Using Multilingual Terms for Biomedical Term Extraction

Ismail Fahmi, Gosse Bouma, and Lonneke van der Plas
{i.fahmi,g.bouma,M.L.E.van.der.Plas}@rug.nl
Fax: +31 (0)50 363 4900
Information Science, University of Groningen
Oude Kijk in't Jatstraat 26, 9712 EK Groningen, The Netherlands

Abstract

The goal of automatic term extraction often is not so much the creation of a new list of domain specific terms, but rather the (semi-) automatic extension of a list of known terms. In this paper, we focus on the use of existing terms from glossaries, thesaurus, or ontologies to extract new terms from a domain specific text. Our new method is used to extract language-specific terms with the help of multilingual terminological resources. Our baseline system combines a linguistic pattern for extracting candidate noun phrases with a statistical method (χ^2) for ranking candidate phrases according to their association strength in a domain-specific corpus. Our scoring method also takes into account the termhood of candidate phrases computed on the basis of a list of known terms. We show that uninterpolated average precision of the resulting term list is improved when tested using human evaluators.

Keywords

Automatic term extraction, association measures, multilingual terminology, multi-word terms

1 Introduction

Automatic extraction of information from text is very important for many applications (e.g. information retrieval, question answering, and for bootstrapping or extending ontologies for the semantic web). The goal of automatic term extraction systems is to identify candidate terms in (more or less) unstructured text. Many different methods have been proposed, using linguistic techniques [3, 1], statistical techniques [6, 17], or a combination of both [14, 21]. Most of the statistical methods use only frequency of occurrence of the candidate terms in text as a parameter to measure the probability of multi-word terms (e.g. using log-likelihood, mutual information, or C-value).

Previous work by Jacquemin *et al.* [16] suggests that the use of an initial term set can improve the accuracy of term extraction. In a particular domain where the availability of a list of known terms is rare, we can try to use available multilingual terms. In a medical domain, for example, the UMLS (Unified Medical Language System)[1] is a valuable multilingual terminology resource which can be exploited to improve the extraction of terms in a specific language (e.g. Dutch).

We have developed a method for ranking candidate terms, extracted from medical corpora in Dutch, with the help of the UMLS as an external knowledge source. The use of multilingual terminology to extract Dutch medical terms is motivated by the fact that Dutch and English words often have the same stems, for example, *leukemische/leukemic* → *leukemic*, *cellen/cells* → *cell*, and *bacterie/bacterium* → *bacteri*. In general, the existing words from other resources are modified through affixation, compounding, creation of phrasal terms, conversion (e.g., a verb used as a noun) and compression (e.g., abbreviation, acronym, etc). The modification of existing resources is considered the most productive method in creating new terms [24].

We concentrate on extraction of phrasal terms. Our method (UT-Score) combines frequency of occurrence of candidate terms in a corpus (*unithood*) and information on how the candidate terms are formed computed from existing multilingual terms (*termhood*).

The method is evaluated through the extraction of new terms from two Dutch medical corpora (an encyclopedia and a handbook, approximately 1M words in total). We use the best setting of our experiments to rank the candidate terms, and improve it with the help of a subset of multilingual terminologies from the UMLS as our external knowledge.

Our approach for extracting Dutch medical terms is summarized in the following steps:

- Select the best statistical association measure (unithood) that will be used as a baseline system (section 3).
- Apply the best method to extract language-specific terms with the help of external knowledge from multilingual terminology resources (section 4).

2 Previous work

The use of external knowledge to enhance the performance of traditional statistical methods of term extraction has been reported in several previous studies. Maynard & Ananiadou [22] compute the similarity of a candidate term and the context terms it occurs with, with terms found in the UMLS Metathesaurus and Semantic Network. Information Weight is combined with statistical information obtained using the NC-

[1] UMLS http://umlsinfo.nlm.nih.gov

value method [14, 13] and is used to rank candidate terms.

Mukherjae *et al.* [23] develop the BioAnnotator system which uses three dictionaries (UMLS, LocusLink, and GeneAlias) to discover biological terms among noun phrases extracted from shallow-parsed documents. After removing stop words from the beginning and end of the phrases, the system searches the dictionaries to find the stripped phrases. It then uses a rule engine to generate phrases missing from the dictionaries.

The use of the UMLS as a multilingual thesaurus for multilingual term extraction has been reported by Déjan *et al.* [9]. They extract new lexicons from the UMLS, which is dominated by English language, to enrich a German thesaurus, by exploiting a bilingual dictionary and the hierarchical information contained in the thesaurus.

Valderrabanos *et al.* [27] also reported the use of initial term sets to extract multilingual terminologies in English, Spanish, French and German. They used as initial terms the keywords in the description of each document in their corpora. New terms are generated from the initial terms through a set of derivation rules (67 different rules for each language). For example, given an initial term *head and neck neoplasms*, the rules will generate a new term *neck neoplasms*. The generated terms are then validated by checking their occurrence in the corpora.

Our method is different from the previously mentioned methods, in which a set of multilingual terms are exploited to measure the *termhood* of candidate terms. A small set of rules, similar to stemming rules in [27], is created to normalize the multilingual terms and the candidate terms (section 3.4).

3 Term extraction baseline

In this section we discuss the merits of two linguistic filters for identifying candidate terms, using POS-tags and syntax respectively. Furthermore, we select a baseline from the eight different statistical measures for ranking multiword candidate terms based on their distribution in a corpus. Our baseline system combines a linguistic filter based on a regular expression over POS-tags with χ^2 as the best measure.

3.1 Corpora used

We used two Dutch medical corpora for experiments: Elseviers medical encyclopedia (379K words), a medical encyclopedia intended for the general audience,[2] and the Dutch edition of the Merck Manual (780K words)[3], a general-purpose medical handbook intended for professionals. Both corpora were parsed syntactically using the Alpino parser [28, 19].

[2] The encyclopedia was made available to us by Spectrum b.v., and can also be found online at `www.kiesbeter.nl/medischeinformatie/`

[3] `www.merckmanual.nl`

3.2 Creating a gold standard and a list of multilingual terms

We create a gold standard to automatically annotate the results of the statistical measures in section 3.5. The standard is collected from a list of known Dutch terms from the medical domain of various sources such as Gezondheid.nl[4], Elsevier's Medical Encyclopedia, ICD-9 DE (International Classification of Diseases, 9th revision, Dutch Edition)[5], terms in a manually annotated medical corpus, and the titles of medical lemmas from Wikipedia (NL)[6]. In total we compiled a list of 27,621 unique terms.

To measure the termhood of candidate terms in section 4.3, we use the UMLS Specialist Lexicon[7] which contains 286.998 terms. And to avoid the same-language bias, we subtract all Dutch terms which are found in our known terms above from the UMLS. Thus, we assume that the resulting subset of the UMLS lexicon (284.706 terms) does not contain any Dutch terms.

3.3 Extraction of candidate terms

Terms typically consist of a nominal head and one or more adjectival or PP modifiers. Thus, they can be extracted by the following regular expression over POS-tags (proposed by Justeson & Katz [17], adapted by us for Dutch):

```
((Adj|N)+|(((Adj|N)*(N Prep)?)(Adj|N)*))N
```

Using a regular expression over POS-tags is robust, but it also has a number of potential disadvantages: as a longest match is applied, terms within terms are not extracted, some strings will be extracted that are not part of a single NP, and finally, some linguistic structures (most importantly, coordination) are not taken into account. As we have syntactic structures at our disposal, one might therefore also consider using a syntactic filter. Using the XML-query language XQuery,[8] we extracted all NPs from the corpus, with the exception of temporal NPs and NPs containing a relative clause or clausal complement. After removal of initial determiners and adverbial phrases, the terminal strings of the extracted NPs were returned as candidate terms. Table 1 gives some examples of terms extracted using the POS-tag filter and the syntactic filter. The term *ziekte* 'disease' is among the most frequent terms from both extraction filters.

We evaluated the recall of both extraction methods by computing the overlap between the list of extracted terms and the list of known terms described in section 3.2 above. Figure 1 shows the number of known terms extracted by the POS-tag and syntactic method, respectively. Note that there is considerable overlap between the two. As noted above, the POS-tag filter misses some terms (e.g. *Actinomyces israeli*) within terms (e.g. *bacterie Actinomyces israeli*), and also fails to extract coordinations (e.g. *aandoening*

[4] `www.gezondheid.nl`

[5] `icd9cm.chrisendres.com`

[6] `nl.wikipedia.org/wiki/Gezondheid_van_A_tot_Z`

[7] www.nlm.nih.gov/research/umls/meta4.html

[8] `www.w3.org/TR/xquery/`

Example	POS	Syn
ziekte	1008	925
aandoening van de nieren of urinewegen	0	1
belemmering van de beweeglijkheid	1	0
bacterie Actinomyces israeli	1	1
Actinomyces israeli	0	1
afschilfering van de vingertoppen	1	0

Table 1: *The number of times a given candidate term was extracted using the POS-tag and the syntactic filter.*

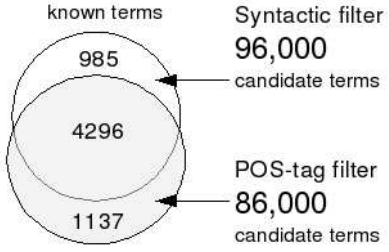

Fig. 1: *The numbers of known terms extracted by the POS-tag and syntactic filter.*

van de nieren of urinewegen 'disorder of the kidneys or the urinary systems'). The syntactic filter on the other hand, suffers from attachment errors (typically involving coordination and PP-modifiers, e.g. *afschilfering van de vingertoppen* 'desquamation of the finger tops' in a longer string *roodheid en afschilfering van de vingertoppen* 'redness and desquamation of the finger tops') and misses subphrases which do not correspond to a full NP (i.e. phrases consisting of a noun and one of its modifiers, e.g. *belemmering van de beweeglijkheid* 'obstruction of the movability' in a longer string *belemmering van de beweeglijkheid in de vinger voor de patiënt* 'obstruction of the movability in the finger for the patient').

As the POS-tag filter currently has a somewhat higher recall while extracting less candidate terms, we decided to work with the candidate terms extracted by this filter in the rest of the experiments.

3.4 Preposition filtering and suffix stripping

To reduce noise from the regular expression filter, we apply a preposition filtering in which candidate terms containing particular prepositions–e.g. *als* 'such as', *zoals* 'such as', *tegen* 'against', and *naar* 'to'–will be discarded. Candidate terms with one of these prepositions, for example, *middelen als cocaïne* 'drug such as cocaine', *voedingsadditieven zoals conserveermiddelen* 'food additives such as conservatives', and *bloedstroom naar hart* 'blood flow to heart', are usually not terms.

Especially for the equation 5 and 6 (section 4.1) or step 3 (section 4.2), we apply suffix stripping. The main goal of this stripping is to get the stem form of a word in a term, and then to make it possible that a word form matches with other word forms. For example, the word *genetisch* will match with the word *genetic*, after the suffix -*isch* is replaced with -*ic* by our stripping rules. This process could be considered as a form of stemming.

The suffix stripping will later help matching words of multilingual terms in a specific domain. This is motivated by the fact that some terms in both Dutch and English use the same stems. For example, the term *chemisch element* becomes *chemic element* by a stemming rule for Dutch. The same stem will also be produced from the term *chemical element* by a stemming rule for English. We make use of these regularities to create a small set of rules for suffix stripping: 13 rules for Dutch, and 6 rules for English.

	y	$\bar{y}$	
x	n_{11}	n_{12}	n_{1p}
$\bar{x}$	n_{21}	n_{22}	n_{2p}
	n_{p1}	n_{p2}	n_{pp}

Table 2: *Contingency table of frequency data for a word pair xy.*

3.5 Comparing statistical measures

In this section we compare eight commonly-used approaches (defined in Table 3) to measure the strength of association of bigram word strings. Formulae in Table 3 are defined in terms of the contingency table given in Table 2. In this table, n_{11} is the frequency of the bigram xy, n_{12} is the frequency of x followed by any word other than y, and n_{1p} is the total frequency all bigrams with x as the first word. m_{11} is the expected value of the bigram xy ($m_{11} = \frac{n_{p1}n_{1p}}{n_{pp}}$). And particularly for the C-value, a is the candidate string (xy), $|a|$ is the length of a, $f(a)$ is the frequency of a in the corpus, T_a is the set of extracted candidate terms that contain a, $P(T_a)$ is the number of these candidate terms, and $f(b)$ is the frequency of a candidate term b that contains a.

Given a list of candidate term bigrams with their associated frequency, the statistical measures above will rank the terms according to their association scores.

3.6 Pseudo-bigrams

Since most of the statistical algorithms were originally designed to measure the association of two-word collocations (bigrams), and our candidate noun phrases are of any length, we need to expand the algorithms from identifying bigrams to identifying n-grams ($n >= 2$). One attractive solution – as has been reported in [26] and [25] – is to think of any n-gram as a pseudo-bigram XY where X is its left part and Y is its right part.

Given an n-gram:

$$C = w_1 w_2 ... w_n \tag{1}$$

which can be generalized into:

$$C = w_1 ... w_i w_{i+1} ... w_n \tag{2}$$

we can construct a pseudo-bigram $C = XY$, where:

$$X = w_1 ... w_i \quad \text{and} \quad Y = w_{i+1} ... w_n \tag{3}$$

Method	Formula
Frequency [15]	n_{11}
T-Score [5]	$\frac{n_{11}-\frac{n_{1p}n_{p1}}{n_{pp}}}{n_{11}^2}$
Log-likelihood [11, 7]	$2(n_{11}log\frac{n_{11}}{m_{11}}+n_{12}log\frac{n_{12}}{m_{12}}+n_{21}log\frac{n_{21}}{m_{21}}+n_{22}log\frac{n_{22}}{m_{22}})$
Chi-squared (χ^2) [4]	$2(\frac{n_{11}-m_{11}}{m_{11}}^2+\frac{n_{12}-m_{12}}{m_{12}}^2+\frac{n_{21}-m_{21}}{m_{21}}^2+\frac{n_{22}-m_{22}}{m_{22}}^2)$
Dice [10]	$2\frac{n_{11}}{n_{p1}+n_{1p}}$
Pointwise Mutual Information (PMI) [12, 5]	$log\frac{n_{11}}{m_{11}}$
True Mutual Information (TMI) [20]	$\frac{n_{11}}{n_{pp}}log\frac{n_{11}}{m_{11}}+\frac{n_{12}}{n_{pp}}log\frac{n_{12}}{m_{12}}+\frac{n_{21}}{n_{pp}}log\frac{n_{21}}{m_{21}}+\frac{n_{22}}{n_{pp}}log\frac{n_{22}}{m_{22}}$
C-value [14]	$\begin{cases}\log_2\lvert a\rvert.f(a) & \text{if } a \text{ is not nested,}\\ \log_2\lvert a\rvert(f(a)-\frac{1}{P(T_a)}\sum_{b\in T_a}f(b)) & \text{otherwise}\end{cases}$

Table 3: *Statistical algorithms used to measure the association strength of a word pair xy.*

For each n-gram, we compute all pseudo-bigrams and choose which i maximizes its association score. To illustrate this strategy, take as an example the candidate term *Body Mass Index* which is found three times in the corpus of total 193,123 candidate terms. This term can be approximated by two pseudo-bigrams: *Body_Mass Index* and *Body Mass_Index*. To find the one that will represent the candidate term, we count the frequencies of their substrings on the left and the right parts. After normalizing (uppercasing) the substrings, we get the frequencies as shown in Table 4.

Pseudo-bigram (XY)	$f(XY)$	$f(X)$	$f(Y)$
BODY_MASS INDEX	3	3	5
BODY MASS_INDEX	3	3	3

Table 4: *Pseudo-bigrams of the candidate term Body Mass_Index and their substring frequencies.*

When measured using the χ^2 method, *BODY_MASS INDEX* gets 64.8 score while *BODY MASS_INDEX* gets 71.5 score. Thus, the last pseudo-bigram will be selected to represent the candidate term.

This bigram approximation will be applied to all of the association measures except for the C-value and the Frequency method. The last two methods are not calculated based on the bigram model, but based on the frequency of occurrences of any n-gram in the corpora.

3.7 Local-rank ordering

It is important to reorder the χ^2 output since this method will give the same value for bigrams with particular frequencies. For example, the candidate term *vicieuze cirkel* 'vicious circle' which occurs 10 times (*vicieuze_* 10 times, *_cirkel* 10 times), and the candidate term *euthyroid sick syndrome* which occurs 2 times (*euthyroid_* 2 times, *_sick_syndrome* 2 times), both will have the same χ^2 value. In our experiment, the first 78 candidate terms in the χ^2 output have the same normalized value of 1. These candidate terms need to be ordered locally using some criterion.

This case also happens when we evaluate the candidate terms using some other methods. To improve the ranking of candidate terms of the same score, we use the following heuristic:

- If several candidate terms have the same score, order them by their frequency of occurrence in the corpus.
- If a subset of the previous candidate terms has the same frequency, order them by total frequency of their words in the known terms.
- If a subset of the previous candidate terms has the same total frequency, order them alphabetically.

3.8 Selecting the best measure

We evaluate the statistical algorithms against the medical encyclopedia corpus (section 3.1). The POS-tag filter extracted 86,000 unique candidate terms, of which 64,000 were multi-word terms. In this paper, only multi-word terms are taken into account.

We apply various settings of frequency cut-offs, and compute the association scores of the candidate multi-word terms in each setting using the NSP package [2]. We take from each produced ranking a set of K (=100) true terms that match with our gold standard, and calculate the area under the associated precision-recall curve using uninterpolated average precision (UAP) as shown in equation 4 [25]:

$$UAP = \frac{1}{K}\sum_{i=1}^{K} P_i \quad (4)$$

where P_i (precision at i) equals i/H_i, and H_i is the number of hypothesized terms required to find the i^{th} true term.

For each of the frequency cut-off settings, the UAP (*precision*) at every i^{th} true term (*recall*) is calculated. The precision-recall curves of the statistical methods at the frequency cut-off of 8 is provided in Figure 2. In this figure, the curves of χ^2 and Dice are overlapped on the top, followed by the overlapping curves of the Log-likelihood and TMI, T-score curve, the overlapping curves of the C-value and Frequency, and the last curve at the beginning of the recall is the PMI curve.

Performance of these methods at various frequency cut-offs for $K = 500$ is shown in Table 5. Note that at most of the frequency cut-offs, χ^2 and Dice outperform other methods. These results are consistent with those reported in [8] and [25], where the information-theoretic measures (e.g. χ^2, Dice, and PMI) are shown to outperform frequency-based measures (e.g. Frequency, T-Score, Log-likelihood, and C-value).

Based on the above results, we chose χ^2 as the association measure to rank multi-word terms in the following experiments. Besides taking into account the

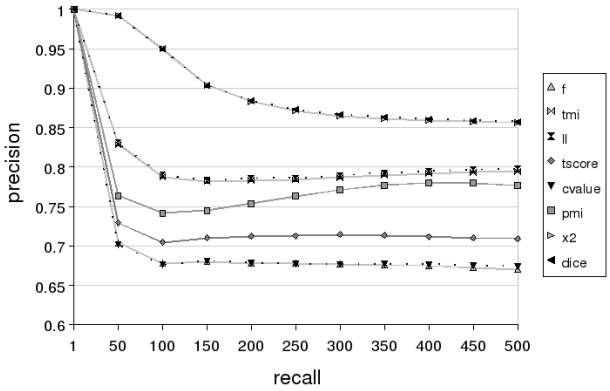

Fig. 2: *The precision-recal curves of the statistical methods at the frequency cut-off 2 and with the maximum recall of 500 terms.*

Method	Frequency cut-off			
	2	4	6	8
Freq	0.670	0.670	0.670	0.670
TMI	0.794	0.794	0.794	0.791
Loglike	0.798	0.798	0.797	0.793
Tscore	0.709	0.709	0.709	0.709
C-value	0.674	0.673	0.674	0.672
PMI	0.777	0.885	0.869	**0.877**
X2	0.856	**0.912**	0.880	0.872
Dice	**0.857**	0.911	**0.881**	0.873

Table 5: *Performance of statistical measures at different frequency cut-offs for $K = 500$.*

frequency of occurrence of a bigram, χ^2 also takes into account the frequency of each individual part of the bigram in measuring the deviation between observed and expected data. This concept is important for our improvement strategy. This is because in the known terms that will become our external knowledge for the improvement, only individual parts of bigrams exist and none of the bigrams occurs. In this case, Dice will not help because it requires the bigram to occur in the known terms in order to have a Dice value.

4 Using multilingual terminologies

Up to this section, we only use information from the corpus to assign association scores to candidate multiword terms. Several ideas have been explored previously to come up with more accurate methods. Schone & Jurafsky [25], for instance, used Latent Semantic Analysis (LSA) to improve the output of their best algorithm, and Maynard & Ananiadou [22] used semantic information from the UMLS to improve the NC-value output. Jacquemin *et al.* [16] have emphasized that the use of an initial term set will improve the performance of an ATR system. In this section we discuss the idea of using a set of known terms, especially the multilingual terms, to improve the ranking of candidate terms.

4.1 Hypothesis and formulae

Our hypothesis, an extension of [16], is that a set of multilingual terms can help improving the ranking of new (unknown) terms in a special language.

To prove the hypothesis, we propose a new scoring formula which combines two essential aspects of the nature of terms, namely *unithood* and *termhood* [18]. The unithood of a term can be measured as the association strength or degree of syntagmatic combination stability. On the other hand, the termhood of a term can be measured as relatedness to domain-specific known terms.

We use the χ^2 test (section 3.8) to measure the first aspect (unithood), while for the second aspect (termhood), we experiment with two methods: χ^2 and a term overlap heuristic. Thus, given a corpus and a list of known terms from multilingual terminologies to calculate the collocational strength and the domain-specific relatedness, we calculate each part of the combined measure as follows:

- *unithood*: calculate the χ^2 score of a candidate term given frequencies taken from the corpus.
- *termhood* (alternative 1): calculate the χ^2 score of a candidate term given frequencies taken from the list of known terms (domain-specific concepts).
- *termhood* (alternative 2): calculate the ratio between the number of words in a candidate term (which matched with the words in the known terms) and the length of the candidate term.

Equation 5 and 6 below show how both aspects of terms are combined to obtain an improved score called the *Unithood and Termhood Score* (UT-Score or UTS). The formula for the first alternative, which is based on an association mesure (UTS_α), is defined as follows:

$$\text{UTS}_\alpha(xy) = \frac{\chi^2(xy|C)}{max[\chi^2(XY|C)]} + \alpha.\chi^2(xy|T) \quad (5)$$

and the formula for the second the alternative, which is based on a bag of words (UTS_β), is defined as follows:

$$\text{UTS}_\beta(xy) = \frac{\chi^2(xy|C)}{max[\chi^2(XY|C)]} + \frac{|w(xy|T)|}{|w(xy)|} \quad (6)$$

where xy is a bigram of a candidate term extracted from a corpus C, XY is a set of xy bigrams, $\chi^2(xy|C)$ is the association strength of xy given its frequencies taken from the corpus C, $max[\chi^2(XY|C)]$ is the maximum value of the association strength observed for any element in the set XY, $\chi^2(xy|T)$ is a relatedness degree of xy given its frequencies taken from a set of known terms T, α is a weighting factor determined experimentally, $|w(xy|T)|$ is the number of words in xy which are matched with words in T, and $|w(xy)|$ is the number of words in xy. We normalize the unithood score because its maximum value is usually much higher than one. Since our aim is to enrich an existing terminology, only candidate terms which do not occur in the set of known terms will be evaluated. Therefore, it is not necessary to normalize the termhood

score because if there is no evidence for the new (unknown) term xy in T, its χ^2 score will always be less than one.

Our intuition behind these formulae is: if the words x and y of a candidate term xy are also used by terms in a domain, we can expect that the candidate term xy will represent a concept related to the domain. For example, the terms *chronische ziekte* 'chronic disease' and *chronische infectie* 'chronic infection' are at the 4684th and 5690th ranks when evaluated solely using the χ^2 method given their frequencies of occurrences taken from the corpus. However, the words *chronische*, *ziekte*, and *infectie* are well known words for the medical domain, and in fact they are frequently used to generate terms in this domain. Thus, their frequencies of occurrence in the set of known terms should help to improve their rankings.

In the first alternative, the set of known terms can be seen as a pseudo-corpus, in which all of its lines represent candidate terms. Thus, for a candidate term xy, we can count the frequencies of its bigram parts x and y in the set of known terms, and use the frequencies to calculate the χ^2 value with respect to the domain.

In the second alternative, we simply convert the set of known terms into a bag of words, count the number of words in the candidate term xy which are also found in the bag of words, and then compare this number to the length of the term. For example, if the words *chronische* and *ziekte* are used anywhere in the known terms, regardless of their positions and frequencies, then the matched ratio for the term *chronische ziekte* is 1.

4.2 The algorithm

The following steps, provided with walkthrough examples, summarize our method:

Step 1 Tag the corpus with part-of-speech information.

Step 2 Create a *unithood* model. Extract a set of candidate terms XY from the tagged corpus using the linguistic filter (section 3.3) and count their frequency. Generate pseudo-bigrams for candidates of more than 2 words. For each bigram, compute its association score, $\chi^2(xy|C)$. For example, the trigram *tropische spastische paraparese* (occures 2 times) will be represented by two pseudo-bigrams: *tropische_spastische paraparese* (occur 2 and 2 times) and *tropische spastische_paraparese* (occur 35 and 2 times). The χ^2 values of both pseudo-bigrams are 107738 and 6155, respectively. Thus, *tropische_spastische paraparese* will become the final representation for the trigram.

Step 3, alternative 1 Create a *termhood* model. Given a set of known terms T, compute frequencies for each element in the set XY of extracted terms. Generate pseudo-bigrams for candidates of more than 2 words. For each bigram, compute its termhood value, $\chi^2(xy|T)$. Run this step for two settings: with stemming (suffix stripping) and without stemming (see section 3.4). Take the previous trigram as an example. In this model, its pseudo-bigram *tropische_spastische paraparese* (Dutch) has no occurrence in the multilingual terms. Therefore its termhood value is zero.

Step 3, alternative 2 Create a *termhood* model. Given a set of known terms T, create a bag of words B of the terms. For each candidate term in C, count the number of its words which overlap with the words in B, and compute its termhood value ($\frac{|w(xy|T)|}{|w(xy)|}$) by comparing that number to the length of the candidate term. Run this step for two settings: with stemming (suffix stripping) and without stemming (see section 3.4). For example, the term *tropical spastic paraparesis* (English) is found in the multilingual terms. To create a bag of stem words, we stem this term into *tropic*, *spastic*, and *parapares*. On the other hand, our Dutch stemming rules will stem the previous trigram *tropische spastische paraparese* into the same stem words. Thus, its termhood is 1.

Step 4 Compute the combined scores (UT-Score) of every candidate term xy in the set XY using equation 5 and 6, and sort them using the method as described in subsection 3.7. Optionally, remove terms below a certain frequency threshold.

4.3 Experiment and results

We use the corpora described in section 3.1 to create a *unithood* model, the set of multilingual terms described in section 3.2 to create a *termhood* model. Since our aim is to enrich the existing terminologies, we subtract all candidate terms (extracted in section 3.3) which overlap with all known terms collected in section 3.2. This reduction provides us with 54,000 candidate terms. Applying a frequency threshold of 2 and 4 will left us with 5,333 and 1,233 candidate terms. To get a higher recall, we decided to process with the first threshold. The candidate terms are evaluated using the formulae 5 and 6.

To compare the results, we run the following measurement settings: χ^2, UTS_α (equation 5), and UTS_β (equation 6). For UTS_α, we set $\alpha = 10$, the best value according to our experiments. At both equations, we run two settings: *with stemming* and *without stemming*.

Examples of candidate terms in several rank positions for those settings are shown in Table 6. All of the settings show promising results. At a glance, most of the top ranked candidate terms look like medical terms. There are some english candidate terms in that list, such as *evoked potentials*, *case management*, and *undetermined significance*. Most of them are incorrectly tagged by our parser. For example, the last candidate term is tagged as *noun noun* because *noun* is the default tag for unknown words.

All of the settings except χ^2 give higher weights to candidates which have high values for both *unithood* and *termhood*. For example, *cystosarcoma phylloides* (28th rank by χ^2) is placed at the 2nd and 4th ranks by $\text{UTS}_\alpha|\alpha = 10$ and UTS_β since both of its word elements are in B. The effect of the stemming is shown by the term *tropische spastische paraparese* (70th rank by χ^2). Although its word elements are not found in B, it has an equivalent form in the multilingual terms, namely *tropical spastic paraparesis*. The stemming has normalized both forms into the same stems. As a result, the termhood of the candidate term will be

Rnk	χ^2	UTS$_\alpha$|$\alpha = 10$	UTS$_\alpha$|$\alpha = 10$ **(stem)**	UTS$_\beta$	UTS$_\beta$ **(stem)**
1	*vicieuze cirkel*	evoked potentials	*tropische spastische paraparese*	evoked potentials	evoked potentials
2	*ziekte van von willebrand-jrgens*	*cystosarcoma phyl-loides*	*vasovagale syncope*	*morbus haemolyticus neonatorum*	*morbus haemolyticus neonatorum*
3	*allergische bron-chopulmonale as-pergillose*	*vicieuze cirkel*	*cholinergische ur-ticaria*	*case management*	*case management*
4	*cardiopulmonale resuscitatie*	*ziekte van von willebrand-jrgens*	*genetisch defect*	*cystosarcoma phyl-loides*	*mastopathia fibrosa cystica*
5	*cellen per microliter bloed*	*allergische bron-chopulmonale as-pergillose*	*toxische shock*	*mastopathia fibrosa cystica*	*mental disorders*
6	*thoracaal aorta-aneurysma*	*cardiopulmonale resuscitatie*	*cerebrale malaria*	*mental disorders*	*tropische spastische paraparese*
7	*gecomputeriseerd to-mografisch onderzoek*	*cellen per microliter bloed*	*normale cellen*	undetermined signifi-cance	undetermined signifi-cance
8	*subduraal empyeem*	*thoracaal aorta-aneurysma*	*case management*	*engelse sudden infant death syndrome*	*processus mastoideus*
9	*endoscopische retro-grade pancreatografie*	*gecomputeriseerd to-mografisch onderzoek*	evoked potentials	*processus mastoideus*	*endoscopische retro-grade pancreatografie*
10	evoked potentials	*subduraal empyeem*	*vicieuze cirkel*	*acute herpetische gingivostomatitis*	*acute herpetische gingivostomatitis*
32	*erytroplasie van queyrat*	*pericarditis constric-tiva*	*ulcus oesophagi*	*thoracaal aorta-aneurysma*	*multiform glioblas-toom*
64	*solitaire agressieve gedragsstoornis*	portugese oorlogss-chip	*humorale (antistof) therapie*	diagnose allergische alveolitis	*bacterie bartonella henselae*
128	*minimaal normaal lichaamsgewicht*	enkelvoudige dosis azitromycine	*lucide interval*	g vers paddenstoel-gewicht	*chemische dampen*

Table 6: *Examples of candidate terms in several rank positions for different experimental settings. Candidate terms printed in italics are proposed new terms. The terms are not translated into English due to the space limitation.*

Setting	**UAP**			
	strict	*lenient*	*Ann1*	*Ann2*
χ^2	0.608	0.802	0.735	0.676
UTS$_\alpha$	0.612	0.825	0.734	0.702
UTS$_\alpha$ (stem)	0.626	0.826	0.758	0.694
UTS$_\beta$	0.677	0.865	0.782	0.757
UTS$_\beta$ (stem)	**0.769**	**0.918**	**0.845**	**0.843**

Table 7: *The uninterpolated average precision values for $K = 100$ of several measurement settings in two agreement modes and by two annotators (Ann*1 *and Ann*2*). For UTS$_\alpha$ settings, the value of α is 10.*

high. This term is placed at the 1st and 6th ranks by UTS$_\alpha$\|$\alpha = 10$ *(stem)* and UTS$_\beta$ *(stem)*, respectively.

4.4 Evaluation

To evaluate the results quantitatively, we asked two human annotators to annotate a list of candidate terms extracted from the experiments. From each setting, we take the first 200 candidate terms in its rank, and then they annotate the selected candidate terms with *yes* or *no*. To compare the performance of the settings, we compute the uninterpolated average precision (section 3.8) at $K = 100$.

The results of this evaluation are shown in Table 7. The settings are evaluated using both a **strict mode**, in which a candidate is counted as a term only if both evaluators agreed, and a **lenient mode**, in which a candidate is counted as a term if one of the evaluators annotated it as a *yes*. We also present scores for the annotated lists produced by each of the annotators (*Ann1* and *Ann2*).

In both modes, the annotators agree that combining the unithood and termhood values improves the baseline (χ^2). Applying stemming rules also results in a slightly better result for the UTS$_\alpha$ method (from 0.612 to 0.626), and a higher improvement is achieved by the UTS$_\beta$ method (from 0.677 to 0.769). Both annotators agree that the last mentioned method, either with stemming or without stemming, outperforms other methods. And the best result is achieved by the UTS$_\beta$ method when stemming is applied.

The use of the association measure to calculate the termhood in the UTS$_\alpha$ method apparently is not a good strategy. Since most of the overlapping words–between word elements in a candidate terms and in B–do not form a cooccurrence (note that we have subtracted all candidate terms that are found in the existing term list), the association measure will not give a high termhood value to the candidate. However, the stemming shows some improvements.

Computing the termhood value using a matching ratio in the UTS$_\beta$ method solves the problem faced by the association measure. This method does not rely on the cooccurrence of the overlapping words but on the number of the overlapping words. Combined with the stemming, this method shows a promising result in using multilingual terminologies to improve the extraction of multiword terms in a particular language, especially in a medical domain. In this experiment, the stemming rules we construct are very simple and not exhaustive. We only need to take some productive suffixes for a particular domain by investigating candidate terms extracted from the corpus. We expect that this method can be adapted to other languages and domains with little effort.

5 Conclusion and future work

We have presented two experiments in this paper: (1) selecting the best statistical association measure for multi-word term extraction, (2) creating a new method which combines unithood and termhood values to extract new terms. We use χ^2 as the best measure for the second experiment. However, one can use other methods such as log-likelihood which shows good performance. In that experiment, we use a set of multilingual terms as a source to compute the termhood values.

Our new methods (UT-Scores) have shown to outperform the baseline system, and worked well in exploiting multilingual terminologies to induce new terms from a text in another language. A simple matching ratio in UTS_β shows good performance in calculating the termhood values. This approach is useful especially in a domain where the use and modification of terminologies from other languages is very productive, such as in the medical domain.

Although we evaluated on a domain-specific corpus, we expect that our method will be useful in other settings as well. For instance, given a general corpus (say, a newspaper corpus), one might use our method to identify the use of terminology from a given domain (say, medical terms) in this general corpus. We are in the process of building a medical ontology from a Dutch medical text to see the structure of the domain in the text. The use of existing terminologies such as the UMLS is an important strategy.

Our method is still open for improvement. We have the intuition that using a bilingual dictionary, instead of or together with stemming, will improve the results. Our future work will be focused on this method.

References

[1] S. Ananiadou. A methodology for automatic term recognition. In *Proceedings of the 15th conference on Computational linguistics*, pages 1034–1038, Morristown, NJ, USA, 1994. Association for Computational Linguistics.

[2] S. Banerjee and T. Pedersen. The design, implementation, and use of the ngram statistics package. In *CICLing*, pages 370–381, 2003.

[3] D. Bourigault. Surface grammatical analysis for the extraction of terminological noun phrases. In *Proceedings of the 14th conference on Computational linguistics*, pages 977–981, Morristown, NJ, USA, 1992. Association for Computational Linguistics.

[4] K. W. Church and W. Gale. Concordances for parallel text. In *Proceedings of the Seventh Annual Conference of the UW Center for the New OED and Text Research*, pages 40–62. Association for Computational Linguistics, 1991.

[5] K. W. Church and P. Hanks. Word association norms, mutual information, and lexicography. In *Proceedings of the 27th. Annual Meeting of the Association for Computational Linguistics*, pages 76–83, Vancouver, B.C., 1989. Association for Computational Linguistics.

[6] I. Dagan and K. W. Church. Termight: Identifying and translating technical terminology. In *ANLP*, pages 34–40, 1994.

[7] B. Daille. Study and implementation of combined techniques for automatic extraction of terminology. In J. Klavans and P. Resnik, editors, *The Balancing Act: Combining Symbolic and Statistical Approaches to Language*, pages 49–66. The MIT Press, Cambridge, Massachusetts, 1996.

[8] P. Deane. A nonparametric method for extraction of candidate phrasal terms. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 605–613, Ann Arbor, Michigan, 2005. Association for Computational Linguistics.

[9] H. Déjean, E. Gaussier, and F. Sadat. Bilingual terminology extraction: an approach based on a multilingual thesaurus applicable to comparable corpora. In *Proc. of COLING*, Tapei, Taiwan, 2002.

[10] L. Dice. Measures of the amount of ecological association between species. *J. Ecology*, 26:297–302, 1945.

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

[12] R. Fano. *Transmission of Information*. MIT Press and John Wiley and Sons, 1961.

[13] K. T. Frantzi and S. Ananiadou. The c-value/nc-value domain independent method for multiword term extraction. *Journal of Natural Language Processing*, 6(3):145–180, 1999.

[14] K. T. Frantzi, S. Ananiadou, and J.-i. Tsujii. The c-value/nc-value method of automatic recognition for multi-word terms. In *ECDL '98: Proceedings of the Second European Conference on Research and Advanced Technology for Digital Libraries*, pages 585–604, London, UK, 1998. Springer-Verlag.

[15] V. Giuliano. The interpretation of word associations. In M. Steven, editor, *Statistical Association Methods for Mechanical Documentation, Symposium Proceedings*, Washington, D.C., 1964. NBS Miscellaneous Publication No. 269.

[16] C. Jacquemin, J. L. Klavans, and E. Tzoukermann. Expansion of multi-word terms for indexing and retrieval using morphology and syntax. In P. R. Cohen and W. Wahlster, editors, *Proceedings of the Thirty-Fifth Annual Meeting of the Association for Computational Linguistics*, pages 24–31, Somerset, New Jersey, 1997. Association for Computational Linguistics.

[17] J. Justeson and S. Katz. Technical terminology: some linguistic properties and an algorithm for identification in text. *Natural Language Engineering*, 1:9–27, 1995.

[18] K. Kageura and B. Umino. Methods of automatic term recognition: a review. *Terminology*, 3(2):259–289, 1996.

[19] R. Malouf and G. van Noord. Wide coverage parsing with stochastic attribute value grammars. In *IJCNLP-04 Workshop Beyond Shallow Analyses - Formalisms and statistical modeling for deep analyses*, Hainan, 2004.

[20] C. Manning and H. Schutze. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA., 1999.

[21] D. G. Maynard. *Term Recognition Using Combined Knowledge Sources*. PhD thesis, Manchester Metropolitan University, Manchester, UK, 2000.

[22] D. G. Maynard and S. Ananiadou. Identifying terms by their family and friends. In *Proceedings of the 18th conference on Computational linguistics*, pages 530–536, Morristown, NJ, USA, 2000. Association for Computational Linguistics.

[23] S. Mukherjea, L. Subramaniam, G. Chanda, S. Sankararaman, R. Kothari, V. Batra, D. Bhardwaj, and B. Srivastava. Enhancing a biomedical information extraction system with dictionary mining and context disambiguation. *IBM J. Res. Dev.*, 48(5/6):693–701, 2004.

[24] J. C. Sager. *Term Formation*, volume 2 of *Handbook of Terminology Management*, pages 25–41. John Benjamins Publishing Company, 1997.

[25] P. Schone and D. Jurafsky. Is knowledge-free induction of multiword unit dictionary headwords a solved problem? In L. Lee and D. Harman, editors, *Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing*, pages 100–108, 2001.

[26] J. Silva and J. G. P. Lopes. Extracting multiword terms from document collections. In *Proceedings of the VExTAL, Venezia per il Trattamento Automatico delle Lingu*, Venezia, 1999. Universiá Cá Foscari.

[27] A. S. Valderrabanos, A. Belskis, and L. I. Moreno. Multilingual terminology extraction and validation. In *LREC 2002 (Third International Conference on Language Resources and Evaluation)*, 2002.

[28] L. van der Beek, G. Bouma, and G. van Noord. Een brede computationele grammatica voor het Nederlands. *Nederlandse Taalkunde*, 7(4):353–374, 2002.