Determining Termhood for Learning Domain Ontologies in a Probabilistic Framework

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

Many existing techniques for term extraction are heuristically-motivated and criticised as ad-hoc. The definitions and assumptions critical to set the boundary for the effectiveness of the techniques are often implicit and unclear. Here we present a probabilistic framework for measuring termhood to address the lack of mathematical foundation in existing techniques.

1 Introduction

Term extraction, also known as *automatic term recognition* and *terminology mining*, is essential to many text mining applications such as ontology learning. The aim of term extraction is to identify content-bearing lexical units (i.e. terms) from text, which can either be individual or group of words. Term extraction consists of two fundamental steps: 1) identifying term candidates from text, and 2) filtering through the candidates to separate terms from non-terms. The first step involves the determination of *unithood*, which concerns with whether sequences of words can be combined to form stable lexical units (Wong, Liu & Bennamoun 2007*b*). On the other hand, *termhood* characterises the second step, which is to determine to what extent a stable lexical unit is related to a certain domain-specific concept. This paper focuses on developing a probabilistic framework for measuring termhood.

The tasks of termhood determination is different from the two well-known problems of named-entity recognition and information retrieval. The biggest dissimilarity between named-entity recognition and termhood determination is that the former is a deterministic problem of classification whereas the latter involves the subjective measurement of relevance and ranking. Hence, unlike the availability of various platforms for the evaluation of named-entity recognition such as the *BioCreAtIvE Task 1* (Hirschman, Yeh, Blaschke & Valencia 2005) and the *Message Understanding Conference (MUC)* (Chinchor, Lewis & Hirschman 1993), determining the performance of term extraction remains an extremely subjective problem domain. While appearing more similar to information retrieval in that both involves relevance ranking, the determination of termhood does have its unique requirements in processing text. Most importantly, the determination of termhood does not have user queries as evidences for deciding on relevance. As such, the only source of evidence for determining termhood is a set of heuristically-motivated term characteristics.

The surveys by (Cabre-Castellvi, Estopa & Vivaldi-Palatresi 2001) and (Kit 2002) show that existing term extraction methods rely on the above-mentioned hypothetical term characteristics to devise empirical measures for termhood. Consequently, these methods are often criticised for their lack of theorisation and mathematical validity. Such criticisms become obvious when one poses simple but crucial questions on the ways certain measures are derived, for example, *"Why taking different bases for logarithm?"*, or *"Why combining two weights using addition and not multiplication?"*.

To address the lack of proper theorisation in term extraction, we present a probabilistic framework for scoring and ranking term candidates to measure termhood. This measure is founded on Bayes Theorem and the Zipf-Mandelbrot model (Tullo & Hurford 2003) for computing the evidences. Our new measure is adaptable in that new or obsolete evidences can be added or removed based on different requirements of the system. Section 2 summarises some prominent methods in term extraction. Section 3 develops the probabilistic framework and so-derived measure of termhood. Section 4 presents the results of a comparative study and the paper concludes in Section 5 with an outlook to future works.

2 Related Work

Surveys (Cabre-Castellvi et al. 2001, Kageura & Umino 1996) on term extraction approaches revealed that most of the existing methods were based on ad-hoc statistical measures combined with linguistics information. These measures are usually put together using term or document frequency, and are modified as per need as the observation of immediate results progresses. As such, the significance of the different weights that compose the measures usually assume an empirical viewpoint. Obviously, such methods are at most inspired by, but not derived from formal models. Many critics claim that such methods are unfounded and the results that were reported using these methods are merely coincidental. In the words of (Kageura & Umino 1996), *"As for the validity of statistical methods or models, we have seen that many use intuitively reasonable by mathematically unfounded measures."*.

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

Copyright ©2007, Australian Computer Society, Inc. This paper appeared at Sixth Australasian Data Mining Conference (AusDM 2007), Gold Coast, Australia. Conferences in Research and Practice in Information Technology (CRPIT), Vol. 70. Peter Christen, Paul Kennedy, Jiuyong Li, Inna Kolyshkina and Graham Williams, Ed. Reproduction for academic, not-for profit purposes permitted provided this text is included.

is employed as a guide to determine if a word should be considered as an index term. This model is derived from the fact that single Poisson distribution is only a good fit for functional words while content words tend to deviate from it (vanRijsbergen 1979, Church & Gale 1995, Manning & Schutze 1999). Such variation from the Poisson distribution or colloquially known *"non-poissonness"* can then be employed as a predictor of whether a lexical unit is a content word or not, and hence as an indicator of possible termhood.

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).

(Basili, Moschitti, Pazienza & Zanzotto 2001) proposed a *TF-IDF* inspired measure for assigning terms with weights quantifying their specificity to the target domain. A *Contrastive Weight* is defined for a simple term candidate a in a target domain d 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 domain corpora, and $\sum_j f_{aj}$ is the sum of the frequencies of the term candidate a in all domain corpora. For complex term candidates, the frequencies of their heads are utilised to compute their weights:

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

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, a^h of the complex term candidate.

There is also the use of contextual evidence to assist in the identification of terms. One of the works is *NCvalue* by (Frantzi & Ananiadou 1997). Given that TC is the set of all term candidates and c is a noun, verb or adjective (i.e. context words) appearing with the term candidates, $weight(c)$ is defined as:

$$weight(c) = 0.5 \left(\frac{|TC_c|}{|TC|} + \frac{\sum_{e \in TC_c} f_e}{f_c} \right)$$

where TC_c is the set of term candidates that have c as a context word, $\sum_{e \in TC_c} f_e$ is the sum of the frequencies of term candidates that appear with c, and f_c is the frequency of c in the corpus. After calculating the weights for all possible context words, the sum of the weights of context words appearing with each term candidate can be obtained. Formally, for each term candidate a that has a set of accompanying context words C_a, the cumulative context weight is defined as:

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

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

$$NCvalue(a) = \frac{1}{\log F} Cvalue(a) cweight(a) \quad (3)$$

where F is the size of the corpus in terms of the number of words. $CValue(a)$ is given by:

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

where $|a|$ is the number of words in a, L_a is the set of potential longer term candidates that contain a, g is the longest n-gram considered, and f_a is frequency of occurrences of a in the corpus.

(Wong, Liu & Bennamoun 2007*a*) proposed a *Discriminative Weight (DW)* as part of a scoring and ranking scheme called *Termhood (TH)*. *DW* is a product of *Domain Prevalence (DP)* and *Domain Tendency (DT)*. If a is a simple term candidate, the domain prevalence 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 term candidates $j \in TC$ 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 a is complex, DP is defined as:

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

DT is employed to determine the extent to which term candidate a is used for domain purposes. It is defined as:

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

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. The adjustment of contextual evidence is also introduced to ensure that the weights of non-terms are not inflated by domain-relevant context. The *Adjusted Contextual Contribution (ACC)* is defined as:

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

$ACDW$ is simply the average DW of the context words of candidate a adjusted according to the context's relatedness to a:

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

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 use of NGD overcomes the problems associated with the use of semantic information. The final weight of each term candidate a is given by:

$$TH(a) = DW(a) + ACC(a) \tag{4}$$

3 A Probabilistic Framework for Determining Termhood

Terms are lexical realisations of their abstract counterparts (i.e. concepts) that are relevant to some domains of interest. As such, the aim of determining termhood in term extraction is to identify terms that are relevant to the same domain as are the concepts they represent. As pointed out before, the only source of evidence in term extraction is the characteristics of terms embedded in the domain corpora. From here on, notation d is used to denote the domain corpus and the target domain it represents, and $\bar{d}$ denotes the contrastive corpora, that is, all corpora other than d. In probabilistic terms, we can describe our aim of termhood determination as:

Aim 1 *What is the probability of candidate a being relevant to domain d, given the evidence candidate a has?*

3.1 A General Probabilistic Model

Definition 1 outlines the primary characteristics of terms (Kageura & Umino 1996). These characteristics are considered as ideal because they rarely exist in real-world situations as we will discuss later.

Definition 1 *The primary characteristics of terms in ideal settings:*

1. A term should not have any synonyms.

2. The meaning of a term is independent of context.

3. The meaning of a term should be precise and related directly to a concept.

In addition, there are other characteristics that are equally important in the determination of termhood. Some of these characteristics are inherent general properties of words. This list is not a standard and by no means exhaustive. They are:

Definition 2 *Extended characteristics of terms*

1. Terms are properties of domains, not documents (Basili et al. 2001).

2. Terms tend to clump together (Bookstein, Klein & Raita 1998) the same way as content-bearing words do (Zipf 1949).

3. Terms of longer length are rare in a corpus since the usage of phrases of shorter length are more predominant (Zipf 1935).

4. Simple terms are often ambiguous and modifiers are required to reduce the number of possible interpretations. Hence complex terms are usually preferred in terminology (Frantzi & Ananiadou 1997).

Definition 1 states that a term is unambiguously relevant to a domain. For example, when one encounter the term *"bridge"*, there should be one and exactly one meaning: *"a device that connects multiple network segments at the data link layer"*. As such, an *ideal term* cannot be relevant to more than one domain. Therefore, determining termhood is simplified to just measuring the extent to which a term candidate is relevant to a domain regardless of its relevance to other domains. Aim 1 can thus be formulated as a conditional probability between two events using Bayes Theorem.

$$P(R_1|A) = \frac{P(A|R_1)P(R_1)}{P(A)} \tag{5}$$

where R_1 is the event that a is relevant to domain d and A is the event that a is a candidate term supported by an evidence vector $V =< E_1, ..., E_m >$. The details of the evidence vector will be presented in Section 3.2. $P(R_1|A)$ is the posterior probability of candidate a being relevant to d given the evidence vector V. $P(R_1)$ is the prior probability of candidate a being relevant without any evidence, and $P(A)$ is the prior probability of a being a candidate with evidence V. As we shall see later, these two prior probabilities will be immaterial in the final computation of the weights for the candidates. Since the reliability of the evidences of terms is dependent on the representativeness of the corpus, the following Assumptions 1 and 2 are also necessary:

Assumption 1 *Corpus d is a balanced, unbiased and randomised sample of the population text representing the corresponding domain.*

Determining the representativeness of a corpus is important but beyond the scope of this paper.

Assumption 2 *Contrastive corpora $\bar{d}$ is the set of balanced, unbiased and randomised sample of the population text representing approximately all major domains other than d.*

Given Assumption 2, let us define R_2 as the event that candidate a is relevant to other domains $\bar{d}$. Following this and based on Definition 1, we have $P(R_1 \cap R_2) = 0$. In other words, R_1 and R_2 are mutually exclusive in ideal settings. Ignoring the fact that a term may appear in certain domains by chance, any candidate a can either be relevant to d or to $\bar{d}$, but not both.

Unfortunately, according to (Loukachevitch & Dobrov 2004), *"An impregnable barrier between words of a general language and terminologies does not exist."*. For example, the word *"bridge"* has multiple meanings and is relevant to more than one domain. In other words, $P(R_1 \cap R_2)$ is rarely 0 in reality. However, words like *"bridge"* are regarded as poor choice of terms because they are simple terms and inherently ambiguous as defined in Definition 2. Instead, a better term for denoting the concept which the candidate *"bridge"* attempts to represent would be *"network bridge"*. This is usually the case in writings where authors first introduce new concepts using unambiguous complex terms and later, reiterating the same concepts with shorter terms. As such, we assume that:

Assumption 3 *Each concept represented using a polysemous simple term in corpus has a corresponding unambiguous complex term representation occurring in the same corpus.*

From Assumption 3, since all important concepts of a domain have unambiguous manifestation in the corpus, the possibility of the ambiguous counterparts being inappropriately ranked during our termhood measurement will have no effect on the overall term extraction output. As such, polysemous simple terms can be considered as insignificant in our determination of termhood. Based on Definition 1 and Assumption 3, the probability of relevance of candidate a to both d and $\bar{d}$ is approximately 0, i.e.

$P(R_1 \cap R_2) \approx 0$. Following this, we have $P(R_1 \cup R_2) = P(R_1)+P(R_2) \approx 1$. This approximation of the sum of the probability of relevance without evidence can be extended to the conditional probability of relevance given evidence vector V:

$$P(R_1|A) + P(R_2|A) \approx 1 \tag{6}$$

without violating the axioms of probability.

Since $P(R_1 \cap R_2)$ only approximates to 0 in reality, determining the probability of relevance of candidate a to d alone may not be enough. We need to calculate the odds of relevance to demonstrate that candidate a is more relevant to d than to $\bar{d}$:

Aim 2 What are the odds of candidate a being relevant to d given the evidence candidate a has?

Since $Odds = P/(1-P)$, we can obtain the odds of relevance given the evidence candidate a has by applying $(1 - P(R_1|A))^{-1}$ to Equation 5:

$$\frac{P(R_1|A)}{1 - P(R_1|A)} = \frac{P(A|R_1)P(R_1)}{P(A)(1 - P(R_1|A))} \tag{7}$$

and since $1 - P(R_1|A) \approx P(R_2|A)$ from Equation 6, and by applying the multiplication rule $P(R_2|A)P(A) = P(A|R_2)P(R_2)$ to the left side of Equation 7, we have:

$$\frac{P(R_1|A)}{P(R_2|A)} = \frac{P(A|R_1)}{P(A|R_2)} \frac{P(R_1)}{P(R_2)} \tag{8}$$

Equation 8 can also be called as the odds of relevance of candidate a to d given the evidence a has. This odds can be used to rank the term candidates. Taking the log of odds (i.e. logit) gives us

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

$P(A|R_1)$ and $P(A|R_2)$ are the class conditional probabilities for a being a candidate with evidence vector V given its different state of relevance. Since the chance of any candidate being relevant to d and to $\bar{d}$ without any evidence is the same (i.e. $P(R_1)/P(R_2) = 1$), we can safely ignore the second term (i.e. the odds of relevance without evidence) in Equation 8. This gives us

$$\log \frac{P(A|R_1)}{P(A|R_2)} = \log \frac{P(R_1|A)}{P(R_2|A)} \tag{9}$$

To score and rank the term candidates $a \in TC$ based on the evidences they have, we define the *Odds of Termhood (OT)* as

$$OT(a) = \log \frac{P(A|R_1)}{P(A|R_2)} \tag{10}$$

Since we are only interested in the relative ranking, ranking candidates using OT, according to Equation 9, is the same as ranking the candidates according to our Aim 2 as formulated in Equation 8. Obviously, from Equation 10, our initial predicament on not being able to empirically determine prior probabilities $P(A)$ and $P(R_1)$ is no longer a problem.

Assumption 4 *Independence between evidences in V.*

Next we can decompose the evidence vector V associated with each candidate a to enable the assessment of the class conditional probabilities $P(A|R_1)$ and $P(A|R_2)$. Given Assumption 4, $P(A|R_1)$ and $P(A|R_2)$ can be expanded as

$$P(A|R_1) = \prod_i P(E_i|R_1) \tag{11}$$
$$P(A|R_2) = \prod_i P(E_i|R_2)$$

where $P(E_i|R_1)$ and $P(E_i|R_2)$ is the probability of a as a candidate associated with evidence E_i given its different state of relevance. Substituting Equation 11 in 10 will give us

$$OT(a) = \sum_i \log \frac{P(E_i|R_1)}{P(E_i|R_2)} \tag{12}$$

Lastly, to simplify the notation, individual scores are defined for each evidence E_i, and we call them *evidential weights (O_i)*

$$O_i = \frac{P(E_i|R_1)}{P(E_i|R_2)} \tag{13}$$

and substituting Equation 13 in 12 gives us

$$OT(a) = \sum_i \log O_i \tag{14}$$

The purpose of OT is similar to many other functions for scoring and ranking term candidates such as those reviewed in Section 2. However, what differentiates our new function from the existing ones is the fact that OT is founded upon and derived in a probabilistic framework with explicit assumptions. Moreover, as shown in the following Section 3.2, the individual evidences themselves are formulated based on probability theory and the necessary term distributions are derived from formal distribution models.

3.2 Formalising Evidences in a Probabilistic Framework

Commonly adopted characteristics for determining the relevance of terms (Kageura & Umino 1996) are highlighted below in Definition 3.

Definition 3 *Characteristics of term relevance*

1. A term candidate is relevant to a domain if it appears relatively more frequent in that domain than in others.

2. A term candidate is relevant to a domain if it appears only in one domain.

3. A term candidate relevant to a domain may have biased occurrences in that domain.

4. A complex term candidate is relevant to a domain if its head is specific to that domain.

We propose seven evidences for the evidence vector V to capture the characteristics presented in Definition 2 and 3. They are as follow:

Evidence 1: Occurrence of candidate a

Evidence 2: Existence of candidate a

Evidence 3: Specificity of the head a^h of a

Evidence 4: Uniqueness of candidate a

Evidence 5: Exclusivity of candidate a

Evidence 6: Pervasiveness of candidate a

Evidence 7: Clumping tendency of candidate a

However, due to space limitations, here we present Evidence 3 and 4 as a proof of concept. It is worthwhile to note that evidences can always be introduced or removed depending on the goal or constraints imposed upon the applications implementing OT of Equation 14. The various evidences contribute to the computation of the corresponding evidential weights O_i, which in turn are summed to produce the final ranking of OT. Since OT serves as a probabilistically-derived formulaic realisation of our Aim 2, the various O_i can be considered as manifestations of sub-aims derivable from Aim 2. Each sub-aim is formulated into its corresponding evidential weight using the probability distributions of the occurrences of term candidates in d and in $\bar{d}$:

- P(occurrence of a in d)=$P(a, d)$ is the probability of occurrence of a in the domain corpus d.
- P(occurrence of a in $\bar{d}$)=$P(a, \bar{d})$ is the probability of occurrence of a in the contrastive corpora $\bar{d}$.

There are a few possible models for such distribution. One of them is the Zipf-Mandelbrot model which have been rigorously discussed by (Tullo & Hurford 2003). We use the Zipf-Mandelbrot model to obtain the distributions for both $P(a, d)$ and $P(a, \bar{d})$.

3.2.1 Odds of Specificity

The evidential weight O_3 focuses specifically on Definition 3.4 for complex term candidates. O_3 is meant for capturing the odds of whether the inherently ambiguous head a^h of a complex term a is specific to d. If the head a^h of a complex terms is found to occur individually without a in large numbers across different domains, then the specificity of the concept represented by a^h and a in d is doubtable. O_3 can be formally stated as:

> **Sub-Aim 3** What are the odds that the head a^h of a complex term candidate a is specific to d?

The head of a complex term candidate is considered as specific to a domain if the head and the candidate itself both have higher tendency of occurring together in that domain. The higher the chances of co-occurrence of a and a^h in a domain, the more specific is a^h to that domain. For example, if the event of both *"bridge"* and *"network bridge"* occurring together in the *"computer networking"* domain is very high, this means the possibly ambiguous head *"bridge"* is used in a very specific context in that domain. In such cases, when *"bridge"* is encountered in *"computer networking"*, one can safely deduce that it refers to the same domain-specific concept as *"network bridge"*. Consequently, the more specific the head a^h is with respect to d, the less ambiguous its occurrence is in d. From Definition 3.4, the less ambiguous a^h is, the higher are the chances of its complex counterpart a being relevant to d.

To proceed further, we assume that the occurrences of candidate a and its head a^h within the same domain (i.e. either d or $\bar{d}$) are independent. Even though the independence assumption may not always be the case in reality, it does remove many complications related to the non-trivial formulation of O_3 and other evidential weights not presented here. As such,

- P(occurrence of a in $d \cap$ occurrence of a^h in d) = $P(a, d)P(a^h, d)$

Based on the assumptions above, we define O_3 for complex term candidates as:

$$
\begin{aligned}
O_3 &= \frac{P(\text{specificity of } a|R_1)}{P(\text{specificity of } a|R_2)} \\
&= \frac{P(\text{specificity of } a \text{ to } d)}{P(\text{specificity of } a \text{ to } \bar{d})} \\
&= \frac{P(\text{occurrence of } a \text{ in } d \cap \text{occurrence of } a^h \text{ in } d)}{P(\text{occurrence of } a \text{ in } \bar{d} \cap \text{occurrence of } a^h \text{ in } \bar{d})} \\
&= \frac{P(a, d)P(a^h, d)}{P(a, \bar{d})P(a^h, \bar{d})}
\end{aligned}
$$

3.2.2 Odds of Uniqueness

This evidential weight O_4 attempts to realise Definition 3.2. O_4 captures the odds of whether a is unique to d or to $\bar{d}$. A term candidate is considered as unique if it occurs only in one domain and not the other. Formally, O_4 is described as

> **Sub-Aim 4** What are the odds of term candidate a being unique to d?

Based on the following intuitively reasonable independence and complementary assumption of the events of occurrence and non-occurrence of candidate a,

- P(non-occurrence of a in d) $= 1 - P(a, d)$
- P(occurrence of a in $d \cap$ non-occurrence of a in $\bar{d}$) $= P(a, d)(1 - P(a, \bar{d}))$

we can mathematically formulate O_4 as:

$$
\begin{aligned}
O_4 &= \frac{P(\text{uniqueness of } a|R_1)}{P(\text{uniqueness of } a|R_2)} \\
&= \frac{P(\text{uniqueness of } a \text{ to } d)}{P(\text{uniqueness of } a \text{ to } \bar{d})} \\
&= \frac{P(\text{occurrence of } a \text{ in } d \cap \text{non-occurrence of } a \text{ in } \bar{d})}{P(\text{occurrence of } a \text{ in } \bar{d} \cap \text{non-occurrence of } a \text{ in } d)} \\
&= \frac{P(a, d)(1 - P(a, \bar{d}))}{P(a, \bar{d})(1 - P(a, d))}
\end{aligned}
$$

4 Experiments

To evaluate the effectiveness of our probabilistic framework for determining termhood using *Odds of Termhood (OT)*, here we carry out a comparative study with three existing scoring and ranking scheme, namely, *Contrastive Weight (CW)*, *NCvalue (NCV)* and *Termhood (TH)*. The implementation of CW, NCV and TH are in accordance to Equation 1 and 2, 3, and 4 respectively. Although only two evidences are presented in this paper, we have included all seven evidences during our implementation to obtain the new measure OT in this experiment. The summary of the data sets is presented in Table 1. The data sets consist of three distinct domain corpora d and a collection of contrastive corpora $\bar{d}$. The domain corpora consist of three distinct collections of text gathered from `BioMedCentral.com` in the area of *"musculoskeletal disease"* (denoted by d_{MUS}), *"cancer"* (denoted by d_{CAN}), and *"cardiovascular disease"* (denoted by d_{CAR}). The contrastive corpora is a single collection of news articles across a wide range of genres gathered from various sources such as `Reuters.com`, `CNet.com` and `ABC.com` between the period of February 2006 and April 2007. The term candidates and context words are extracted as instantiated sub-categorisation frames (Wong 2005) from

Table 1. Summary of the datasets employed throughout this paper for experiments and evaluations. For simplicity reasons, two notations are adopted: d to represent the domain corpus, and $\bar{d}$ to represent the contrastive corpora.

	Notation	Source	Domain	No. of documents (N)	No. of words (F)	Average no. of words per document
Contrastive corpora	$\bar{d}$	Reuters	Business	2,691	987,305	
			Sports	2,306	792,902	
			Politics	2,187	871,788	
			United States local news	612	223,588	
			Entertainment	2,280	862,233	
			Health	1,039	387,905	
		CNet	Technology	3,375	1,626,849	
		ABC	Australian local news	1,394	349,301	
		Discovery	Travel	291	138,178	
			History	65	31,253	
			Wildlife	247	117,016	
		AllRecipes	Cooking recipe	552	89,644	
			Total	***17,039***	***6,477,962***	***380.18***
Domain corpus 1	d_{MUS}	BioMed Central	Musculoskeletal diseases	302	860,601	
			Total	***976***	***2,533,717***	***2,596.02***
Domain corpus 2	d_{CAN}	BioMed Central	Cancer	453	1,043,070	
			Total	***453***	***1,043,070***	***2,302.58***
Domain corpus 3	d_{CAR}	BioMed Central	Cardiovascular diseases	282	607,973	
			Total	***282***	***607,973***	***2,155.93***

the three domain corpora. As a result, three sets of term candidates are produced, one for each of the three domains (i.e. d_{MUS}, d_{CAN} and d_{CAR}). The experiments on three different sets of terms from three different domain corpora are necessary to demonstrate the consistency of any trends exhibited by the four measures.

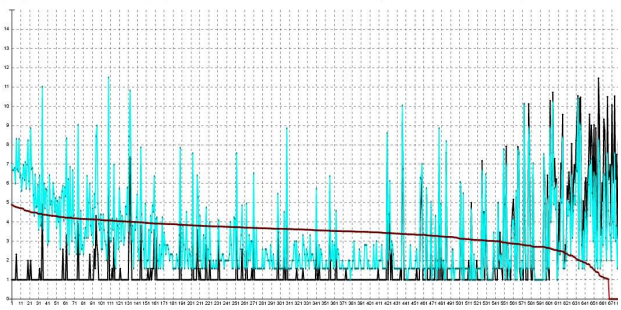
The four measures are used to score and rank the three sets of term candidates. The frequency distributions of the ranked candidates from d_{MUS}, d_{CAN} and d_{CAR} are shown in Figures 1, 2 and 3. The candidates are ranked in descending order according to their scores assigned by the respective measures. The first half of the graphs by CW, prior to the sudden surge of frequency consisted of only complex terms. Complex terms tend to have lower word counts compared to simple terms and hence, the disparity in the frequency distributions are clearly shown in Figures 1(c), 2(c) and 3(c). This is attributed to the biased treatment given to complex terms evident in Equation 2. However, priority is also given to complex terms by TH and OT, but as one can see from the distributions of candidates by TH in Figures 1(b), 2(b) and 3(b) and those by OT in Figures 1(a), 2(a) and 3(a), such undesirable trend does not occur. One of the explanation is CW relies heavily on frequency while TH and OT attempt to diversify the evidences. Even though frequency is a reliable source of evidence, the use of it alone is definitely inadequate (Cabre-Castellvi et al. 2001). As for the NCV measure, Figures 1(d), 2(d) and 3(d) show that scores for term candidates are calculated solely based on their domain frequencies. In other words, NCV is not suitable for performing contrastive analysis and hence, cannot be employed for term extraction to identify between domain-specific terms. Another advantage of TH and OT is their ability to assign higher weights to terms that occur relatively more frequent in d than in $\bar{d}$. This is evident through the gap between f_d and $f_{\bar{d}}$, especially at the beginning of the x-axis. Candidates along the end of the x-axis are those with $f_{\bar{d}} > f_d$. However, the discriminating power of OT is better since the gap between f_d and $f_{\bar{d}}$ is wider and lasted longer.

In addition to the absence of evaluation methods and datasets, the subjective nature of termhood assessment makes the tasks of objectively evaluating

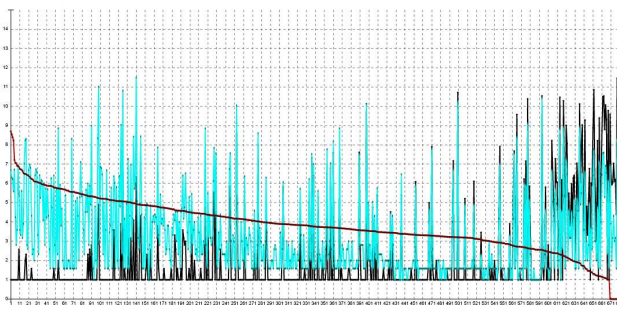
Table 2. A summary of the mean (μ) and standard deviation (σ) of the scores assigned by the four measures (i.e. NCV, CW, TH and OT) for the three sets of term candidates extracted from three different domain corpora (i.e. d_{MUS}, d_{CAN} and d_{CAR}). The sum of the domain frequencies and of the contrastive frequencies of the three sets of term candidates are also shown in this table.

		μ *of weights*	σ *of weights*	$\sum f_d$	$\sum f_{\bar{d}}$	$\frac{\sum f_d}{\sum f_{\bar{d}}}$
d_{MUS}	***NCV***	2105.71	35435.75	33286	65396	0.51
	CW	43.10	132.91			
	TH	20.98	32.63			
	OT	10.44	5.93			
		μ *of weights*	σ *of weights*	$\sum f_d$	$\sum f_{\bar{d}}$	$\frac{\sum f_d}{\sum f_{\bar{d}}}$
d_{CAN}	***NCV***	112886.02	1209270.22	120820	134936	0.90
	CW	49.80	241.98			
	TH	25.71	38.05			
	OT	10.77	5.22			
		μ *of weights*	σ *of weights*	$\sum f_d$	$\sum f_{\bar{d}}$	$\frac{\sum f_d}{\sum f_{\bar{d}}}$
d_{CAR}	***NCV***	1778.82	29670.35	38740	88790	0.44
	CW	34.78	183.13			
	TH	15.90	17.06			
	OT	10.83	4.98			

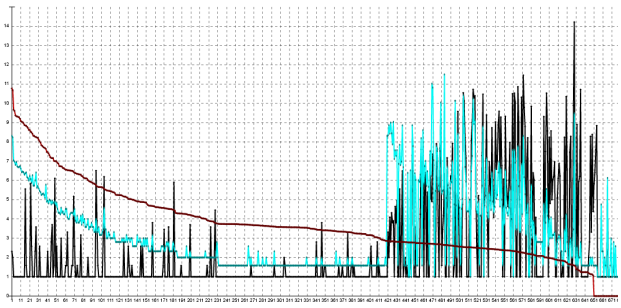
and comparing our new measure with existing approaches a problem domain by itself. In the words of (Damle & Uren 2005), *"Unfortunately, there is no objective evaluation method reported in the literature for term extraction..."*. However, our subjective assessments of OT and three other existing measures offer promising insights into probabilistically-derived measures for termhood. Table 2 summarises the mean and standard deviation of the weights generated by the various measures. One can notice the extremely high dispersion from the mean of the weights generated by CW and also NCV. We speculate that such trends are due to the erratic assignments of weights, heavily influenced by domain frequencies. This is further enforced by the visible increase of the means and standard deviations of the weights produced by CW, NCV and even TH as the domain frequency f_d increases. On the other hand, our probabilistically-motivated measure OT appeared to be unaffected by



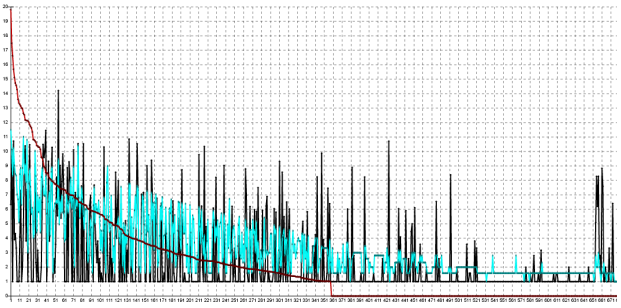
(a) Candidates ranked by OT



(b) Candidates ranked by TH

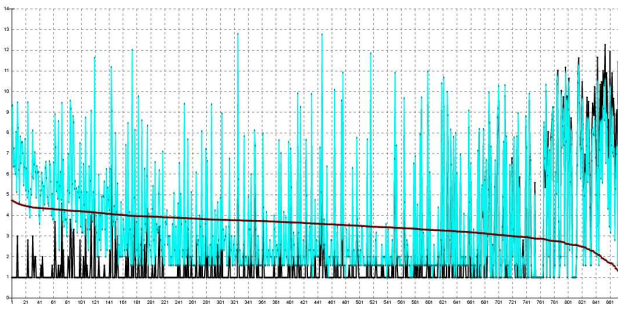


(c) Candidates ranked by CW

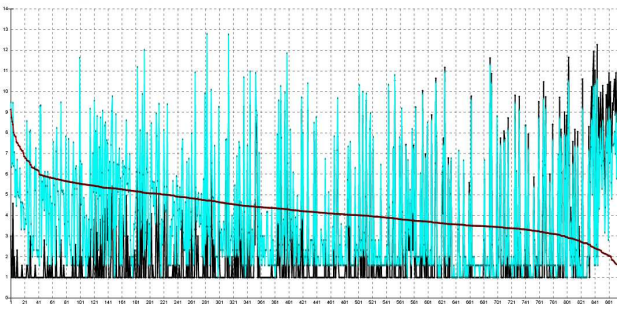


(d) Candidates ranked by NCV

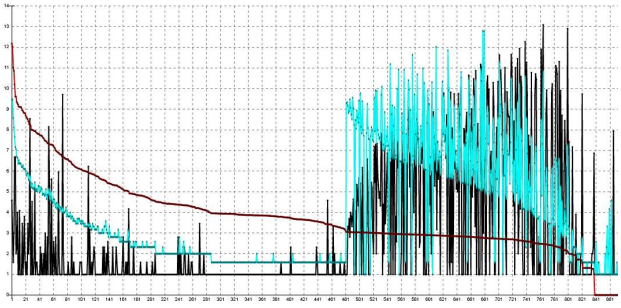
Figure 1: This graph shows the frequency distributions of 709 candidates extracted from d_{MUS} ranked in descending order according to the scores assigned by the respective measures. The black, oscillating line represents the frequency of the candidates in the contrastive corpora $f_{\bar{d}}$ while 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.



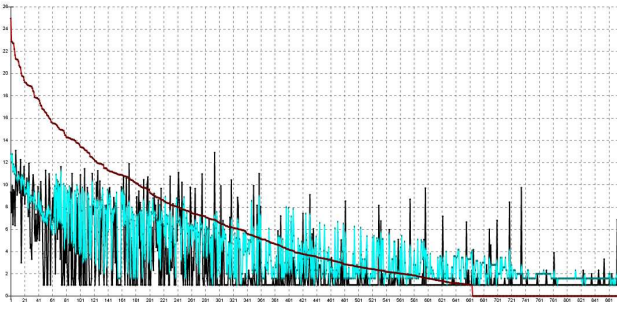
(a) Candidates ranked by OT



(b) Candidates ranked by TH

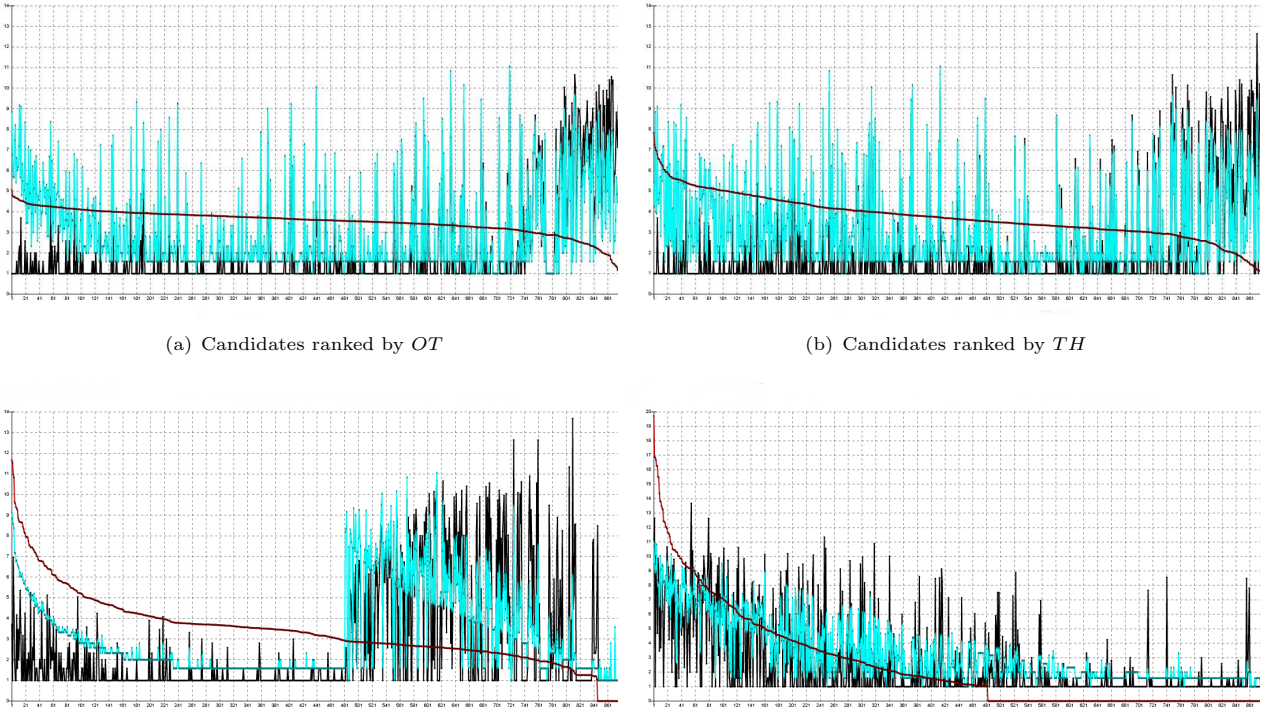


(c) Candidates ranked by CW



(d) Candidates ranked by NCV

Figure 2: This graph shows the frequency distributions of 709 candidates extracted from d_{CAN} 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.



(a) Candidates ranked by OT

(b) Candidates ranked by TH

(c) Candidates ranked by CW

(d) Candidates ranked by NCV

Figure 3: This graph shows the frequency distributions of 709 candidates extracted from d_{CAR} 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.

Table 3. The Spearman rank correlation coefficients ρ between all pairs of measures over the three sets of term candidates extracted from d_{MUS}, d_{CAN} and d_{CAR}.

	ρ	*NCV*	*CW*	*TH*	*OT*
d_{MUS}	*NCV*	1	0.2527	0.3421	0.3321
	CW	0.2527	1	0.5958	0.6333
	TH	0.3421	0.5958	1	0.8565
	OT	0.3321	0.6333	0.8565	1
	ρ	*NCV*	*CW*	*TH*	*OT*
d_{CAN}	*NCV*	1	0.0053	0.2138	0.1457
	CW	0.0053	1	0.4637	0.5313
	TH	0.2138	0.4637	1	0.8289
	OT	0.1457	0.5313	0.8289	1
	ρ	*NCV*	*CW*	*TH*	*OT*
d_{CAR}	*NCV*	1	0.1221	0.1985	0.1467
	CW	0.1221	1	0.4737	0.5129
	TH	0.1985	0.4737	1	0.8211
	OT	0.1467	0.5129	0.8211	1

the changes in frequencies. We also employ the Spearman rank correlation coefficient to study the possibility of any correlation between the four ranking schemes under evaluation. Table 3 summarises the coefficients between the various measures. Note that there is a strong correlation between the ranks produced by our new probabilistic measure OT and the ranks by the ad-hoc measure TH. This correlation is consistent throughout all the experiments using different sets of term candidates. The correlation of TH with OT reveals the possibility of providing mathematical justifications for the former's heuristically-motivated ad-hoc approach using a general probabilistic framework. We believe by adjusting the inclusion or exclusion of various evidences, other ad-hoc measures can be captured as well.

5 Conclusions

In this paper, we presented a probabilistically-derived measure termed as the *Odds of Termhood (OT)* for scoring and ranking term candidates for term extraction. We have also introduced seven evidences, founded on formal models of word distribution, to facilitate the calculation of OT. The evidences are motivated by characteristics of terms in a domain, which are made explicit. The fact that evidences can be added or removed makes OT a highly flexible framework that is adaptable to the applications' requirements and constraints. Our experiments comparing OT with three other existing ad-hoc measures, namely CW, NCV and TH have demonstrated the effectiveness of the new measure and the new framework.

More research is required for introducing new evidences to realise more term characteristics. Due to the difficulty in objectively determining the performance of termhood measures, we intend to assess OT within the scope of a larger application such as document retrieval to establish its precision, recall and accuracy.

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 extrac-

tion, *in* 'Proceedings of the 4th Terminology and Artificial Intelligence Conference (TIA)', France.

Bookstein, A., Klein, S. & Raita, T. (1998), 'Clumping properties of content-bearing words', *Journal of the American Society of Information Science* **49**(2), 102–114.

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.

Chinchor, N., Lewis, D. & Hirschman, L. (1993), 'Evaluating message understanding systems: An analysis of the third message understanding conference (muc-3)', *Computational Linguistics* **19**(3), 409–449.

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.

Hirschman, L., Yeh, A., Blaschke, C. & Valencia, A. (2005), 'Overview of biocreative: Critical assessment of information', *BMC Bioinformatics* **6**(1), S1.

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.

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

Loukachevitch, N. & Dobrov, B. (2004), Sociopolitical domain as a bridge from general words to terms of specific domains, *in* 'Proceedings of the 2nd International Global Wordnet Conference'.

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

Tullo, C. & Hurford, J. (2003), Modelling zipfian distributions in language, *in* 'Proceedings of the ESSLLI Workshop on Language Evolution and Computation', Vienna.

vanRijsbergen, C. (1979), Automatic text analysis, *in* 'Information Retrieval', University of Glasgow.

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 termhood for learning domain ontologies using domain prevalence and tendency, *in* 'Proceedings of the 6th Australasian Conference on Data Mining (AusDM)', Gold Coast.

Wong, W., Liu, W. & Bennamoun, M. (2007*b*), 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.

Zipf, G. (1935), The psycho-biology of language, Houghton Mifflin, Boston, MA.

Zipf, G. (1949), Human behaviour and the principle of least-effort, Addison-Wesley, Cambridge, MA.