to the practical focus on multiword terms. Multiword terms are by definition semantically more specified than their single word counterparts. The semantic scope of the head narrows down due to semantic restrictions imposed by its modifier. Bourigault and Jacquemin (1999) claim that “single-word terms are too polysemous and too generic” whereas multi-word terms “represent finer concepts in a domain”.

Thirdly, also more practical considerations played a role initially. The lack of easily available, extensive general language corpora in the early nineties meant that probabilistic, frequency based techniques could not be readily applied to decide on the TH of simplex words through comparison of in and out-of-domain corpora. The detection of multiword units on the other hand relies solely on technical documents supplied by the interested parties. In this type of research, termhood was considered to be implied by unithood (e.g. Kit 2002). However,

recent approaches to term extraction consider the detection of UH as a separate step from the assessment of TH and most term extractors also extract simplex TC's.

In most terminological approaches, multiword combinations constitute a terminological expression if and only if they refer to a conceptual unit. However, access to the conceptual level is not straightforward. Therefore, the degree of unithood is determined on the basis of linguistic and statistical properties observable in linguistic surface forms. So-called linguistic approaches use morpho-syntactic patterns as evidence for unithood. Statistical approaches rely on corpus frequency information about word combinations. Current term extractors combine the strengths of both methods, in what are called hybrid approaches (e.g. Vivaldi and Rodriguez 2001, Pazienza, Pennacchiotti and Zanzotto 2005). Here, we discuss these two basic approaches separately.

3.2 Linguistic Approaches

Linguistic approaches to ATE are based on the property that multiword terms tend to follow specific morpho-syntactic patterns. They rely on this templatic behaviour to determine the validity of a word combination as a linguistic unit, and if so, as a TC. The advent of more powerful corpus pre-processing methods enables the inclusion of linguistic information in a semi-automatic detection process: Part-of-speech (POS) taggers automatically process large quantities of text and provide words with their POS-tags. This allows the ATE-process to incorporate a templatic extraction of admissible surface forms, which are called *syntactic templates* and which consist of a sequence of POS patterns.

The domain expert defines the relevant syntactic templates based mainly on linguistic criteria (French combines words into units differently than English does) and domain relevance. For a car manufacturer interested in the names of car parts there is an obvious point to focus on

objects, and therefore on noun phrases. A lawyer interested in which subjects are involved in legal processes and how these subjects interact with each other, benefits from including templates that allow for verb phrases. Table 1 shows the patterns of the POS filter that Justeson and Katz (1995) proposed for English[2] with examples from the domain of Mathematics. However, also practical considerations play a role in the selection of valid templates. The amount of templates and the leniency with which they are applied directly influences the accuracy the ATE. More specifically, differences in accuracy motivate the choice for what are called open class, or closed class filters. Open class filters allow many optional POS elements and result in more surface forms. This has the advantage of allowing the extraction to include many TC's, but the disadvantage of yielding many false candidates. The manual correction of lists provided by open class filters are therefore more labour-intensive. However, if coverage is the expert's main concern this method is preferred. Closed class filters are more restrictive in the choice of allowed patterns. This has the clear benefit of boosting precision, but comes at the cost of coverage of possible TC's.

Adjective Noun	linear function
Noun Noun	regression coefficients
Adj. Adj. Noun	Gaussian random variable
Adj. Noun Noun	cumulative distribution function
Noun Adj. Noun	mean squared error
Noun Noun Noun	class probability function
Noun Preposition Noun	degrees of freedom

Table 1. Examples of POS-templates from Justeson and Katz (1995)

The candidates of a POS-tagged corpus are straightforwardly compared and matched to the final selection of syntactic templates. For instance, if an *Adjective Noun* combination is defined as a valid surface form, and this filter is applied to the candidates occurring in a

[2] Taken from Schütze (1999: 154)

fashion corpus, it will yield instances such as “high heels”, “high expectations” and “high building”.

Because the use of templates does not differentiate between general, everyday words and technical words, a list of unwanted words is used as a secondary filter to eliminate non-technical candidates. Such a list, which is often compiled on the basis of a list of high frequency words obtained from a general corpus, is called a stopword list. For instance, by including the word “high” on such a stopword list, the TC’s extracted from the fashion magazines would be filtered out. This procedure can boost precision, but it can also falsely remove valid TC’s, as is the case with “high heels”.

3.3 Statistical Approaches

Statistical approaches make use of two properties that are typical of multiword terms and that, in principle, require no linguistic information: Firstly, multiword terms are relatively fixed word combinations, and secondly, they occur with relatively high frequency. Because most multiword terms exhibit a high degree of syntagmatic stability, without variation in word order, statistical approaches can in principle limit themselves to analysing n-grams, i.e. continuous word sequences, without taking into account the underlying linguistic structure. Unithood of the n-grams is measured as some function of their corpus frequency. Again, this requires no linguistic analysis, only a corpus of sufficient size. Additionally, n-gram extraction and quantitative analysis are highly amenable to computer processing, making this approach very scalable to large document collections. Note however that there are few examples of pure statistical approaches (Pantel and Lin 2001). Most state-of-the art term extractors combine the strengths of statistical methods with the knowledge advantage of the linguistic approach.

3.3.1 Collocation Measures

Basic frequency information is obtained from corpora by counting words and words co-occurring together. The frequent co-occurrence of two or more words in sequence is an indication that these words belong together and form a multiword term. However, raw frequency counts are only used in combination with linguistic filters, as with the Justeson and Katz (1995) POS-filter cited above. In purely statistical approaches, raw co-occurrence frequencies are typically rescaled based on some measure of informativeness. These *collocation measures* compare the frequency of a word combination with the frequencies of the individual words making up that combination. Whereas the regular co-occurrence of two frequent words (e.g. “new” and “thing”) is not very surprising, a frequent co-occurrence of two not so frequent words (e.g. “diesel” and “engine”) does indicate that the word combination could be a fixed expression and potentially a term. More formally, these collocation measures quantify how much the observed co-occurrence of two words deviates from what is expected by chance given the individual frequencies of the words. Table 2 shows for a toy example how the observed co-occurrence of “diesel” and “engine” (60) is considerably higher than the expected frequency by chance (24.76). The latter is calculated by multiplying the individual frequencies of “diesel” (258) and “engine” (96), and dividing that product by the corpus size (1000).

	diesel	¬ diesel	
engine	60	36	96
¬ engine	198	706	904
	258	742	1000

	diesel	¬ diesel
engine	24.76	71.23
¬ engine	233.23	670.76

with: $E_{ij} = (R_i * C_j) / N$

Table 2. Observed Frequencies (left) and Expected Frequencies (right) of the collocation "diesel engine"

There are many ways to quantify the divergence between observed and expected frequencies and there is plethora of collocation measuresavailable.. A well-known example is the X^2 statistic, used in term extraction experiments by a/o Drouin (2006) and Matsuo and Ishizuka (2004). It measures the difference between observed (O_{ij}) and expected frequencies (E_{ij}) in the rows (r) and columns (c) of a contingency table according to the following formula:

$$X^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

For our "diesel engine" example, this gives a X^2 value of 74.7156. Doing this calculation for every word combination in the corpus then allows ranking all word combinations by unithood and selecting only those above a certain threshold. Other collocation measures for unithood include t-score, log likelihood ratio (Dunning 1993), mutual information (Church and Hanks 1990 and the phi coefficient. Manning and Schütze (1999) offer an introduction to collocation measures and a more comprehensive overview and mathematical background can be found in Evert (2004) and Wiechmann (2008). Pecina and Schlesinger discuss how different collocation measures can be combined. Again, it should be noted that these co-occurrence measures are usually combined with a linguistic approach and they are then only calculated for word combinations that have passed a linguistic filter first.

3.3.2 Paradigmatic Modifiability

As a consequence of their relative fixedness, the constituting parts of multiword terms cannot easily be replaced by other words. Whereas "new" in the non-term combination "new things" can be easily replaced by "novel", this is not the case for "diesel" in "diesel engine". Wermter

and Hahn (2005) use this property of reduced paradigmatic modifiability to determine the unithood of a word combination. For each candidate multiword combination that has come out of an initial linguistic filtering step, they collect the frequencies of all word combinations that have the same length and share at least one word with the candidate, but that also have one or more constituting parts replaced by another word. The accumulated frequency of these modified versions is then compared with the frequency of the actual multiword term candidate, resulting in the P-Mod measure of paradigmatic modifiability for unithood. Wermter and Hahn (2005) show that their P-Mod measure outperforms C-value and t- score in the task of term candidate extraction from biomedical texts.

3.3.3 Lexical Bundles.

A number of approaches focus on the detection of longer sequences of words, with no a priori limitation of length or restriction to predefined POS patterns like noun phrases. This is especially important for domains that are characterised by phraseological expressions like the legal domain and its formulae like “Do you swear to tell the truth, the whole truth, and nothing but the truth?” In their analyses of register-specific expressions, Biber and Conrad (1999) refer to such longer word sequences as lexical bundles and use relative frequency per million words as a selection criterion. Simpson-Vlach and Ellis (2010) build upon Biber’s lexical bundles for the extraction of formulaic expressions, but to reduce the list of candidates, they combine a frequency cut-off of 10 occurrences per million with a collocation measure, viz. Mutual Information (MI). Based on psycholinguistic judgments of unithood, a regression analysis then determines the contribution of relative frequency and MI to final unithood measure.

Da Silva et al. (1999) propose a more complex algorithm to detect lexical bundles that uses nestedness information next to relative frequency and information measures.

4. Termhood

4.1 Introduction

By the late 1990s, the notion of Termhood (TH) was introduced into ATE to refer to “the degree to which a stable lexical unit is related to some domain-specific concepts” (Kageura and Umino 1996). Termhood and unithood are considered to be separate properties of a TC and unithood does not necessarily imply termhood: A multiword expression like “most of the time” has a high degree of unithood but low termhood in any specialised domain. On the other hand, a single word expression like “hypoglycæmia” lacks the unithood associated with multiword units, but it does have high termhood in the medical domain.

The earliest and simplest approach to measure TH is the use of domain internal frequency as an indicator of a TC’s importance within a given domain and hence its likelihood to be a valid term (e.g. Daille 1994 and Daille, Gaussier and Langé 1994). However, while domain internal frequency is to some extent correlated with TH -certainly when longer multiword units are involved- it is not informative enough to decide on the termhood of single words or of highly frequent word combinations: General language words and word combinations are among the most frequent elements in any corpus, specialised or not, but they are not very interesting from a terminological point of view.

A second approach therefore looks at the distributional properties of TCs within the domain, and more specifically the dispersion over different documents. A third approach goes beyond pure frequency by looking at the contextual usage of TCs. A fourth method is specifically intended for single word TCs and analyses the internal morphological structure of a TC.

Finally, a fifth family of methods contrast domain-internal with domain-external information. Below we discuss in more detail the four approaches that go beyond mere frequency.

4.2 Distributional Approach: TF-IDF

A distributional approach looks at the dispersion of term candidates across the different documents that make up a domain-specific corpus. Words or word combinations that occur in almost every text are assumed to be not very specific and probably general language elements that also happen to occur frequently in the specialised corpus. On the other hand, TCs that only occur in a limited subset of documents are assumed to be truly domain specific. Formally, this termhood property is measured as Term Frequency multiplied by Inverse Document Frequency (TF-IDF: Salton, Wong, and Yang 1975; Evans, Milic-Frayling, and Lefferts 1995; Medelyan and Witten 2006): If the TC's frequency is spread over many documents TF-IDF will be low, whereas a high TC frequency in a limited number of documents results in a high TF-IDF.

4.3 A Contextual Approach to TH: C/NC value

Maynard and Ananiadou (1999) and Frantzi, Ananiadou and Mima's (2000) widely used contextual approach starts from TCs coming out of a linguistic filter and then analyses how these co-occur with additional context words. The C/NC-approach works in two steps. First, the C-value analyses to what extent TCs occur in the context of other TCs. More specifically, the C-value quantifies to what extent multiword TCs are nested. Nested terms appear as substrings of longer terms (whether or not they appear as a standalone term as well). TCs that only occur nested, i.e. as part of longer terms that were also extracted with the linguistic filter,

are deemed to be incomplete term fragments that do not occur independently and hence receive a low C-value. For example in an ophthalmological corpus, “contact lens” occurs independently from “soft contact lens” and is considered a term, whereas “soft contact” does not. Additionally, some nested terms occur in many different longer sequences and this also is an indication of termhood. For example, “floating point” occurs nested in “floating point arithmetic”, “floating point constant”, “floating point operation”, “floating point routine”, “floating point number” etc. and can be considered a term even though it does not occur solely by itself. “Point arithmetic” on the other hand occurs only nested in one longer sequence and is an incomplete fragment. Formally, the C-value[3] is calculated as follows:

$$C-value(a) = \begin{cases} \log_2|a| f(a) & if\ a\ is\ not\ nested \\ \log_2|a| (f(a) - \frac{1}{P(T_a)} \sum_{b \in T_a} f(b)) & if\ a\ is\ nested \end{cases}$$

With a is the candidate string

$f(.)$ is the corpus frequency

T_a is the set of extracted TC’s that contain a

$P(T_a)$ is the number of these TC’s

In a second step, Frantzi, Ananiadou and Mima (2001) exploit another characteristic of how terms typically co-occur with context words. The NC-value models the importance of certain context words as indicators of termhood. More specifically, the NC-measure relies on the fact that terms are generally quite strict about the modifiers they accept:

> Extended term units are different in type from extended word units in that they cannot be freely modified. There is a very limited range of qualifiers which can be used with the term “heat transfer”; the word “heat wave” can be modified by such hyperbolic

[3] Note that C-value by itself is sometimes considered a unithood measure (Foo 2012) because it only measures TC independence.

> expressions as “suffocating” or “never ending” and a great number of other qualifiers. Extended terms are linguistic representations of essential characteristics whereas in words such collocations are inessential in that they can be omitted without affecting the denotation of the head of the nominal group as a lexeme (Sager 1978).

The criterion they use for considering a word as a term-indicative context word is the number of different terms it appears with, divided by the total number of terms that are identified. Frantzi, Ananiadou and Mima (2001) use this NC-value as a complement to C-value to co-determine the TH of a given string: they combine both measures using a different weight, 0.8 for C-value and 0.2 for NC-value, resulting in a termhood measure that also attributes context a certain role in the ATE-process.

4.4 Morphological Approaches

A more linguistically informed approach to termhood analyses the internal morphological structure of TCs (Aubin and Hamon 2006). Some domains, like the medical domain, make heavy use of neoclassical terminology with terms derived from Latin or Greek. This characteristic can be used as an indication of termhood. Ananiadou (1994) provides a morphological description of medical terms and focuses on typical Latin or Greek affixes that are indicative of termhood. A second morphological approach is specifically designed for compounding languages like German, Dutch, Swedish or Japanese, which typically make new terms by combining existing words into one orthographic unit. De-compounding approaches do the opposite of unithood detection and try to split up complex terms in their constituting parts (Nakagawa 2000). Although the mere property of being a compound might already increase the termhood of a TC (Foo and Merkel 2010), these approaches typically try to infer the termhood of the compound as a whole from the termhood properties of the constituting

parts. The later can be any type of termhood information, for example productivity of the compound's semantic head (Kageura 2009, Assadi and Bourigault 1996, Nakagawa and Mori 2002).

4.5 Contrastive Approaches to TH

So-called Contrastive Term Extraction (CTE) approaches come in a wide variety of flavours but all methods rely on the fact that terms are per definition domain-specific, and as a consequence are hypothesised to occur more frequently in their proper domain than they do in other domains or in general language use. These approaches therefore compare the frequency of a TC in a domain-specific corpus with its frequency in a reference corpus (either a balanced, general language corpus, or a corpus from another domain). A number of approaches use measures that are very similar to the collocation measures from Section 2.2. In this case, the observed domain-internal frequency is compared to the expected frequency if a TC would have an equal probability of occurrence in the domain-specific and the reference corpus. Table 4 shows single word TCs in Dutch and their association with a legal corpus by using the X^2 statistic as a contrastive termhood measure and a general newspaper corpus as reference corpus. This approach has much in common with keyword extraction (Scott 1997) in corpus linguistics.

Dutch	**English**	**X^2**
uitstsellen	to delay	166,75
ontvangstbericht	acknowledgment	114,83
hoofdfunctie	principal function	94,75
Staatsblad	Official Gazette	34,92
inrichtingskosten	costs of setting up	16,62
validiteitsperiode	period of validity	8,56

Table 4. Ranked TC's from a Belgian Dutch legal corpus

Many other approaches in ATE (see Drouin 2003, Drouin and Doll 2008 for an overview) use the same underlying idea of association to the proper domain but come to different operationalizations. The *contrastive weight* method by Basili et al.(2001) is an adaption TF-IDF where the dispersion over different documents (as indication of non-termhood) is replaced by dispersion over different domains. Ahmad, Gillam and Tostevin. (1999) use a measure they refer to as the *weirdness* of a word, which is defined as the result of the comparison of the word's normalised frequencies between a specialised corpus and a general language corpus. In this manner they "identify signatures of a specialism". Those words which combine high frequency and high weirdness are of most interest when it concerns term identification. Kit and Liu (2008) quantify the termhood of a term candidate as its rank difference between a domain and a reference corpus. This rank is based on the word's frequency for both types of corpora and is normalised by the total number of types in the corpus' vocabulary. Chung (2003) uses a normalised frequency ratio to decide on termhood. Wong, Liu and Bennamoun (2007) propose a similar technique that uses distributional behaviour of a word in opposing corpora to measure what he calls intra-domain distribution and cross-domain distributional behaviour. The first distribution is then used to calculate a *domain prevalence* score, which measures the extent of the term's usage in the target domain. The second distribution is the basis for a *domain tendency* score, which measures the extent of term usage towards the target domain. Drouin (2006) compares precision and recall for the ranking of different hypothesis testing methods, trying to determine which method works best.

5. Term Variation

Ident	Base Term	Variant
NAInsAv	Noun1 Adj2	Noun1 ((Adv? Adj)0-3 Adv) Adj2
NAInsAj	Noun1 Adj2	Noun1 ((Adv? Adj)1-3 Adv?) Adj2
NAInsN	Noun1 Adj2	Noun1 ((Adv? Adj)? (Prep? Det? (Adv? Adj)? Noun) (Adv? Adj)? Adv?) Adj2
ANInsAv	Adj1 Noun2	(Adv) Adj1 Noun2
NPNSynt	Noun1 Prep2 Noun3	Noun1 ((Prep Det?)? Noun3
NPDNSynt	Noun1 Prep2 Det4 Noun3	Noun ((Prep Det?)?) Noun3

Table 5. Transformational rules for the detection of term "variants"

The classical approach to terminology defines a term as a domain-specific concept that ideally has a one on one relationship with a linguistic expression. However, this ideal situation of univocity is more complicated in reality because of term variation, i.e. the expression of a single concept by means of several linguistic surface forms. Daille (1996) states that "a variant of a term is an utterance which is semantically and conceptually related to an original term" and Daille (2005) reports that between 15% to 35% of the concepts are variants of each other. In order to meet classical terminology in its theoretical assumptions, one subtask of TE is therefore the identification and clustering of term variants after the extraction process. Daille (2005) proposes a typology of term variants and focuses on typical patterns of deletion, insertion or adjective-PP modifier alternations. Similarly, Bourigault and Jacquemin (1999) in their FASTR system for French use transformational rules exploiting shallow syntactic information to detect term variations. Table 5 exemplifies these transformational rules which can be classified in two families: internal insertion of modifiers and/or preposition switch, and determiner insertion. Instead of grouping variants post hoc,Nenadic, Ananiadou

and McNaught 2004 integrate pattern-based variant detection in the extraction step to enhance performance.

6. Evaluation and Validation

The final subtask of ATE is the evaluation step that assesses how well an ATE method performs relative to Manual Term Extraction. Lists of TC's are evaluated according to an established gold-standard glossary of domain terminology, or ad hoc, by a domain expert and/or a terminologist engaged in a specific project. Based on this gold standard or the expert's judgements, the ATE process is evaluated with several measures. *Precision* of the TC list is the percentage of correctly identified terms out of a total of all proposed TCs. Often top n-thousand lists are chosen to show the extractor's performance. Note that gold standards term glossaries are rarely exhaustive in their coverage and often this type of evaluation underestimates the real precision because some correctly identified TCs are incorrectly classified as mistakes. If an exhaustive manually compiled list of terms for a document collection or an exhaustive gold standard glossary for a domain is available, it is also possible to calculate *recall*, i.e. the proportion of terms identified out of all terms that appear in the specialised corpus. Note that manual validation of a list of TC's by definition excludes the possibility to calculate recall. In general, high precision comes at the expense of recall and vice versa. In practice, choices have to be made in the engineering process of the ATE whether to favour high recall or high precision, and often the latter is preferred. Zhang et al. (2008) evaluate the precision (but not the recall) of a number of methods to extract terms from large corpora. Precision is measured in tiers or n-best lists, i.e. precision for the top 100, 1000, 5000 etc. 10k. Korkontzelos, Klapaftis, and Manandhar (2008) compare the performance of

unithood and termhood measures and conclude that termhood measures achieve superior results. Vivaldi and Rodriguez (2007) point out that despite many years of research, generally accepted gold standards and evaluation methods are not readily available and this still complicates an objective and qualitative comparison of the performance of different ATE systems.

7. Conclusion

Automatic Term Extraction is well-established discipline within Natural Language Processing and many different approaches and systems have been developed. Yet for all, a number of recurrent subtasks can be distinguished: corpus compilation, unithood, termhood and variant detection, and evaluation. The earliest systems only used linguistic information to identify terms but gradually, increasingly sophisticated statistical methods have been developed to extract terms from large corpora. Most state-of-art systems are hybrids, combining both types of information (e.g. Sclano and Velardi 2007. Despite the large body of research, there is no generally agreed standard of what a good automatic term extractor should achieve. This depends both on the specific application that the term extraction is intended for, be it terminography, translation support or information retrieval, as well as on the specific language, domain and corpus.

References

Ahmad, Khurshid, Lee Gillam, and Lena Tostevin. 1999. “Weirdness Indexing for Logical Document Extrapolation and Retrieval (WILDER).” In *The 8th Text Retrieval*

Conference, edited by Ellen Voorhees and Donna Harman, 717-724. Washington: National Institute of Standards and Technology.

Ananiadou, Sophia. 1994. "A methodology for automatic term recognition." In *Proceedings of the 15th conference on Computational linguistics (COLING'94)*, 1034-1038. Kyoto, Japan.

Assadi, Houssem and Didier Bourigault. 1996. "Acquisition et modélisation des connaissances à partir de textes: outils informatiques et éléments méthodologiques." In *Actes du 10ème congrès Reconnaissance des Formes et Intelligence Artificielle*, 505-514. Rennes: Association Française pour la Cybernétique Economique et Technique.

Aubin, Sophie and Thierry Hamon. 2006. "Improving term extraction with terminological resources." In *Proceedings of the 5th international conference on Advances in Natural Language Processing*, edited by Tapio Salakoski, Filip Ginter, Sampo Pyysalo and Tapio Pahikkala, 380-387. Berlin/Heidelberg: Springer-Verlag.

Baroni, Marco and Silvia Bernardini. 2004. "BootCaT: Bootstrapping Corpora and Terms from the Web." In *Proceedings of the Fourth International Conference On Language Resources And Evaluation*, edited by Maria Teresa Lino et al., 1313-1316. Lisbon, Portugal: European Language Resources Association.

Basili, Roberto, Alessandro Moschitti, Maria Teresa Pazienza, and Fabio Massimo Zanzotto. 2001. "Modelling Syntactic Context in Automatic Term Extraction." In *Proceedings of Recent Advances in Natural Language Processing*, edited by Nicolas Nicolov and Ruslan Mitkov, 28-34. Amsterdam/Philadelphia: John Benjamins.

Biber, Douglas. 1993. "Representativeness in Corpus Design." *Literary and Linguistic Computing* 8(4):243-257.

Biber, Douglas and Susan Conrad. 1999. "Lexical bundles in conversation and academic prose." *Language and Computers* 26:181-190.

Bourigault, Didier. 1992. "Surface grammatical analysis for the extraction of terminological noun phrases." In *Proceedings of 14th International Conference on Computational Linguistics*, edited by Christian Boitet, 977-981. Stroudsburg, PA, USA: Association for Computational Linguistics.

Bourigault, Didier and Christian Jacquemin. 1999. "Term extraction + term clustering: An integrated platform for computer-aided terminology." In *Proceedings of the ninth conference on European Chapter of the Association for Computational Linguistics (EACL)*, Bergen, 15-22. Stroudsburg, PA, USA: Association for Computational Linguistics.

Cabré Castellví, M. Teresa, Rosa Estopà, and Jordi Vivaldi 2001. "Automatic term detection: a review of current systems." In *Recent Advances in Computational Terminology*, edited by Didier Bourigault, Christian Jacquemin and Marie-Claude L'Homme, 53-88. Natural Language Processing, vol. 2. Amsterdam/Philadelphia: John Benjamins Publishing Company.

Chung, Teresa Mihwa. 2003. "A corpus comparison approach for terminology extraction." *Terminology* 9(26):221-246.

Church, Kenneth and Patrick Hanks. 1990. “Word association norms, mutual information, and lexicography.” *Computational Linguistics* 16(1):22-29.

Da Silva, Joaquim, Gaël Dias, Sylvie Guilloré, and José Pereira Lopes. 1999. “Using LocalMaxs Algorithm for the Extraction of Contiguous and Non-contiguous Multiword Lexical Units.” In *Proceedings of the 9th Portuguese Conference on Artificial Intelligence: Progress in Artificial Intelligence*, edited by Pedro Barahona and José Júlio Alferes, 113-132. London, UK: Springer-Verlag.

Daille, Béatrice. 1994. “Study and Implementation of Combined Techniques for Automatic Extraction of Terminology.” In *The Balancing Act: Combining Symbolic and Statistical Approaches to Language. Workshop at the 32nd Annual Meeting of the Association for Computational Linguistics*, 29-36. Stroudsburg, PA, USA: Association for Computational Linguistics.

Daille, Béatrice. 1996. “Study and Implementation of Combined Techniques for Automatic Extraction of Terminology.” In *The Balancing Act: Combining Symbolic and Statistical Approaches to Language*, edited by Philip Resnik and Judith L. Klavans, 49-66. Cambridge, MA, USA: MIT Press.

Daille, Béatrice. 2005. “Variations and application-oriented terminology engineering.” *Terminology* 11(1):181-197.

Daille, Béatrice, Eric Gaussier, and Jean-Marc Langé. 1994. “Towards automatic extraction of monolingual and bilingual terminology.” In *Proceedings of the 15th International Conference on Computational Linguistics*, 515-521. Stroudsburg, PA, USA: Association for Computational Linguistics.

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

Drouin, Patrick. 2006. “Termhood: Quantifying the Relevance of a Candidate Term.” *Linguistic Insights. Studies in Language and Communication* 36:375-391.

Drouin, Patrick and Frédéric Doll. 2008. “Quantifying Termhood Through Corpus Comparison”, In *Terminology and Knowledge Engineering (TKE-2008)*, 191-206. Copenhagen, Denmark: Copenhagen Business School.

Dunning, Ted. 1993. “Accurate methods for the statistics of surprise and coincidence.” *Computational Linguistics* 19(1):61-74.

Evans, David, Natasa Milic-Frayling, and Robert Lefferts. 1995. “Clarit TREC-4 Experiments.” In *NIST Special Publication 500-236*, edited by Donna Harman, 305-322.

Evert, Stefan. 2004. “The Statistics of Word Cooccurrences: Word Pairs and Collocations.” PhD diss., University of Stuttgart.

Frantzi, Katerina, Sophia Ananiadou, and Hideki Mima. 2000. “Automatic recognition of multi-word terms: The C-value/NC-value method.” *International Journal on Digital Libraries* 3(2):115-130.

Foo, Jody. 2012. “Computational Terminology: Exploring Bilingual and Monolingual Term Extraction.” PhD diss., Linköping University.

Foo, Jody and Magnus Merkel. (2010). “Computer aided term bank creation and standardization: Building standardized term banks through automated term extraction and advanced editing tools.” In *Terminology in Everyday Life,* edited by Marcel Thelen and Frieda Steurs, 163-180. New York: John Benjamins.

Groc, Clément de. 2011. “Babouk: Focused Web Crawling for Corpus Compilation and Automatic Terminology Extraction.” In *Proceedings of the International Conference on Web Intelligence and Intelligent Agent Technology,* edited by Olivier Boissier, Boualem Benatallah, Mike P. Papazoglou, Zbigniew W. Ras and Mohand-Said Hacid, 497-498. IEEE Computer Society.

Justeson, John S. and Slava M. Katz. 1995. “Technical terminology: some linguistic properties and an algorithm for identification in text”. *Natural Language Engineering* 1(1):9-27.

Kageura, Kyo. 2009. “Computing the potential lexical productivity of head elements in nominal compounds using the textual corpus”. *Progress in Informatics*, (6):49-56.

Kageura, Kyo and Umino, Bin 1996. “Methods of automatic term recognition: a review”. *Terminology* 3(2):259-289.

Kit, Chunyu. 2002. “Corpus tools for retrieving and deriving termhood evidence.” In *5th East Asia Forum of Terminology*, 69-80. Haikou, China.

Kit, Chunyu and Xiauyue Lui. 2008. “Measuring mono-word termhood by rank difference via corpus comparison.” *Terminology* 14(2):204-229.

Korkontzelos, Ioannis, Ioannis Klapaftis, and Suresh Manandhar. 2008. “Reviewing and Evaluating Automatic Term Recognition Techniques.” In *Proceedings of the 6th International Conference on Natural Language Processing*, edited by Bengt Nordström and Aarne Ranta, 248-259. Berlin/Heidelberg, Germany: Springer.

Liu, Xiaoyue and Chunyu Kit. 2009. “Statistical termhood measurement for mono-word terms via corpus comparison.” In *Proceedings of the Eighth International Conference on Machine Learning and Cybernetics*, 3499-3504. IEEE Computer Society.

Manning, Christopher and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. Cambridge, MA, USA: MIT Press.

Matsuo, Yutaka and Mitsuru Ishizuka. 2004. “Keyword extraction from a single document using word co-occurrence statistical information.” *International Journal on Artificial Intelligence Tools* 13(1):157-169.

Maynard, Diana and Sophia Ananiadou. 1999. “Identifying Contextual Information for Multi-Word Term Extraction.” In *Proceedings of the TKE ‘99 International Congress on Terminology and Knowledge Engineering*, edited by Peter Sandrini, 212-221. Vienna, Austria: Termnet.

McEnery, Tony, Richard Xiao, and Yukio Tono, editors. 2006. *Corpus-based Language Studies: An Advanced Resource Book*. London, UK: Routledge.

Medelyan, Olena and Ian H. Witten. 2006. "Thesaurus based automatic keyphrase indexing." In *Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries*, edited by Gary Marchionini, Michael L. Nelson and Catherine C. Marshall, 296-297. New York, USA: Association for Computer Machinery.

Nakagawa, Hiroshi. 2000. "Automatic Term Recognition based on Statistics of Compound Nouns." *Terminology* 6(2):195-210.

Nakagawa, Hiroshi and Tatsunori Mori. 1998. "Nested collocation and compound noun for term recognition." In *Proceedings of the First Workshop on Computational Terminology*, edited by Didier Bourigault, Christian Jacquemin, and Marie-Claude L'Homme, 64-70. Montreal, Canada: Université de Montréal.

Nakagawa, Hiroshi and Tatsunori Mori. 2002. "A simple but powerful automatic term extraction method." In *Proceedings of the Second International Workshop on Computational Terminology*, 1-7. Stroudsburg, PA, USA: Association for Computational Linguistics.

Nenadic, Goran, Sophia Ananiadou, and John McNaught. 2004. "Enhancing automatic term recognition through recognition of variation." In *Proceedings of the 20th international Conference on Computational Linguistics*. Stroudsburg, PA, USA: Association for Computational Linguistics.

Pantel, Patrick and Lin, Dekang. 2001. "A Statistical Corpus-Based Term Extractor". In *Proceedings of the 14th Biennial Conference of the Canadian Society on Computational Studies of intelligence: Advances in Artificial intelligence*, edited by Eleni Stroulia and Stan Matwin, 36-46. Lecture Notes In *Computer Science*, vol. 2056. London: Springer-Verlag.

Pazienza, Maria Teresa, Marco Pennacchiotti, and Fabio Massimo Zanzotto. 2005. "Terminology extraction: an analysis of linguistic and statistical approaches." In *Knowledge Mining*, edited by Spiros Sirmakessis. Series: Studies in Fuzziness and Soft Computing, Vol.185. Springer-Verlag.

Pecina, Pavel and Pavel Schlesinger. 2006. "Combining association measures for collocation extraction." In *Proceedings of the COLING/ACL on Main Conference Poster Sessions Annual Meeting of the ACL*, 651-658. Morristown, NJ: Association for Computational Linguistics.

Rizzo, Camino R. 2010. "Getting on with corpus compilation: from theory to practice." *English for Specific Purposes World*, Issue 1(27), vol. 9. http://www.esp-world.info.

Sager, Juan C. 1978. Commentary by Prof. Juan Carlos Sager. In *Actes Table Ronde sur les Problèmes du Découpage du Terme* , edited by G. Rondeau, 39-74. Montréal: Commission de Terminologie de l'AILA.

Salton, Gerard, Andrew Wong, and Chung-Su Yang. 1975. "A vector space model for automatic indexing." *Communications of the ACM* 18:613-620.

Sclano, Francesco, Paola Velardi. 2007. “Termextractor: a web application to learn the common terminology of interest groups and research communities.” In *Proceedings of the 7th Conference on Terminology and Artificial Intelligence (TIA-2007)*, Sophia Antipolis.

Scott, Mike. 1997. “The Right Word in the Right Place: Key Word Associates in Two Languages.” *AAA - Arbeiten aus Anglistik und Amerikanistik,* 22 (2):239-252.

Simpson-Vlach, Rita and Nick Ellis. 2010. “An Academic Formulas List: New Methods in Phraseology Research.” *Applied Linguistics* 31:487-512.

Thurmair, Gregor. 2003. “Making Term Extraction Tools Usable.” In *Proceedings of the Joint Conference of the 8th Workshop of the European Association for Machine Translation and the 4th Controlled Language Applications Workshop.* Dublin: European Association for Machine Translation.

Vivaldi, Jordi and Horacio Rodriguez. 2007. “Evaluation of terms and term extraction systems - A practical approach.” *Terminology* 13(2):225-248.

Vivaldi, Jordi, Lluis Màrquez, and Horacio Rodríguez. 2001. “Improving Term Extraction by System Combination Using Boosting.” In *Machine Learning ECML 2001*, edited by Luc de Raedt and Peter Flach, 515-526. Series: Lecture Notes in Computer Science, vol. 2167. Springer.

Wermter, Joachim and Udo Hahn. 2005. “Paradigmatic Modifiability Statistics for the Extraction of Complex Multi-Word Terms.” In *Proceedings of the Human Language Technology Conference and the Conference on Empirical Methods in Natural Language Processing*, 843-850. Association for Computational Linguistics.

Wiechmann, Daniel. 2008. “On the Computation of Collostruction Strength: Testing Measures of Association as Expressions of Lexical Bias.” *Corpus Linguistics and Linguistic Theory* 4 (2):253-290.

Wong, Wilson, Wei Liu, and Mohammed Bennamoun. 2007. “Determining termhood for learning domain ontologies using domain prevalence and tendency.” In *Proceedings of the Sixth Australasian Conference on Data Mining and Analytics*, edited by Peter Christen, Paul Kennedy, Jiuyong Li, Inna Kolyshkina and Graham Williams, 47-54. Australian Computer Society.

Zhang, Ziqi, José Iria, Christopher Brewster, and Fabio Ciravegna. 2008. “A Comparative Evaluation of Term Recognition Algorithms.” In *Proceedings of the Sixth Language Resources and Evaluation Conference (LREC 2008)*, Marrakech, Morocco.