The Bare Necessities: Increasing Lexical Coverage for Multi-Word Domain Terms with Less Lexical Data

Branimir Boguraev, Esme Manandise, Benjamin Segal
IBM Thomas J. Watson Research Center
1101 Kitchawan Road, Yorktown Heights, NY 10598, USA.
{bran, esme, bpsegal}@us.ibm.com

Abstract

We argue that many multi-word domain terms are not (and should not be regarded as) strictly atomic, especially from a parser's point of view. We introduce the notion of Lexical Kernel Units (LKUs), and discuss some of their essential properties. LKUs are building blocks for lexicalizations of domain concepts, and as such, can be used for compositional derivation of an open-ended set of domain terms. Benefits from such an approach include reduction in size of the domain lexicon, improved coverage for domain terms, and improved accuracy for parsing.

1 Introduction

Knowledge about collocations and multi-word expressions (MWEs) can be beneficial for parsing, ultimately improving a parser's accuracy (Nivre and Nilsson, 2004; Korkontzelos and Manandhar, 2010; Wehrli, 2014). Typically such knowledge is made present by treating collocations and MWEs as single lexical and syntactic units (Baldwin and Kim, 2010; Escartín et al., 2013; Fotopoulou et al., 2014). This practice is also reflected in domain adaptation, where domain-specific lexicons hold collocations and MWEs with 'domain terms' status.

In the medical domain, terminological and lexical resources list collocations and MWEs as varied as *history of trauma to toes of both feet* and *morning after pill* as single "words with spaces" (Sag et al., 2002). As mandated by the lexicon-parser interface, such domain terms parse as single lexical units, which improves parser performance by reducing the lexical, structural, and distributional complexity of these noun phrases. This simplification is intuitively appealing. However, when closely-related, or similar, multi-word domain terms such as *day after pill* or *history of trauma to toes of left foot* are unlisted in the terminology lexica, the potential for parse error resurfaces. Relying on explicitly listed terms alone compromises parser accuracy.

We present here an approach to lexicon enrichment, which mitigates the inherent incompleteness of such lists, inevitably arising during processes of populating domain term banks. In our work on extracting domain-specific terms from a medical terminology resource,[1] we observe certain compositional properties of a large subset of such domain-specific terms.[2] In particular, this subset is open-ended: through generative patterns, even if some such domain terms are not in the lexicon, a mechanism can be construed which can license them as terms (virtual entries in the lexicon). These patterns operate on smaller expressions, which exhibit a much more atomic status than the terms proper, and enable—through compositionality—the dynamic generation and interpretation of the longer domain terms. Such smaller expressions we call *lex-*

[1] Proper domain multi-word terms are derived from the Unified Medical Language System (UMLS) (NIH, 2009) knowledge bases (KBs), which contain medical concepts, relations, and definitions, spread over millions of concepts and terms from over 160 source vocabularies. Not all entries in UMLS qualify for 'term' status; term extraction proper is, however, outside of the scope of this paper. The UMLS-derived terminology lexicon—close to 56 million tokens comprising over 8 million terms—is the source data of our analysis.

[2] We focus on noun phrases of varying structural complexity.

Proceedings of NAACL-HLT 2015, pages 60–64,
Denver, Colorado, May 31 – June 5, 2015.

ical kernel units (LKUs).

For example, in the set of medical domain terms *history of spastic paraplegia, spastic paraplegia with retinal degeneration*, and *family history of spastic paraplegia with Kallmann's syndrome*, we see repeated patterns of behavior of the same multi-word expression, *spastic paraplegia*: it can be governed by *history of*; it co-occurs with the preposition *with*; an instance of *history of* is pre-modified by the noun *family*.

Spastic paraplegia is a lexical kernel unit. Regarding it as a 'kernel' for an open-ended set of expressions like the ones above—and deploying appropriate generative patterns and devices—we argue that a newly-encountered word grouping like *family spastic paraplegia with neuropathy* can be licensed as a domain-specific term, available to a parser, even if the term is a virtual one, absent from the static domain-dependent terminology lexicon.

This paper discusses the nature and some practical consequences of LKUs. Given that ours is very much work in progress, the intent is to hint at an algorithmic procedure for the identification and extraction of LKUs from an externally provided terminology lexicon. Additionally, the paper aims to show the ability afforded by LKUs to transform a finite, static, lexicon of domain collocations and multi-word expressions to an open-ended, dynamic (or virtual) lexicon which can better support parsing. While not in a position to present a formal evaluation of the benefit of an LKU lexicon, we offer examples of how such a lexicon benefits a parser.

2 Mining LKUs from terminology lexica

The essence of what makes lexical kernel units atomic can be illustrated by an analysis of sample subsets[3] of domain terms from which LKU status for certain word sequences can be inferred.

Consider the subset of term entries containing (not necessarily consecutively) the words in the multi-word expression *spastic paraplegia*:

a. *spastic paraplegia syndrome*,
b. *spastic congenital paraplegia*,
c. *infantile spastic paraplegia*,
d. *familial spastic paraplegia with Kallmann's syndrome*,
e. *familial spastic paraplegia with neuropathy and poikiloderma*,
f. *familial spastic paraplegia, mental retardation, and precocious puberty*,
g. *slowly progressive spastic paraplegia*,
h. *hereditary x linked recessive spastic paraplegia*,
i. *onset in first year of life of spastic paraplegia*.

Spastic and *paraplegia* appear in domain terms of varying length and with different noun phrase structures; additionally, the two words may or may not be adjacent. In the entries d.–f., *spastic paraplegia* shares the adjective *familial* on its left; but it can also co-occur with other adjectives as pre-modifiers, *infantile, hereditary,* and *progressive* among them (b.–c. and g.–h.) Further, the phrases to the right of *spastic paraplegia* in entries d.–f. are of different phrase types. For instance, in entries d.–e., *spastic paraplegia,* is immediately adjacent to the preposition *with*.

Looking at the examples together, it is intuitively clear that variability around (an LKU) phrase exists across all the elements of the term subset; furthermore, this variability can be captured by a relatively small number of patterns.

To reinforce the confidence with which *spastic paraplegia* can be putatively assigned lexical kernel unit status, these patterns can be put to the test by a broader search, against the terminology lexicon. A pattern like `[LKU [with NP]]` (see Section 3) inspired by the domain entries d. and e., can be tested with the query string *spastic paraplegia with*.

Such search returns, among many, the domain terms *spastic paraplegia with amyotrophy of distal muscle wasting, spastic paraplegia with mental handicap, spastic paraplegia with mental retardation,* and *spastic paraplegia with amyotrophy of hands and feet* (although the terminology lexicon does not list either *spastic paraplegia with amyotrophy of hands* or *spastic paraplegia with amyotrophy of feet* as domain terms).

As another example, of an LKU with different profile and distributional properties, consider the *[noun]-of* collocation instantiated by *history of.* In the terminology lexicon, there are 7,087 domain terms with the anchor *history of.* A few are:

a. *current social history of patient*,
b. *current history of allergies*,

[3] We will not discuss here the process of deriving such subsets from the terminology lexicon.

c. *family history of alcohol abuse*,
d. *history of current illness*,
e. *history of domestic violence at home*,
f. *history of falling into a swimming pool*,
g. *history of freckles*,
h. *past medical history of drug abuse*,
i. *personal history of alcohol abuse*,
j. *past personal history of allergy to other anti-infective agents*.

A cursory analysis of the semantic and syntactic composition of the above domain terms reveals that they are unexceptional (even though statistically salient in the domain). Looking at all examples together, an important question to consider is whether the variability in the terms is expressed by the contexts around *history of* or around *history* alone. A search keyed off the word *history* and based on a pattern with prepositional placeholder `[history [prep]]` returns, among many, the domain terms *history in family of hypertension, medical history relating to child,* and *current history with assessment of changing moles.* Clearly, in addition to *of*, *history* sanctions prepositional collocations *in, relating to,* and *with.* This is supporting evidence that LKU status can be attributed to *history*; it is also indicative of the kind of lexical (collocational) knowledge that needs to be associated with the LKU entry for *history*.

3 Capturing the essence of domain terms

The examples above suggest that *spastic paraplegia* and *history* function as building blocks from which an open-ended set of larger domain terms can be compositionally built, and interpreted. The many multi-word domain terms found in the terminology lexicon that contain *spastic paraplegia* and *history* can be informally represented with the following patterns:

a. `[[adjective* and/or noun*] spastic paraplegia [with [noun]]]`
b. `[[adjective* and/or noun*] history [[in | of | with] [noun]]]`

These capture the essence of multi-word expressions and collocations that can have many domain term instantiations—including ones beyond the closed sets which prompted the patterns (Section 2). The free slots, `noun` and `adjective`, must be filled by collocations with the appropriate part of speech, some of which can be LKUs themselves.

The many variations—potentially an open-ended set—of domain terms are thus collapsed into a single pattern, anchored by a putative LKU, and augmented with linguistic and usage information (part-of-speech, semantic types, collocation preferences, etc...) extracted from the terminology lexicon.

It may be tempting to collapse the patterns, and seek generalizations covering sets of LKUs: replacing the kernel units *spastic paraplegia* and *history* with a place-holder would have pattern a. to be subsumed by pattern b. This would be counterproductive, however: it would allow for overgeneration, as well as fail to distinguish between frame-specific lexical knowledge to be associated with the individual LKUs (e.g. we would not want *spastic paraplegia* to allow for the full set of prepositional complements compatible with *history*).

The lexical knowledge discovered during this LKU extraction and captured in the domain terms patterns eventually ends up in LKU entries. For instance, from the patterns above, the preposition collocations would induce appropriately specified lexical frames. These would allow for uniform treatment, by a parser, of similar noun phrases—even if some of them lack 'domain term' status: e.g. both *spastic paraplegia with retinal degeneration* (a term, and therefore a single syntactic unit) and *spastic paraplegia with no retinal degeneration* (not designated a term, but inferred as such), would keep the *with-* PP attached to *spastic paraplegia*.

4 Some characteristics of LKUs

Lexical kernel units can be single- or multi-word sequences, as exemplified by the earlier analyses of *spastic paraplegia* and *history*. The degree to which LKUs by themselves are representative of a domain varies. However, what is more important is that through composition, they combine with other words or LKUs to construct larger, domain-specific, terms (consider, for example, *history of spastic paraplegia*). It is through analysis of such terms that an LKU lexicon is compiled.

Multi-word LKUs tend to be invariable and function as domain-specific, atomic, language units. A large subset of such LKUs have some of the linguis-

tic features of MWEs. Two characteristics are particularly descriptive.

First, LKUs can display various degrees of semantic and syntactic opaqueness (e.g. *popcorn lung* or *airway morbidity*), as well as transparency (*small intestine* or *airway passage*).

Second, substitutability of a word within the LKU word sequence by another of the same or similar category may be barred. *Popcorn lung* and *popcorn disease symptom* cannot be substituted by **maize lung* or **edible corn disease symptom.*

As atomic units at the kernel of larger, compositionally built domain terms, it is much more revealing to look at what determines the exocentric pull (or valency) of LKUs, than analyzing their internal structure. LKUs determine the range and type of the larger phrases that can be construed around them.

While they serve as atoms for the creation of novel, longer domain expressions which can reflect a more general property of grammar as in *popcorn lung disease symptom* and *airway morbidity disorder*, the pool of words which LKUs can use to create longer domain units can be small and is highly domain-specific. Many collocations and novel creations are constrained by the semantics of the domain. In the medical and clinical domains, we do not see **popcorn lung morbidity, *popcorn lung rehabilitation,* or **popcorn lung remission.*

Finally, they need not operate in text as standalone words. For instance, the LKU *Silver Russell* does not function in domain texts as an individual noun compound. *Silver Russell* only functions as an LKU in longer domain terms as in *Silver Russell dwarfism* or *Silver Russell syndrome.*

5 Parsing with LKUs

The LKU notion allows for the creation of a domain-specific lexicon with a minimal number of entries that describe the nature of a given domain. As we saw in Section 2, there are thousands of domain terms anchored by collocations of the LKU *history* with prepositions *of, in,* or *with,* with variations both to the left (*family history of ..., medical history of ...,* and so forth) and right (*history of panic disorder ..., history of falling into ..., history of drug and alcohol abuse ...,* and so forth) of the anchor. Even so, it is unrealistic to expect that *all* instances of similar terms can be discovered for capture in a terminology lexicon. We also saw, however, that very simple patterns can be very expressive. Leveraging the contextual information captured in, for instance, pattern (b.; Section 3), as part of the lexical representation of the LKU entry for *history*, makes such discovery unnecessary, even for terms as complex in structure as the examples above.

When a lexical kernel unit becomes a part of the domain-dependent lexicon, none of the terms which were analyzed to derive it needs to be listed in that lexicon. Thus the 7,087 domain terms anchored by *history+of* (Section 2) can be replaced by a single, one-token, LKU entry (*history*) in the domain lexicon. This same entry would also account for the extra domain terms anchored by *history+in* and *history+with*.

While not in any way a formal evaluation, a preliminary, small scale experiment to determine impact of LKUs on parser[4] performance shows improvements, in particular in the area of coordination (itself a long-standing challenge to parsing). We created two domain lexicons (DLs): DL1 included all well-formed terms from the terminology lexicon with the words *history* and *spastic paraplegia,*; DL2 listed *history* and *spastic paraplegia* as LKU entries, while it eliminated the 7,000+ domain terms from the lexicon. Randomly extracted segments from medical corpus were parsed, in alternative regimes,[5] with DL1, and then with DL2.

Consider the segment *Bupropion has two absolute clinical contraindications (i.e., current or past history of seizures).* DL1 contains an entry for *past history of seizures* (but not one for *current history of seizures*). The parse derived with DL1 is wrong: *current* gets a 'noun' analysis, coordinated with the noun phrase *past history of seizures*. The correct analysis—a coordinated node joining *current* and *past*, and pre-modifying *history*—is achieved, however, with DL2, whose atomic *history* LKU allows a granular structured interpretation of what DL1 declares to be a single multi-word unit.

Another example illustrates the benefits of capturing the word-specific collocations within the repre-

[4]We use the English Slot Grammar (ESG) parser (McCord, 1990; McCord et al., 2012).

[5]We skip over how the parser interprets LKU entries, dynamically creating virtual domain terms anchored by the LKUs.

sentation of an LKU. In the DL1 parse of segment *Familial hereditary spastic paraplegias (paralyses) are a group of single-gene disorders*, the adjective *familial* pre-modifies both the noun *hereditary spastic paraplegia* (listed as a term in DL1) and the material in parentheses.[6] With the LKU-enabled DL2, ESG is instructed by the lexical information associated with the LKU *spastic paraplegia* (pattern (a.); Section 3) to treat both *familial* and *hereditary* as sister pre-modifiers to *spastic paraplegia* in particular.

6 Conclusion

Lexical kernel units give an embodiment to an intuition concerning the compositional aspects of domain terms in a conventional terminology lexicon. To the best of our knowledge, no attempts have been made to question the 'term entries are atomic' assumption.

We propose a view where lexical kernel units provide a more uniform partitioning of a terminology lexicon, teasing out its prominent lexical collocations. Once captured into an LKU lexicon, lexical kernel units allow for a granular view into that domain; this, in itself, is beneficial to a parser. Also, by virtue of being relevant to domain concepts, they allow for a degree of open-endedness of such a lexicon: in effect, they underpin a compositional mechanism to domain term identification and interpretation. Thanks to a pattern-driven generative device, instead of parsing with a fixed size terminology lexicon, we leverage a process aiming to license domain terms 'on demand'.

Pilot experiments to date show that LKUs have a positive impact on parsing. Future work will articulate an algorithm and heuristics for identifying and extracting LKUs from terminological lexica and other resources. In particular, we will address the questions of generating the sets of terms indicative of LKUs, abstracting the pattern specifications for LKU-to-term derivations, and deriving fully instantiated (canonical) LKU lexicon entries. We will also conduct an extensive contrastive evaluation of LKU-based parsing of medical corpora.

[6] ESG analyzes most parenthetical, appositive, constructions as coordinations around the opening parenthesis.

Acknowledgments

We thank three anonymous reviewers for in-depth, and helpful, comments.

References

Timothy Baldwin and Su Nam Kim. 2010. Multiword expressions. *Handbook of Natural Language Processing, Second Edition. Morgan and Claypool*.

Carla Escartín, Gyri Losnegaard, Gunn Samdal, and Pedro García. 2013. Representing multiword expressions in lexical and terminological resources: an analysis for natural language processing purposes. In *Electronic lexicography in the 21st century: thinking outside the paper: Proceedings of the eLex 2013 Conference, Tallinn, Estonia*, pages 338–357.

Aggeliki Fotopoulou, Stella Markantonatou, and Voula Giouli. 2014. Encoding MWEs in a conceptual lexicon. *EACL 2014*, page 43.

Ioannis Korkontzelos and Suresh Manandhar. 2010. Can recognising multiword expressions improve shallow parsing? In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL*, pages 636–644. Association for Computational Linguistics.

Michael C. McCord, J. William Murdock, and Branimir Boguraev. 2012. Deep parsing in Watson. *IBM Journal of Research and Development*, 56(3):3.

Michael C. McCord. 1990. Slot grammar: A system for simpler construction of practical natural language grammars. In R. Studer, editor, *Natural Language and Logic: Proc. of the International Scientific Symposium, Hamburg, FRG*, pages 118–145. Springer, Berlin, Heidelberg.

NIH. 2009. Unified Medical Language System (UMLS). `http://www.nlm.nih.gov/research/umls/`, July. US National Library of Medicine, National Institute of Health.

Joakim Nivre and Jens Nilsson. 2004. Multiword units in syntactic parsing. *Proceedings of Methodologies and Evaluation of Multiword Units in Real-World Applications (MEMURA)*.

Ivan A Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword expressions: A pain in the neck for NLP. In *Computational Linguistics and Intelligent Text Processing*, pages 1–15. Springer.

Eric Wehrli. 2014. The relevance of collocations for parsing. *EACL 2014*, page 26.