Determining Termhood for Learning Domain Ontologies using Domain Prevalence and Tendency

Wilson Wong, Wei Liu and Mohammed Bennamoun

School of Computer Science and Software Engineering
University of Western Australia
Crawley WA 6009
{wilson,wei,bennamou}@csse.uwa.edu.au

Abstract

In the course of reviewing existing automatic term recognition techniques for applications in ontology learning, we came across four issues which can be improved upon. We proposed a new mechanism that incorporates both statistical and linguistic evidences for the computation of a final weight defined as *Termhood (TH)* for ranking term candidates. The analysis of the frequency distributions of the term candidates during our initial experiments revealed three advantages for higher quality term recognition.

1 Introduction

Automatic term recognition, also known as *term extraction* or *terminology mining*, is an integral part of many applications that deal with natural language such as document retrieval (Teevan & Karger 2003), automatic thesaurus construction (Grefenstette 1994), and ontology learning (Wong, Liu & Bennamoun 2007*b*). It involves the extraction and filtering of term candidates for the purpose of identifying domain-relevant terms. The main aim in automatic term recognition is to determine whether a word or a sequence of words is a term that characterises the target domain. The key question can be further decomposed to reveal two critical notions in this area, namely *unithood* and *termhood*. Formally, (Kageura & Umino 1996) defines unithood and termhood as the *"degree of strength or stability of syntagmatic combinations and collocations"* and *"degree that a linguistic unit is related to domain-specific concepts"*, respectively. Unithood is only relevant to *complex terms* (i.e. multi-word terms), while termhood deals with both *simple terms* (i.e. single-word terms) and complex terms.

The determination of unithood and of termhood inevitably requires the use of frequency of occurrence or co-occurrence of words and documents. This is clearly demonstrated through surveys of term extraction approaches by (Kit 2002, Cabre-Castellvi, Estopa & Vivaldi-Palatresi 2001, Kageura & Umino 1996). The difference between unithood and termhood measures lies in how the frequency is employed as evidence. Unithood measurements rely heavily on statistical tests or information-theoretic measures for determining if a sequence of words has strong collocational strength (Wong, Liu & Bennamoun 2007*a*), while termhood determination mostly employ measures of relevance such as those in information retrieval.

This paper is the result of our attempt to look at how advances in automatic term recognition can assist in the term extraction phase of ontology learning. Surveys (e.g. (Gomez-Perez & Manzano-Macho 2003)) have shown that many existing approaches in ontology learning merely employ isolated and non-comprehensive techniques that were not designed to address the various requirements and peculiarities in terminology. During our review of the start-of-the-art in automatic term recognition, we have identified four issues and open problems that remain unaddressed:

- *Inadequate attention to the difference between prevalence and tendency*: One of the main issues in automatic term recognition is that many frequency-oriented techniques adapted from *term frequency inverse document frequency (TF-IDF)* and others alike from information retrieval fail to comprehend that terms are properties of domains and not documents (Basili, Moschitti, Pazienza & Zanzotto 2001). Existing weights do not reflect the tendency of term usage across different domains. They merely measure the prevalence of the term in a particular target domain.

- *Oversimplication of the role of heads and modifiers*: Many approaches have attempted to utilise the head as the representative of the entire complex term in various occasions. Such move is an oversimplication of the relation between a complex term and its head. For example, the assumption that *"term sense is usually determined by its head"* by (Basili, Pazienza & Zanzotto 2001) is not entirely true since heads are inherently ambiguous and modifiers are required to narrow down their possible interpretations.

- *How to determine the relatedness between terms and their context?*: For approaches that attempt to use contextual information, they have realised the prior need to examine the relatedness of context words with the associated term candidate. The existing solutions to this requirement have yet to deliver overall satisfactory results. For example, (Maynard & Ananiadou 1999) rely on the use of rare and static resources for computation of similarity while others such as (Basili, Pazienza & Zanzotto 2001) require large corpora to enable the extraction of context as features to enable the use of feature-based similarity measures.

- *The overemphasis on the role of contextual evidence*: Many researchers have repeatedly stressed on the significance of the old cliché *"you shall know a word by the company it keeps"*. Unless one has the mechanism of identifying the

companies which are truly in the position to describe a word, the overemphasis on contextual evidence may results in negative effects.

To address the highlighted issues above, we propose a new scoring and ranking mechanism that incorporates a series of weights that lead to a final new score known as *Termhood (TH)*. The main aim of this new mechanism is to utilise as much evidences as possible in order to improve the approximation of termhood. The mechanism consists of two new base measures which capture the statistical evidence, and four new derived measures that employ the statistical evidence to quantify the linguistic evidences. In Section 2, we will have an elaborate review on the existing techniques for measuring termhood. In Section 3, we will present our new mechanism, the measures involved and the justification behind every aspect of these measures. In Section 4, we will summarize some findings from our initial experiments. Finally, we conclude this paper with an outlook to future works in Section 5.

2 Related Works

Commonly, the mechanism for assessing termhood will require a ranking scheme, similar to that of relevance ranking for information retrieval, where each term is assigned a score. Such ranking scheme will assist in the selection of *"true"* terms from less likely ones. Existing measures based on formal probabilistic models for determining the relevance of words with respect to certain topics or documents are mainly studied within the realm of document retrieval and automatic indexing. In probabilistic indexing, one of the first few detailed quantitative models was proposed by (Bookstein & Swanson 1974). In this model, the differences in the distributional behavior of words is employed as a guide to determine if a word should be considered as an index term. This model is founded upon works on how function words can be closely modeled by a Poisson distribution whereas content words deviates from it (Church & Gale 1995, Manning & Schutze 1999). An even larger collection of literature on probabilistic models can be found in a related area of document retrieval. The simplest of all the retrieval models is the binary independence model (Fuhr 1986, Lewis 1998). As with all other retrieval models, the binary independence model is designed to estimate the probability that a document j is considered as relevant given a specific query k. Let $T = \{t_1, ...t_n\}$ be the set of terms in the collection of documents (i.e. corpus). We can then represent the set of terms T_j occurring in document j as a binary vector $v_j = \{x_1, ..., x_n\}$ where $x_i = 1$ if $t_i \in T_j$ and $x_i = 0$ otherwise. This way, the odds of document j, represented by a binary vector v_j being relevant to query k can be computed as (Fuhr 1992):

$$O(R|k, v_j) = \frac{P(R|k, v_j)}{P(\bar{R}|k, v_j)} = \frac{P(R|k)}{P(\bar{R}|k)} \frac{P(v_j|R, k)}{P(v_j|\bar{R}, k)}$$

and based on the assumption of independence between the presence and absence of terms,

$$\frac{P(v_j|R, k)}{P(v_j|\bar{R}, k)} = \prod_{i=1}^{n} \frac{P(x_i|R, k)}{P(x_i|\bar{R}, k)}$$

Other more advanced models that take into considerations other factors such as term frequency, document frequency and document length have also been proposed (Jones, Walker & Robertson 1998).

Besides formal models for term weighting and ranking, a more straightforward and commonly-adopted method is *term frequency inverse document frequency (TF-IDF)* and its variants (Salton & Buckley 1988). (Basili, Moschitti, Pazienza & Zanzotto 2001) proposed a TF-IDF inspired measure for assigning terms with more accurate weight that reflects their specificity with respect to the target domain. This contrastive analysis is based on the heuristic that general language-dependent phenomena should spread similarly across different domain corpus and special-language phenomena should portray odd behaviors. This *contrastive weight* for simple term candidate a in target domain d is defined as:

$$CW(a) = \log f_{ad} \left(\log \frac{\sum_j \sum_i f_{ij}}{\sum_j f_{aj}} \right) \quad (1)$$

where f_{ad} is the frequency of the simple term candidate a in the target domain d, $\sum_j \sum_i f_{ij}$ is the sum of the frequencies of all term candidates in all domains, and $\sum_j f_{aj}$ is the sum of the frequencies of term candidate a in all domains. For complex term candidates, the frequency of their heads are utilised to compute their weights. This is necessary as the low frequencies among complex terms make estimations difficult. Consequently, the weight for complex term candidate a in domain d is defined as:

$$CW(a) = f_{ad} CW(a^h) \quad (2)$$

where f_{ad} is the frequency of the complex term candidate a in the target domain d, and $CW(a^h)$ is the contrastive weight for the head of the complex term candidate, a^h. The use of heads by (Basili, Moschitti, Pazienza & Zanzotto 2001) for computing the contrastive weights $CW(a)$ for complex term candidates reflects the head-modifier principle (Hippisley, Cheng & Ahmad 2005). The principle suggests that the information being conveyed by complex terms manifest itself in the arrangement of the constituents. The head acts as the key that refers to a general category to which all other modifications of the head belong. The modifiers are responsible for distinguishing the head from other forms in the same category.

Besides contrastive analysis, the use of contextual evidence to assist in the correct identification of terms has become popular. There are currently two dominant approaches to extract context words: the use of fixed-size windows (Maynard & Ananiadou 1999), and the use of grammatical relations (Basili, Pazienza & Zanzotto 2001, LeMoigno, Charlet, Bourigault, Degoulet & Jaulent 2002). One of the works along the line of incorporating contextual evidence is *Cvalue* and *NCvalue* by (Frantzi & Ananiadou 1997). *Cvalue* can be regarded as a unithood measure that contributes to the calculation of *NCvalue*. Discussions on *Cvalue* is beyond the scope of this paper. It suffices to know that given a simple or complex term candidate a to be examined for unithood, the *Cvalue* is defined as:

$$Cvalue(a) = \begin{cases} \log_2 |a| f_a & \text{if } |a| = g \\ \log_2 |a| \left(f_a - \frac{\sum_{l \in L_a} f_l}{|L_a|} \right) & \text{otherwise} \end{cases} \quad (3)$$

where $|a|$ is the number of words in a, L_a is the set of potential longer term candidates that contain a, and g is the longest n-gram considered, f_a is frequency of occurrences of a in the corpus. As for *NCvalue*, this measure involves the assignment of weights to context words (in the form of nouns, adjectives and verbs) located within a fixed-size window from the term candidate. Given that TC is the set of all term candidates

and c is a noun, verb or adjective which appears with term candidates, the $weight(c)$ is defined as:

$$weight(c) = 0.5 \left(\frac{|TC_c|}{|TC|} + \frac{\sum_{a \in TC_c} f_a}{f_c} \right) \quad (4)$$

where $TC_c \subset TC$ is the set of term candidates that appear with c, $\sum_{a \in TC_c} f_a$ is the total frequency of c appearing with term candidates, and f_c is the frequency of c. After calculating the weights for all possible context words, the sum of weights for context words appearing with each term candidate can be obtained. Formally, for each simple or complex term candidate a that has a set of accompanying context words C_a, the cumulative context weight is defined as:

$$cweight(a) = \sum_{c \in C_a} weight(c) + 1 \quad (5)$$

Eventually, the *NCvalue* for a term candidate is defined as:

$$NCvalue(a) = \frac{1}{\log F} Cvalue(a) cweight(a) \quad (6)$$

where F is the size of the corpus in terms of number of words.

There has also been an increasing interest in incorporating semantic information for measuring termhood. The use of semantic measures is mainly to gauge the relatedness of context words with the associated term candidates in the process of measuring termhood. Maynard & Ananiadou (Maynard & Ananiadou 1999) employ the *Unified Medical Language System (UMLS)* to compute two weights, namely, positional $pos(a, b)$ and commonality $com(a, b)$ where a and b are term candidates. The *UMLS* is organised as a hierarchical structure of concepts. Each concept has a set of related terms. Positional weight is obtained based on the combined number of concepts belonging to each term, while commonality is measured by the number of shared common ancestors multiplied by the number of times the term occurs. The similarity of two term candidates is simply $sim(a, b) = com(a, b)/pos(a, b)$. The authors then modified the *NCvalue* discussed in Equation 6 by incorporating the similarity measure as part of a *context factor (CF)* (Maynard & Ananiadou 2000) defined as:

$$CF(a) = weight(c) \sum_{c \in C_a} f_c + sim(a, b) \sum_{b \in CT_a} f_b$$

where C_a is the set of context words of a, f_c is the frequency of c as a context word of a, $weight(c)$ is the weight for context word c as defined in Equation 4, CT_a is the set of context words of a which also happens to be term candidates (i.e. context terms), f_b is the frequency of b as a context term of a, and $sim(a, b)$ is the similarity between term candidate a and its context term b using *UMLS*. The new *NCvalue* is defined as:

$$NCvalue(a) = 0.8Cvalue(a) + 0.2CF(a)$$

(Basili, Pazienza & Zanzotto 2001) commented that the use of extensive and well-grounded semantic resources by (Maynard & Ananiadou 1999) faces the issue of portability to other domains. Instead, (Basili, Pazienza & Zanzotto 2001) combine the use of contextual information and the head-modifier principle to capture term candidates and their context words on a feature space for computing similarity. Given the term candidate a, the feature vector for a is $\tau_a = (v_1, ..., v_n)$ where v_i is the value of the attribute F_i and n is the number of features. F_i comprises of the tuple (T_i, h_i) where T_i is the type of grammatical relations (e.g. *subj*, *obj*) and h_i is the head of the relation. In other words, only the head of complex terms will be used as referent to the entire structure. According to the authors, *"the term sense is usually determined by its head."*. This statement opposes the fundamental fact, not only in terminology but in general linguistic, that simple terms are polysemous and modification of such terms are necessary to narrow down their possible interpretations (Hippisley et al. 2005). The authors chose the cosine measure for computing similarity, $sim(\tau_a, \tau_b)$ between two terms over the syntactic feature space. To assist in the ranking of the term candidates using their heads, a hand-crafted controlled terminology *CTO* is employed as evidence of *"correct"* terms during the computation of similarity. Accordingly, given that $\tau_{CTO} = \sum_{e \in CTO} \tau_e$ the measure for assigning weight to term candidate a is defined as:

$$ext(a) = sim(\tau_a, \tau_{CTO}) f_a \quad (7)$$

where f_a is the frequency of a in the corpus.

3 The Proposed Approach for Combining Termhood Evidences

We propose a new mechanism for scoring and ranking that employs distributional behaviors of term candidates within the target domain and also across different domains as statistical evidence to quantify the linguistic evidences in the form of candidate, modifier and context. The evidences are gathered from two types of corpus, namely, *domain corpus* d which contains text in the target domain, and *contrastive corpus* $\bar{d}$ which contains text across different genre other than the target domain. Since the quality of the evidences of terms with respect to the domain is dependent on the issue of representativeness of the corresponding corpus, we will assume that both d and $\bar{d}$ are balanced, unbiased and randomised samples of the population text representing the corresponding domain. The actual discussion on corpus representativeness is nevertheless important but the issue is beyond the scope of this paper. Next, we introduce two base measures for capturing the statistical evidence based on the cross-domain and intra-domain distribution:

- **Intra-domain distribution** of term candidates and context words are employed to compute the basic *domain prevalence (DP)*. DP measures the extent of term usage in the target domain.
- **Cross-domain distributional behavior** of term candidates and context words are employed to compute the *domain tendency (DT)*. DT measures the extent of inclination of term usage toward the target domain.

A high DP means that the term is frequently encountered (i.e. prevalent) in the target domain. It is a sign of high domain relevance if and only if the frequent usage of that term is exclusive to the target domain (i.e. high DT). The three types of linguistic evidences, which are essential to the estimation of termhood are quantified using new measures derived from the prevalence and tendency measures described above. The linguistic evidences and the corresponding derived measures are:

- **Candidate evidence**, in the form of *discriminative weight (DW)*, is measured as the product of the domain tendency and the domain prevalence of term candidates. This evidence constitutes the first step in an attempt to isolate domain-relevant from general candidates.
- **Modifier evidence**, in the form of *modifier factor (MF)* is obtained by computing DT using the cross-domain cumulative frequency of modifiers of complex terms.
- **Contextual evidence**, in the form of *average contextual discriminative weight (ACDW)*, is computed using the cumulative DW of context words, scaled according to their semantic relatedness with the corresponding term candidates. $ACDW$ is later adjusted with respect to the DW of the term candidate to obtain the *adjusted contextual contribution (ACC)* to reflect the reliability of the contextual evidence.

Given that we have a list of term candidates (both simple and complex) $TC = \{a_i, ..., a_n\}$, the aim of this mechanism is to assign scores to term candidates to assist in the ranking and identification of the most suitable candidates as terms $t \in T$. Each complex term, a will comprise of a head a^h and modifiers $m \in M(a)$. Each term candidate is assigned a weight depending on its type (i.e. simple or complex). We refer to this new CW-inspired weight as *domain prevalence (DP)* because of its ability to capture the extent of occurrences of terms in the target domain. If a is a simple term, its DP is defined as:

$$DP(a) = \log_{10}(f_{ad} + 10) \log_{10}\left(\frac{F_{TC}}{f_{ad} + f_{a\bar{d}}} + 10\right)$$

where $F_{TC} = \sum_j f_{jd} + \sum_j f_{j\bar{d}}$ is the sum of the frequencies of occurrences of all $a \in TC$ in both domain and contrastive corpora, while f_{ad} and $f_{a\bar{d}}$ are the frequencies of occurrences of a in the domain corpus and contrastive corpora, respectively. If the term candidate is complex, we define its DP as:

$$DP(a) = \log_{10}(f_{ad} + 10) DP(a^h) MF(a)$$

Please note the use of the DP of the head a^h for the computation of DP for complex terms. Motivated by CW, rarity of appearances of complex terms does not allow a proper computation of the weight. Nonetheless, we have noticed from the original CW that the direct multiplication of f_{ad} of extremely common and general complex terms will distort the weights and give a false impression of their importance in the domain. Consequently, unlike the original CW, we add a constant 10 and later log the domain frequency of complex terms. This modification has shown to eliminate the biased ranking of CW as demonstrated in Section 4. Besides the addition of the constant 10 to f_{ad} prior to log, we introduce another new measure called *modifier factor (MF)* to:

- provide relevant complex terms with higher weights than their head;
- penalise those potentially deceiving domain-unrelated complex terms that have domain-related heads. For example, the head *"virus"* will yield high weight in the *"technology"* domain. If we did not take into consideration the fact that the head was modified by *"H5N1"* to form the complex term *"H5N1 virus"*, the complex term makes its way into the list of terms for the *"technology"* domain; and
- compensate for the low weight of domain-related complex terms that have domain-unrelated heads. For example, upon looking at the head *"account"*, one would be tempted to immediately rule the corresponding complex term out as irrelevant for the *"technology"* domain. With MF, we can take into consideration the modifiers *"Google"* and *"Gmail"* to safely assign higher weights to the complex term *"Google Gmail account"*.

The MF of a complex term a is measured using the cumulative domain frequency and cumulative contrastive frequency of modifiers which also happen to be term candidates, $m \in M_a \cap TC$. Formally, the MF of a complex term a is defined as:

$$MF(a) = \log_2\left(\frac{\sum_{m \in M_a \cap TC} f_{md} + 1}{\sum_{m \in M_a \cap TC} f_{m\bar{d}} + 1} + 1\right)$$

MF is actually a derived measure modelled after our second new base measure *domain tendency (DT)*. The only difference between the two is that MF works with modifiers while DT works with the entire term candidate, both simple and complex. MF and DT are two powerful discriminating measures that help to differentiate between candidates which are truly relevant to the target domain from generally-prevalent candidates. Formally, we can determine the extent of the inclination of the usage of term candidate a for domain and non-domain purposes through:

$$DT(a) = \log_2\left(\frac{f_{ad} + 1}{f_{a\bar{d}} + 1} + 1\right)$$

where f_{ad} is the frequency of occurrences of a in the domain corpus, while $f_{a\bar{d}}$ is the frequency of occurrences of a in the contrastive corpora. If term candidate a is equally common in both domain and non-domain (i.e. contrastive domain), $DT = 1$. If the usage of a is more inclined toward the target domain, $f_{ad} > f_{a\bar{d}}$, then $DT > 1$, and $DT < 1$ otherwise. Next, this new base measure DT together with DP will contribute to a new weight known as *discriminative weight (DW)*. A term that appears frequently in the target domain (i.e. high DP) will still have a low overall weight DW if its usage is more inclined toward the contrastive domains (i.e. low DT). DW is simply the product of DP and the corresponding DT of the term candidate:

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

Assuming that term candidate a has the set of context words C_a, the *average contextual discriminative weight (ACDW)* is defined as:

$$ACDW(a) = \frac{\sum_{c \in C_a} DW(c) sim(a, c)}{|C_a|}$$

where $sim(a, c) = 1 - NGD(a, c)\theta$, $NGD(a, c)$ is the *Normalized Google Distance* (Cilibrasi & Vitanyi 2007) between term candidate a and c, and θ is a constant for scaling the distance value of NGD. The $ACDW$ weight allows us to take into consideration the company a term candidate keeps. Nonetheless, not all context words describe or are related to the terms they appear with. Unlike other applications that can completely rely on contextual information, we cannot allow $ACDW$ to have direct contribution to the overall termhood. In this regard, we employ two measures to adjust the contribution of the contextual weight to the overall termhood. First, we

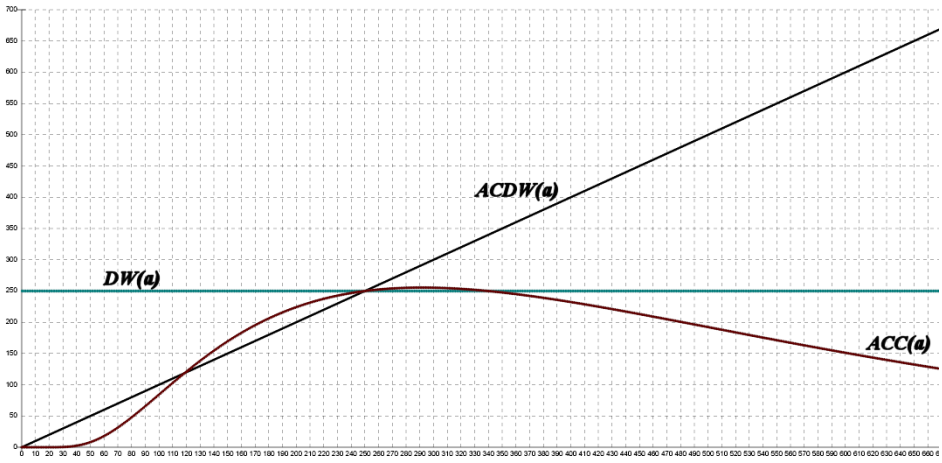


Figure 1: ACC experiences an increase for $ACDW < DW$. The distribution of ACC is reflected at the meeting point of DW and $ACDW$ and experiences subsequent decrease more or less inversely proportional to $ACDW$.

utilise NGD to quantify the relatedness between a term and its context words during the computation of $ACDW$. So far, NGD has only been successfully adopted for use with clustering (Wong et al. 2007*b*). Contextual words which are more related to the term candidate will have higher contribution to the overall $ACDW$. NGD is at present the most ideal solution for the problems introduced by the use of static and restricted semantic resources faced by many researchers. Secondly, we adjust $ACDW$ according to its ratio with the corresponding DW to produce the *adjusted contextual contribution (ACC)* as shown in Figure 1. Formally, we define ACC as:

$$ACC(a) = ACDW(a)\frac{e^{\left(1-\frac{ACDW(a)+1}{DW(a)+1}\right)}e^{\left(1-\frac{DW(a)+1}{ACDW(a)+1}\right)}}{\log_2 \frac{ACDW(a)+1}{DW(a)+1}+1}$$

In the end, we define the final weight known as *Termhood (TH)* for each term candidate as:

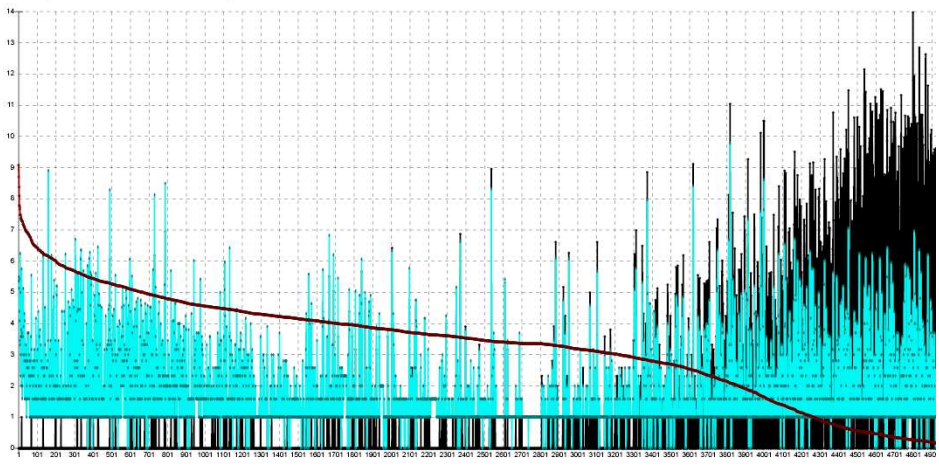
$$TH(a) = DW(a) + ACC(a) \tag{8}$$

4 Experiments

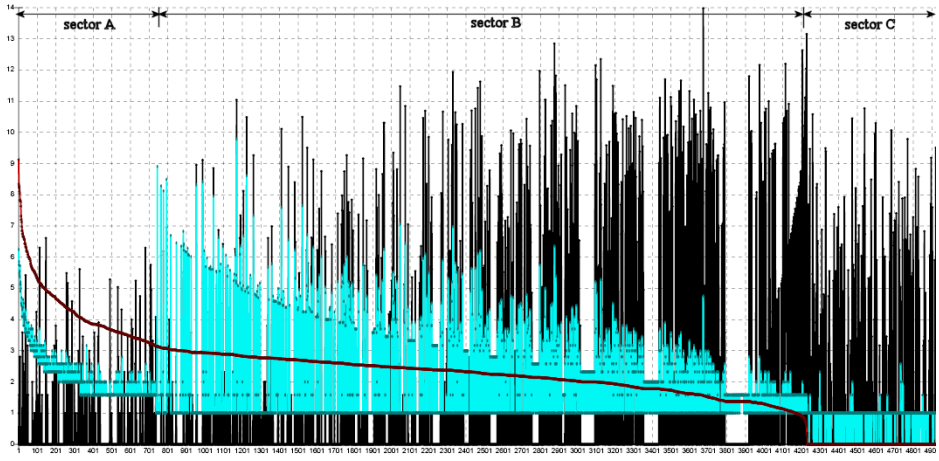
We employ two text sources: a domain corpus containing 24 documents (with 51,289 word count) in *"liver cancer"* from `BioMedCentral.com`, and a contrastive corpora consisting of 11,115 news articles (with 4,378,210 word count) in various domains such as *"technology"*, *"business"*, *"politics"* and *"sports"*. The news articles are gathered from `Reuters.com`, `CNet.com` and `ABC.com` between the period of February 2006 and April 2007. The implementation of *Contrastive Weight (CW)*, *NCvalue (NCV)* and *Termhood (TH)* are in accordance to Equation 1 and 2, 6, and 8 respectively. The source of the term candidates and context words is a list of 6,000 instantiated subcategorisation frames (Wong 2005) extracted from the *"liver cancer"* domain corpus. By selecting only noun phrases from the first and second arguments of the frames, we obtained 5,156 term candidates.

Like other evaluations in ontology learning, determining the quality of the extracted terms is a lengthy and challenging process due to the subjective nature of the task and the efforts required. The quantitative evaluations required for a numerical analysis and comparative study of CW, NCV and TH is beyond the scope of this paper. Nonetheless, we have decided to assess CW, NCV and TH using an equally effective method, namely, the analysis of the frequency distributions. The role of term frequency as the main source of termhood evidence makes this evaluation method (i.e. analysis of frequency distribution) highly applicable. We will discuss the reasons and ramifications of the various interesting characteristics displayed in the graphs, with reference to the related measures. For this experiment, the performance of the termhood measures are judged solely based on their ability to provide higher ranks to candidates that have domain frequencies higher than contrastive frequencies. Ideally, candidates with higher domain frequencies should congregate along the left end of the x-axis in the graphs while those with higher contrastive frequencies are pushed to the far right of the x-axis. The graphs are based on the logarithmic scale to accommodate the frequencies which can become very large.

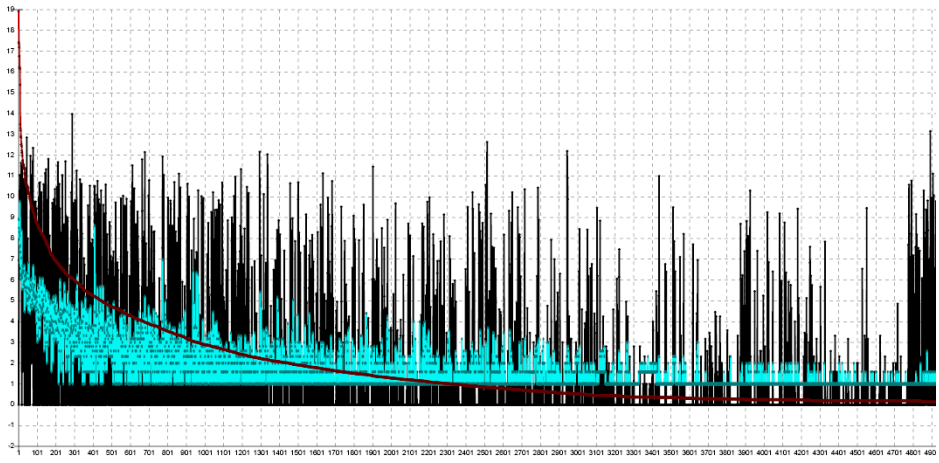
Firstly, the frequency distribution of the candidates sorted according to CW displays an interesting trend that reflects the different treatment given to simple and complex terms. Despite exhibiting some characteristics of being able to discriminate and identify domain-relevant candidates but certain peculiarities in Figure 2(b) call for deeper analysis. If we recall what has been discussed regarding Equations 1 and 2, we noticed that sector A in Figure 2(b) is the direct result from the use of heads to compute the contrastive weights of complex terms. The listing of the ranked terms which corresponds to Figure 2(b) clearly shows that all candidates in sector A are complex terms. Despite the much lower domain frequency of complex terms f_{ad} in sector A, these terms are ranked among the highest by CW due to the high domain frequency of their corresponding heads $f_{a^h d}$. In addition, instead of $\log_{10} f_{ad}$, the direct multiplication of f_{ad} to the contrastive weight of the head of complex terms $CW(a^h)$ further increases the contrastive weight of the complex terms $CW(a)$. In sector B, the distributions behave as intended due to the design of the contrastive weight that places emphasis on the domain frequency. This sector contains a fair mix of simple and complex terms. In this sector, one will notice that the candidates are ordered from left to right in such a way that their f_{ad} decreases as $f_{a\bar{d}}$ increases. A point worth noting about this section is the somewhat periodic clusters of candidates



(a) Candidates ranked according to the scores by TH



(b) Candidates ranked according to the scores by CW



(c) Candidates ranked according to the scores by NCV

Figure 2: This graph shows the frequency distributions of $5,156$ candidates ranked in descending order according to the scores assigned by the respective measures. The black, oscillating line represents the frequency of the candidates in the contrastive corpora $f_{\bar{d}}$ while the greyish, oscillating line is the domain frequency f_d. The single dark line that spans from the left to the right of the graph in descending order represents the scores assigned by the respective measures.

characterised by the sudden drop and surge in frequency. This phenomena can be explained by the use of heads for the computation of weights for complex terms. Due to the increasingly less significant contribution of f_{ad} to $CW(a)$ in sector B, complex terms having similar heads are inevitably grouped together. Once in sector C, the contrastive weights of term candidates $CW(a)$ drop to zero. Most terms in this sector are single-word and have $f_{ad} = 1$. Without the support of $CW(a^h)$, the single-occurrence simple terms in this sector are weighted 0.

Secondly, Figure 2(c) illustrates why NCV, the version discussed in Equation 6, is not suitable for recognizing terms for domain-specific uses. The weight helps to rank those candidates in the order of decreasing domain frequency f_d, regardless of the candidates' prevalence and tendency in other domains. Firstly, looking at the list of ranked terms, one would notice that those candidates which are highly ranked are accompanied by more context words. As formulated in Equation 5, more context words will contribute to an increasingly higher contextual weight. As we have pointed out repeatedly, contextual evidence does contribute to the determination of termhood but indiscriminate use will results in undesirable effects. Secondly, the highly-ranked candidates are also those that have extremely high $Cvalue$ as defined in Equation 3. In other words, a candidate that demonstrates better independence from the longer term candidates which it is part of will have higher rank. This results in the emphasis and high ranking for mostly simple terms or shorter complex terms that are extremely popular and at the same, commonly used to construct other longer complex terms.

Lastly, Figure 2(a) shows the discriminating power of our new scoring and ranking mechanism using termhood TH. Based on the distribution of f_{ad} and $f_{a\bar{d}}$, one will notice that candidates with high f_{ad} and near-zero $f_{a\bar{d}}$ are highly ranked. This trend progresses until all candidates with high $f_{a\bar{d}}$ and comparatively lower f_{ad} are pushed to the far right end. It is worth pointing out that the candidates in Figure 2(a) are not ordered in a smooth descending flow according to their domain frequencies f_{ad}. This can be explained by the diversified evidences that our new measures rely on. For example, using our measures, some candidates with high f_{ad} may appear in lower ranks than those with lower f_{ad}. This is in sharp contrast to the other two existing weights namely CW and NCV where one can observe from sector B in Figure 2(b) and the whole of Figure 2(c) that, despite minor fluctuations, the ranking of the candidates are largely influenced by frequencies.

In addition to the absence of evaluation methods and datasets, the subjective nature of termhood assessment makes the tasks of objectively evaluating and comparing our new measure with existing approaches a problem domain by itself. In the words of (Damle & Uren 2005), *"Unfortunately, there is no objective evaluation method reported in the literature for term extraction..."*. However, our initial experiments using the frequency distributions alone have revealed three positive traits present in our new weight TH that will assist in improving the quality of automatic term recognition:

- Unlike CW and NCV, the diversification of evidence to cover both statistical and linguistic has allowed TH to be all-inclusive and not purely dependent on frequency. Despite the subjectivity of the notion termhood, such diversification provides opportunity for more accurate approximation to reflect the *"actual"* termhood.
- Unlike NCV, the increase in the number of context words does not indiscriminately propel our new weight TH. The selective use of contextual evidence allows its contribution in TH to be adjusted depending on the relationship between the context words and the term candidates.
- In NCV, shorter term candidates have better chances of gaining higher rank, while complex term candidates always top the list in CW. The size of candidates should not have any influence on their weights. Unlike the two measures which impose baseless, intrinsic prejudice on the candidates during the scoring process, both simple and complex term candidates are given equal opportunities using TH, based purely on observable evidences.

5 Conclusion and Future Work

In this paper, we have presented a new mechanism consisting of a series of base and derived measures for recognising terms. The base measures, namely, *domain prevalence (DP)* and *domain tendency (DT)* capture the statistical evidence that appear in the form of intra-domain and cross-domain term distributional behavior. Using these base measures, four additional measures, namely *discriminative weight (DW)*, *modifier factor (MF)*, *average contextual discriminative weight (ACDW)*, and *adjusted contextual contribution (ACC)* were derived to quantify linguistic evidences in the form of candidates, modifiers and context words. Together, these base and derived measures contribute to the computation of a final weight known as *Termhood (TH)* that is used for the ranking of candidates and selection of terms. Our initial experiments revealed three advantages exhibited by our new weight TH. Such revelation has prompted us to plan for future works to evaluate our new mechanism using larger datasets and established measures such as precision and recall to enable numerical analysis and comparative study. In addition, to demonstrate the applicability of our new approach in real-word, domain-specific scenarios, we intend to have domain experts assisting in future evaluations.

Acknowledgement

This research was supported by the Australian Endeavour International Postgraduate Research Scholarship, and the Research Grant 2006 by the University of Western Australia.

References

Basili, R., Moschitti, A., Pazienza, M. & Zanzotto, F. (2001), A contrastive approach to term extraction, *in* 'Proceedings of the 4th Terminology and Artificial Intelligence Conference (TIA)', France.

Basili, R., Pazienza, M. & Zanzotto, F. (2001), Modelling syntactic context in automatic term extraction, *in* 'Proceedings of the International Conference on Recent Advances in Natural Language Processing', Bulgaria.

Bookstein, A. & Swanson, D. (1974), 'Probabilistic models for automatic indexing', *Journal of the American Society for Information Science* **25**(5), 312–8.

Cabre-Castellvi, T., Estopa, R. & Vivaldi-Palatresi, J. (2001), Automatic term detection: A review of current systems, *in* D. Bourigault, C. Jacquemin & M. LHomme, eds, 'Recent Advances in Computational Terminology', John Benjamins.

Church, K. & Gale, W. (1995), Inverse document frequency (idf): A measure of deviations from poisson, *in* 'Proceedings of the ACL 3rd Workshop on Very Large Corpora'.

Cilibrasi, R. & Vitanyi, P. (2007), 'The google similarity distance', *IEEE Transactions on Knowledge and Data Engineering* **19**(3), 370–383.

Damle, D. & Uren, V. (2005), Extracting significant words from corpora for ontology extraction, *in* 'Proceedings of the 3rd International Conference on Knowledge Capture', Alberta, Canada.

Frantzi, K. & Ananiadou, S. (1997), Automatic term recognition using contextual cues, *in* 'Proceedings of the IJCAI Workshop on Multilinguality in Software Industry: the AI Contribution', Japan.

Fuhr, N. (1986), Two models of retrieval with probabilistic indexing, *in* 'Proceedings of the 9th ACM SIGIR International Conference on Research and Development in Information Retrieval'.

Fuhr, N. (1992), 'Probabilistic models in information retrieval', *The Computer Journal* **35**(3), 243–255.

Gomez-Perez, A. & Manzano-Macho, D. (2003), A survey of ontology learning methods and techniques, Deliverable 1.5, OntoWeb Consortium.

Grefenstette, G. (1994), Explorations in automatic thesaurus discovery, Kluwer Academic Publishers, MA, USA.

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

Jones, K., Walker, S. & Robertson, S. (1998), 'A probabilistic model of information retrieval: Development and status', *Information Processing and Management* **36**(6), 809–840.

Kageura, K. & Umino, B. (1996), 'Methods of automatic term recognition: A review', *Terminology* **3**(2), 259–289.

Kit, C. (2002), Corpus tools for retrieving and deriving termhood evidence, *in* 'Proceedings of the 5th East Asia Forum of Terminology', Haikou, China.

LeMoigno, S., Charlet, J., Bourigault, D., Degoulet, P. & Jaulent, M. (2002), Terminology extraction from text to build an ontology in surgical intensive care, *in* 'Proceedings of the ECAI Workshop on Machine Learning and Natural Language Processing for Ontology Engine'.

Lewis, D. (1998), Naive (bayes) at forty: The independence assumption in information retrieval, *in* 'Proceedings of the 10th European Conference on Machine Learning'.

Manning, C. & Schutze, H. (1999), Foundations of statistical natural language processing, MIT Press, MA, USA.

Maynard, D. & Ananiadou, S. (1999), Term extraction using a similarity-based approach, *in* 'Recent Advances in Computational Terminology', John Benjamins.

Maynard, D. & Ananiadou, S. (2000), Identifying terms by their family and friends, *in* 'Proceedings of the 18th International Conference on Computational Linguistics', Germany.

Salton, G. & Buckley, C. (1988), 'Term-weighting approaches in automatic text retrieval', *Information Processing & Management* **24**(5), 513–523.

Teevan, J. & Karger, D. (2003), Empirical development of an exponential probabilistic model for text retrieval: Using textual analysis to build a better model, *in* 'Proceedings of the 26th ACM SIGIR International Conference on Research and Development in Information Retrieval'.

Wong, W. (2005), Practical approach to knowledge-based question answering with natural language understanding and advanced reasoning, Master's thesis, National Technical University College of Malaysia, arXiv:cs.CL/0707.3559.

Wong, W., Liu, W. & Bennamoun, M. (2007*a*), Determining the unithood of word sequences using mutual information and independence measure, *in* 'Proceedings of the 10th Conference of the Pacific Association for Computational Linguistics (PACLING)', Melbourne, Australia.

Wong, W., Liu, W. & Bennamoun, M. (2007*b*), 'Tree-traversing ant algorithm for term clustering based on featureless similarities', Journal on Data Mining and Knowledge Discovery, doi: 10.1007/s10618-007-0073-y.