

Term extraction: A Review

Draft Version 091221

Lars Ahrenberg

Linköping University
Department of Computer and Information Science
E-mail: lah@ida.liu.se

1 Introduction

This is a review of the current state-of-the-art in term extraction. We consider both mono-lingual and multi-lingual term extraction, and single-word as well as multi-word terms. The review is restricted to methods, however, and so does not make any attempt to cover commercial systems. Readers interested in commercial systems are referred to Zielinski & Ramirez Safar (2005) and the continuously updated Compendium of Translation Software¹.

First, however, we delimit the scope of the paper and propose a working definition of term extraction.

2 On terms and term extraction

In classical terminology a *term* is defined as the expression (or label, or representation) of a concept². This seems to imply that we cannot approach the problem of term extraction without knowing what a concept is, or whether a certain text segment is the expression of a concept or not. However, this is a situation we certainly wish to avoid as we only seldom have access to concepts and don't want to adhere to any particular concept definition. Following Jacquemin & Borigault (2003) we prefer a more pragmatic approach leaving definitions to end users.

In corpus-based computational terminology the output from a term extraction process may serve different purposes: construction of ontologies, document in-

¹Available at <http://www.hutchinsweb.me.uk/Compendium.htm>

²"a term is the designation of a defined concept in a special language by a linguistic expression. A term may consist of one or more words." (ISO 1087)

| | Prior terminological data | No prior terminological data |
|-------------------------|----------------------------------|-------------------------------------|
| Term discovery | Term enrichment | Term acquisition |
| Term recognition | Controlled indexing | Free indexing |

Table 1: A view of term-based NLP according to Jacquemin & Borigault (2003).

dices, validation of translation memories, and even classical terminology work. Thus, the exact definition of terms must be subordinated to the purpose at hand. What is common to the different applications, however, is the constructive nature of the activity, and the need to distinguish terms from non-terms, or, if we prefer, domain-specific terms from general vocabulary (Justeson & Katz, 1995).

Usually, the automatic process of term extraction will be followed by a manual, though often computer-aided, process of *validation*. For this reason, the outputs of a term extraction process are better referred to as *term candidates* of which some, after the validation process, may be elevated to term status. To support the validation process the output from extraction is not just an unordered list of candidates, but a list where each candidate has received a score for *termhood*. Candidates that receive a score above some threshold can then be sent for validation as a ranked list.

Besides ranking, it is often useful to *sort* the term candidates according to some measure of similarity. In the monolingual case complex terms can be sorted in terms of common parts, in particular sharing of heads, while in the bilingual case sorting can be done on the basis of term pairs that share the same source term, or the same target term. Starting with a set of term pairs formed on this basis, the process may be continued until a cluster of related term pairs have been obtained.

Instead of *term extraction* terms such as *term recognition*, *term identification* and *term acquisition* are also in common use. We see all these as synonymous. They should all be distinguished from *term checking* and *term spotting* which assumes that a list of (validated) terms is available, possibly including prohibited ones, which are searched for in a set of documents.

Jacquemin & Borigault (2003) proposes a division of term-based NLP into four sub-domains, as depicted in Table 1. Using that division this review is concerned with *term acquisition*. Note also that *term recognition* is used differently by these authors as restricted to indexing. If prior terminological data exists, it can be used in various ways to aid or constrain the extraction process (Valderrabanos et al., 2002), or be applied as a filter in validation, but we see the existence of prior terminological data as a minor parameter in the process. An overview of the different parts of the term acquisition process, as used in this review, is given in Figure 1.

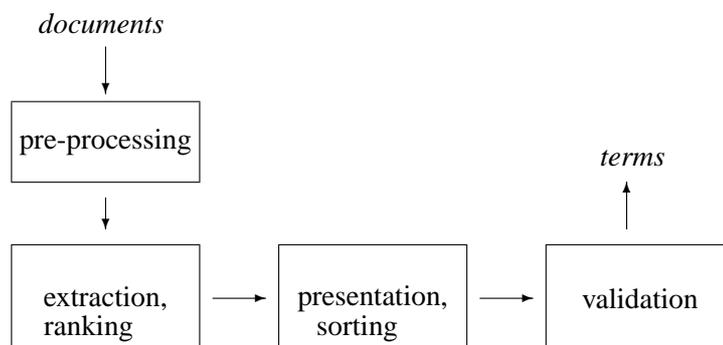


Figure 1: Four modules of the term extraction process

3 Term characteristics

While the defining properties of terms are dependent on the intended use, there are nevertheless a number of properties that algorithms for term extraction can exploit. These properties can be classified as linguistic, statistical and distributional.

Properties may also be classified in terms of their target concept. It is argued that a term is a linguistic unit of some sort that enters into syntactic relations with other units and is subject to inflections. Thus, one class of properties are those that can be used to define *unithood*. However, there are other types of units than terms so another class of properties are those that are more specifically geared towards the recognition of *termhood*, by Kageura & Umino (1996) defined as “the degree to which a linguistic unit is related to domain-specific context”. When we speak of a term candidate as having recognizable translational correspondents in the other half of a parallel corpus, we may further introduce a notion of *translatability*, meaning that the unit can consistently be associated with units of the other language. This property is arguably a reflection of the possibility that the candidate designates a well-defined and more or less language-independent concept. The notion of translatability naturally extends to the multilingual case.

A notion not covered in this review is *centrality*, the degree to which a term can represent a group of terms that belong to a specific sub-domain. This notion is relevant when one wants a list of terms that together cover all or most of the sub-domains of a document (Zhang et al. (2009)).

3.1 Linguistic properties of terms

Most work on term extraction focus on noun phrases or noun groups and for good reasons: most terms tend to be nominal. However, verbs and adjectives, though they have received much less attention, can also be domain-specific.

Thus, for a given language (and application) one wants to capture the linguistic structure of terms. For instance, Justeson & Katz (1995) structure English terms as in (1), while (2) has been used for French (Daille, 2000):

- (1) $((Adj|Noun)^+|((Adj|Noun)^*(NounPrep)^?)(Adj|Noun)^*)Noun$
(2) $Noun1(Adj|(Prep(Det)^?)^?Noun2|VInf)$

Obviously, such linguistic patterns can only be applied to a corpus that has been tagged for parts-of-speech in a pre-processing phase.

While phrase patterns such as these help to capture terms, they also capture many non-terms. For this reason, they are indicators of unithood rather than termhood. For instance, it has been observed that some words, in spite of being nouns or adjectives, seldom appear in terms. Examples are adjectives such as *following* or *interesting* and nouns such as *thing*, *kind*. These words are collected in *stop lists*.

In addition to patterns referring to parts-of-speech the relations between the different parts of a pattern can be relevant. A complex term can very often be analysed in terms of a head with one or more modifiers (Hippisley et al., 2005). In English, for example, the head is usually the last noun not following a preposition as in *database manager, non-European Union national* or *flush bolt with fliplock*. In the French patterns in (2) Noun1 is usually the head while other parts are modifiers.

Another linguistic aspect of interest is morphological structure. In Swedish, unlike in English, compounds are written as one orthographic word and since it is common to use compounds for terms we need to recognize compounds as such and preferably also the inner structure in terms of head and modifier. Such compounds may also have more than two parts as in Swedish *hjärnhinneinflammation* (meningitis), *katodstrålerör* (cathode ray tube). Moreover, words of Latin or Greek origin occur more often in terms than in ordinary words and for some domains affixes such as *meta-*, *orto-*, *-graf*, *-logi* are characteristic morphemes for terms.

Base forms, or *lemmata*, are of limited use for extraction, but are quite useful for presentation and validation, as they reduce the number of candidates that the validator need to inspect or else can be used to sort inflected variants. As with parts-of-speech, the recognition of base forms requires linguistic preprocessing.

Apart from their inner structure of words and word groups, terms have context associations. The context associations may concern words in the neighborhood of a term or so called rich contexts which are typical patterns of clauses and sentences

where terms are commonly used such as definitions and exemplifications. The first kind of context is often represented by vector spaces whereas the second kind is captured by regular expressions. An example, from Pearson (1998), is given in (3). An instance of this formula, with *is* as the connective verb, is *a cucumber is a vegetable used for...*

(3) $indef\ Article? + term + connective_verb + (indef\ Article|def\ Article)? + (term|classword) + past_participle$

To ease term standardization tasks it can be important to recognize morpho-syntactic patterns of variation in the linguistic structure of synonymous term candidates. With the aid of such patterns all variant forms of a term candidate can be presented to the validator as a cohort. For instance, Daille (2000) shows that what she calls relational adjectives, i.e., adjectives that have a derivational relation to a noun taken to be a term, are quite commonly used as alternatives to the nouns in term formation. She gives examples such *produit laitier* (diary product) vs. *produit de lait* (product of milk), and *débit horaire* (hourly rate) vs. *débit par heure* (rate per hour). We note in these examples that the head is constant while the variation resides in the modifier.

3.2 Statistical properties of terms

A basic statistical property of terms is their frequency in a corpus. This frequency may then be compared to the frequency of the term in other corpora, such as balanced corpora or corpora from other domains. We will use the representation f_A for the frequency of some unit or event A occurring in a corpus, and N for the size of the corpus.

Basic frequency counts are combined to compute co-occurrence measures for words. Co-occurrence measures are used both to estimate the propensity for words to appear together as multi-word units in documents, and to estimate the likelihood that units on either side of a bilingual corpus correspond under translation. Common co-occurrence measures are *the Dice coefficient*, *Pointwise Mutual Information* (PMI) and *Log-Likelihood Ratio* (LLR), as defined below:

$$Dice = \frac{2 \times f_{AB}}{f_A + f_B}$$

$$PMI = \log f_{AB} - (\log f_A + \log f_B)$$

$$LLR = \log L(f_{AB}, f_A, f_B/N) + \log L(f_B - f_{AB}, N - f_A, f_B/N) - \log L(f_{AB}, f_A, f_{AB}/f_A) - \log L(f_B - f_{AB}, N - f_A, (f_B - f_{AB})/N - f_A)$$

Here, AB is used for the simultaneous occurrence of A and B . $L(k, n, x)$ is the function $x^k(1 - x)^{n-k}$.

Other relevant frequency information concerns *nested terms* whereby a term candidate overlaps with other, longer or shorter, term candidates. Such situations are actually quite common. For example, Frantzi et al. (1998) notes expressions such as *floating point* and *floating point arithmetic* which are both terms, while the bigram *point arithmetic* is not. For cases such as these the following frequencies are relevant:

- The frequency of a term candidate as a substring of another candidate,
- The frequency of a term candidate as a modifier substring or a head,
- The number of longer candidate terms of which a candidate is a part,
- The length $|a|$ of a term candidate a (in number of words).

As we will see, these basic frequencies can be combined into measures of termhood in various ways. For instance, a high frequency for a candidate as a substring of other candidates is a good indicator of termhood, while a low frequency is not.

3.3 Distributional properties of terms

There are a number of distributional properties of terms that are of interest. One concerns their distribution within documents where Justeson & Katz (1995) points out that they tend to occur by spurts and, when repeated, are repeated as wholes. Another distributional property concerns their distribution across documents in a corpus (prevalence), and a third their distribution in a domain-specific corpus as compared to their distribution in a general corpus or in a contrastive corpus, i.e., a corpus with documents from other domains (specificity).

Distributional properties may be seen as a special case of statistical properties since their measurement tends to be based on counts. A common measure, much used in information retrieval is *tf-idf*. Here, *tf* stands for *term frequency* (in a given document) and *idf* stands for *inverse document frequency* measuring the spread of a term through the entire document collection. *idf* is the inverse of the proportion of documents where the term t_i occurs, usually scaled to logarithmic scale:

$$idf = \log\left(\frac{N}{n_i}\right)$$

Here, N is the total number of documents in a collection, and n_i the number of documents in which t_i occurs. Combining term frequency and inverse document frequency the *term weight* of a word for a document is defined as the product

$tf \times idf$. While, in information retrieval, the $tf-idf$ is primarily used to rank documents, it can also be used to rank words and word sequences of a document as term candidates for (the domain of) the document. Arguably, a high frequency and a high degree of concentration of a term to a given document speaks in favour of its being document-specific.

A simple metric that directly compares the distribution of a term in a domain-specific corpus to its distribution in a general corpus is *weirdness*. Weirdness is the ratio of the relative frequencies of the term in the two corpora, with the relative frequency of the domain-specific corpus appearing as numerator Ahmad et al. (2007). Using D for the domain-specific corpus, G for the general corpus, N for corpus size, and f for absolute frequency, the definition is

$$\text{Weirdness} = \frac{f_D \times N_G}{f_G \times N_D}$$

4 Monolingual term extraction

Early work on term extraction from monolingual corpora was generally based on linguistic properties and restricted to noun phrases. Noun-phrases could be identified from part-of-speech patterns, from phrase border identification or be parsed by a grammar of some sort. An example is the LEXTER system for multi-word noun phrases (Bourigault, 1993). Single-word terms have also been recognized on the basis of patterns or grammars, e.g., Ananiadou (1994) for analysis of English single-word terms. In addition, stop-lists would be used for filtering, i.e., words or morphemes that are assumed never to be part of a term.

Justeson & Katz (1995) in the TERMS system combined their noun phrase patterns with stop lists and a simple frequency threshold set to 2 by default. The ACABIT system (Daille (1996)) was among the first to rank term candidates on the basis of statistical measures. Daille found absolute frequency as well as log-likelihood ratio to be good measures for ranking candidates.

4.1 The C-value/NC-value method

A more elaborate combinations of linguistic and statistical features were used by Frantzi et al. (1998) in their so called *C-value/NC-value Method* for multi-word terms. The C-value is a basic measure of termhood for a candidate a with frequency f_a and length $|a|$ that in addition refers to the set of candidate terms that contains a as a proper sub-part (T_a), and the cardinality of that set ($|T_a|$).

$$C = \log_2 |a| f_a - \frac{1}{|T_a|} \sum_{b \in T_a} f_b$$

The idea is to subtract from the basic, candidate-specific score based on frequency and unit length the average frequency of longer candidates of which the given candidate is a part. Note that if a is maximal the second term will be zero.

The C-value is only applied to multi-word strings that have passed the linguistic filter. The linguistic filter is based on part-of-speech tags and a stop-list; it can be varied so as to balance precision against recall.

We can observe that the C-value will be high for long non-nested strings that have a high absolute frequency in the studied corpus. On the other hand, non-maximal candidates that are part of longer candidates with high frequencies will get a low C-value.

The C-value ranking is but a first stage of the method, however. In a second stage a set of *term context words* are identified. The idea is that terms are not only characterized by their internal structure and frequencies but by occurring in typical term contexts. Each term context word is given a weight depending on the number of terms it appears with and this weight is defined as $\frac{t_w}{T}$ where t_w is the number of terms appearing with the word and T is the total number of terms considered. On the basis of the term context words and their weights a *cumulative context weight*, CCW , can be calculated for each candidate term, as below³:

$$CCW = \sum_{b \in C_a} b_w f_{b/[a]}$$

Here C_a is the set of context words with a positive weight for a , and $f_{b/[a]}$ is the frequency of b as a context word for a with weight b_w .

To identify term context words the method requires validation of a set of term candidates from the top of the list ranked by C-value. The term context words are then selected by looking at verbs, adjectives, and nouns that can be found in a small window around the validated terms. When a set of term context words are at hand, the set of term candidates ranked by C-value can be re-ranked. This is done by considering, for each term candidate, its NC-value, defined as follows:

$$NC(a) = 0.8C(a) + 0.2CCW(a)$$

Thus, the NC-value is a modification of the C-value that replaces part of the C-value with a score based on context words. The weights 0.8 and 0.2 have been set empirically according to Frantzi et al. (1998).

The C-value/NC-value method was first applied to a corpus of eye-pathology medical records of some 810,000 words. On this corpus the NC-value was shown

³The definitions of the cumulative context weight as well as of the weight of a term context word are different in other papers by the same authors.

to give higher precision than the mere C-value and both of them had higher precision than absolute frequency. In later work (Maynard & Ananiadou, 2000) attempts have been made to integrate semantic similarity as part of the contextual information.

4.2 Contrastive weight

The C-value/NC-value approach uses evidence from a single domain only. If a contrastive corpus is available other measures can also be employed. Basili et al. (2001) proposed a measure called *contrastive weight* (CW) to take the distribution of a term candidate with respect to in-domain and out-of-domain into account. Before introducing the definitions some notation is needed:

- D is the target domain,
- TC is the set of term candidates $\{a_1, a_2, \dots, a_n\}$,
- f_{TC} is the sum of the frequencies of all $a_i \in TC$ in all domains,
- $f_{a|D}$ is the frequency of candidate a in domain D ,
- $f_{a|\neg D}$ is the frequency of candidate a in the contrastive domains for D ,
- For a complex term candidate, a^h is its head and M_a its set of modifiers,
- m_a is the set of modifiers of a complex term a that also happen to be a candidate term (i.e. $m_a = M_a \cap TC$),
- MF_a is a *modifier factor* applied to modifiers in the set m_a , which will be defined below.

For simple terms contrastive weight is defined as follows:

$$CW(a) = \log f_{a|D} \left(\log \frac{f_{TC}}{\sum_J f_{a|J}} \right)$$

The contrastive weight of a complex term candidate is computed with the aid of its frequency and the contrastive weight of its head:

$$CW(a) = f_{a|D} \times CW(a^h)$$

Thus, for different complex terms sharing the same head, the contrastive weight of the head will be the same and the ranking will depend on the candidate's own

frequency. For complex term candidates of the same in-domain frequency, the contrastive weight of their heads will determine their rank. Wong et al. (2007b) notes that this approach takes the linguistic structure of a complex term into account, but criticizes it for not going far enough. In particular, the distribution of modifiers on the target domain and contrastive domains is not considered.

4.3 Discriminative weight

Just as Basili et al. (2001), Wong et al. (2007b) considers contrastive measures to be of importance, but they propose a finer measure, called *discriminative weight* (*DW*). This measure, in turn, is a product of two factors called *domain prevalence* (*DP*) that applies to in-domain usage of a term, and *domain tendency* (*DT*) that applies to the extent of inclination of term candidate usage towards the target domain. The idea is that a high in-domain frequency is not sufficient as an indicator of termhood in the domain, while having both a high domain prevalence and a high domain tendency is. For a simple term candidate a the domain prevalence is defined as:

$$DP(a) = \log_{10}(f_{a|D} + 10) \log_{10} \left(\frac{f_{TC}}{f_{a|D} + f_{a|\neg D}} + 10 \right)$$

This definition guarantees that $DP(a)$ is always larger than 1, that it grows with higher in-domain frequencies and is reduced with high frequencies in contrastive domains. For a complex term candidate the DP is defined as:

$$DP(a) = \log_{10}(f_{a|D} + 10) \log_{10} DP(a^h) MF_a \text{ where}$$

$$MF_a = \log_2 \left(\frac{\sum_{m \in m_a} f_{m|D} + 1}{\sum_{m \in m_a} f_{m|\neg D} + 1} + 1 \right)$$

This definition also give values larger than 1, grows with candidate in-domain frequency, with the DP of the head, and with the modifier factor. The idea of the modifier factor is to give higher scores to complex terms where the head is ambiguous while the modifier is disambiguating, e.g. in examples such as *H5N1 virus*, *computer virus*. Here, *H5N1* is a stronger signal of the medical domain than the word *virus*.

The modifier factor is actually very similar to the *domain tendency* measure, $DT(a)$, defined as

$$DT(a) = \log_2 \left(\frac{f_{a|D} + 1}{f_{a|\neg D}} + 1 \right)$$

We can see that for a term candidate that has the same frequency in the target

domain as in a contrastive domain, the DT will be close to 1, while it will be much larger than 1 if the in-domain frequency is higher than the contrastive frequency. For measuring termhood, Wong et al. (2007b) applies the DT as a weight on the DP . This combined measure is the *discriminative weight* (DW) of a term candidate a , and is their basic measure of termhood:

$$DW(a) = DP(a)DT(a)$$

To allow contextual factors to enter the measures they modify the DW by a measure called the *adjusted contextual contribution* (ACC). This measure refers to a set of context words, C_a , and employs a similarity measure between words in C_a and the candidate a itself based on the so called *Normalized Google Distance* (NGD). Basically the similarity $sim(a, c) = 1 - wNGD(a, c)$ where w is a factor to scale the distance. The exact definition of ACC is omitted here. The final *termhood measure* (TH) is:

$$TH(a) = DW(a) + ACC(a)$$

The contribution of contextual factors is treated in a very similar fashion to the C-value/NC-value approach. It should be noted, though, that the impact of context words is much smaller for TH than it is for the NC-Value. Wong et al. (2007b) end their proposal with a comparison of the behavior of three measures, NC-Value, Contrastive weight, and TH, on the same set of 5,156 term candidates generated from some medical documents. They note the following interesting differences:

- NC-Value does not use a contrastive corpus and so place candidates high irrespective of domain tendency. Ranking is also much influenced by context words and domain frequency giving high ranks to simple terms and short complex terms.
- Contrastive weight tends to place low frequency candidates high. These are often complex term candidates that derive their high score from their heads and the fact that they tend not to occur in the contrastive corpus. Also, terms with the same head are grouped together irrespective of the modifier.
- TH demotes candidates with a low domain tendency and get high frequency candidates at the top of the ranking. This latter fact is in accordance with early findings that high frequency is a good indicator of termhood. Context words are taken into account but has a smaller impact on the score than they have for the NC-Value.

In addition, they show that the two measures NC-Value and the Contrastive weight have a higher standard deviation than TH , a fact which they take to speak in favour of the TH .

4.4 Odds of termhood

Wong et al. (2007a) is an attempt to derive a probabilistic measure for termhood on a principled mathematical basis. The measure, *Odds of termhood*, OT , is proposed as an alternative to TH .

The derivation of OT uses the notion of *relevance* from information retrieval. In the case of term extraction we can't speak of relevance in relation to a query, but we can speak of relevance in relation to evidence, where evidence refers to the typical properties of terms, as described in section 3. Evidence can be represented by an evidence vector $V = \langle E_1, E_2, \dots, E_m \rangle$. The event that a term candidate a is relevant to a domain d is called R_1 while the event that it is relevant to a contrastive domain is called R_2 . The event that a candidate a is supported by an evidence vector is called A . What we then first want to know is the probability that A implies R_1 which can be expressed using Bayes' rule as

$$P(R_1|A) = \frac{P(A|R_1)P(R_1)}{P(A)}$$

Next, the authors set out to answer the question "What are the odds of a candidate a being relevant to d given its evidence?", which, by definition, is

$$O = \frac{P(R_1|A)}{1-P(R_1|A)} = \frac{P(A|R_1)P(R_1)}{P(A)(1-P(R_1|A))}$$

Now, if the contrastive corpus can be taken as a representative collection of documents from domains where a is not relevant, we are allowed to replace $(1 - P(R_1|A))$ with $P(R_2|A)$ which gives

$$O = \frac{P(R_1|A)}{P(R_2|A)} = \frac{P(A|R_1)P(R_1)}{P(R_2|A)P(A)} = \frac{P(A|R_1)P(R_1)}{P(A|R_2)P(R_2)}$$

This is actually a measure that can be used to rank term candidates, a measure that expresses a kind of odds of relevance. Taking logarithms of both sides and moving the non-contingent odds factor to the other side, we derive

$$\log \frac{P(A|R_1)}{P(A|R_2)} = \log \frac{P(R_1|A)}{P(R_2|A)} - \log \frac{P(R_1)}{P(R_2)}$$

Without any evidence the chance that a is relevant to d can be assumed to be equal to the chance that it is not relevant. Thus the non-contingent odds factor

$P(R_1)/P(R_2)$ is set to 1, and the last logarithm above is zero. This yields the *Odds of termhood* as

$$OT(a) = \log \frac{P(A|R_1)}{P(A|R_2)}$$

In order to make $OT(a)$ computable we assume independence of the elements in the evidence vector so that for $k = 1, 2$

$$P(A|R_k) = \prod_i P(E_i|R_k)$$

Substituting this into the definition of $OT(a)$ we derive:

$$OT(a) = \sum_i \log \frac{P(E_i|R_1)}{P(E_i|R_2)}$$

Thus, given training data in the form of a domain corpus and a contrastive corpus, and a set of individual evidences (features), $OT(a)$ can be estimated for any term candidate.

Finally, for ease of writing, each individual evidence E_i can be taken to supply an individual score, S_i , where

$$S_i = \frac{P(E_i|R_1)}{P(E_i|R_2)}$$

so that $OT(a) = \sum_i \log S_i$.

Wong et al. (2007a) claims several benefits for this measure in comparison with earlier ones. First, it is derived in a sound probabilistic framework with explicit assumptions (most of which have not been recapitulated here). Second, the individual evidences can also be formulated in a similar probabilistic framework, and, third, the empirical ranking correlates well with their earlier best ad-hoc-measure, TH . Moreover, it has an even smaller standard deviation than TH , about one half of the mean, while TH has a standard deviation in the interval 1.5-3 times the mean.

In the paper the correlation of OT with TH is shown only for two types of evidence, the specificity of the head of a complex term candidate with respect to the domain, and the uniqueness of a term candidate with respect to the domain. The argument for the first property is that, if the head of a complex term is specific to a domain, all complex candidates with that head are more likely to be terms. The argument for the second property is that a candidate that does not occur at all in the contrastive domain is more likely to be a term.

4.5 TermExtractor

TermExtractor (Sclano & Velardi, 2007b,a) is a system primarily developed for the purposes of ontology creation. It is actually available online⁴.

After pre-processing of documents TermExtractor performs chunking and proper name recognition, and then proceeds to identify typical terminological structures based on linguistic patterns, including stop words, detection of misspellings and acronyms. These units are then filtered based on three distributional factors. The first, *Domain Pertinence*, or *Domain Relevance*, requires a contrastive corpus and compares the occurrence of a candidate in the documents belonging to the target domain to its occurrence in other domains. The measure as such only depends on the contrastive domain where the candidate has the highest frequency, though:

$$DR_{D_i}(t) = \frac{tf_i}{\max_j(tf_j)}$$

This measure should be compared with the domain tendency measure *DT* defined in the previous section, which, however, is more elaborate. The next factor, *Domain Consensus*, assumes that a domain is represented by several documents. It measures the extent to which the candidate is evenly distributed on these documents by considering normalised term frequencies (ϕ):

$$DC_{D_i}(t) = - \sum_{d_k \in D_i} \phi_k \log \phi_k$$

Here, we assume k distinct documents for the domain D_i .

The third factor is called *Lexical Cohesion* and compares the in-term distribution of words that make up a term with their out-of-term distribution. The definition is:

$$LC_{D_i}(t) = \frac{n \cdot tf_i \cdot \log tf_i}{\sum_j tf_{w_{j^i}}}$$

In computing these three different measures, the counting of frequencies may be modified by taking into account factors such as orthographic emphasis, occurrence in the document title, and others. Whether these are used or not, the final weight of a term is computed as a weighted average of the three filters above:

$$score(t, D_i) = \alpha \cdot DR + \beta \cdot DC + \gamma \cdot LC$$

The default is that $\alpha = \beta = \gamma = \frac{1}{3}$.

⁴Downloads and access possible at the web page <http://lcl2.uniroma1.it/termextractor/>

5 Evaluation

Evaluation of term extraction systems is quite problematic since the exact definition of a term depends on external criteria. Gold standards are hard to find, and even when they can be found, their relevance is often limited to a single application (cf. Thurmair (2003)). Still, existing dictionaries and ontologies are often used for evaluation, or else human judges are called in to help.

Evaluations of term recognition systems tend to focus on precision. This is a necessity when human judges are used, since it is simply too much work to identify all terms in the test data. Moreover, it is rarely the case that the full list of outputs from a system is evaluated. Instead evaluation is restricted to some sections of the output. Precision is then measured as the proportion, or percentage, of the examined candidates that are judged as terms, or could be found in the resource.

$$P = \frac{\text{no.of.terms}}{\text{no.of.candidates}}$$

Since precision tends to vary with the frequency of the term candidate it is also common to calculate precision with respect to a certain frequency interval.

An alternative measure which to some extent can capture recall, given a gold standard, is *Un-interpolated Average Position (UAP)* attributed to Schone & Jurafsky (2001), and defined as follows:

$$UAP = \frac{1}{K} \sum_{i=1}^K P_i$$

Here, P_i denotes “precision at i ”, where i is a certain number of correct terms. It is computed as the ratio $1/H_i$ where H_i is the number of candidates required to reach i correct terms. It is common that P_i will decrease with increasing i , but the UAP will form an average over all i :s considered. In case a gold standard is given, we can speak of the i -th correct term as a given.

Other common measures are *noise* and *silence*. Noise is the dual of precision, but measuring the proportion of non-usable candidates rather than the proportion of useful ones. Silence is the dual of recall and is equally difficult to measure.

In a comparative evaluation, Zhang et al. (2008) used two corpora, one on animals extracted from Wikipedia, and the other the GENIA corpus that contains 200,000 abstracts from the MEDLINE database (Kim et al., 2003). For the Wikipedia corpus three human judges were asked to review the top 300 candidates proposed by each system; for the GENIA corpus the existing annotations in the corpus were used to extract a gold standard of terms.

Zhang et al. (2008) compared six different systems/algorithms capable of recognizing both single and complex term candidates with *tf-idf* as the baseline. The

other measures were weirdness, cf. 3.3, C-value, cf. 4.1, Glossex⁵, and TermExtractor, cf. 4.5. They also defined an algorithm of their own based on voting, which computed a new ranking as a weighted average of the ranks of the other systems. Results were different on the two corpora. On the Wikipedia corpus, the voting algorithm gave a clear improvement, while TermExtractor outperformed the other algorithms on both counts. On the GENIA corpus, however, results were mixed with both *C-value* and *tf-idf* having better results than TermExtractor.⁶

It seems to be a general truth that results vary a lot with the corpora and evaluation methods used. For a different Wikipedia corpus, Hjelm (2009) found precision values as low as 12-13% while in Zhang et al. (2008) they are around and above 90%.

6 Multilingual term extraction

Multilingual term extraction is term extraction from two or more languages with the purpose of creating or extending a common resource such as a bilingual term bank, or a domain ontology. Sometimes, the term 'multilingual term extraction' is even used for knowledge-poor methods that are applied to several languages simultaneously with no intention to create a common resource, e.g., Hippisley et al. (2005). This, we think, is not quite appropriate.

As in the monolingual case, there may be different purposes to term extraction when it is multilingual, and the purpose will ultimately determine the criteria for termhood. As a matter of fact, a purpose may also be just to support monolingual term extraction. Then translatability is used as evidence for termhood in the target language. A general difference, though, is that the task is not just to extract term candidates, but to pair terms that correspond across languages.

Multilingual term extraction require that documents are available in the languages of interest. In the worst case these documents may be unrelated to one another, but they may also be related in systematic ways. If the documents consist of originals and translations, or of translations from the same source, we say that they are *parallel*; if they can be related through topics, we say that they are *comparable*. In this review we will primarily be concerned with extraction from parallel bilingual corpora.

⁵Glossex, Kozakov et al. (2004), is similar but not as elaborate as TermExtractor.

⁶As weirdness is just measuring unithood and Glossex is similar to TermExtractor, we have only described the latter here.

6.1 Translatability

Translatability may be determined for a given list of terms in a process usually called *term spotting* or *translation spotting*. The assumption is then that we know the relevant terms for one language, but wish to find equivalent terms in the other language.

Translatability may be determined for any linguistic unit, whether a single word or a complete phrase. The common term for this type of process in natural language processing is (*word*) *alignment*. The currently most widely used system for word alignment is Giza++ (Och & Ney, 2003), based on the EM-algorithm. Phrasal alignments, i.e., alignments of connected word sequences, can be generated from word alignments in a straight-forward manner.

As a consequence, there are two basic approaches to bilingual term extraction from parallel corpora. You can either first determine term candidates for one or both languages using monolingual methods and then pair them together, or you can align words and word sequences of the parallel corpus to determine correspondents and then classify the set of corresponding pairs into terms and non-terms. These two approaches are illustrated in Figure 2.

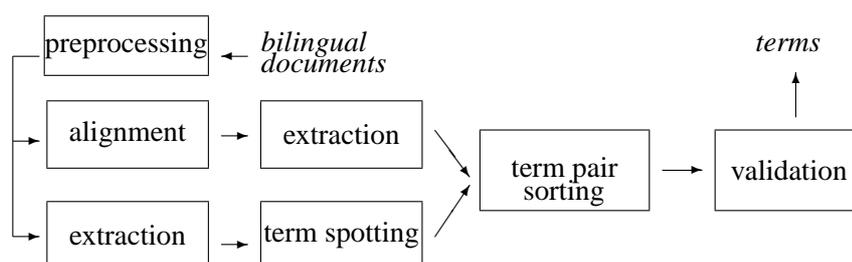


Figure 2: Variants of bilingual term acquisition

6.2 Extraction + Translation spotting

Given a parallel corpus we can start by applying any method of our choice to identify term candidates on both halves of the corpus, and then try to align them. With access to an existing monolingual term bank we can even assume that the terms are known and just mark them in the corpus. The alignment task is then reduced to the task of *term spotting*, i.e., identify the most probable equivalents in the other language to the terms that we know. Hjelm (2007) uses this method in a study where he wanted to compare the performance of statistical alignment to distributional association measures for this task. For statistical alignment he used Giza++ (Och &

Ney, 2003) with standard settings. Two association measures were tried: cosine and mutual information and the former was also used with dimensionality reduction (random indexing). The parallel corpus used was JRC-Acquis (Steinberger et al., 2006) for four languages, English, Spanish, German, Swedish, and thus for twelve translation directions. The term bank used was Eurovoc V4.2 where terms exist for almost all EU languages.

The methods were evaluated on terms covering all frequency ranges, from 1 to more than 50,000 occurrences in the training data. A general conclusion was that Giza++ outperformed all distributional measures for almost all frequency ranges. However, it was shown that an ensemble method combining both types of measures could perform better than Giza++ alone.

Thurmair (2003) gives a short overview a system called **BiExtract**, apparently proprietary software of the German firm Compendium GmbH. Input to the system is a file of monolingual terms and a translation memory in either Ascii-format or TMX. Linguistic normalisation is applied to both halves of the translation memory. The system is largely based on cooccurrence statistics but treats multiword candidates on a par with single word candidates and takes position as well as orthography into account. The output is a ranked list of alternatives, including examples in context and a range of information parameters that the user can set.

6.3 Alignment + Extraction

The examples of this kind of approach are few, but there are some. Merkel & Foo (2007) describes a method for generating term candidate pairs from linguistically analysed sentence-aligned parallel texts. The linguistic analysis includes lemmatization, part-of-speech tagging as well as dependency-based syntactic analysis. Moreover, to improve quality of alignment, it is tailored to the data at hand in an interactive pre-process. This process generates a core lexicon and positive as well as negative alignment data for correspondences at the levels of lemmas and parts-of-speech that are (or can be) used by the automatic word aligner that performs the bulk of the work.

A distinctive factor of this approach is that a full alignment of the given parallel text is performed before any term candidate data is extracted. Extraction is based on pairs, including pairs of multi-word units, and may be restricted to certain linguistic patterns. Ranking of the pairs is based on a statistical score. Merkel & Foo (2007) compares three different scores: (1) a baseline based on absolute frequency of the pair, (2) the Dice coefficient, and (3) a measure introduced by Deléger et al. (2006) called the Q-value. In contrast to common co-occurrence measures the Q-value refers to the fanout of a candidate, i.e., the number of candidates of the other language to which it is found to correspond under alignment. We will denote the

fanout of a unit U as g_U .

$$Q = \frac{f_{ST}}{g_S + g_T}$$

Here, S and T represents a source unit and a target unit, respectively, and ST aligned pairs of them. Dice, in comparison, is

$$Dice = \frac{2 \times f_{ST}}{f_S + f_T}$$

Thus, while Dice is inversely proportional to the sum of the individual frequencies for S and T , the Q-value is inversely proportional to the sum of their fanouts in the corpus. The paper shows that the Q-value and Dice both outperform the baseline and are largely equivalent for high-frequency candidates. Dice, however, ranks low-frequency candidate pairs too high as it fails to distinguish pairs for which $f_S = f_T = f_{ST}$. Given that two pairs have the same fanout, the Q-value, however, will always rank high-frequency pairs higher.

The (inverse) fanout and the Q-value can both be seen as measures of consistency in translation for a term candidate in either language. It is not clear, however, that Q-value would outperform more elaborate measures combining co-occurrence with absolute frequency.

Lefever et al. (2009) also align before testing for termhood. Monolingual processing uses a lemmatizer, a part-of-speech tagger, and a chunker. Chunking exploits the part-of-speech tags and distituecy rules. The alignment is performed in phases where the first phase uses an implementation of IBM model 1 to create alignments for single words. Safe links of content words are stored in a bilingual lexicon. For the alignment of chunks a lexical link matrix is first constructed based on the bilingual lexicon and word tokens that are identical for the two languages. With information from the lexical link matrix the alignments of single tokens are extended in both directions thus yielding chunk pairs. The chunk pairs are then subject to similiraity checks that controls for lexical links and correspondences at the part-of-speech level. These chunks are referred to as *anchor chunks*.

In the next phase function words that are adjacent to anchor chunks are linked and finally chunks that have aligned anchor chunks as left and right neighbors are linked. The similarity test was applied to these chunks as well with a somewhat lower threshold for the percentage of lexical links.

For the experiment a French automotive corpus and its translation into three different languages, English, Italian, and Dutch was used. Thus there were three bilingual parallel corpora to test on.

The generation of term candidates also involves several steps. In a first step all anchor chunks and there lexically linked parts are considered term candidates. In a

following step chunks satisfying the part-of-speech pattern NP+PP are considered to be candidates, and then, chunks that can be constructed from the anchor chunks by stripping off adjectives. These candidates are then filtered by the application of several tests. As a test for termhood, the Log-Likelihood metric was applied comparing the distribution of a candidate in the project corpus with its distribution in a general corpus, in this case French newspaper text.

To measure unithood of multiword chunks they applied a metric called the Normalised Expectation, NE, defined as

$$NE = \frac{p(n\text{-gram})}{\frac{1}{n} \sum p(n-1\text{-gram})}$$

The idea is that NE “expresses the cost, in terms of cohesiveness, of the possible loss of one word in an n-gram” (*op.cit.* p. 501). NE is then multiplied with the frequency of the n-gram to obtain what is called the Mutual Expectation, ME. The threshold for the ME-value was set empirically on a development corpus.

In the study both the quality of the alignment and the quality of the term extraction was evaluated. Alignment error rates were generally low, as the corpus is evidently fairly well-behaved, though figures for French-Dutch were much worse than for the other languages. This is explained by the fact that Dutch has closed compounds, i.e. compounds are written as one word.

The term recognition performance was compared with that of SDL MultiTerm Extract and was found to be superior in recall on all three corpora, even after filtering. It was also superior on precision for the two language pairs French-English and French-Dutch, while there was no difference on French-Italian test data. Test data output was reviewed by human reviewers in two categories, translational quality and relevance, and with three values: OK, not OK, and maybe. For English and Italian the system’s precision varied from 77,5% OK (Italian) to 84,5% OK (English) with Dutch in between on 79,5%. Interestingly, the system was better on multiword terms than single terms for Dutch with opposite results for Italian, while the figures for English were quite close.

The authors make the point that their method avoids the need of using predefined linguistic patterns to recognize terms. This is true to some extent as they generate candidates on the basis of contexts and use lexical links to test for similarity. However, part-of-speech patterns enter into their definition of anchor chunks, and part-of-speech tagging is a necessary pre-process.

6.4 Word alignment supporting monolingual term extraction

Very few studies have been done comparing monolingual term extraction with and without access to translations. However, Tiedemann (2001) investigated whether

good quality word alignment could help phrasal term recognition. A basic finding was that when word alignment was used to filter term candidates generated by a simple statistical term generator, precision increased significantly while only few correct term candidates were eliminated. He also showed that the alignment filter increased the portion of candidates that matched a list of nominal part-of-speech patterns whereas those that were eliminated matched the patterns significantly less often, suggesting that word alignment could be used in place of linguistic patterns when taggers are not available.

Also, Lefever et al. (2009) made a comparison of their bilingual system with the five systems for monolingual term extraction used by Zhang et al. (2008). All systems had the input text pre-processed in the same way. The evaluation was performed through a human evaluation of the top 300 terms proposed by each system, where the ranking of the bilingual system was based on Log Likelihood and Mutual Expectation. The authors describe their results as competitive but in terms of a simple ranking the bilingual system came out as the third best system both for single word terms and multi-word terms, slipping 10 and 6% respectively, behind the best system which turned out to be the one using weirdness (Ahmad et al. (2007)) for this test.

7 Conclusions

Computational terminology, and term extraction in particular, has been a field of research for some twenty years. As in other areas of language technology, methods based purely on linguistic analysis and pattern matching have given way to hybrid methods where statistical methods and machine learning are central. However, since the outputs of these systems constitute inputs to processes of human validation linguistic properties of candidates as well as access to context are still important.

For the same reason, evaluation is a critical, but difficult issue for term extraction systems. There are very few reports on extrinsic evaluation of term extraction systems, although this would seem to be what is required, given that term definitions are different from one application to the other. Moreover, results from intrinsic evaluations, such as those reported in this review vary from one corpus to the other, suggesting that the properties of test corpora and their relation to the training corpus are of central importance.

Methods of increased sophistication for monolingual term extraction continue to be developed. Judging from the studies reported in this review, the best results are obtained when linguistic, statistical, and distributional criteria are combined. For the reason just given, however, there is no guarantee that a measure that has

been found to be superior on a single corpus will be equally superior on another corpus. Moreover, it is not known whether a gain in precision by a few percent will yield a similar improvement in the validation process.

Most work so far has been done on monolingual term extraction. Proposals for bilingual term extraction have generally favoured a process whereby monolingual term extraction precedes candidate pairing. While existing studies seem to support (rather than contradict) the hypothesis that recognition of corresponding units under translation helps termhood recognition, there are so far only few studies that compare the two main workflows for bilingual term extraction, i.e., extraction+spotting vs. alignment+extraction, and none of them in any real depth.

References

- K. Ahmad, et al. (2007). 'University of Surrey participation in TREC8: Weirdness indexing for logical document extrapolation and retrieval (WILDER)'. In *Proceedings of the Eighth Text Retrieval Conference (TREC-8)*.
- S. Ananiadou (1994). 'A methodology for automatic term recognition'. In *Proceedings of the 15th Conference on Computational Linguistics (Coling 94)*, pp. 1034–1038.
- R. Basili, et al. (2001). 'A contrastive approach to term extraction'. In *Proceedings of the 4th Terminological and Artificial Intelligence Conference (TIA2001)*.
- D. Bourigault (1993). 'An endogenous corpus-based method for structural noun phrase disambiguation'. In *Proceedings of the 6th Conference of the European Chapter of the ACL (EACL'93) Utrecht, The Netherlands*, pp. 81–86.
- B. Daille (1996). 'Study and implementation of combined techniques for automatic extraction of terminology'. In J. L. Klavans & P. Resnik (eds.), *The Balancing Act: Combining Symbolic and Statistical Approaches to Language*, pp. 49–66. Cambridge, MA: MIT Press.
- B. Daille (2000). 'Morphological rule induction for terminology acquisition'. In *Proceedings of the 18th Conference on Computational Linguistics (Coling 2000)*, pp. 215–221.
- L. Deléger, et al. (2006). 'Enriching Medical Terminologies: an Approach Based on Aligned Corpora'. In *Proceedings of the 20th International Congress of the European Federation for Medical Informatics (MIE 2006)*.

- K. Frantzi, et al. (1998). 'The C-value/NC-value Method of Automatic Recognition for Multi-Word Terms'. *Lecture Notes in Computer Science* **1513**:585–604.
- A. Hippiisley, et al. (2005). 'The head-modifier principle and multilingual term extraction'. *Natural Language Engineering* **11(2)**:129–157.
- H. Hjelm (2007). 'Identifying Cross Language Term Equivalents Using Statistical Machine Translation and Distributional Association Measures.'. In *Proceedings of the 16th Nordic Conference of Computational Linguistics (Nodalida 2007)*, Tartu, Estonia.
- H. Hjelm (2009). *Cross-Language Ontology Learning*. Ph.D. thesis, Stockholm University.
- C. Jacquemin & D. Borigault (2003). 'Term extraction and automatic indexing'. In R. Mitkov (ed.), *The Oxford Handbook of Computational Linguistics*, pp. 599–615. Oxford University Press.
- J. S. Justeson & S. M. Katz (1995). 'Technical terminology: some linguistic properties and an algorithm for identification in text'. *Natural Language Engineering* **1(1)**:9–27.
- K. Kageura & B. Umino (1996). 'Methods of automatic term recognition: a review'. *Terminology* **3(2)**:259–289.
- J.-D. Kim, et al. (2003). 'GENIA corpus–semantically annotated corpus for bio-textmining'. *Bioinformatics* **19 Suppl 1**:i180–i182.
- L. Kozakov, et al. (2004). 'Glossary extraction and knowledge in large organisations via semantic web technologies'. In *Proceedings of the 6th International Semantic Web Conference and the 2nd Asian Semantic Web Conference*.
- E. Lefever, et al. (2009). 'Language-independent bilingual terminology extraction from a multilingual parallel corpus'. In *Proceedings of the 12th Conference of the European Chapter of the ACL*, pp. 496–504, Athens, Greece.
- D. Maynard & S. Ananiadou (2000). 'Identifying terms by their family and friends'. In *Proceedings of the 18th International Conference on Computational Linguistics (Coling 2000)*, pp. 530–536.
- M. Merkel & J. Foo (2007). 'Terminology extraction and term ranking for standardizing term banks'. In *Proceedings of the 16th Nordic Conference of Computational Linguistics (Nodalida 2007)*, Tartu, Estonia.

- F. J. Och & H. Ney (2003). 'A Systematic Comparison of Various Statistical Alignment Models'. *Computational Linguistics* **29**(1):19–51.
- J. Pearson (1998). *Terms in Context*. Amsterdam: John Benjamins.
- P. Schone & D. Jurafsky (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 (EMNLP'01)*, Pittsburg, PA.
- F. Sclano & P. Velardi (2007a). '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.
- F. Sclano & P. Velardi (2007b). 'Termextractor: a web application to learn the shared terminology of emergent web communities'. In *Proceedings of the 3rd International Conference on Interoperability for Enterprise Software and Applications (I-ESA 2007)*.
- R. Steinberger, et al. (2006). 'The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages.'. In *Proceedings of the 5th Language Resources and Evaluation Conference (LREC 2006)*, Genoa, Italy.
- G. Thurmair (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 (EAMT/CLAW 2003)*, Dublin City University, Ireland.
- J. Tiedemann (2001). 'Can bilingual word alignment improve monolingual phrasal term extraction?'. *Terminology* **7**(2):199–215.
- A. S. Valderrabanos, et al. (2002). 'Multilingual terminology extraction and validation'. In *Proceedings of the Fourth Language Resources and Evaluation Conference (LREC 2002)*, Las Palmas de Gran Canaria, Spain.
- W. Wong, et al. (2007a). 'Determining Termhood for Learning Domain Ontologies in a Probabilistic Framework'. In *Proceedings of the 6th Australasian Conference on Data Mining (AusDM)*, pp. 55–63.
- W. Wong, et al. (2007b). 'Determining Termhood for Learning Domain Ontologies using Domain Prevalence and Tendency'. In *Proceedings of the 6th Australasian Conference on Data Mining (AusDM)*, pp. 47–54.

- Z. Zhang, et al. (2008). 'A Comparative Evaluation of Term Recognition Algorithms'. In *Proceedings of the Sixth Language Resources and Evaluation Conference (LREC 2008)*, Marrakech, Morocco.
- Z. Zhang, et al. (2009). 'Too many mammals: Improving the diversity of automatically recognized terms'. In *International Conference on Recent Advances in Natural Language Processing 2009 (RANLP'09)*, Borovets, Bulgaria.
- D. Zielinski & Y. Ramirez Safar (2005). 'Research meets practice: t-survey 2005: An online survey on terminology extraction and terminology management'. In *Proceedings of Translating and the Computer (ASLIB 27)*, London.