

Generating Multiwords from MEDLINE in the SPECIALIST Lexicon

Chris J. Lu, Ph.D.^{1,2}, Destinee Tormey¹, Lynn McCreedy, Ph.D.¹ and Allen C. Browne¹

¹National Library of Medicine, Bethesda, MD

²Medical Science & Computing, LLC, Rockville, MD

Abstract

Multiwords are vital to better NLP systems for more effective and efficient parsers, refining information retrieval searches, enhancing precision and recall in NLP applications, etc. The Lexical Systems Group (LSG) enhanced the coverage of multiwords in the Lexicon to provide a more comprehensive resource. This paper describes a new systematic approach to lexical multiword acquisition from MEDLINE through filters and matchers based on empirical models. The design goal, function description, and performance test of these filters and matchers are discussed. The result includes: 1) Generating the distilled MEDLINE n-gram set with better precision and similar recall to the MEDLINE n-gram set; 2) Establishing a model for generating high precision multiword candidates for a faster Lexicon build. We anticipate an accelerated growth of multiwords in the Lexicon with this system. Hence, improvement in recall or precision can be anticipated in NLP projects using the SPECIALIST Lexicon and its applications.

Introduction

The SPECIALIST Lexicon, distributed in the Unified Medical Language System (UMLS) Knowledge Sources by the National Library of Medicine (NLM), is a large syntactic lexicon of biomedical and general English, designed and developed to provide the lexical information needed for the SPECIALIST Natural Language Processing (NLP) System [1]. Lexical records are used for Part-of-Speech (POS) tagging, indexing, information retrieval, concept mapping, etc. in many NLP projects, such as Lexical Tools [2], MetaMap [3-4], cTAKES [5], Sophia [6], gSpell [7], STMT [8], SemRep, UMLS Metathesaurus, ClinicalTrials.gov, etc. It has been one of the richest and most robust NLP resources for the NLP/Medical Language Processing (MLP) community since its first release in 1994. It is important to keep the Lexicon up to date with broad coverage to ensure the success of NLP applications that use it.

Each lexical entry in the Lexicon records the syntactic, morphological, and orthographic information needed by the SPECIALIST NLP System. Terms must meet 3 requirements to qualify as lexical entries: 1) part-of-speech (POS), 2) inflections, and 3) a special unit of lexical meaning by themselves. Linguists in the LSG look at the usage of candidate terms from various sources to add terms into the Lexicon if the above three requirements are met. Terms (base forms and inflectional variants) may be single words or multiwords - namely words that contain space(s). If it is a multiword, such as “ice cream” or “hot dog”, it is called a LexMultiWord (LMW). Single words in the Lexicon have increased over 2.5 times from 180,468 in 2002 to 464,781 in 2015. These Lexicon single words cover only about 12.17% of unigrams (single words) from titles and abstracts in MEDLINE.2015. However, single-word Lexicon terms comprise 99.03% of MEDLINE unigrams if the word count (WC) is taken into consideration. In other words, the current Lexicon has a very high recall rate of single words in MEDLINE, because most frequently used single words in MEDLINE are covered. As for LMWs, we observe a continuous growth in the Lexicon from 88,324 (32.86%) in 2002 to 431,432 (48.14%) in 2015. Both the high coverage of existing single words and the trend of increasing growth of LMWs in the Lexicon lead to our position that multiword acquisition is key for future Lexicon building.

Multiwords are also vital to the success of high quality NLP applications [9-10]. First, multiwords are ubiquitous. Technical terminologies in many specialized knowledge domains, particularly in areas like medicine, computer science and engineering, are often created as Multiword Expressions (MWEs) [11-13]. Second, MWEs are hard to deal with in NLP tasks, such as identification, parsing, translation, and disambiguation, not only because MWEs have a large amount of distinct phenomena, but also due to the absence of major syntactic theories and semantic formalisms. Our Lexicon with multiwords remedies these issues. For example, most word segmentations are word-oriented (tokenization), relying on POS taggers, stemmers, and chunkers to segment each MWE as a phrasal unit from the sentence. This process can be improved if multiwords can be identified as a phrasal unit directly (such as through a Lexicon lookup) and not processed further by taggers, e.g. phrasal preposition (because of, due to), and adverbs (on time). Thus, part-of-speech ambiguity can be reduced through identifying the part-of-speech of these MWEs. Third, non-decomposable MWEs, such as fixed phrases (kingdom come, by and large) and idioms (kick the bucket, shoot the breeze), are very challenging tasks for NLP syntactically as well as semantically. While syntactic aspects of idiom usage necessitates a beyond-Lexical-level solution to those non-decomposable MWEs, fixed phrases are handled well as LMWs in our Lexicon. NLP techniques, such as Query Expansion, do not work well on fixed-phrase MWEs for

concept mapping, unless they are seen as LMWs. For example, “hot dog” should not be expanded as “high temperature canine” to find its concept. Instead, a direct Lexicon look up of “hot dog” (E0233017) without further query expansion resolves issues caused by fixed-phrase MWEs. Furthermore, the Metathesaurus concept associated with a sentence often coincides with the longest multiword in the sentence. This idea is used in MetaMap by identifying the longest LMWs in sentences for mapped concept ranking. Accordingly, a comprehensive lexicon with a rich resource of MWEs is an essential component to a more precise, effective, and efficient NLP system.

Research on Multiword Expressions (MWEs) has been growing since the late 1990s [13]. State of the art methods including statistical association measures [14-16], machine learning [17-19], syntactic patterns [12, 20-21], web queries [22], semantic analysis [23-24], and a combination of the above methods [11, 25-27] are used on MWE research for acquisition, identification, interpretation, disambiguation and applications. Despite a great deal of research on MWEs, there is no approach that fits perfectly for building LMWs in the SPECIALIST Lexicon. LMWs are a subset of MWEs due to our requirements that a legitimate Lexical entry have a POS, inflections, and be a unit of meaning. LMWs are distinguished from the broader notion of MWEs in four ways. First, a collocation¹ is not necessarily a LMW because it is not necessarily qualified as a Lexical entry. For example, “undergoing cardiac surgery” occurs frequently (2,799 hits in 3,416 documents in MEDLINE n-gram set, 2015 [28]), but it is not a LMW because it is not functioning as a special unit of meaning by itself². On the other hand, its subterm, “cardiac surgery”, which occurs frequently (20,978 hits in 34,465 documents) in MEDLINE, is a LMW. In other words, frequency alone is not sufficient to determine if a term is a LMW. For the same reason, some phrases are not LMWs. For example, “in the house” is not a LMW while “in house” is³. Second, verb particle constructions are handled by complementation types [29] in Lexical records to coordinate lexical meaning with syntactic characteristics of the verb. For example, “beat someone up” can be constructed from the Lexical record of “beat”, as shown in Figure 1. Similarly, light verbs that are covered within Lexical records, such as “make love” and “give birth”, are included in the Lexical records of make (E0038623) and give (E0029785), respectively. The information on these types of MWEs is stored inside the Lexical records and they are not considered LMWs (not a base form or inflectional variants of a Lexical entry). However, they can be retrieved/identified by a parser based on the Lexicon. Third, non-decomposable idioms are beyond the scope of the Lexicon⁴. Fourth, due to the complicated nature of MWEs, many methods only focus on bi-gram or tri-grams, which do not meet the requirements of including up to five-grams to reach an estimated recall value of 99.47% of multiwords [30].

Previously, an element word approach [31] was used to build the Lexicon by linguists through a web-based computer-aided tool, LexBuild [32]. Unigrams with high frequency (WC) from the MEDLINE n-gram set that are not in Lexicon were used as element words for finding new candidate terms in Lexicon. There are several issues with this approach: 1) it is time consuming; 2) multiwords associated with high frequency element words do not necessarily have high frequency; 3) new multiwords associated with processed element words are missed. According to our estimate, it will take more than 21 years for current LSG staff to add all multiwords to the Lexicon by this approach⁵. Thus, developing a new system for multiword acquisition is imperative to build the Lexicon.

Objective and Approach

Our goal is to develop a system to add LMWs to the SPECIALIST Lexicon efficiently and systematically by

¹ Collocation is an arbitrary statistically significant association between co-occurring items.

² Moreover, this collocation is sometimes, but not always, a single part of speech. In MEDLINE, it is usually a reduced relative clause, e.g. “patients undergoing cardiac surgery” (reduced from “patients who are undergoing cardiac surgery”), which puts it outside part-of-speech categories. However, it could also be a subject, which makes it a MWE noun phrase, e.g. “Undergoing cardiac surgery is taxing.”

³ For example, PMID 17987364: “An in house library of more than 400 compounds with systematically varied structures is available...” This adjective has the spelling variant in-house (E0555310).

⁴ Idioms, such as “kick the bucket” and “shoot the breeze” are another type of non-decomposable MWE. Aligning the syntactic analysis of idiomatic phrases with their semantic interpretations is beyond the scope of what a lexicon can accomplish. Thus, they are not under consideration here.

⁵ Empirical measurements show the ratio of multiwords is between 50~80% in a corpus of scientific biomedical abstracts [13, 33]. Let’s use the mean value, 65%, as an estimated ratio. So, the estimated amount of multiwords is around 870K based on the latest Lexicon, 2016 (468,655 single words and 446,928 multiwords). Multiwords that are not in the Lexicon (423K) shall take 21 years for LSG staff to complete (averagely, 20,000 terms are added by LSG staff annually). And this estimate does not account for the fact that many new multiword terms are continuously being created by biomedical researchers and English users in general.

generating a high precision LMW candidates list. MEDLINE is chosen as our corpus for the initial phase. The MEDLINE n-gram set distributed annually by NLM [30] contains all possible LMWs in MEDLINE. Two types of computer programs, filters and matchers, are developed to extract LMWs from n-grams as described below.

```
{base=beat
entry=E0012175
  cat=verb
  variants=irreg|beat|beats|beat|beaten|beating|
  intran
  intran;part(about)
  tran=np
  tran=np;part(back)
  tran=np;part(up)
  tran=np;part(down)
  tran=np;part(in)
  link=advbl
  cplxtran=np,advbl
  nominalization=beating|noun|E0219216
}
```

Figure 1. Lexical record of the verb “beat”.

Filters

Filters (exclusive filters) are designed to retrieve LMWs from MEDLINE n-grams by trapping invalid LMWs. The performance of a filter can be measured by passing rate (pass-through terms/total terms), efficiency (trapped terms/total terms), and accuracy (see the section below, “Accuracy Test of Filters”). The design goal of these filters focuses on high accuracy so they do not trap valid LMWs unintentionally and result in dropping recall. Ideally, no valid multiwords should be trapped by these filters. Such filters do not necessarily have high efficiency. Nevertheless, the efficiency can be improved significantly by applying a series of high accuracy filters. Exclusive filters are developed based on empirical models with heuristic rules in this task. They are categorized into four types as described below:

1. General Exclusive Filters:

This type of filter is intuitive and based on surface features of terms. Terms composed merely of certain characters/words, such as punctuation, digit, number, etc., do not meet the requirement of having a special unit of lexical meaning to themselves. They can be used for general purpose to filter out n-grams that are not valid LMWs.

- Pipe Filter: A term that contains pipe(s) is trapped because a pipe is used as a field separator in most NLP systems. Trapped examples include: “(|r|”, “Ag|AgCl”, etc.
- Punctuation or Space Filter: A term that contains nothing but punctuation or space(s) is trapped. Trapped examples include: “=”, “+/-”, “<”, “(%)”, and “-->”.
- Digit Filter: A term that contains nothing but digit(s), punctuation, and space(s) is trapped. Trapped examples include: “2”, “10”, “95%”, “2000”, “3-5”, “\$1,500”, “(+/10.05)”, “192.168.1.1”, and “[192, 168]”.
- Number Filter: A term that contains nothing but number(s) is trapped. This filter can be considered as a domain filter (described in the section of domain filters) because all numbers are already recorded in the Lexicon. Trapped examples include: “two”, “first and second”, “one third”, “twenty-eight”, “Four hundred and forty-seven”, and “half”.
- Digit and Stopword Filter: A term that contains nothing but digit(s) or stopword(s) is trapped. Trapped examples include: “50% of”, “of the”, “1, 2, and”, “2003 to 2007”, “for >=50%”, and “OR-462”.

2. Pattern Exclusive Filters:

This type of filter looks for certain matching patterns in a term for trapping. Computer programs are implemented based on observed empirical patterns. Some filters involve sophisticated computer algorithms to work.

- Parenthetic Acronym (ACR) Pattern Filter: A parenthetic acronym is a conventional way of representing an acronym expansion with the associated acronym. The pattern is an acronym expansion followed by an acronym within a closed parenthesis, e.g., [acronym-expansion (ACR)]. A term that contains such a parenthetic acronym

pattern is trapped because it contains a potential multiword plus the associated acronym and thus cannot be a valid term. This pattern can be used as a matcher (discussed in the section of matchers) to retrieve multiword candidates because expansions of acronyms are usually valid multiwords. Trapped examples include: “magnetic resonance imaging (MRI)”, “imaging (MRI)”, “magnetic resonance (MR) imaging” and “(CREB)-binding protein (CBP)”.

- Indefinite Article Filter: A term that starts with an indefinite article and a space, “a “ or “A “, without other n-grams that match as its spelling variants (spVar) pattern in the corpus (n-gram set) is trapped. Patterns of [a-XXX] and [aXXX] are used as the spVar pattern of indefinite articles of [a XXX], where XXX represents any term. Trapped examples include: “a significant”, “a case”, “a case of”. “a dose-dependent” and “a delivery rate per”.
- UPPERCASE Colon Filter: A term that contains the pattern of [UPPERCASE:] is trapped. In MEDLINE, this is a conventional usage for this pattern, such as “RESULTS:”, “CONCLUSION:”, “BACKGROUND:”, “OBJECTIVE:”, “METHODS:” and “MATERIALS AND METHODS:”. In addition, trapped examples include “MATERIALS AND METHODS: The”, “95% CI:” and “PHPT:”, “vs N:”.
- Disallowed Punctuation Filter: A term that contains disallowed punctuation is trapped. Disallowed punctuation includes: { } _ ! @ # * ; " ' ? ~ = | < > \$ % ^ . Trapped examples include: “(n =”, “(P < 0.05)”, “N^N”, “group (n=6) received”, and “CYP3A7*1C”.
- Measurement Pattern Filter: A term that contains a measurement pattern is trapped. A measurement pattern is [number + unit], including age (4-year-old, 4 year-old, four year-old, 4 year-olds, 65 years or older with), time (four months, 1 January 1991, from May 2002, 6 hours plus), range (2-3 days, 1-2 tablets), temperature (at -5 degrees), dosage (10 cigarettes per day, 0.1-2.3 mg/day), and others (e.g. 60 inches, 0.5 mg, 3 mg/EE, 10 mg/kg and 50 mg/kg/day).
- Incomplete Pattern Filter: A term that contains an incomplete pattern is trapped. A valid multiword should have completed parentheses or brackets. Incomplete patterns are terms that do not have an even number of left and right parentheses or square brackets or they are not closed. Trapped examples include: “II (Hunter syndrome”, “0.05 higher”, “bond]C-C[triple”, “(chi(2)” and “interval [95%”.

3. Lead-End-Terms Exclusive Filters:

LMWs don't start with certain terms, such as auxiliaries (be, do, etc.), complementizers (that), conjunctions (and, or, but, etc.), determiners (a, the, some, etc.), modals (may, must, can, etc.), pronouns (it, he, they, etc.), and prepositions (to, on, by, etc.). They are called invalid lead terms. Similarly, multiwords don't end with invalid end terms. They are used in exclusive filters to exclude invalid multiwords from the n-grams. Terms from the Lexicon with any of the above seven categories are used as invalid lead end term (ILET) candidates. ILETs only comprise 0.05% (488) of total forms in Lexicon 2016 (915,583). Furthermore, ILET candidates are considered static because no new terms in the above 7 categories have been added since 2010. Please refer to LSG documents on deriving invalid lead-terms and end-terms at the Lexicon web site for detail [34].

- Absolute Invalid Lead-Term Filter: A term that leads with an absolute invalid lead-term (AILT) is trapped. There are 382 AILTs derived from the Lexicon, such as “the”, “about”, “aka”, “as to”, “as well as”, “isn't”, etc. [35]. Trapped examples include: “The results”, “from the”, “is a”, and “of a”.
- Absolute Invalid End-Term Filter: A term that ends with an absolute invalid end-term (AIET) is trapped. There are 407 AIETs derived from the Lexicon, such as “the”, “W/O”, “with”, “along with”, “i.e.” and “such as” [36]. Trapped examples include: “patients with”, “associated with”, “at the”, suggest that” and “between the”.
- Lead-End-Term filter: A term that leads with an ILET and also ends with an ILET is trapped. Trapped examples include: “in a”, “to be”, “with a” and “as a”.
- Lead-Term No SpVar Filter: A term that leads with a valid lead term (VLT) without any other n-gram matching its spelling variants (spVar) pattern in the corpus (n-gram set) is trapped. There are 52 VLTs derived from the Lexicon, such as “to”, “as”, “as if”, “on board” and “on-board” [37]. Trapped examples include: “to determine”, “as a result”, “to evaluate”, “for example” and “plus LHRH-A”.
- End-Term No SpVar Filter: A term that ends with a valid end term (VET) without any other n-gram matching its spelling variants (spVar) pattern in the corpus (n-gram set) is trapped. There are 27 VETs derived from the Lexicon, such as “a”, “be”, “being”, “of” and “off” [38]. Trapped examples include: “effects of”, “presence of”, “comparison

of”, “was used to”, “(HPV) in” and “loss of two or more”.

4. Domain Exclusive Filters:

This type of filter is developed to exclude terms that are in a certain domain, such as single word, frequency, and existing in the current Lexicon.

- **Lexicon Domain Filter:** Our task is to find new terms that are not in the Lexicon. A term that is already in the Lexicon (our domain) is trapped. This filter includes exact matches as well as matches after different levels of normalization (such as lowercase and punctuation). For example, “skin disease”, “Skin disease”, “:skin disease”, “:SKIN DISEASE”, “:skin-disease,”, “:SKIN-DISEASE,” are all trapped because “skin disease” is in the Lexicon.
- **Single Word Filter:** A term that does not contain a space is trapped. This filter is used as a domain filter if multiwords are the only interest.
- **MEDLINE Frequency Filer:** A term with a document count (DC) or word count (WC) in the MEDLINE within a specified range is trapped.

Accuracy Test of Filters

An accuracy test model is established for testing the above developed filters. Accuracy is defined as: $(TP + TN) / (TP + TN + FP + FN)$. Terms (875, 890) in the Lexicon 2015 release are used to test exclusive filters. The accuracy is simplified to the passing rate (also recall) in this test because all Lexicon terms are valid (relevant) and thus both TN (not retrieved, not relevant) and FP (retrieved, not relevant) are 0, as shown in equation 1.

$$\text{Accuracy} = (TP + TN) / (TP + TN + FP + FN) = TP / (TP + FN) = \text{recall} = \text{passing rate} \quad \text{Eq. (1)}$$

Columns 3 and 4 in Table-1 show the accuracy rate and number of trapped terms (FN) from Lexicon for this accuracy test. The results show that all filters have high accuracy rates above 99.9887%, which is at filter-15, Lead-Term No SpVar Filter.

Distilled MEDLINE N-gram Set

The LSG has distributed the MEDLINE n-gram set annually since 2014. This n-gram set provides comprehensive raw data from titles and abstracts of MEDLINE [30]. Due to its large-scale size (18,148,692 n-grams), it is difficult to be handled by computer programs with complicated algorithms. A distilled MEDLINE n-gram set, with higher precision and similar recall in terms of LMWs, is derived by applying a series of high accuracy filters. Let’s say X filters are applied to all MEDLINE n-grams. The number of valid LMWs (TP) and number of invalid LMWs (FP) of the filtered MEDLINE n-gram set are TP_i and FP_i , respectively, where $i = 0, 1, 2, \dots, X$. The number of valid LMWs are about the same ($TP_0 \cong TP_1 \cong TP_2 \cong \dots \cong TP_X$) because they have high accuracy while the number of invalid LMWs is reduced ($FP_0 > FP_1 > FP_2 > \dots > FP_X$) from the original MEDLINE n-gram set to the final distilled MEDLINE n-gram set. Accordingly, the distilled MEDLINE n-gram set (X) has higher precision and similar recall to the MEDLINE n-gram set (0), as shown in equations 2 and 3, respectively, where the number of FN_i (not retrieved, relevant) is a constant.

$$\text{Precision:} \quad P_X = TP_X / (TP_X + FP_X) \cong TP_0 / (TP_0 + FP_X) > TP_0 / (TP_0 + FP_0) \quad \text{Eq. (2)}$$

$$\text{Recall:} \quad R_X = TP_X / (TP_X + FN_X) = TP_X / (TP_X + FN_0) \cong TP_0 / (TP_0 + FN_0) \quad \text{Eq. (3)}$$

Sixteen high accuracy filters are applied to the MEDLINE n-gram set in the sequential order as shown in Table 1 to filter out invalid terms. Columns 5, 6, and 7 in Table 1 show the passing rate, number of trapped terms from MEDLINE n-grams, and cumulative passing rate for all filters. As a result, the distilled MEDLINE n-gram set, after filtering out the majority of invalid LMWs, contains about 37.31% (6,793,561) n-grams of the MEDLINE n-gram in the 2015 release. Theoretically, the distilled MEDLINE n-gram set has higher precision and similar recall compared to the MEDLINE n-gram set. The size reduction (to about 1/3) also makes it possible for computer programs with complicated algorithms (such as the SpVar Pattern matcher) to work in a reasonable time frame in practice (see discussion in the sections “Matchers” and “Practice and Evaluation”).

Matchers

Matchers (inclusive filters) are designed to retrieve LMWs from MEDLINE n-grams by trapping valid multiwords that match valid LMW patterns. In other words, terms trapped by matchers should be valid LMWs. The design goal

of matchers is set to generate high precision LMW candidates. Accordingly, not all valid LMWs might be trapped and results might decrease in recall. Four matchers are developed based on empirical models as described below.

Table 1. Exclusive filters: accuracy, passing rate, and number of trapped terms.

ID	Filter Name	Accuracy	NTT-L	Pass Rate	NTT-M	Cum. P.R.
1	Pipe	100.0000%	0	100.0000%	6	100.0000%
2	Punctuation or Space	100.0000%	0	99.9977%	386	99.9977%
3	Digit	99.9999%	1	99.3141%	116,772	99.3118%
4	Number	99.9953%	41	99.9760%	4,056	99.2879%
5	Digit and Stopword	99.9993%	6	99.1595%	142,067	98.4534%
6	Parenthetic Acronym - (ACR)	100.0000%	0	99.0232%	163,714	97.4917%
7	Indefinite Article	99.9985%	13	98.1703%	303,679	95.7079%
8	UPPERCASE Colon	99.9999%	1	99.4302%	92,841	95.1625%
9	Disallowed Punctuation	99.9978%	19	99.3020%	113,073	94.4983%
10	Measurement	99.9967%	29	98.1947%	290,421	92.7924%
11	Incomplete	99.9999%	1	97.8470%	340,109	90.7945%
12	Absolute Invalid Lead-Term	99.9947%	46	73.0945%	4,158,702	66.3658%
13	Absolute Invalid End-Term	99.9997%	3	78.8984%	2,384,059	52.3615%
14	Lead-End-Term	99.9992%	7	99.9741%	2,312	52.3480%
15	Lead-Term No SpVar	99.9887%	99	85.6678%	1,277,229	44.8454%
16	End-Term No SpVar	99.9975%	22	83.1945%	1,283,001	37.3089%

1. Parenthetic Acronym Pattern Matcher

Acronym expansions are good candidates for LMWs. They are retrieved by the following steps. First, apply Parenthetic Acronym Filter on the MEDLINE n-gram set to retrieve terms matching the pattern of [acronym expansion (ACR)]. For example, “computed tomography (CT)”, “magnetic resonance imaging (MRI)”, “Unified Health System (SUS)”, etc. are retrieved from the n-gram set. Second, retrieve expansions if they match the associated acronym. Heuristic rules are implemented, such as checking the initial characters of first and last words of the expansion to match the first and last characters of the associated acronym. For example, the expansion of “Unified Health System (SUS)” is not a valid candidate because the first initial of the expansion (U) does not match the first character of acronym, (S). Third, remove terms if the expansion is a subterm of other expansions in the list. For example, both “cell sarcoma (CCA)” and “clear cell sarcoma (CCA)” pass the first two steps. The invalid LMW of “cell sarcoma” is removed in this step because it is a subterm of the valid LMW “clear cell sarcoma”. This matcher is used to retrieve LMW candidates as discussed in the section of practice and evaluation.

2. Spelling Variant Pattern Matcher

In general, an n-gram is a good LMW candidate if it has spelling variants existing in the same corpora (n-gram set). A sophisticated computer algorithm is developed to identify all n-grams that have potential spVars. First, a special normalization program is developed to normalize spVars into their canonical forms by converting non-ASCII Unicode to ASCII (e.g. “Labbé” to “Labbe”), synonym substitution (e.g. “St. Anthony's fire” to “Saint Anthony's fire”), rank substitution (e.g. “Vth nerve” to “5th nerve”), number substitution (e.g. “12-lead” to “twelve-lead”), Roman numeral substitution (e.g. “BoHV-I” to “BoHV-1”), strip punctuation (e.g. “lamin-A” to “lamin A”), stripping genitive (e.g. “Laufe's forceps” to “Laufe forceps”), converting to lowercase, and removing any space(s). All terms that have the same normalized spVar canonical form are potential spVars to each other. As shown in Table 2, 88.30% of 260,431 spVars in Lexicon 2015 are identified by spVar normalization. Second, a M.E.S. model is developed to improve the recall. The M.E.S. model is composed of an algorithm of Metaphone phonetic code [39], edit distance⁶, and minimum

⁶ Edit distance is the minimum number of operations required to transform one term into the other. It is used to measure the similarity of two terms.

sorted distance⁷. All terms that have the same phonetic code and edit distance less than a specified value are collected and sorted. The pair that has the minimum sorted distance (the closest pair) is identified as spVars. For example, “yuppie flu” and “yuppy flu” have different spVar canonical forms of “yuppieflu” and “yuppyflu”, respectively, and thus are not identified as spVars in the normalization step. They are identified as spVars in the M.E.S. model because they have the same Metaphone code of [YPFL], edit distance of 2, and the minimum sorted distance. This step identifies more spVars that can’t be identified by the normalization in the previous step. Third, an E.S. model is developed for further improvement on recall. Terms that have an edited distance less than a specified value are collected and sorted. The pair that has the minimum sorted distance is identified as spVars. For example, “zincemia” and “zinkaemia” are identified as spVars by the E.S. model with an edit distance of 1 while they were not identified as spVars in the previous steps because they have different spVar canonical forms of “zincemia” and “zinkaemia”; different Metaphone codes of [SNSM] and [SNKM], respectively. By relaxing the value of edit distance in both models, our program reaches 99.90% recall on spVar identification in six steps in this test, as shown in Table 2. Due to the complexity of this algorithm, further purification processes of core-term⁸ normalization and frequency threshold restriction (WC > 150) are applied to reduce the size of the n-gram set for better performance in practice. As a result, 650,518 spVars in 233,651 spVar classes are identified from the purified MEDLINE n-gram set by this approach.

Table 2. Recall test for processes in the Spelling Variant Pattern matcher.

Step	Model	Edit Dist.	Recall
1	SpVar Norm	N/A	88.30%
2	M.E.S.	2	98.53%
3	E.S.	1	99.49%
4	M.E.S.	3	99.59%
5	E.S.	2	99.87%
6	M.E.S.	4	99.90%

3. Metathesaurus CUI Pattern Matcher

A term with valid concept(s) has a better possibility of being an LMW candidate. The Synonym Mapping Tool (SMT) in STMT [8] is used to retrieve Metathesaurus concepts (CUIs) in this model to generate LMW candidates. The SMT is set up to find concepts within 2 subterm substitutions by their synonyms. The default synonym list in SMT is used.

4. EndWord Pattern Matcher

In the biomedical domain, multiwords often end with certain words (EndWords), such as “syndrome” (e.g. “migraine syndrome”, “contiguous gene syndrome”, etc.), “disease” (e.g. “Fabry disease”, “Devic disease”, etc.), and so on. An EndWord candidate list composed of the top 20 frequency EndWords for LMWs has been derived from the Lexicon [40]. These EndWords are used in this matcher to retrieve LMW candidates.

Practice and Evaluation

The first attempt in our practice is to apply the Parenthetic Acronym Pattern matcher on the MEDLINE n-gram set to retrieve acronym expansions. The lowercased core-terms of these acronym expansions are used for LMW candidates. 14,400 LMW candidates are retrieved by this process. 13,170 candidates in this list are tagged as valid LMWs to reach 91.46% precision, as shown in case 1 (Table 3). The recall can’t be found because all LMWs from MEDLINE cannot be identified in real practice. Case 1 is used as the gold standard for the further analysis of other filters and matchers.

⁷ Sorted distance is the distance between two terms in a sorted list in an alphabetic order for a set of terms. It is used to measure the similarity of two terms compared to other terms in the list. All terms in the sorted list should have similar characteristics. Sorted distance is a relative measurement and usually requires other algorithms to find similar terms for the sorted list.

⁸ A core-term normalization is to normalize an n-gram to its core form by stripping the leading and ending punctuation. For example, “in details,”, “- in details”, “- in details,” have the same core-term form of “in details”. Core-terms might have punctuation at the end or internally, such as “clean room(s)” or “in (5) details”. It is a useful normalization to cluster terms with same core together from the n-gram set in multiword acquisition.

Recall in Table 3 is calculated based on case 1.

Table 3. Precision, recall, F1 analysis for matchers and filters.

Case	Test Case - Model	TP	FP	Precision	Recall	F1
1	Parenthetic Acronym - gold standard	13,170	1,230	0.9146	1.0000	0.9554
2	Distilled MEDLINE N-gram Set (16 filters)	13,165	795	0.9431	0.9996	0.9705
3	Spelling Variant Pattern matcher	6,609	283	0.9589	0.5018	0.6589
4	Metathesaurus CUI Pattern matcher	8,678	512	0.9443	0.6589	0.7762
5	EndWord Pattern matcher	1,587	108	0.9363	0.1205	0.2135
6	SpVar + CUI + Distilled	4,993	127	0.9752	0.3791	0.5460
7	SpVar + CUI + EndWord + Distilled	690	5	0.9928	0.0524	0.0995

In the second attempt, we applied the 16 filters in the same sequential order of deriving the distilled MEDLINE n-gram set (Table 1) to the baseline (case 1). The results (case 2) show an improvement on F1 score with better precision and almost the same recall. This confirms the theoretic conclusion from the result of the accuracy test on these filters that the distilled MEDLINE n-gram contains almost the same amount of valid multiwords as the MEDLINE n-gram set while its size is reduced to 37.31%. Furthermore, we observed that the cumulative recall rates of these 16 filters on the Lexicon (0.999671, multiple product of column 3 in table 1) and baseline (0.999620) are almost identical. This implies that the Lexicon and the baseline derived from MEDLINE have similar characteristics on multiwords. The Lexicon could be concluded to be a good representative subset of MEDLINE in terms of multiwords if this number is the same for a bigger test set. Further study of this possibility is needed.

The models of Spelling Variants Pattern matcher, Metathesaurus CUI Pattern matcher, and EndWord Pattern matcher are applied to the baseline, as shown in test cases 3-5 in Table 3. As expected, the results show improvement in precision while the recall dropped. SpVar Pattern seems have the best improvement on precision while CUI Pattern seems have the best F1 score among these 3 matchers. Precision can be further improved by combining these matchers. Case 6 combines SpVar and CUI patterns to reach a higher precision. This model is applied to the distilled MEDLINE n-gram set to generate LMW candidates for building the Lexicon. Furthermore, the precision can further be improved in case 7 if terms ending with certain EndWords are the only interest. In short, combining of filters and matchers improves precision even on a baseline with high precision. This work focuses on generating high precision LMW candidates. On the other hand, the recall of the matchers is not emphasized because there are too many multiwords yet to be found.

Conclusion and Future Work

A set of high accuracy filters has been developed and tested. The filters are used to derive the distilled MEDLINE n-gram set with better precision and similar recall to the MEDLINE n-gram set. Four matchers have also been developed and evaluated. Combinations of filters and matchers are used to generate high precision LMW candidates for building the Lexicon. The LSG plans to continuously enhance filters and matchers for further improvement. The filters and matchers we have developed are generic and can be used independently or in combination for different research purposes. Most importantly, this approach provides a modular framework that serves as a platform for collaboration on extending more and better filters and matchers for LMW acquisition and NLP research.

Multiwords are pervasive, challenging and vital in NLP. The LSG attempts to provide a lexicon with high coverage (recall) of multiwords matching that of single words. We believe the impact of enriched multiword acquisition will enhance the precision, recall, and naturalness of NLP applications. The SPECIALIST Lexicon, the MEDLINE n-gram set and the distilled MEDLINE n-gram set are distributed by the National Library of Medicine (NLM) annually via an Open Source License agreement.

Acknowledgement

This research was supported by the Intramural Research Program of the NIH, National Library of Medicine. The authors would like to thank Mr. Guy Divita for his valuable discussions and suggestions.

References

1. McCray AT, Aronson AR, Browne AC, Rindfleisch, TC, Razi, A, Srinivasan S. UMLS Knowledge for Biomedical Language Processing. Bull. Medical Library Assoc. 1993;81(2):184-94.

2. McCray AT, Srinivasan S, Browne AC. Lexical Methods for Managing Variation in Biomedical Terminologies. Proceedings of the 18th Annual Symposium on Computer Applications in Medical Care; 1994. p. 235-9.
3. Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. Proceedings of AMIA Annual Symposium; 2001. p. 17-21.
4. Aronson AR and Lang FM. An Overview of MetaMap: Historical Perspective and Recent Advances. JAMIA. 2010;17:229-36.
5. Savova GK, Masanz JJ, Ogren PV, et al. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. JAMIA. 2010;17(5):507-13.
6. Divita G, Zeng QT, Gundlapalli AV, et al. Sophia: A Expedient UMLS Concept Extraction Annotator. Proceedings of AMIA Annual Symposium; 2014 Nov. 15-19; Wash., DC; 2014. p. 467-76.
7. Divita G, Browne AC, Tse T, Cheh ML, Loane RF, Abramson M., A Spelling Suggestion Technique for Terminology Servers. Proceedings of AMIA Annual Symposium; 2000 Nov. 4-8; Los Angeles, CA; 2000. p. 994.
8. Lu, CJ and Browne AC. Development of Sub-Term Mapping Tools (STMT). Proceedings of AMIA Annual Symposium; 2012 Nov. 3-7; Chicago, IL; 2012. p. 1845.
9. Sag I, Baldwin T, Bond F, Copestake A, and Flickinger D. Multiword expressions: A pain in the neck for NLP. Proceedings of the 3rd International Conference on Intelligent Text Processing and Computational Linguistics (CICLING 2002); Mexico City, Mexico; 2002. p. 1-15.
10. Fraser S. Technical vocabulary and collocational behaviour in a specialised corpus. Proceedings of the British Association for Applied Linguistics (BAAL); 2009 Sep 3-5; Newcastle University; 2009. p. 43-8.
11. Frantzi K, Ananiadou S, Mima H. Automatic Recognition of Multi-Word Terms: the C-value/NC-value Method. International Journal on Digital Libraries. 2000;3(2):115-30.
12. Green S, de Marneffe MC, Manning CD. Parsing models for identifying multiword expressions. Computational Linguistics. 2013;39(1):195–227.
13. Ramisch C. Multiword Expressions Acquisition: A Generic and Open Framework (Theory and Applications of Natural Language Processing). 2015th Edition. Springer; 2014; p. 4, 9, 37.
14. Silva JF and Lopes GP. A Local Maxima method and a Fair Dispersion Normalization for extracting multi-word units from corpora. Proceeding of the Sixth Meeting on Mathematics of Language (MOL6); 1999; Orlando, FL, USA; 1999. p. 369-81.
15. Fazly A, Cook P, Stevenson S. Unsupervised Type and Token Identification of Idiomatic Expressions. Computational Linguistics. 2009;35(1):61-103.
16. Pecina P. Lexical association measures collocation extraction. Language Resources and Evaluation. 2010;44:137-58.
17. Boukobza R, Rappoport A. Multi-Word Expression Identification Using Sentence Surface Features. Proceedings of the Conference on Empirical Methods in Natural Language Processing; 2009 August 6-7; Singapore; 2009. p. 468–77.
18. Tsvetkov Y, Wintner S. Identification of Multi-word Expressions by Combining Multiple Linguistic Information Sources. Proceedings of the Conference on Empirical Methods in Natural Language Processing; 2011 July 27-31; Edinburgh, Scotland, UK; 2011. p. 836–45.
19. Green S, de Marneffe MC, Bauer J, and Manning CD. Multiword Expression Identification with Tree Substitution Grammars: A Parsing tour deforce with French. Proceedings of EMNLP; 2011 July 27-31; Edinburgh, Scotland, UK; 2011. p. 725–35.
20. Seretan V and Wehrli E. Multilingual collocation extraction with a syntactic parser. Language Resources and Evaluation, 2009 March;43(1):71-85.
21. Kim SN and Baldwin T. How to pick out token instances of English verb-particle constructions. Language Resources and Evaluation. 2010 April;44(1):97-113.
22. Takahashi S and Morimoto T. Selection of Multi-Word Expressions from Web N-gram Corpus for Speech Recognition. Proceedings of International Symposium on Natural Language Processing (SNLP); 2013 Oct. 28-30; Phuket, Thailand; 2013. p. 6-11.
23. Pearce D. Using Conceptual Similarity for Collocation Extraction. Proceedings of the 4th UK Special Interest Group for Computational Linguistics (CLUK4), 2001 January 10-11; Sheffield, U.K., University of Sheffield; 2001. p. 34–42.
24. Baldwin T, Bannard C, Tanaka T, Widdows D. An Empirical Model of Multiword Expression Decomposability. Proceedings of the ACL Workshop on Multiword Expressions: Analysis, Acquisition and Treatment, 2003 July 12; Sapporo, Japan; 2003. p. 89-96.

25. Calzolari N, Fillmore CJ, Grishman R, et al. Towards Best Practice for Multiword Expressions in Computational Lexicon. Proceedings of the Third International Conference on Language Resources and Evaluation (LREC); 2002 May 29-31; Las Palmas, Canary Islands, Spain; 2002. p. 1934-40.
26. Bejček E, Straňák P, Pecina P. Syntactic Identification of Occurrences of Multiword Expressions in Text using a Lexicon with Dependency Structures. Proceedings of the 9th Workshop on Multiword Expressions; 2013 June 13-14; Atlanta, Georgia; 2013. p. 106–15.
27. Sangati F, Cranenburgh AV. Multiword Expression Identification with Recurring Tree Fragments and Association Measures. Proceedings of Annual conference of North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT); 2015 May 31-June 5; Denver, Colorado; 2015. p. 10-8.
28. The Lexical System Group, Lister Hill National Center for Biomedical Communications, National Library of Medicine. The MEDLINE n-gram set, 2015 [Internet]. Available from: <https://lsg3.nlm.nih.gov/LexSysGroup/Projects/nGram/2015/index.html>
29. Browne AC, McCray AT, Srinivasan S. The SPECIALIST LEXICON. Lister Hill National Center for Biomedical Communications, National Library of Medicine, Bethesda, Maryland, 2000 June. p. 30-49.
30. Lu CJ, Tormey D, McCreedy L, Browne AC. Generating the MEDLINE N-Gram Set, Proceedings of AMIA Annual Symposium, 2015 Nov. 14-18; San Francisco, CA; 2015. p. 1569.
31. Lu CJ, Tormey D, McCreedy L, Browne AC. Using Element Words to Generate (Multi)Words for the SPECIALIST Lexicon. Proceedings of AMIA Annual Symposium; 2014 Nov 15-19; Wash., DC; 2014. p. 1499.
32. Lu CJ, McCreedy L, Tormey D, and Browne AC. A Systematic Approach for Automatically Generating Derivational Variants in Lexical Tools Based on the SPECIALIST Lexicon. IEEE IT Professional Magazine. 2012 May/June;36-42.
33. Ramisch C. Multiword terminology extraction for domain-specific documents. Master's thesis, École Nationale Supérieure d'Informatique et de Mathématiques Appliquées, Grenoble, 2009. 79p
34. The Lexical System Group, Lister Hill National Center for Biomedical Communications, National Library of Medicine. Deriving invalid lead terms and invalid end terms from Lexicon [Internet]. Available from: <https://lsg3.nlm.nih.gov/LexSysGroup/Projects/lexicon/current/docs/designDoc/UDF/multiwords/leadEndTerms/invalidLeadEndTermsFromLexicon.html>
35. The Lexical System Group, Lister Hill National Center for Biomedical Communications, National Library of Medicine. List of absolute invalid lead terms for filters [Internet]. Available from: <https://lsg3.nlm.nih.gov/LexSysGroup/Projects/lexicon/current/docs/designDoc/UDF/multiwords/exFilters/exFilterLeadTermAbs.html>
36. The Lexical System Group, Lister Hill National Center for Biomedical Communications, National Library of Medicine. List of absolute invalid lead terms for filters [Internet]. Available from: <https://lsg3.nlm.nih.gov/LexSysGroup/Projects/lexicon/current/docs/designDoc/UDF/multiwords/exFilters/exFilterEndTermAbs.html>
37. The Lexical System Group, Lister Hill National Center for Biomedical Communications, National Library of Medicine. List of invalid lead terms for filters [Internet]. Available from: <https://lsg3.nlm.nih.gov/LexSysGroup/Projects/lexicon/current/docs/designDoc/UDF/multiwords/exFilters/exFilterLeadTermPat.html>
38. The Lexical System Group, Lister Hill National Center for Biomedical Communications, National Library of Medicine. List of invalid end terms for filters [Internet]. Available from: <https://lsg3.nlm.nih.gov/LexSysGroup/Projects/lexicon/current/docs/designDoc/UDF/multiwords/exFilters/exFilterEndTermPat.html>
39. Philips L. Hanging on the Metaphone. Computer Language. 1990 December;17(12):39-43.
40. The Lexical System Group, Lister Hill National Center for Biomedical Communications, National Library of Medicine. List of EndWords for EndWord matcher [Internet]. Available from: <https://lsg3.nlm.nih.gov/LexSysGroup/Projects/lexicon/current/docs/designDoc/UDF/multiwords/matchers/matcherEndWord.html>