Detecting Multiword Verbs in the English Sublanguage of MEDLINE Abstracts

Chun Xiao and Dietmar Rösner
Institut für Wissens- und Sprachverarbeitung
Otto-von-Guericke Universität Magdeburg
Universitätsplatz 2, Magdeburg, Germany 39106
xiao|roesner@iws.cs.uni-magdeburg.de

Abstract

In this paper, we investigate the multiword verbs in the English sublanguage of MEDLINE abstracts. Based on the integration of the domain-specific named entity knowledge and syntactic as well as statistical information, this work mainly focuses on how to evaluate a proper multiword verb candidate. Our results present a sound balance between the low- and high-frequency multiword verb candidates in the sublanguage corpus. We get a F-measure of 0.753, when tested on a manual sample subset consisting of multiword candidates with both low- and high-frequencies.

1 Introduction

During the construction of an information extraction (IE) system in the biomedical domain, we found that not only the task of named entity recognition (NER), but also the appropriate handling of verbs in this domain plays an important role. It is very helpful to determine the domain-specific verbs in a specific domain when extracting useful information, because the domain-specific verbs construct semantic relations between named entities (NEs). However, three problems in the handling of verbs in a specific domain are still open:

The first problem is how to determine *domain-specific* verbs. This problem has not received enough notice from most of the researchers yet. Actually, *domain-specific verbs* have been mentioned quite often in biomedical text processing (Thomas et al., 2000; Ono et al., 2001; Xiao and Rösner, 2004b), but often referred to a set of manually or experientially selected verbs. Spasić et al. (2003) briefly presented a method to find domain-specific verbs by filtering the verbs in a stoplist, at the same time, using the co-occurrence of a verb and specific terms in the text. In our experiment, the *domain-specific verbs* are determined through the comparison between different corpora in different domains, or through genre analysis of the sublanguage dominated corpus.

The second problem is how to determine multiword verbs (MWVs). Here we do not make differences between the more detailed classification of multiword verbs, especially the *verb-particle constructions* and *verb-preposition constructions* (Baldwin and Villavicencio, 2002). As a subcategory of multiword expressions (Sag et al., 2002), MWVs raise the complexity of our processing. Because some MWVs share the same verb head but lead to different semantic interpretations, like *result in* and *result from*, considering only verb heads in the processing is obviously not sufficient. A good IE system should deal with such MWVs automatically and appropriately.

The third problem is that there is a need to investigate the inflectional and derivational forms of the verbs. An IE system may have to deal with a set of patterns, in which the inflectional and derivational forms of the verbs should be taken into account. For example, in biomedical texts, the verb *interact* defines a binary relation between two substances, whereas its nominalization morpheme in a pattern such as *the interaction of ... with ...* also constructs such a relation. Note that such patterns often have close relationship with its common verb lemma, which is often a MWV. For instance, the above pattern can map to the MWV *interact with*. Table 1 shows the distribution of all inflectional and derivational forms of the verb *inhibit* in a corpus of 800 MEDLINE abstracts[1] extracted from the GENIA corpus.[2] This verb is a very important domain-specific verb in the biomedical domain. To deal with those inflectional and derivational forms appropriately will improve the performance of the IE system.

The following text focuses on the second problem

[1] MEDLINE/PubMed is a collection of references and abstracts from 4600 biomedical journals from all over the world, available at http://www.ncbi.nlm.nih.gov/PubMed/.

[2] The GENIA Corpus V3.0p consists of 2000 POS-tagged MEDLINE abstracts, V3.01 consists of the same 2000 abstracts annotated semantically with the GENIA ontology, available at http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/.

Form	Times	Typical pattern
inhibitor(s)	161	a/the ... inhibitor ...
inhibition	167	... inhibition of ...
inhibitory	61	... inhibitory effect/factor ...
inhibiting	24	... in inhibiting ...
inhibited	119	... inhibited ...
be inhibited	73	... be inhibited by ...
inhibit	63	... inhibit ...
inhibits	57	... inhibits ...

Table 1: Typical patterns of syntactic structures containing the verb stem *inhibit* and their occurrences in a test corpus with 800 MEDLINE abstracts.

above. Section 2 introduces a set of language processing tools used in the experiment. Detailed description of the approach for the extraction of proper MWVs is presented in section 3. The evaluation of the result and the aspects that have influence on the result are discussed in section 4, as well as our future works. Finally, in the appendix, we list a number of MWVs that have been extracted by our approach.

2 Tokeniser, POS Tagger and Chunker

Our experiment in this paper is carried out mainly on chunk sequences, therefore the following processing components are necessary:

- **Tokeniser:** Following the *whitespace-delimited* tokenisation discipline, the tokeniser determines the segmentation of the non-lexical entries such as tokens with non-alphabet characters or abbreviations. After tokenisation, the sentence boundaries are determined as well.
- **POS tagger:** The maximum entropy POS tagger developed by Ratnaparkhi (Ratnaparkhi, 1996) and the rule-based POS tagger developed by Brill (Brill, 1994) are trained with 1200 abstracts extracted from the GENIA corpus, which achieve accuracies of 97.97% and 98.06% respectively, when testing on the rest 800 abstract of the GENIA corpus. Since our test corpus is directly extracted from the POS tagged GENIA Corpus V3.0p, we do not have to apply the process of tokenisation and POS tagging.
- **Chunker:** In this experiment, unlike the traditional statistical method for collocation extraction, where sentences are treated as word sequences (Manning and Schütze, 2002), a shallow chunking process is first carried out. Then, sentences in our test corpus are treated as chunk sequences.

  Up to now, the chunker consists of two parts, both utilize WordNet 1.7.1[3] (Fellbaum, 1999) as the lexical resource for the lemmatization, i.e., as the verb and noun stemmer.

  – **Verb chunker**, which extracts the smallest verb chunks (not including the MWV structures) with the additional syntactic information such as number (3rd singular present), voice (active/passive), and negation. Since most of the scientific abstracts are written in present or past tense, the temporal information is not extracted especially. The verb chunker returns the common verb lemma of a verb, with the additional syntactic information mentioned above. For example, given an input verb chunk *has not been established*, it returns *[establish, singular, passive, negation]*.

  – **Noun chunker**, which determines the noun chunk boundaries, negation, number (singular/plural), as well as some inner dependencies of the noun chunks containing substructure(s). For example, a noun chunk like [[*the retinoic acid-synthesizing enzyme*] [*aldehyde dehydrogenase 1*]] is actually an apposition structure.[4] In this experiment, the singular stem of a plural noun token is not returned in order to avoid missing of necessary information. For example, although both *take place* and *take places* can map to the same base structure *take place*, they must be treated separately.

3 MWV Extraction

3.1 Analysis of MWVs in the Corpus

The following experiment is carried out on a test corpus consisting of 1800 abstracts from the GENIA Corpus V3.0p, with 14955 sentences and 40.84K tokens (abstract titles are not included).

In general, the methodologies for the extraction of multiword expressions (MWEs, including MWVs) can be classified into syntactic, statistical and hybrid syntactic-statistical (Dias, 2003). Purely syntactic processing of MWEs requires specific linguistic knowledge across the different domains of a language, such as a semantic ontology (Piao et al.,

[3] http://www.cogsci.princeton.edu/∼wn/index.shtml

[4] presented in pairs of matching parentheses.

2003). Purely statistical processing overgenerates the MWE candidates (Gaël Dias, 2002), and is not sensitive enough to the MWE candidates with low frequencies (Piao et al., 2003). It is practical in some cases for a hybrid syntactic-statistical system to pre-define a set of MWE pattern rules, then use statistical techniques to filter proper candidates. But it lacks the flexibility to obtain a comprehensive coverage of possible MWE candidates, especially when a MWV is non-contiguous in our case. In addition, it also suffers from the problem of overgeneration, if the pre-defined syntactic pattern occurs rarely in the corpus. Sag et al. (2002) indicated that it is very necessary to find the balance between the two methods in hybrid systems. This point of view is taken into account in our approach.

3.2 Extraction of Contiguous MWV Candidates

A number of works about MWV extraction from corpora are based on the output of a POS tagger and a chunker (Baldwin and Villavicencio, 2002; Bannard et al., 2003), or the output of a parser (McCarthy et al., 2003). These works extracted mainly the *verb+particle* structures. Similar to those works, the MWV extraction in our experiment is also based on the chunking output. But, since MWVs has various POS tag patterns, it is not practical to assign each pattern an according syntactic rule. Therefore a variation of finite state automaton is considered in our approach for the extraction of MWVs. Let Γ denote this automaton.

Definition $\Gamma = \{S, I, O, F, G, START\}$, where:

- S is the set of automaton states, $S = \{nextOut, stop, nextIn, head, halt\}$;
- I is the input set, namely the chunks in both POS tag sequence and lexical sequence;
- O is the output set, namely the MWV candidates, $O = \{o_i | o_i$ is a successful MWV candidate.$\}$;
- F is the set of output controlling functions;
- G is the set of automaton state transition functions;
- $START$ is the initial state of the automaton, $START = head$.

Controlling functions in F define operations for the output. Controlling functions in G define state transitions of the automaton with respect to the features from both POS tags and lexical entries of an input chunk. An example is given as following.

- **Example sentence:** *The 3'NF-E2/AP1 motif is able to exert both positive and negative regulatory effects on the zeta 2-globin promoter activity in K562 cells*.

- **Chunk sequence of the example:**

Chunk	Chunk tag
The 3'NF-E2/AP1 motif	ENP
be	EVP
able	ADJP
to exert	IVP
both positive and negative regulatory effects	ENPS
on	IN
the zeta 2-globin promoter activity	ENP
in	IN
K562 cells	ENPS
.	SEPR

where ADJP is an *adjective phrase*; ENP is a *singular English noun phrase*; ENPS is a *plural English noun phrase*; EVP is a *singular English verb phrase*; IN is a *preposition*; IVP is an *infinitive verb phrase*; and, SEPR is a *sentence seperator*.

- **Extraction of Contiguous MWV Candidates:** In the following table, the *Input* items are the combination of both lexical sequences and the corresponding chunk tags, but only chunk tags are presented in the table. The output operation ϕ means no operation. For this example, it returns *be able to* as a MWV candidate.

Input	State transition	Output operation (o_i)
Initialization	nextOut	ϕ
ENP	nextOut	ϕ
EVP	head	$o_i =$ "be"
ADJP	nextIn	$o_i =$ "be able"
IVP	stop	$o_i =$ "be able to" (success)
IVP	head	$o_{i+1} =$ "exert"
ENPS	stop	$o_{i+1} =$ "exert" (failure)
IN	nextOut	ϕ
ENP	nextOut	ϕ
IN	nextOut	ϕ
ENPS	nextOut	ϕ
SEPR	halt	ϕ

3.3 Extraction of Non-contiguous MWV Candidates

When a set of new controlling functions are given, the finite automaton mentioned above also extracts non-contiguous MWV candidates. We primarily focuses on non-contiguous MWVs in form of *verb + particle*. As the particles in *verb + particle* MWVs are often intransitive (Baldwin and Villavicencio, 2002; McCarthy et al., 2003), which are different from the transitive prepositions followed by a noun chunk, we use this feature and a *nearest* assumption to extract non-contiguous MWV candidates. In general, we assume that a non-contiguous MWV occurs

in a limited context window.[5]

Because of the specific test corpus in our experiment, the non-contiguous MWV candidates extracted in our experiment are a relative small sub-set of all the candidates.[6] Most of them are not proper candidates. We suppose that the genre of scientific abstracts is an important reason for that: there are much more specific nominal terms as well as specific verbs (not MWVs) in scientific abstracts than in everyday language.

3.4 Solutions to Overgeneration of MWV Candidates

It is not surprising that the finite automaton is also sensitive to the low-frequent MWVs such as "take place" (7 times in the test corpus) and "shed light on" (4 times). But several problems of **overgeneration**[7] are still found, which include:

Case 1. Example: [*take place*]$_{1.1}$, [*take place at*]$_{1.2}$, [*take place in*]$_{1.3}$. In general, we assume that the *short structures* are more reliable, especially when the occurrences of the *short structures* are much more frequent than the *long structures*. But in this example, all three phrases occur with the same frequency in the test corpus, we still choose the more reliable *short structure*, and add up all occurrences of these structures.

Case 2. Example: [*be able to*]$_{2.1}$, [*be important for*]$_{2.2}$. The structure $[2.1]$ is a MWV, but the structure $[2.2]$, which has the same POS tag sequence as $[2.1]$, is actually not a MWV accepted by a lexicon. In this case, the verb head is usually one of the most frequent verbs such as *be*, *take*, *go*, and etc. In a previous experiment, we computed the logarithm likelihood ratio of the two mutual hypotheses[8] for the contiguous MWV candidates extracted from the test corpus, in order to find the reliability of such collocations. But we got some unexpected results, like *be important* in *be important for* was a more reliable structure than *shed light* in *shed light on*. It indicates that this score is still not sensitive enough to extremely sparse samples. In addition, this method is not suitable for non-contiguous MWV candidates.

In our experiment, we suppose that it is neither the verb head nor the preposition that determines the reliability of such MWV structures. Therefore we only focus on the distribution of the rest words (*able* or *important*) in the corpus. Such words, together with the verb head in a MWV pattern like *verb + particle*, in the following parts of paper, are given the name *MWV head*. For instance, we find that 83% occurrences of *able* are in the MWV candidate structure *be able to*, but only 8.4% occurrences of *important* are in *be important for*. Hence the structure $[2.1]$ is a much better candidate of MWV than $[2.2]$. By this means, the low-frequency candidate *shed light on* can also get a better rank than the relative high-frequency candidate *be important for*.

Case 3. Example: [*take place*]$_{3.1}$, [*bind DNA*]$_{3.2}$. $[3.1]$ is a MWV, but $[3.2]$, which also has the same POS tag sequence as $[3.1]$, is not a MWV. In our case, a set of domain-specific terms are available from the NE-annotated GENIA corpus V3.01. Since we suppose that the MWVs contain only *general* words, the word like *DNA* in this corpus can be found in the specific word list, then this structure can be excluded from the list of MWV candidates. However, this method induces also problems. For example, *give rise to* is a MWV, but *rise* is also in the specific word list of this corpus. In this case, the specific word list could be selected according to some criteria (e.g., frequencies in the list of specific terms), so that a much more comprehensive list of MWV candidates can be produced without losing the generality.

Case 4. Example: [*be able to*]$_{4.1}$, [*be unaffected*]$_{4.2}$. The POS tag pattern of $[4.2]$ is a substring of $[4.1]$, i.e., *EVP + ADJP*, but $[4.2]$ is obviously not a proper MWV candidate. We assume that a proper MWV should have *closed* left and right boundaries, which means, the left boundary of a MWV candidate should be a verb, and the right boundary should be a preposition (including *to*) or a noun. Therefore such patterns with *open* right boundaries like $[4.2]$ in this example are deleted from the candidate list.

Case 5. Example: [*be associated with*]$_{5.1}$ and [*associate(d) with*]$_{5.2}$, [*be used to*]$_{5.3}$ and [*used to*]$_{5.4}$. The pair of $[5.1]$ and $[5.2]$ have no semantic differences between the past and present tense, as well as no semantic transition of the MWV itself between the passive and active voice. That means, they can all map to the MWV base form *associate with*. But there are semantic differences between the pair of $[5.3]$ and $[5.4]$. The past tense phrase *used to* is a fixed idiomatic verb phrase (e.g., *He*

[5] Similar assumption or experiment data can be also found in (Baldwin and Villavicencio, 2002; Bannard et al., 2003; Dias, 2003).

[6] 41 non-contiguous MWV candidates extracted in this experiment have the occurrences over ten, but there are 151 contiguous MWV candidates whose occurrences are over ten.

[7] MWV candidates share the same lexical substring/string or POS-tag subsequence/sequence.

[8] Take the bigram (w_1w_2) model as an example, hypothesis H_1: $P(w_2|w_1) = p = P(w_2|\neg w_1)$, hypothesis H_2: $P(w_2|w_1) = p_1 \neq p_2 = P(w_2|\neg w_1)$. The likelihood ratio $-2log\lambda = -2logL(H_1)/L(H_2)$ is more appropriate than χ^2 test, since some MWVs are quite sparse in our test corpus (Manning and Schütze, 2002).

used to smoke a pipe.), like the present tense phrase *according to*, generally they do not occur in forms of other tense. There is no semantic relationship between [5.3] and [5.4] in some cases, although the base forms of both structures are the same, i.e., *use to*. In this experiment, we do not consider the later case. All MWV candidates have the mapping to their base form, but the information about the passive and active voice are reserved, so that some candidates in passive forms (e.g., *be inhibited by*) can be excluded.

3.5 Evaluation of the Reliability of the MWV Candidates

After the above processing on the set of MWV candidates extracted by the finite automaton, the following task is to examine the reliability of the candidates, especially for those candidates that share the same MWV head. To solve this problem, statistical measurement is necessary. First, the frequencies of the MWV candidates in the test corpus are taken into account. For instance, *result in* is the most frequent MWV candidate, which has more than 320 occurrences, it is obviously a proper MWV candidate. From Figure 1, we can find that a large number of MWV candidates occur with relative low frequencies ranged from about 1 to 10. In order to avoid accidental errors during the process (mainly the wrong assignment of POS tags), the MWV candidates with the lowest frequencies from 1 to 4 are out of consideration. Second, the distribution of the MWV head in the MWV candidates is considered. We assume that a verb head of a certain MWV has the inertia (big probability) to construct other MWVs than to be isolated. For instance, 89% of occurrences of the verb head *result* are in *result in*, only 8.5% belongs to *result from*. Although *result from* is not a high frequent MWV candidate, it is still a proper one. Third, the contiguous and non-contiguous MWV candidates are treated as the same structure, so that such structures are not ignored by the statistical measurement. That means, if a MWV candidate occurs in both contiguous and non-contiguous forms, then their occurrences are added up. According to our experiment results, the occurrences of non-contiguous MWV candidates are much less than the contiguous candidates, which leads to a very small number of non-contiguous MWVs successfully extracted from our test corpus.

To evaluate the reliability of a certain MWV candidate c in the candidate set C, following definitions are given.

- $head(c)$, the MWV head of c, $c \in C$;

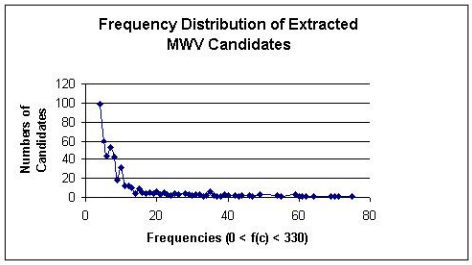

Figure 1: The distribution of frequencies (frequencies between 5 and 80).

- $f(x)$, the frequency of x, x can be c, or $head(c)$;
- $F(c)$, the sum of occurrences of all candidates in C, which share the same MWV head with c;
- $E(c)$, the evaluation score of c,

$$E(c) = c_1 f(c) + c_2 \frac{f(c)}{f(head(c))} + c_3 \frac{F(c)}{f(head(c))} \quad (1)$$

  where c_1, c_2 and c_3 (c_1, c_2, $c_3 \geq 0$) are coefficients;
- t, the threshold of score evaluation,

$$t = a \times \min E(c) + b \quad (2)$$

  where $c \in C$, $a \geq 1$, $b \geq 0$.
  If $E(c) \geq t$, c is a proper candidate.

The flowchart of the process to filter *proper* MWV candidates is shown in Figure 2.

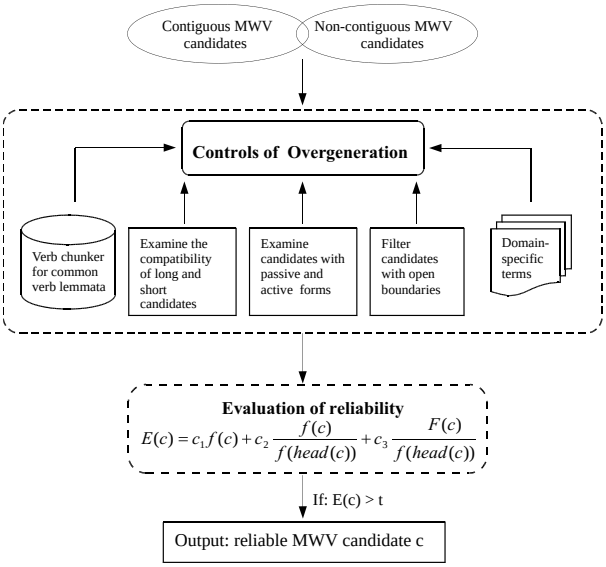

Figure 2: The process to find out *proper* MWVs.

In order to obtain the satisfying values of the coefficients and the threshold, a manual sample is created, so that the values of the coefficients can be tuned. It is not feasible that we give all extracted MWV candidates a human evaluation, therefore we

c_1	c_2	c_3	t	Precision	Recall	$F_{\beta=1}$
0.003	0.5	8	2.27	45.16%	100%	0.622
0.003	**0.5**	**10**	**2.81**	**45.65%**	**100%**	**0.627**
0.003	0.5	12	2.81	45.65%	100%	0.627
0.003	1	10	2.88	45.65%	100%	0.627
0.01	0.5	10	2.86	44.21%	100%	0.613
0.1	0.5	10	3.49	41.58%	100%	0.587

Table 2: Evaluation of the selection of proper MWV candidates according to equation 1, when we set the threshold $t = 1 \times minE(c) + 0$, $c \in C_P$ (baseline).

chose the most frequent 33 candidates ($f(c) \geq 60$), 31 candidates with moderate frequencies ($14 \leq f(c) \leq 19$), and 95 candidates with low frequencies ($6 \leq f(c) \leq 7$) as a manual sample set (C_M, $|C_M| = 159$). Those MWV candidates in the manual sample set are looked up in a dictionary,[9] if there is such a MWV entry in the dictionary, then we assign a *proper* flag to the candidate.

From the manual test sample set C_M, we annotate 42 items as *proper* MWVs (C_P, $|C_P| = 42$). In the experiment, we set the coefficient c_1 to be the reciprocal of the largest occurrence of the MWV candidates ($c_1 = 1/maxf(c), c \in C$), and t is set to be the linear function of the smallest score of the MWV in C_P by the reliability evaluation, e.g., $t = minE(c)$, $c \in C_P$. We use the scores of recall (R), precision(P), and F-measure ($F_{\beta=1}$) to evaluate our result. In the following equations, X denotes the set of candidates in C_P, whose scores of reliability evaluation are greater than t, i.e., $X = \{c|c \in C_P$ and $E(c) \geq t\}$; Y denotes the set of candidates in C_M, whose scores of reliability evaluation are greater than t, i.e., $Y = \{c|c \in C_M$ and $E(c) \geq t\}$.

$$R = |X|/|C_p|,$$
$$P = |X|/|Y|,$$
$$F_{\beta=1} = 2PR/(P+R).$$

4 Result, Discussion, and Future Works

The result in Table 2 indicates that it is neither the frequency of occurrences of a MWV candidate ($c_1 = 0.003$), nor the proportion of a MWV candidate to its head word ($c_2 = 0.5$), especially the verb head, but the *inertia* of a verb to construct MWVs that determines a proper MWV candidate ($c_3 = 10$). The result strongly supports this assumption.

We also found that the initiation of the value of t was very important. In Table 2, the minimum value of $E(c)$ in C_P was set to be the baseline of all test data. But we found that if the value of t was properly increased (according to equation 2), although the precision was therefore reduced, the F-measure was improved remarkably. Figure 3 shows how the changes of value t effect the result, given the same values of the coefficients in equation 1, that $c_1 = 0.003$, $c_2 = 0.5$, and $c_3 = 10$. We got a much better F-measure when we set $a = 2.3$, $b = 0.1$ (or $b = 0.2$), so that $F_{\beta=1} =$ **0.753**, if compared to the data in Table 2, where $a = 1$, $b = 0$, $F_{\beta=1} =$ **0.627**. The reason is that some MWV candidates in C_P, like *use to* and *carry out*, have the MWV heads that seem not to follow our assumption. Such verbs (*use*, *carry*, including *be*, and etc.) are often the most frequent verbs both in specific and general English language. Thus the syntactic and semantic combinations of such verbs and other words are quite rich, which led to a relative low score of $E(c)$ in our experiment. Compared to recent other related works, we found that (Baldwin and Villavicencio, 2002) presented a F-measure of 0.896 by testing on WSJ. But they focused on the single prepositional particle situation, whereas our approach has the special interest in multiple and non-preposition particle cases. Moreover, they used quite a lot of syntactic techniques for more precise extraction of verb-particle constructions (not verb-preposition constructions), which is not the case in ours.

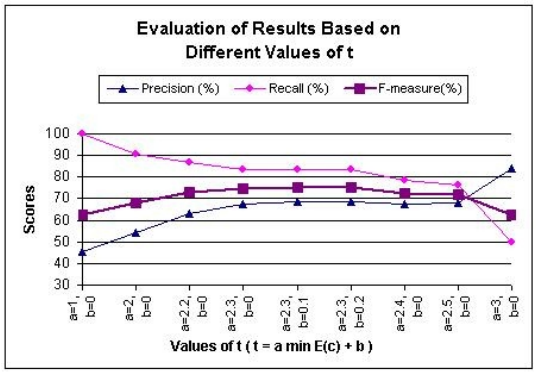

Figure 3: Effect of different t values on P, R and $F_{\beta=1}$, given $c_1 = 0.003$, $c_2 = 0.5$, and $c_3 = 10$.

In addition, several other aspects also have negative effects on the result. First, the sublanguage is anyway specific compared with the general language, therefore some MWV candidates were hard to give an evaluation. For instance, *transfect into/with* can be found in neither dictionaries we used in this experiment, then it is hard to give them human evaluation. Second, the POS tag errors during the processing had also a negative effect. E.g., in MWV candidate *be related to*, *related* was POS

[9] Since WordNet is lacking in MWV entries, we used the Oxford advanced learner's dictionary of current English (Encyclopedic version, 1992), and the online English-German dictionary LEO additionally, available at http://dict.leo.org/.

tagged as an adjective, which led to a reduction of the value of $E(relate\ to)$, since the MWV head of this inflectional structure was set to be the adjective *related* but not the root of the verb *relate*. Third, the language resources used in our experiment provided sometimes not the information we needed. For instance, WordNet was lacking in some specific lexical entries of verbs such as *synergize*, *pretreat*, and etc. Hence the distribution of their inflectional and derivational forms, such as *synergizes* and *pretreated*, could not be analyzed correctly.

Our following work is to combine this work with the domain-specific single verbs determined in the corpus (Xiao and Rösner, 2004a), in order to get a comprehensive understanding of *domain-specific verbs*. And, it will also be investigated if more domain specific resources (e.g., UMLS[10] specialist lexicon, etc.), as well as adaptation of general language resources (e.g., WordNet, etc.) to this specific domain can improve the evaluation in equation 1 or not. Another future work is to examine the distribution of the inflectional and derivational forms of MWVs for both MWV candidate evaluation and other IE tasks.

5 Appendix: Some extracted MWV candidates, ordered by scores of E(c).

Note: what in the parentheses before each candidate is the occurrences. If a MWV is annotated with a *PAS* tag, it means that this MWV is often used in passive form or as past participle phrase in this test corpus. The complete list is available at: http://wwwai.cs.uni-magdeburg.de/Members/xiao/mwvsAppendixTable.pdf

No.	MWV candidate	No.	MWV candidate
1	(8) be subject to	2	(5) subjectPAS to
3	(7) give rise to	4	(7) take place
5	(325) result in	6	(271) lead to
7	(293) associatePAS with	8	(89) fail to

References

Timothy Baldwin and Aline Villavicencio. 2002. Extracting the unextractable: A case study on verb-particles. In *Proceedings of CoNLL-2002*, pages 98–104. Taipei, Taiwan.

Colin Bannard, Timothy Baldwin, and Alex Lascarides. 2003. A statistical approach to the semantics of verb-particles. In *Proceedings of the ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, pages 65–72.

Eric Brill. 1994. Some advances in rule-based part of speech tagging. In *Proceedings of the Twelfth National Conference on Artificial Intelligence*, pages vol(1):722–727.

Gaël Dias. 2003. Multiword unit hybrid extraction. In *Proceedings of the ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, pages 41–48.

Christiane Fellbaum, editor. 1999. *WordNet: An Electronic Lexical Database*. The MIT Press. Cambridge. Massachusetts.

Christopher D. Manning and Hinrich Schütze. 2002. *Foundations of Statistical Natural Language Processing*. MIT Press.

Diana McCarthy, Bill Keller, and John Carroll. 2003. Detecting a continuum of compositionality in phrasal verbs. In *Proceedings of the ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, pages 73–80.

T. Ono, H. Hishigaki, A. Tanigami, and T. Takagi. 2001. Automated extraction of information on protein-protein interactions from the biological literature. *Bioinformatics*, pages 17(2):155–161.

Scott S. L. Piao, Paul Rayson, Dawn Archer, Andrew Wilson, and Tony McEnery. 2003. Extracting multiword expressions with a semantic tagger. In *Proceedings of the ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, pages 49–56.

Adwait Ratnaparkhi. 1996. A maximum entropy part-of-speech tagger. In *Proceedings of the Empirical Methods in Natural Language Processing Conference, University of Pennsylvania*.

Ivan A. Sag, T. Baldwin, F. Bond, A. Copestake, and D. Flickinger. 2002. Multiword expressions: A pain in the neck for nlp. In *Proceedings of the Third International Conference on Intelligent Text Processing and Computational Linguistics (CICLING 2002), Mexico city*, pages 1–15.

J. Thomas, D. Milward, C. Ouzounis, S. Pulman, and M. Carroll. 2000. Automatic extraction of protein interactions from scientific abstracts. In *The Pacific Symposium on Biocomputing'2000, Hawaii*, pages 541–551.

Chun Xiao and Dietmar Rösner. 2004a. Determining domain-specific verb vocabulary through corpora comparison and genre analysis. submitted.

Chun Xiao and Dietmar Rösner. 2004b. Finding high-frequent synonyms of a domain-specific verb in english sub-language of medline abstracts using wordnet. In *The Second Global Wordnet Conference, Brno, Czech Republic*, pages 242–247.

[10] Unified Medical Language System (UMLS), see http://www.nlm.nih.gov/research/umls/.