Generating a Distilled N-Gram Set
Effective Lexical Multiword Building in the SPECIALIST Lexicon

Chris J. Lu[1,2], Destinee Tormey[1,2], Lynn McCreedy[1,2] and Allen C. Browne[1]
[1]*National Library of Medicine, Bethesda, MD, U.S.A.*
[2]*Medical Science & Computing, LLC, Rockville, MD, U.S.A.*

Keywords: MEDLINE N-Gram Set, Multiwords, Medical Language Processing, Natural Language Processing, the SPECIALIST Lexicon.

Abstract: Multiwords are vital to better Natural Language Processing (NLP) systems for more effective and efficient parsers, refining information retrieval searches, enhancing precision and recall in Medical Language Processing (MLP) applications, etc. The Lexical Systems Group has enhanced the coverage of multiwords in the Lexicon to provide a more comprehensive resource for such applications. 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, various tests and applications of filters, matchers, and data are discussed. Results include: 1) Generating a smaller (38%) distilled MEDLINE n-gram set with better precision and similar recall to the MEDLINE n-gram set; 2) Establishing a system for generating high precision multiword candidates for effective Lexicon building. We believe the MLP/NLP community can benefit from access to these big data (MEDLINE n-gram) sets. We also anticipate an accelerated growth of multiwords in the Lexicon with this system. Ultimately, improvement in recall or precision can be anticipated in NLP projects using the MEDLINE distilled n-gram set, SPECIALIST Lexicon and its applications.

1 INTRODUCTION

This section introduces: first, the SPECIALIST Lexicon; second, the importance of multiwords in NLP; third, the background and purpose of developing a new n-gram-based system for building lexical multiwords.

1.1 The SPECIALIST Lexicon

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 System (McCray et al., 1993). Lexical records are used for part-of-speech (POS) tagging, indexing, information retrieval, concept mapping, etc. in many NLP projects, such as Lexical Tools (McCray et al., 1994), MetaMap (Aronson, 2001; Aronson and Lang, 2010), cTAKES (Savova, 2010), Sophia (Divita et al., 2014), gSpell (Divita et al., 2000), STMT (Lu and Browne, 2012), SemRep, UMLS Metathesaurus, ClinicalTrials.gov, etc. It has been one of the richest and most robust NLP resources for the NLP/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, 2) inflections, and 3) a special unit of lexical meaning by themselves. Linguists in the Lexical Systems Group (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 lexical multiword (LexMultiword or LMW). Single words in the Lexicon have increased 2.6 times from 180,468 in 2002 to 468,655 in 2016. These Lexicon single words cover only about 10.62% of unigrams (single words) from titles and abstracts in

Lu C., Tormey D., McCreedy L. and Browne A.
Generating a Distilled N-Gram Set - Effective Lexical Multiword Building in the SPECIALIST Lexicon.
DOI: 10.5220/0006142000770087
In *Proceedings of the 10th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2017)*, pages 77-87
ISBN: 978-989-758-213-4

MEDLINE.2016. However, single-word Lexicon terms comprise 98.42% 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 446,928 (48.81%) in 2016. 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.

1.2 Multiwords in NLP

Multiwords are vital to the success of high quality NLP applications (Sag et al., 2002; Fraser, 2009). 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) (Frantzi et al., 2000; Green et al., 2013; Ramisch, 2014). 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 NLP applications on 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, POS ambiguity can be reduced through identifying the POS 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 implemented 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.

1.3 MWEs and LMWs

Research on Multiword Expressions (MWEs) has been growing since the late 1990s. State of the art methods including statistical association measures (Silva and Lopes, 1999; Fazly et al., 2009; Pecina, 2010), machine learning (Boukobza and Rappoport, 2009; Tsvetkov and Wintner, 2011; Green et al., 2011), syntactic patterns (Seretan and Wehrli, 2009; Kim and Baldwin, 2010; Green et al., 2013), web queries (Takahashi and Morimoto, 2013), semantic analysis (Pearce, 2001; Baldwin et al., 2003), and a combination of the above methods (Calzolari et al., 2002; Bejček et al., 2013; Sangati and Cranenburgh, 2015) are used in MWE research for acquisition, identification, interpretation, disambiguation and other 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 must have a POS, inflections, and be a unit of meaning. In short, the broader notion of MWEs are distinguished from LWMs in four ways. First, a collocation (an arbitrary statistically significant association between co-occurring items) is not necessarily a LMW because it is not necessarily qualified as a Lexical entry. For example, "*undergoing cardiac surgery*" occurs frequently (3,770 hits in 3,062 documents) in the 2016 MEDLINE n-gram set, but it is not a LMW because it is not functioning as a special unit of meaning by itself. Moreover, this collocation is sometimes, but not always, a single POS. On the other hand, its subterm, "*cardiac surgery*", which occurs frequently (37,171 hits in 22,404 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*" is. Second, verb particle constructions are handled by complementation types (Browne et al., 2000) 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 Lexicon, such as "*kick the bucket*" and "*shoot the breeze*". 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. Fourth, due to the complicated nature of multiwords, much previous MWE research only focuses on bi-grams or tri-grams, which do not meet the requirement of including up to five-grams to reach an estimated recall value of 99.47% of multiwords (Lu et al., 2015).

```
{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".

2 MOTIVATION

Previously, an element word approach (Lu et al., 2014) was used to build the Lexicon by linguists through a web-based computer-aided tool, LexBuild (Lu et al., 2012). Unigrams with high frequency (WC) from the MEDLINE n-gram set that are not in the Lexicon were used as element words for finding new LMW candidates through the Essie search engine (Ide et al., 2007). 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. If we use the mean value, 65%, as an estimated multiword ratio based on the empirical measurement (Ramisch, 2014), it will take more than 21 years for current LSG staff to add all multiwords to the Lexicon by this approach (on average, 20,000 terms are added by LSG staff annually). And this estimate does not account for the fact that many new multiwords are continuously being created by biomedical researchers and English users in general. Thus, we decided to develop a new system to effectively build LMWs to the Lexicon by using an n-gram approach. MEDLINE was chosen as the corpus because it is the biggest and most commonly used resource in the biomedical domain. The MEDLINE n-gram set (MNS) was generated by the following steps: 1) English titles and abstracts from MEDLINE documents are collected and then tokenized to sentences and words (tokens); 2) by requirements, the MNS includes up to 5-grams with information of associated document count (DC) and word count (WC); 3) n-gram and DC|WC are used as key and values in Java HashMap class for n-gram retrieval; 4) due to the large scale, the computer program for retrieving n-grams exceeds the maximum keys in the Java HashMap class (2^{30}-1) when n > 3. Thus, a new model is developed to resolve this issue. This model involves processes of splitting, grouping, filtering, combining and sorting (Lu et al., 2015). The MNS is generated by above processes and has been distributed annually to the public since 2014. The MNS provides comprehensive raw n-gram data from titles and abstracts of MEDLINE. Due to its large-scale size (> 19M n-grams), it is difficult to be handled by computer programs with complicated algorithms. So, a distilled MEDLINE n-gram set (DMNS), with reduced size, higher precision and similar recall in terms of LMWs, is required for the multiword acquisition task and useful for NLP applications.

3 APPROACH - FILTERS

Filters (exclusive filters) are designed to be applied on the MNS to generate the DMNS by trapping invalid LMWs. The design goal of these filters is set to keep the similar (high) recall rate by not trapping valid LMWs. Ideally, all valid multiwords should pass through these filters. The precision of the filtered n-gram set can be improved significantly by applying a series of filters with high recall rate. Exclusive filters are developed based on empirical models with heuristic rules in this task. They are categorized into three types as described below. Patterns and trapped

examples are illustrated for each filter in the format of [pattern] and "*example*" in this paper, respectively.

3.1 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, digits, numbers, spaces and stopwords do not meet the requirement of having a special unit of lexical meaning to themselves. They are used for the general purpose of filtering out invalid 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: "*2000*", "*95%*", "*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 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*".

3.2 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 require sophisticated algorithms.

- Parenthetic Acronym Pattern (PAP) 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 (ACRONYM)]. The expansions of acronyms are usually valid multiwords. A term that contains this pattern is trapped because it contains a potential multiword plus the associated acronym and thus cannot be a valid LMW. Trapped examples include: "*magnetic resonance imaging (MRI)*", "*imaging (MRI)*", "*magnetic resonance (MR) imaging*" and "*(CREB)-binding protein (CBP)*".
- Indefinite Article Filter: A lowercased term that starts with an indefinite article and a space, [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 [CONCLUSION:], [RESULTS:], [OBJECTIVE:], [METHODS:], [MATERIALS AND METHODS:], and [BACKGROUND:]. Trapped examples include "*MATERIALS AND METHODS: The*", "*95% CI:*" and "*PHPT:*"
- 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*" and "*4 years or older with*"), time ("*four months*", "*1 January 1991*", "from May 2002" and "*6 hours plus*"), range ("*2-3 days*" and "*1-2 tablets*"), temperature ("*at -5 degrees*"), dosage ("*10 cigarettes per day*" and "*0.1-2.3 mg/day*") and others ("*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%*".

3.3 Lead-End-Term Exclusive Filters

LMWs do not 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 do not end with words in the above-listed categories. N-grams ending in them are invalid LMWs. They are used in exclusive filters to exclude invalid multiwords. 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). Notably, ILET candidates are considered static because no new terms in the above 7 categories have been added since 2010. Please refer to LSG web documents on Lead-End-Term filter models for details (National Library of Medicine, Lexicon: Lead-End-Terms Model, 2015).

- 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], [from], [is] and [of]. 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 [with], [the] and [that]. Trapped examples include: "*patients with*", "*at the*" and "*suggest that*".
- 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 term matching its spVar pattern in the same corpus is trapped. There are 52 VLTs derived from the Lexicon, such as [to], [as], [for] and [plus]. Trapped examples include: "*to determine*", "*as a result*", "*for example*" and "*plus LHRH-A*".
- End-Term No SpVar Filter: A term that ends with a valid end-term (VET) without any other term matching its spVar pattern in the same corpus is trapped. There are 27 VETs derived from the Lexicon, such as [of], [to], [in] and [more]. Trapped examples include: "*effects of*", "*was used to*", "*(HPV) in"* and "*loss of two or more*".

4 TESTS AND RESULTS

The evaluation of each individual filter, the combination of all filters, and the distilled MEDLINE n-gram set are discussed in this section. The 2016 release of the Lexicon and MEDLINE n-gram set are used in this paper, unless specified otherwise.

4.1 Recall Test of Filters

A recall test model has been established for testing each developed filter individually. Recall is defined as: TP / (TP + FN), where T is true, F is false, P is positive, N is negative. Terms (915,583) in the Lexicon are used to test exclusive filters. All Lexicon terms are valid (relevant) and should pass through filters for preserving high recall rate. In this test, the pass-through terms are counted as TP (retrieved, relevant) while the trapped terms are FN (not retrieved, relevant) for the filtered set.

Columns 4 and 5 in Table 1 list the recall rate and number of trapped terms (FN) for this recall test. The results show that all filters meet the design goal to have very high recall rates. The lowest recall rate (99.9913%) is at filter 15, Lead-Term No SpVar Filter.

4.2 The Distilled N-gram Set

The distilled MEDLINE n-gram set is generated by applying these high recall filters to the MEDLINE n-gram set in the same sequential order of the first column (ID) in Table 1. 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 after ith filter are TP_i and FP_i, respectively, where i = 0, 1, 2, … X. The number of valid LMWs are about the same ($TP_0 \cong TP_1 \cong TP_2 \cong \ldots \cong TP_X$) if high recall filters are used. The number of invalid LMWs is reduced ($FP_0 > FP_1 > FP_2 > \ldots > FP_X$) from the original MEDLINE n-gram set to the final distilled MEDLINE n-gram set after applying filters. Accordingly, the distilled MEDLINE n-gram set (X) has higher precision P_X and similar recall R_X to the MEDLINE n-gram set (0), as shown in equations 1 and 2, respectively, where the number of FN_i (not retrieved, relevant) is a constant.

$$P_X = TP_X / (TP_X + FP_X) \cong TP_0 / (TP_0 + FP_X) > TP_0 / (TP_0 + FP_0) \quad (1)$$

$$R_X = TP_X / (TP_X + FN_X) = TP_X / (TP_X + FN_0) \cong TP_0 / (TP_0 + FN_0) \quad (2)$$

Sixteen high recall rate filters are applied to the MEDLINE n-gram set in the same sequential order as the first column in Table 1 to filter out invalid LMWs. Columns 6, 7 and 8 in Table 1 list the number of trapped terms, the passing rate (PR) and cumulative passing rate (cum. PR) for all filters applied on the MEDLINE n-gram set. The passing rate of the ith filter is the pass through terms/total terms when applying the ith filter on the MNS individually. The pass through terms equals the total terms minus the trapped terms. The cum. PR of ith filter is the cumulative passing rate after applying i filters in the sequential order of the first column in Table 1 to the MNS. In other words, the trapped number is the sum of trapped terms by filters that apply before the ith filter. As a result (i = 16), the distilled MEDLINE n-gram set, after filtering out the majority (11,922,490) of invalid LMWs by these 16 filters, contains about 38.31% (7,402,848) n-grams of the MEDLINE n-gram set (19,325,338). Figure 2 shows a schematic diagram for generating the distilled MNS by applying these filters on the MNS. These filters are designed to independently trap invalid lexMultiwords, so the order of filter application does not affect the final results. These filters are generic and can be used by different NLP projects if they meet the project requirements. The Lead-End-Term filters (ID: 12-16) have higher efficiency (trapped terms/total terms) by trapping more n-grams in this process while the recall rate is above 99.99%. Theoretically, the distilled MEDLINE n-gram set, preserves valid terms in the MNS and thus has higher precision and similar recall compared to the MNS. The size of DMNS is reduced to 38% of MNS, making it possible for complicated computer programs to work in a reasonable time frame in practice, such as the SpVar Pattern matcher (please see section 5.1).

Table 1: Results of applying exclusive filters on Lexicon recall test and the MEDLINE n-gram set.

ID	Filter Type	Filter Name	Lexicon Recall Test		Applied on the MEDLINE N-gram Set		
			Recall	Trapped	Trapped	PR	Cum. PR
1	General Filters	Pipe	100.0000%	0	7	100.0000%	100.0000%
2		Punctuation or Space	100.0000%	0	425	99.9978%	99.9978%
3		Digit	99.9999%	1	132,650	99.3136%	99.3114%
4		Number	99.9953%	43	4,326	99.9775%	99.2890%
5		Digit and Stopword	99.9991%	8	157,786	99.1777%	98.4725%
6	Pattern Filters	Parenthetic Acronym - (ACR)	100.0000%	0	197,022	98.9647%	97.4530%
7		Indefinite Article	99.9986%	13	344,403	98.1713%	95.6709%
8		UPPERCASE Colon	99.9999%	1	113,936	99.3838%	95.0813%
9		Disallowed Punctuation	99.9986%	13	135,508	99.2625%	94.3801%
10		Measurement	99.9920%	73	336,112	98.1572%	92.6409%
11		Incomplete	100.0000%	0	166,356	99.0708%	91.7801%
12	Lead-End-Term Filters	Absolute Invalid Lead-Term	99.9943%	52	4,712,162	73.4329%	67.3967%
13		Absolute Invalid End-Term	99.9997%	3	2,710,470	79.1897%	53.3713%
14		Lead-End-Term	99.9992%	7	2,687	99.9739%	53.3573%
15		Lead-Term No SpVar	99.9913%	80	1,450,394	85.9342%	45.8522%
16		End-Term No SpVar	99.9968%	29	1,458,246	83.5433%	38.3064%

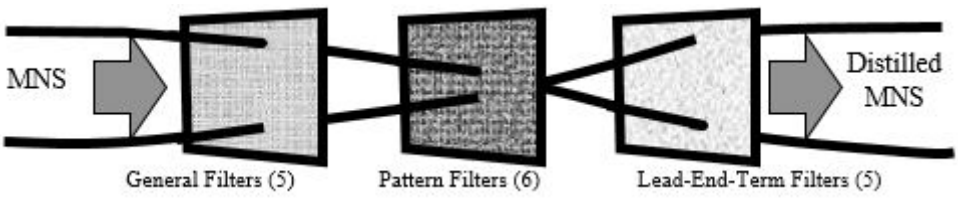

Figure 2: Schematic diagram of the MNS, filters and the distilled MNS.

4.3 Evaluation of DMNS

We further verify the DMNS by comparing the performance of the MNS and the DMNS. A smaller test set is set up by retrieving LMW candidates from the Parenthetic Acronym Pattern (PAP) matcher. 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. On the other hand, the recall of matchers might decrease because not all valid LMWs are trapped.

Acronym expansions are good patterns for a matcher because they have a high possibility of generating valid LMWs. The PAP matcher model is implemented as follows. First, apply Parenthetic Acronym Pattern Filter on the MEDLINE n-gram set to retrieve terms matching the pattern of [acronym expansion (ACRONYM)]. 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 identified as an invalid LMW 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 n-grams of "*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*".

We applied the PAP matcher to the MNS to retrieve LMW candidates. The lowercased core-terms of these candidates are collected as the test set. 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" and "- in details," have the same core-term form of "in details". Core-terms might have punctuation internally, such as "in (5) details". It is a useful normalization to cluster terms with the same core together from the n-gram set in multiword acquisition. As a result, 17,707 LMW candidates are retrieved by this process. They are tagged by LSG linguists and are added to the Lexicon if they are valid LMWs. 15,850 candidates in this set are tagged as valid LMWs to reach 89.51% precision for this PAP matcher, where precision is defined as: TP/(TP+FP), as shown in case 1 in Table 2. The recall cannot be found because all LMWs from MEDLINE cannot be identified in real practice. The result of this PAP matcher is used as the baseline for performance test to compare the results of other filters and matchers. Accordingly, recall in case 1 is set to 1.00 for the purpose of comparison. F1 score is defined as: (2 x precision x recall) / (precision + recall), is calculated and shown in the last column in Table 2.

We repeat the same process by applying the PAP matcher to the DMNS to retrieve LMWs. The results (case 2) show an improvement on F1 score with better precision and almost the same recall. This confirms the theoretic conclusion and the result of the recall 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 38%. Furthermore, the cumulative recall rates of these 16 filters on the recall test (0.9996, multiple product of recall column in table 1) and the recall rate of case 2 in Table 2 (0.9994) are almost identical. This confirms that the approach of applying these filters results in a similarly high recall rate for both the Lexicon and the test set from PAP matcher. Similar results of the Lexicon recall test and DMNS in Table 1 and the performance test of the PAP matcher on the MNS and the DMNS in Table 2 for 3 releases (2014 to 2016) of the Lexicon and MEDLINE are found to confirm the consistency of this approach.

Table 2: Performance comparison of MNS, DMNS and SVP matchers on a test set with 17,707 terms.

Case	Test Case - Model	TP	FP	FN	TN	Precision	Recall	F1
1	PAP matcher on MNS (baseline)	15,850	1,857	0	0	0.8951	(1.0000)	0.9447
2	PAP matcher on DMNS (16 filters)	15,840	1,299	10	558	0.9242	0.9994	0.9603
3	SVP matcher on case 2	8,094	499	7,756	1,358	0.9419	0.5107	0.6623

5 APPLICATIONS ON DMNS

Despite the high precision of the PAP matcher, it only retrieves a small amount of LMW candidates. Other matchers have been developed to retrieve more LMW candidates for Lexicon building.

5.1 Spelling Variant Pattern Matcher

The Spelling Variant Pattern (SVP) matcher model with a complicated algorithm was developed to retrieve large amount of LMW candidates. As we observed, an n-gram is a good LMW candidate if it has spelling variants existing in the same corpus (n-gram set). A sophisticated computer algorithm was developed to identify all n-grams that have potential spVars. First, a special normalization program was 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 identified as spVars to each other. The Lexicon.2015 has 379,269 spVars (including inflectional spelling variants) in 867,728 (unique) inflectional variants, and was used to test this model. As shown in the recall column in Table 3, 80.50% of all spVars in the Lexicon are identified by spVar normalization (step 1). All identified spVars are grouped in spVar classes for further NLP processing. Second, a MES (Metaphone, Edit distance, and Sorted distance) model is developed to improve recall. The MES model is composed of an algorithm of Metaphone phonetic code (Philips, 1990), edit distance (the minimum number of operations required to transform one term into the other), and minimum sorted distance. Sorted distance is the distance between two terms in an alphabetic sorted list of a set of terms. It is used to measure the similarity of two terms compared to other terms in the set. All terms having the same phonetic code and an edit distance (ED) less than a specified value are collected and sorted. The pair with the minimum sorted distance (the closest pair) is identified as spVars to each other. 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 step 1, normalization. They are identified as spVars in step 2 (MES 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 cannot be identified by normalization in step 1. The recall is increased to 97.92% (Table 3). Third, an ES (Edit distance and Sorted distance) model is developed for further improvement of recall. Terms with an edited distance less than a specified value are collected and sorted. The pair with the minimum sorted distance is identified as being spVars. For example, "*zincemia*" and "*zincaemia*" are identified as spVars by the ES 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 "*zincaemia*" and also have different Metaphone codes of [SNSM] and [SNKM], respectively. By relaxing the value of edit distance in both models repeatedly, our program reaches 99.72% recall on spVar identification in six steps in this test, as shown in Table 3. Precision (Prec.), recall, F1, accuracy, and running time (RT) of each step in this SVP matcher model are shown in Table 3, where accuracy is defined as: (TP + TN) / (TP + FP + FN + TN).

For testing purposes, we applied this SVP matcher model to the test set from the PAP matcher (case 2 in Table 2). The results indicate improvement in precision while recall dropped, as shown in case 3 in Table 2. This confirms the design characteristics of matchers.

Table 3: Performance analysis of the SVP matcher model.

Step	Algorithm	ED	TP	FP	FN	TN	Prec.	Recall	F1	Accuracy	RT
1	SpVarNorm	N/A	305,309	3,495	73,960	484,964	0.9887	0.8050	0.8874	0.9107	1 min
2	MES	2	371,385	156,648	7,884	331,811	0.7033	0.9792	0.8187	0.8104	7 hr
3	ES	1	376,646	270,881	2,623	217,578	0.5817	0.9931	0.7336	0.6848	23 hr
4	MES	3	377,004	285,046	2,265	203,413	0.5694	0.9940	0.7241	0.6689	8 min
5	ES	2	378,134	337,461	1,135	150,998	0.5284	0.9970	0.6907	0.6098	26 hr
6	MES	4	378,211	340,105	1,058	148,354	0.5265	0.9972	0.6892	0.6068	2 min

The next step is to apply this SVP matcher model to the MNS to generate LMW candidates from MEDLINE. The running time of this model on the Lexicon took over 56 hours (sum of the RT column in Table 3) even with a powerful computer with 192 GB memory. The running time will be exponentially increased when applying the SVP model on the MNS, which is over 22 times the size of the Lexicon. This is impractical and not feasible in real practice. Thus, the smaller size (38%) DMNS is chosen as input to replace the MNS for reducing the processing time without sacrificing recall. Further purification processes of core-term normalization and frequency threshold restriction (WC > 150) are also applied to reduce the size of the n-gram set for better performance. As a result, 752,920 spVars in 269,871 spVar classes are identified by running this computer program for 20 days and are used for LMWs building in the SPECIALIST Lexicon.

5.2 More Filters and Matchers

Other filters and matchers have also been developed to apply to the DMNS to further improve LMW building. For example, domain filters exclude terms that are in a certain domain, such as single word, frequency, and existing in the current Lexicon.

By requirement, a valid LMW must have a meaning. Thus, a term with valid concept(s) has a better possibility of being a valid LMW. We utilized UMLS Metathesaurus concepts to create one such matcher, the Metathesaurus CUI Pattern (MCP) matcher. The Synonym Mapping Tool (SMT) in STMT (Lu and Browne, 2012) 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. In addition, an End-Word Pattern (EWP) matcher was also developed. In the biomedical domain, multiwords often end with certain words (End-Words), such as [syndrome] (e.g. "*migraine syndrome*", "*contiguous gene syndrome*"), [disease] (e.g. "*Fabry disease*", "*Devic disease*"), and so on. An End-Word candidate list composed of the top 20 frequency End-Words for LMWs has been derived from the Lexicon. These End-Words are used in the EWP matcher to retrieve LMW candidates.

The combining of filters and matchers improves precision. This work focuses on generating high precision LMW candidates for effective LMW building. On the other hand, the recall of the matchers is not emphasized because there are too many multiwords yet to be found.

6 CONCLUSIONS

A set of high recall rate filters has been developed. These filters are used to derive the distilled MEDLINE n-gram set, resulting in reducing its size to 38%, with better precision and similar recall to that of the MEDLINE n-gram set. These filters and the distilled n-gram set have been tested against the Lexicon and a test set of terms retrieved from MNS by PAP matchers. The distilled MEDLINE n-gram set is needed for further NLP processes with complicated algorithms, such as the SVP matcher model, to reduce the running time for retrieving more LMW candidates for Lexicon building.

Other matchers have also been developed and evaluated. Combinations of filters and matchers have been used to generate high precision LMW candidates for effectively building the Lexicon. The LSG plans to continuously enhance and develop 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. The approach of generating the distilled MEDLINE n-gram set is also generic and can be applied to other n-gram sets for reducing size and better precision without sacrificing recall. Most importantly, this approach provides a modular and extendable framework for more and better filters and matchers for LMW acquisition and NLP research.

Multiwords are pervasive, challenging and vital in NLP. The LSG aims to provide a lexicon with high coverage 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 (National Library of Medicine, Lexicon: The MEDLINE n-gram set, 2016) are distributed by the National Library of Medicine (NLM) annually via an Open Source License agreement.

ACKNOWLEDGEMENTS

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

Aronson, A.R., 2001. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. In proceedings of AMIA 2001 Annual Symposium, Wash., DC, USA, Nov. 3-7, pages 17-21.

Aronson, A.R. and Lang, F.M., 2010. An Overview of MetaMap: Historical Perspective and Recent Advances. JAMIA, vol. 17, pages 229-236.

Baldwin, T., Bannard, C., Tanaka, T., Widdows, D., 2003. An Empirical Model of Multiword Expression Decomposability. In proceedings of the ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment, Sapporo, Japan, July 12, pages 89-96.

Bejček, E., Straňák, P., Pecina, P., 2013. Syntactic Identification of Occurrences of Multiword Expressions in Text using a Lexicon with Dependency Structures. In proceedings of the 9th Workshop on Multiword Expressions, Atlanta, Georgia, USA, June 13-14, pages 106–115.

Boukobza, R., Rappoport, A., 2009. Multi-Word Expression Identification Using Sentence Surface Features. In proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, Singapore, August 6-7, pages 468–477.

Browne, A.C., McCray, A.T., Srinivasan, S., 2000. The SPECIALIST LEXICON. Lister Hill National Center for Biomedical Communications, National Library of Medicine, Bethesda, Maryland, USA, June, pages 30-49.

Calzolari, N., Fillmore, C.J., Grishman, R., Ide, N., Lenci, A., MacLeod, C., Zampolli, A., 2002. Towards Best Practice for Multiword Expressions in Computational Lexicon. In proceedings of the Third International Conference on Language Resources and Evaluation (LREC), Las Palmas, Canary Islands, Spain, May 29-31, pages 1934-1940.

Divita, G., Browne, A.C., Tse, T., Cheh, M.L., Loane, R.F., Abramson, M., 2000. A Spelling Suggestion Technique for Terminology Servers. In proceedings of AMIA 2000 Annual Symposium, Los Angeles, CA, USA, Nov. 4-8, page 994.

Divita, G., Zeng, Q.T., Gundlapalli, A.V., Duvall, S., Nebeker, J., and Samore, M.H., 2014. Sophia: An Expedient UMLS Concept Extraction Annotator. In proceedings of AMIA 2014 Annual Symposium, Wash., DC, USA, Nov. 15-19, pages 467-476.

Fazly, A., Cook, P., Stevenson, S., 2009. Unsupervised Type and Token Identification of Idiomatic Expressions. Computational Linguistics, vol. 35, no. 1, pages 61-103.

Frantzi, K., Ananiadou, S., Mima, H., 2000. Automatic Recognition of Multi-Word Terms: the C-value/NC-value Method. International Journal on Digital Libraries, vol. 3, no. 2, pages 115-130.

Fraser, S., 2009. Technical vocabulary and collocational behaviour in a specialised corpus. In proceedings of the British Association for Applied Linguistics (BAAL), Newcastle University, Sep. 3-5, pages 43-48.

Green, S., de Marneffe, M.C., Bauer, J., and Manning, C.D., 2011. Multiword Expression Identification with Tree Substitution Grammars: A Parsing tour deforce with French. In proceedings of EMNLP, Edinburgh, Scotland, UK, July 27-31, pages 725–735.

Green, S., de Marneffe, M.C., Manning, C.D., 2013. Parsing models for identifying multiword expressions. Computational Linguistics. vol. 39, no. 1, pages 195–227.

Ide, N.C., Loane, R.F., Fushman, D.D., 2007. Essie: A Concept-based Search Engine for Structured Biomedical Text. JAMIA, vol. 14, no. 3, May/June, pages 253-263.

Kim, S.N. and Baldwin, T., 2010. How to pick out token instances of English verb-particle constructions. Language Resources and Evaluation, April, vol. 44, no. 1, pages 97-113.

Lu, C.J. and Browne, A.C., 2012. Development of Sub-Term Mapping Tools (STMT). In proceedings of AMIA 2012 Annual Symposium, Chicago, IL, USA, Nov. 3-7, page 1845.

Lu, C.J., McCreedy, L., Tormey, D., and Browne, A.C., 2012. A Systematic Approach for Automatically Generating Derivational Variants in Lexical Tools Based on the SPECIALIST Lexicon. IEEE IT Professional Magazine, May/June, pages 36-42.

Lu, C.J., Tormey, D., McCreedy, L., Browne, A.C., 2014. Using Element Words to Generate (Multi)Words for the SPECIALIST Lexicon. In proceedings of AMIA 2014 Annual Symposium, Wash., DC, USA, Nov. 15-19, page 1499.

Lu, C.J., Tormey, D., McCreedy, L., Browne, A.C., 2015. Generating the MEDLINE N-Gram Set, In proceedings of AMIA 2015 Annual Symposium, San Francisco, CA, USA, Nov. 14-18, page 1569.

McCray, A.T., Aronson, A.R., Browne, A.C., Rindflesch, T.C., Razi, A., Srinivasan, S., 1993. UMLS Knowledge for Biomedical Language Processing. Bull. Medical Library Assoc., vol. 81, no. 2, pages 184-194.

McCray, A.T., Srinivasan, S., Browne, A.C., 1994. Lexical Methods for Managing Variation in Biomedical Terminologies. In proceedings of the 18th Annual Symposium on Computer Applications in Medical Care, pages 235-239.

National Library of Medicine, Lexicon, 2016. Lead-End-Terms Model. Available from: <https://lsg3.nlm.nih.gov/LexSysGroup/Projects/lexicon/current/docs/designDoc/UDF/multiwords/leadEndTerms/index.html>.

National Library of Medicine. Lexicon, 2016. The MEDLINE n-gram set. Available from: <https://lsg3.nlm.nih.gov/LexSysGroup/Projects/nGram/index.html>.

Pearce, D., 2001. Using Conceptual Similarity for Collocation Extraction. In proceedings of the 4th UK Special Interest Group for Computational Linguistics (CLUK4), University of Sheffield, Sheffield, UK, January 10-11, pages 34–42.

Pecina, P., 2010. Lexical association measures collocation extraction. Language Resources and Evaluation, vol. 44, pages 137-158.

Philips, L., 1990. Hanging on the Metaphone. Computer Language, December, vol. 17, no. 12, pages 39-43.

Ramisch, C., 2014. Multiword Expressions Acquisition: A Generic and Open Framework (Theory and Applications of Natural Language Processing). Springer, 2015th Edition, pages 4, 9, 37.

Sag, I., Baldwin, T., Bond, F., Copestake, A., and Flickinger, D., 2002. Multiword expressions: A pain in the neck for NLP. In proceedings of the 3rd International Conference on Intelligent Text Processing and Computational Linguistics (CICLING 2002), Mexico City, Mexico, pages 1-15.

Sangati, F., Cranenburgh, A.V., 2015. Multiword Expression Identification with Recurring Tree Fragments and Association Measures. In proceedings of Annual conference of North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), Denver, Colorado, May 31-June 5, pages 10-8.

Savova, G.K., Masanz, J.J., Ogren, P.V., Zheng, J., Sohn, S., Kipper-Schuler, K.C., Chute, C.G., 2010. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. JAMIA, vol. 17, no. 5, pages 507-513.

Seretan, V. and Wehrli, E., 2009. Multilingual collocation extraction with a syntactic parser. Language Resources and Evaluation, March, vol. 43, no. 1, pages 71-85.

Silva, J.F. and Lopes, G.P., 1999. A Local Maxima method and a Fair Dispersion Normalization for extracting multi-word units from corpora. In proceedings of the Sixth Meeting on Mathematics of Language (MOL6), Orlando, FL, USA, pages 369-381.

Takahashi, S. and Morimoto, T., 2013. Selection of Multi-Word Expressions from Web N-gram Corpus for Speech Recognition. In proceedings of International Symposium on Natural Language Processing (SNLP), Phuket, Thailand, Oct. 28-30, pages 6-11.

Tsvetkov, Y. and Wintner, S., 2011. Identification of Multi-word Expressions by Combining Multiple Linguistic Information Sources. In proceedings of the Conference on Empirical Methods in Natural Language Processing, Edinburgh, Scotland, UK, July 27-31, pages 836–845.