Journal of Biomedical Informatics 46 (2013) 212–227

Contents lists available at SciVerse ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Methodological Review

Approaches to verb subcategorization for biomedicine

Thomas Lippincott [a,*], Laura Rimell [a], Karin Verspoor [b], Anna Korhonen [a]

[a] Computer Laboratory, University of Cambridge, 15 JJ Thomson Avenue, Cambridge CB3 0FD, UK
[b] National ICT Australia, Victoria Research Lab, Melbourne VIC 3010, Australia

ARTICLE INFO

Article history:
Received 14 March 2012
Accepted 6 December 2012
Available online 28 December 2012

Keywords:
Verb subcategorization
Lexical resources
Natural language processing
Biomedical text processing

ABSTRACT

Information about verb subcategorization frames (SCFs) is important to many tasks in natural language processing (NLP) and, in turn, text mining. Biomedicine has a need for high-quality SCF lexicons to support the extraction of information from the biomedical literature, which helps biologists to take advantage of the latest biomedical knowledge despite the overwhelming growth of that literature. Unfortunately, techniques for creating such resources for biomedical text are relatively undeveloped compared to general language. This paper serves as an introduction to subcategorization and existing approaches to acquisition, and provides motivation for developing techniques that address issues particularly important to biomedical NLP. First, we give the traditional linguistic definition of subcategorization, along with several related concepts. Second, we describe approaches to learning SCF lexicons from large data sets for general and biomedical domains. Third, we consider the crucial issue of linguistic variation between biomedical fields (subdomain variation). We demonstrate significant variation among subdomains, and find the variation does not simply follow patterns of general lexical variation. Finally, we note several requirements for future research in biomedical SCF lexicon acquisition: a high-quality gold standard, investigation of different definitions of subcategorization, and minimally-supervised methods that can learn subdomain-specific lexical usage without the need for extensive manual work.

1. Introduction

Text mining of the biomedical literature has an ever-increasing importance in biomedical informatics and systems biology due to the double-exponential growth in research publications in the biomedical domain [1,2]. Natural language processing (NLP) involves development of computational algorithms for analysis of natural language; it is essential for managing such vast amounts of unstructured text, and facilitates access to information and data extraction that would be intractable as a manual task. A number of core NLP technologies used in biomedical informatics could benefit from knowledge of *verb subcategorization*, i.e. the tendency of verbs to "select" the syntactic phrase types they co-occur with: for example, the fact that the verb *decrease* can be intransitive (*The contribution decreased*), while *compare* cannot (*We compared the predictions*, but not simply *We compared*). Technologies such as syntactic and semantic parsing, event identification, relation extraction, and entailment detection all have the potential to make use of subcategorization information to improve the reliability of linguistic analyses of text, and ultimately the correctness of extracted information derived from natural language texts, by improving the identification of participants associated with the events named by verbs in the text. For example, Refs. [3,4] used *subcategorization frames* (SCFs) in event extraction from UKPubMedCentral documents.

Manually constructing subcategorization resources is an expensive and time-consuming task, and those resources may fail to translate when applied to new language domains. It is therefore important to explore data-driven approaches that require less supervision and can be rapidly deployed for arbitrary text. While automatic subcategorization acquisition techniques are relatively well-developed for general English text, and several SCF lexicons have been produced [5–7], there are few comparable techniques or resources for biomedicine. Studies of the lexical characteristics of text such as word and part-of-speech frequencies have shown substantial variation, both between general and biomedical text and across subdomains of biomedicine [8,9] Table 1 illustrates this phenomenon with sentences from Education and Embryology using the verb "develop" and two simplified SCFs, the transitive and intransitive.

It has not been determined how much variation exists in subcategorization behavior, or whether this variation follows the same patterns as other lexical variation.

This paper has two goals. The first is to provide the necessary background for future work on SCF acquisition in biomedical NLP. To this end we present the traditional definition of subcategorization, and describe the typical state-of-the-art approach to SCF

* Corresponding author.
E-mail addresses: Thomas.Lippincott@cl.cam.ac.uk (T. Lippincott), Laura.Rimell@cl.cam.ac.uk (L. Rimell), karin.verspoor@nicta.com.au (K. Verspoor), Anna.Korhonen@cl.cam.ac.uk (A. Korhonen).

1532-0464/$ - see front matter © 2012 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.jbi.2012.12.001

Table 1
Example sentences for the verb "develop" in the Education and Embryology subdomains, illustrating how verb behavior can dramatically shift. In this case, the transitive usage has the highest frequency in Education (as in most subdomains), and the intransitive is far less frequent. In Embryology, the opposite is the case.

Subdomain	Frame	Frequency	Example
Education	Transitive	0.33	We developed a questionnaire to measure knowledge and attitudes
	Intransitive	0.12	Training programmes to support doctors in these summative assessments are developing simultaneously
Embryology	Intransitive	0.51	How the complex TM develops and how spaces form in the initially continuous cellular tissue is not clear
	Transitive	0.12	VK developed a concept of the project and wrote the manuscript

Table 2
Sample SCFs for *decrease* and *compare*. Note that *compare* does not occur as an intransitive, represented by the asterisk. All examples adapted from the PubMed Open Access (PMC OA) [11] corpus.

SCF	Example
NP	The retraction screw and blade decreased [$_{NP}$ the risks of vessel injuries]
NP-PP	Heterozygosity for twine also decreases [$_{NP}$ the frequency of precocious NEB] [$_{PP}$ to less than 10%]
ø	The contribution of cardiovascular diseases as cause of death decreased
NP	We compared [$_{NP}$ the performance of the Charlson and the Elixhauser comorbidity measures]
NP-PP	We compared [$_{NP}$ the predictions] [$_{PP}$ to the known interaction signs]
∗ ø	∗ We compared

Table 3
Sample SCFs with examples from the PMC OA corpus.

SCF	Example
NP-AS-NP	Perception of complex stimuli occurs too rapidly to support rate coding as a reliable mechanism
NP-TOBE	The larger, unsaturated propyne group has been shown to be a useful modification for antisense oligonucleotides
PP-PP	Threshold values ranged from 0.01 to 0.99
THAT-S	Experiments with PTEN-null PGCs in culture revealed that these cells had greater proliferative capacity
TO-INF	Administration of DA agonists to the rat PFC acts to enhance working memory in these animals

ADJUNCT:
... and operated [$_{PP}$ on a pre-warmed operation table] ...
ARGUMENT:
HW provided clinical care for this patient and operated [$_{PP}$ on the patient].

Fig. 1. Example adjunct and argument PPs from the PMC OA corpus for the verb *operate*.

```
(NP
    (NP (VBG mutating) (NN serine) (NNS 209))
    (PP (IN on)
      (NP (NN mouse) (NNS Wnt3a))))
```

Fig. 2. Parse fragment from running the Stanford Parser on example sentence (3): "mutating serine 209" has been incorrectly labeled as a single noun phrase.

acquisition, with examples from general and biomedical language. The second goal is to determine the degree of variation in SCF behavior within biomedicine, which could have major implications for the success of the approach.

2. Background

In this section we present a basic introduction to verb subcategorization, which will be required as background for the rest of the paper. We then describe the typical interpretation of subcategorization in biomedical text, and how subcategorization information can improve NLP and text mining applications in biomedicine.

2.1. Introduction to verb subcategorization

The traditional linguistic notion of subcategorization refers to the syntactic arguments of a verb, that is, the syntactic phrase types which occur obligatorily or with high probability for any given verb. Some common syntactic phrase types which can serve as arguments to a verb include noun phrases, prepositional phrases, subordinate clauses, adjectives and adverbs.

Some basic examples of subcategorization frames (SCFs) can be seen in Table 2. For the SCF names we use COMLEX Syntax notation [10], which includes an abbreviation for each phrase type in the SCF. Thus the SCF for a transitive verb (taking one direct object noun phrase) is NP, and for a verb taking a direct object and a prepositional phrase NP-PP.[1] Most verbs take several SCFs. In Table 2, it can be seen that *decrease* may occur with the following SCFs: NP, NP-PP, or ø (intransitive). On the other hand, *compare* occurs with the NP and NP-PP but not as an intransitive. In addition to presence and absence, SCF frames occur with different verb-specific frequencies.

Additional examples of SCFs are shown in Table 3. Here the COMLEX SCF names include mnemonics for some additional information beyond the simple phrasal types. For example, the frame NP-AS-NP is a subclass of NP-PP, where the preposition is lexicalized as *as*. The frame NP-TOBE represents a direct object and a predicate using *to be*. The frame THAT-S represents a sentential complement introduced by the complementizer *that*, and TO-INF is an infinitival complement that uses the *to* form of the verb in the lower clause.

Comparing SCFs to another argument structure representation sometimes used in biomedicine, SCFs are more general than

[1] Note that we do not specify the subject NP as part of the SCF, since subjects are obligatory in English.

Predicate-Argument Structures (PASs), which have been used in Semantic Role Labeling [12–14]. PASs include very specific per-verb roles such as, for the verb *delete*, "entity doing the removing", "thing being removed", and "removed from". SCFs also do not identify thematic roles such as Agent and Patient nor functional roles such as Subject and Object (though these types of roles can often be inferred from the SCF), but simply the syntactic phrase types that are selected by the verb (NP, PP, etc.). SCFs thus provide a basic level of argument structure information which can aid in event identification, but are general enough to be automatically acquired for a large number of verbs, compared to PASs which must be defined on a per-verb basis and thus can only practically be identified for a small number of very frequent biomedical verbs.

An important concept for subcategorization is that of the *argument-adjunct* distinction, with the linguistic notion of subcategorization – and the one typically used in general language – involving only arguments. The hallmark of a syntactic *argument* is that it is obligatory or very strongly selected by the verb.[2] Arguments are distinguished from *adjuncts*, which are phrases that elaborate on an event and are generally optional. This distinction is often relevant for classifying prepositional phrases. In particular, PPs describing location, manner, or time tend to be adjuncts.

In Fig. 1, the PP *on a pre-warmed operation table* is optional, elaborating on the event description by describing the location at which it took place. The PP *on the patient* is obligatory and exhibits a special, idiomatic meaning in the context of the verb *operate*. The argument-adjunct distinction is sometimes fuzzy, because the judgement of optionality can be difficult to make, especially when a phrase type occurs with high frequency for a given verb. However, Fig. 1 illustrates another criterion, namely that the meaning of arguments often depends on the particular verb, while adjuncts maintain their interpretation (e.g. locative, temporal, manner) across a wide variety of verbal heads [15,16]. See [17,18] for computational approaches to distinguishing arguments and adjuncts.

2.2. Verb subcategorization variation in biomedicine

Traditional linguistics has long recognized that language in a specialized area like biomedicine behaves differently than general language [19], and scientific languages in particular are known to vary at the syntactic level [20]. In biomedicine, subcategorization is often defined more broadly than for general English, to include adjuncts that are less strongly selected but nevertheless important for the complete description of an event, from the point of view of information extraction. Cohen et al. [21] state that "knowledge representation in this [biomedical] domain requires that we *not* make a distinction between adjuncts and core arguments". The use of a more semantic criterion for distinguishing arguments and adjuncts in biomedicine has become common. A common implementation is to relax the definition of "argument" from *obligatory* to *high probability*, e.g. using log-likelihoods [22]. The semantic definition then corresponds to a lower threshold for acceptance.

Within a PAS annotation scheme, for example, [12] includes the location PP in sentence (1) and the manner adverb in sentence (2) as core arguments, neither of which would be considered arguments in general language.

(1) Apparently HeLa cells either initiate transcription at multiple sites within RPS14 exon 1 ... [12].
(2) Mice have previously been shown to develop normally ... [12].

[2] However, most verbs take multiple SCFs which may involve different obligatory arguments. Therefore, the argument is properly considered to be obligatory with regard to the verb–SCF pair, not just the verb.

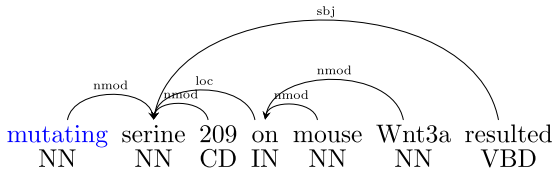

Fig. 3. Parse fragment from running the ClearParser on example sentence (3): "mutating serine 209" has again been incorrectly labeled as a single noun phrase, with "serine" as the sentence's subject.

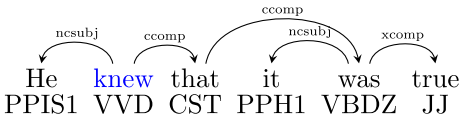

Fig. 4. Example RASP output for the sentence "He knew that it was true." Each arc represents a grammatical relation between two words. This is an example of the verb "know" taking a sentential complement frame (THAT-S).

```
((|ncsubj| ?x ?y _) (|ccomp| ?a ?x ?v) (|ncsubj| ?v ?n _))
     (and (word-value (quote ?a) "that")
          (strict '?x '?patterns '?grs)
          (pos-start (quote ?x) "VV")
          (pos-start (quote ?v) "V"))
     (?x THAT-S))
```

Fig. 5. Cambridge frame rules use a special notation to describe the grammatical relations and part-of-speech tags of arguments that must be present.

Note that even under the broader definition, not every phrase type that co-occurs with the verb is an argument; [12] still consider aspectual or frequency adverbs such as *still* or *always* to be adjuncts.

As Cohen et al. note, the tradeoff of this more semantic definition is a loss of some ability to generalize about adjuncts across verbs, but they argue that this loss is outweighed by the "biological integrity in the knowledge representation". This translates to improvements in the ability of biomedical NLP systems to extract relations and events that reflect biological intuitions about those relations and events.

Consider the use of the verb *mutating* in sentence (3). Parsing this sentence with the Stanford Parser [23] online tool results in the syntactic structure for the subject shown in Fig. 2, with a flat compound noun phrase for *mutating serine 209* and the attachment of the prepositional phrase to that noun phrase. This fails to capture the structure of the mutation event correctly, where *serine 209* is the mutated residue and *mouse Wnt3a* the location of the mutation (the mutated gene). The ClearParser dependency parser [24] result in Fig. 3 makes arguably an even poorer analysis, with *serine* serving as the subject of the sentence, and *mutating* and the prepositional phrase analysed as modifiers. Knowing that the verb "mutate" takes SCF NP-PP with a certain frequency might help these parsers more correctly treat these two elements as arguments of the verb. This in turn will lead to an accurate extraction of the *mutation* event.

(3) Second, mutating serine 209 on mouse Wnt3a resulted in a loss of Wnt3a secretion [PMC2427328].

There have been several studies confirming the importance of relaxing the argument-adjunct distinction in biomedicine, and of using SCF lexicons that adopt this alternative definition for biomedical NLP. For example, a study of biomedical information extraction by [3,4] found that 9.7% of verb arguments in their gold standard were correctly detected in prepositional phrases using a biomedical SCF lexicon, and would have been missed entirely based on the

Table 4
Common subdomain clusters when considering lexical features.

Microscopic		System-specific	Clinical	Social
Cellular	Biochemical			
Cell Biology	Biochemistry	Endocrinology	Geriatrics	Ethics
Virology	Molecular Biology	Rheumatology	Pediatrics	Education
Microbiology	Genetics	Pulmonary Medicine	Psychiatry	
Embryology			Obstetrics	

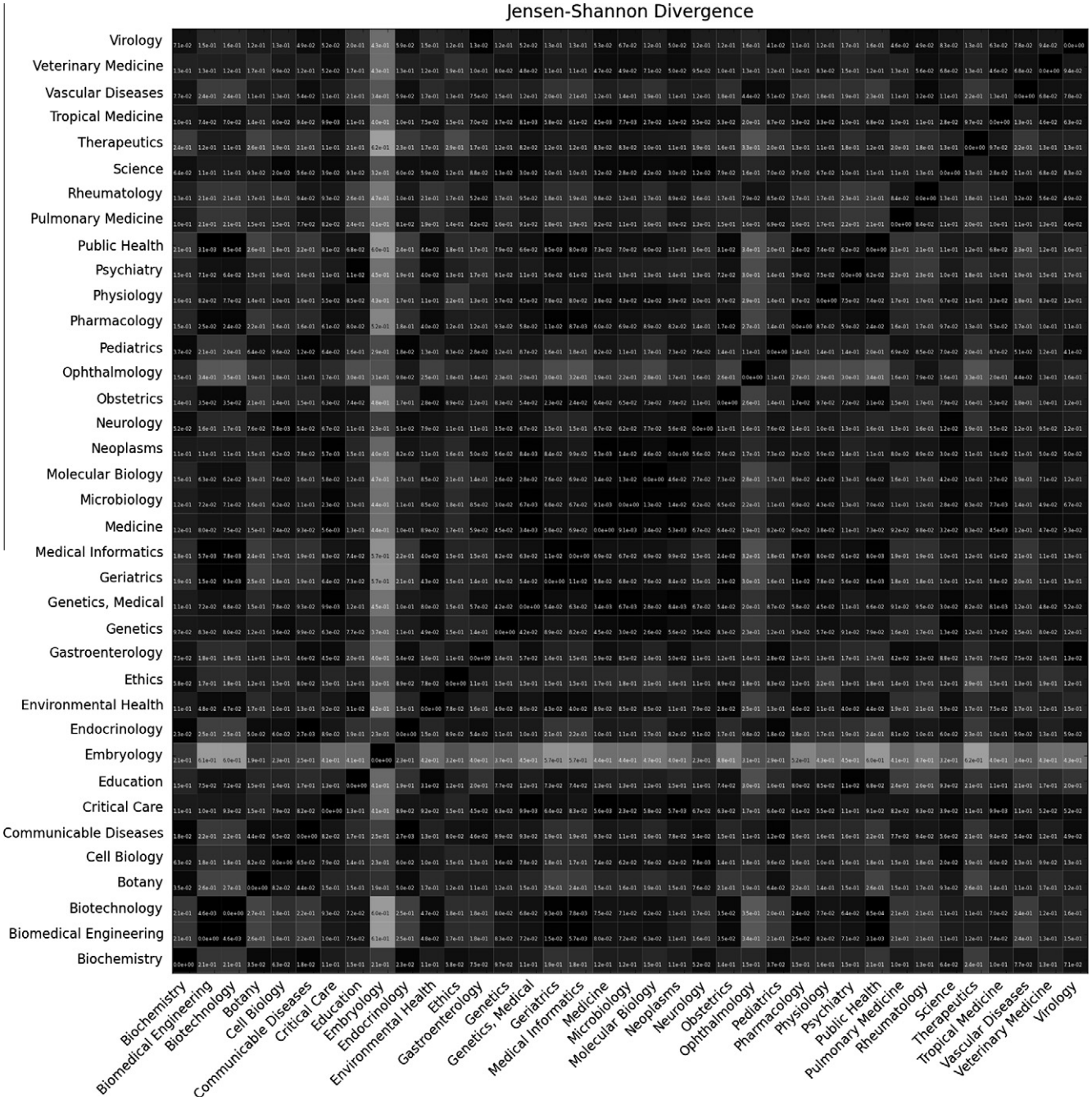

Fig. 6. Heat map of Jensen–Shannon divergence between subdomains for the SCF distributions of *develop*.

parser output alone. Despite these observations, there has been no comprehensive study to date of how specialized definitions of subcategorization, like that used for biomedical text, interact with gold standards annotated using the general language definition.

3. Verb subcategorization frame lexicons

In this section we describe existing SCF resources for general language and biomedicine. In Section 3.1, we describe existing

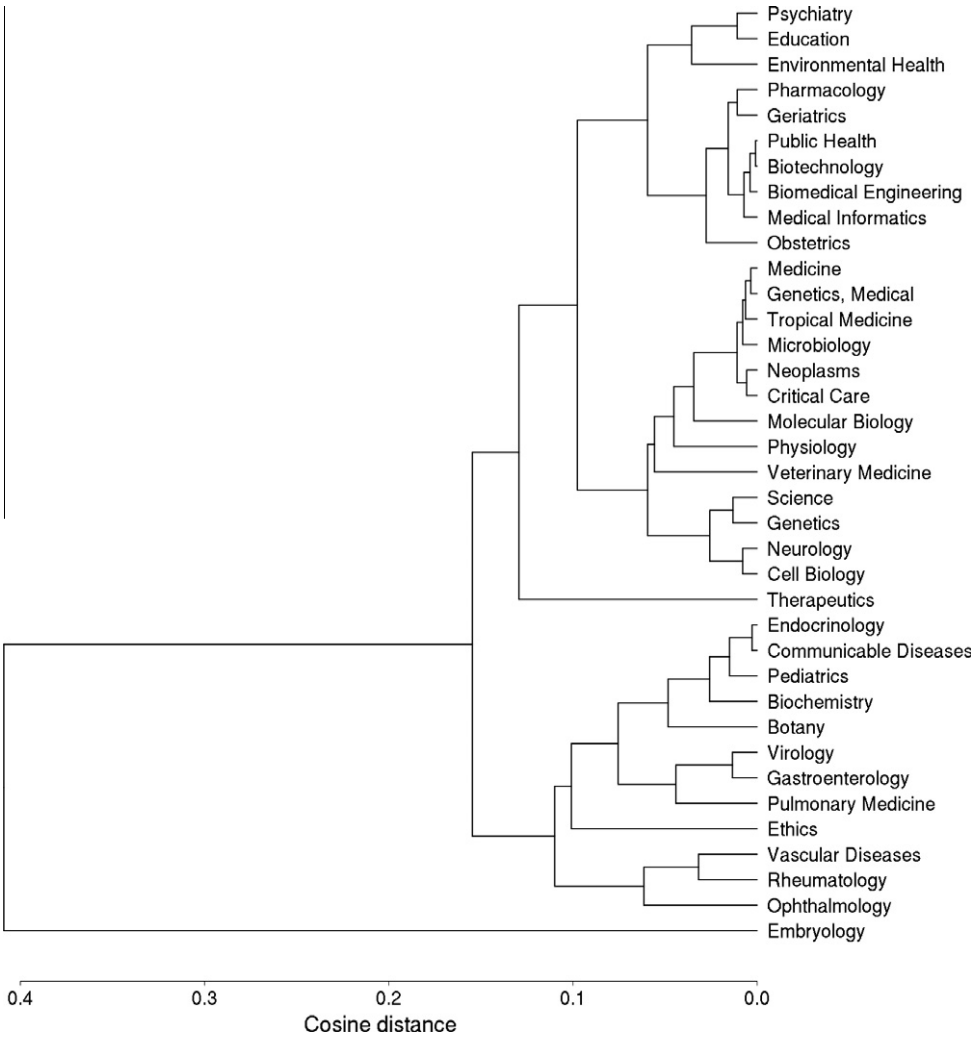

Fig. 7. Hierarchical clustering of subdomains via average-linking for the SCF distributions of *develop*.

lexicons and acquisition methodologies used for general language, and present one state-of-the-art system in detail. None of these resources, however, address the specific needs of verbs in the biomedical domain. It has been demonstrated that biological language can be construed as a sublanguage of general language, and further that alternations in the argument structure of verbs and their nominalizations are both common and diverse [25]. There is therefore a need for lexical resources specific to this sublanguage and we will introduce a few such resources in Section 3.2. As we have suggested, automated SCF acquisition methodologies provide the most resource efficient strategies for creation of these domain-specific resources, and therefore we will describe those methodologies in detail. Finally, we will present the only biomedical-specific SCF acquisition system which exists to date.

3.1. General language SCF resources

3.1.1. Existing lexicons

There are several existing computational verb lexicons that provide syntactic and/or semantic information for general language. For example, the COMLEX lexicon [10] provides subcategorization information for c. 6000 general language verbs. FrameNet [26] and VerbNet [27] provide both syntactic and semantic information about predicate argument structure for c. 3000 and c. 4000 verbs, respectively. PropBank [28] is an extension of the Penn TreeBank [29] with information about predicate-argument relationships for c. 5600 verbs.

The VALEX [6] verb lexicon is the largest SCF resource available for general language. It contains SCF and frequency information for c. 6400 verbs learned from up to 10,000 sentences per verb. In contrast to the aforementioned resources, VALEX is built automatically from large amounts of data, rather than via manual annotation. Automatic SCF acquisition has an advantage over manual SCF lexicon development in terms of significantly lower resource requirements, i.e. time and human effort, and since it is empirically based, allows domain- or genre-specific lexicons to be developed more straightforwardly.

3.1.2. Acquisition methodology and the Cambridge system

Automatic SCF acquisition systems typically consist of two major components: hypothesis generation and hypothesis selection. As a pre-processing step, a corpus of text is processed with a natural language parser to produce a syntactic analysis for each sentence. The hypothesis generator uses the parser output to decide which SCF is taken by each verb in each sentence. These hypotheses are then amalgamated into a lexicon, which consists of each verb occurring in the corpus with its relative frequencies for each SCF.

The larger the corpus, the more likely it is that the lexicon will capture a comprehensive set of SCFs for each verb. However, the output of the hypothesis generation step is typically noisy, due to the difficulty of the task (e.g. parsing errors). Thus a filtering step

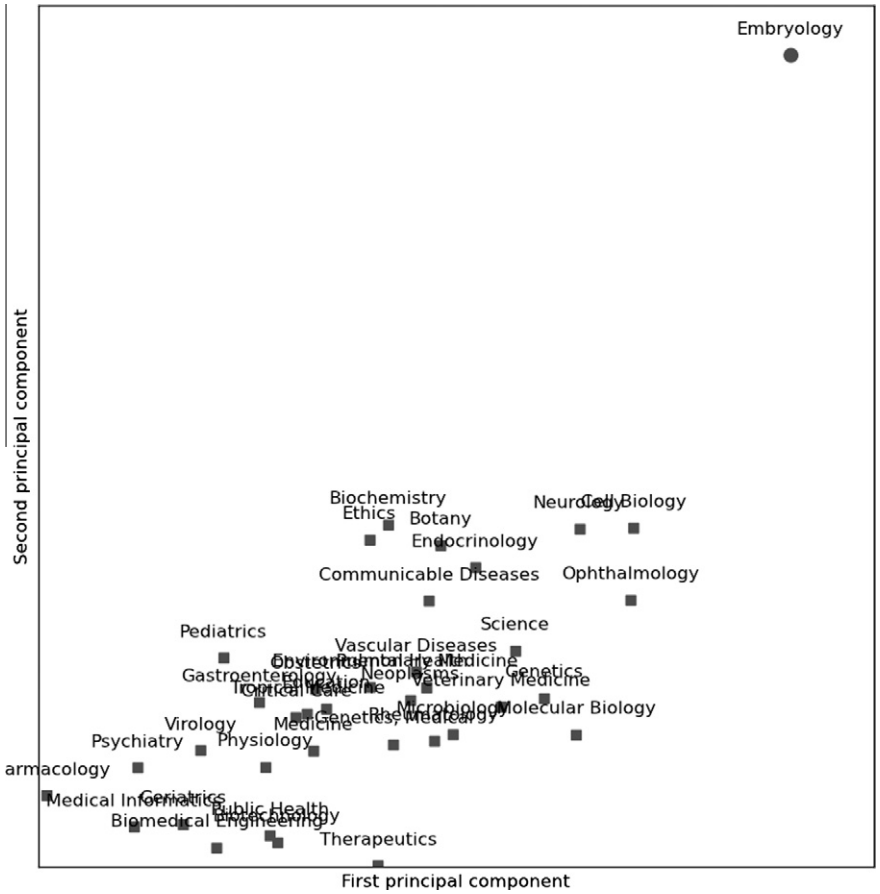

Fig. 8. Two-dimensional PCA reduction with Gap-statistic-optimal clustering for the SCF distributions of *develop*.

is required to select from among the hypotheses those that are most reliable. Filtering is a challenging task, since some SCFs are inherently rare; infrequent attestation does not always mean an SCF should be filtered out of the lexicon. Ideally the filtering process does not make use of lexical information such as verb semantic classes or SCF dictionaries, as this introduces a circular dependency, although such resources are routinely used in real-world systems.

Within these broad outlines, approaches vary along several dimensions; see [30] for an overview. Hypothesis generation may involve a shallow parser/chunker that simply groups adjacent words into abstract phrase-types (e.g. noun phrases) or a deep grammatical parser that fully specifies the sentence's hierarchical structure. The SCF inventory may be manually defined, in which case the task of hypothesis generation involves matching the syntactic analyses to the pre-defined SCFs; or the SCF inventory may be learned directly from the corpus. The size of SCF inventories can vary widely between systems, from only a few to some two hundred SCFs, although more recent state of the art systems for general language tend to use relatively large inventories. There are a number of mechanisms for generating hypotheses, as well, using a variety of cues in the parsed text to identify the SCFs.

There are several SCF acquisition systems for English as well as other languages [31–34]. These typically rely on some form of parsed input and language-specific knowledge, either directly through heuristics, or indirectly through parsing models trained on treebanks. Furthermore, some require labeled training instances for supervised [35] or semi-supervised [34] learning algorithms.

We now describe an example of an SCF acquisition system for general language: the state-of-the-art system used to produce the VALEX lexicon, hereafter referred to as the *Cambridge system*. The Cambridge system operates on output from the RASP parsing suite [36]. RASP is a modular statistical parsing suite which includes a tokenizer that splits a sentence into tokens, a tagger that associates each word form with a part of speech tag based on its context and internal features, a lemmatizer that reduces each token to a canonical form, and a wide-coverage unification-based tag-sequence parser that assigns a tree structure to the sentence where nodes correspond to words and edges correspond to dependency relations. The parser is unlexicalized, which means it considers a sentence's sequence of part-of-speech tags (and not the words themselves). It therefore cannot learn verb-specific behavior (like SCFs) and bias the system towards a pre-existing notion of subcategorization. The parser's output is a dependency tree of *grammatical relations*. Fig. 4 shows the tree structure assigned to the sentence "He knew that it was true."

The Cambridge system defines an SCF inventory of 163 frames. Each frame is specified in terms of the grammatical relations connecting the verb to its arguments, the POS tags of the arguments, and some basic lexical information. Continuing with the example sentence from Figs. 4, 5 shows the definition of the sentential complement frame that would match its dependency tree. It specifies that the lexical item *x* takes SCF THAT-S if (1) it is a verb, (2) it is the head in subject and complement relations, and (3) the dependent of the complement relation is also a verb with a subject.

Verb instances are thus matched to SCFs, and aggregated into preliminary lexical entries for each verb, containing the raw and relative frequencies of SCFs. Finally, these entries are filtered to obtain a more accurate lexicon. The most basic approach simply removes verb–SCF pairs with a relative frequency less than a given threshold: previous work has found a threshold of 0.02 to produce optimal results.

Table 5
Top three SCFs, by subdomain, for *develop*.

Subdomain	Top three SCFs					
Psychiatry	NP	0.399905	NP-PRED-RS	0.141902	NP-FOR-NP	0.137602
Education	NP	0.328025	NP-FOR-NP	0.140127	INTRANS	0.121019
Environmental Health	NP	0.309671	INTRANS	0.138097	NP-FOR-NP	0.128797
Pharmacology	NP	0.441249	NP-FOR-NP	0.118324	NP-PRED-RS	0.115859
Geriatrics	NP	0.390192	NP-PRED-RS	0.140725	NP-FOR-NP	0.115139
Public Health	NP	0.361242	NP-FOR-NP	0.158063	NP-PP-PRED	0.101749
Biotechnology	NP	0.356888	NP-FOR-NP	0.173096	NP-PRED-RS	0.098217
Biomedical Engineering	NP	0.385159	NP-FOR-NP	0.169611	NP-PP-PRED	0.111307
Medical Informatics	NP	0.410649	NP-FOR-NP	0.168911	NP-PP-PRED	0.083231
Obstetrics	NP	0.315455	INTRANS	0.152435	NP-PRED-RS	0.120678
Medicine	NP	0.345473	NP-PRED-RS	0.137849	NP-PP-PRED	0.091899
Genetics, Medical	NP	0.303856	NP-PRED-RS	0.143445	NP-FOR-NP	0.114139
Tropical Medicine	NP	0.345211	INTRANS	0.116705	NP-PRED-RS	0.114743
Microbiology	NP	0.293089	NP-FOR-NP	0.127123	NP-PRED-RS	0.095342
Neoplasms	NP	0.304064	NP-PRED-RS	0.147233	NP-PP-PRED	0.099857
Critical Care	NP	0.340197	NP-PRED-RS	0.182325	INTRANS	0.099528
Molecular Biology	NP	0.245846	NP-FOR-NP	0.156345	NP-PP-PRED	0.100831
Physiology	NP	0.366467	NP-FOR-NP	0.131138	NP-PRED-RS	0.100599
Veterinary Medicine	NP	0.287117	NP-PRED-RS	0.117791	INTRANS	0.099387
Science	NP	0.263721	INTRANS	0.128445	NP-PP-PRED	0.109314
Genetics	NP	0.261829	NP-FOR-NP	0.142401	INTRANS	0.107713
Neurology	NP	0.231093	INTRANS	0.207683	NP-PP-PRED	0.103842
Cell Biology	NP	0.223591	INTRANS	0.200704	PP-PRED-RS	0.084507
Therapeutics	NP	0.350314	NP-FOR-NP	0.155172	NP-PRED-RS	0.101097
Endocrinology	NP	0.273525	INTRANS	0.161085	NP-PRED-RS	0.137959
Communicable Diseases	NP	0.287262	NP-PRED-RS	0.149480	INTRANS	0.144714
Pediatrics	NP	0.361596	NP-PRED-RS	0.194514	INTRANS	0.124688
Biochemistry	NP	0.285505	INTRANS	0.231332	NP-FOR-NP	0.120059
Botany	NP	0.281346	INTRANS	0.189602	NP-FOR-NP	0.128440
Virology	NP	0.379412	NP-PRED-RS	0.136275	NP-FOR-NP	0.109804
Gastroenterology	NP	0.334848	NP-PRED-RS	0.210606	NP-PP-PRED	0.127273
Pulmonary Medicine	NP	0.300429	NP-PRED-RS	0.158798	NP-PP-PRED	0.115880
Ethics	NP	0.274298	INTRANS	0.228942	NP-PP-PRED	0.155508
Vascular Diseases	NP	0.318367	NP-PRED-RS	0.155102	INTRANS	0.101224
Rheumatology	NP	0.306562	NP-PRED-RS	0.159647	NP-PP-PRED	0.119491
Ophthalmology	NP	0.245421	NP-PRED-RS	0.146520	INTRANS	0.124542
Embryology	INTRANS	0.510504	INTRANS-RECIPSUBJ-PL	0.172269	NP	0.120798

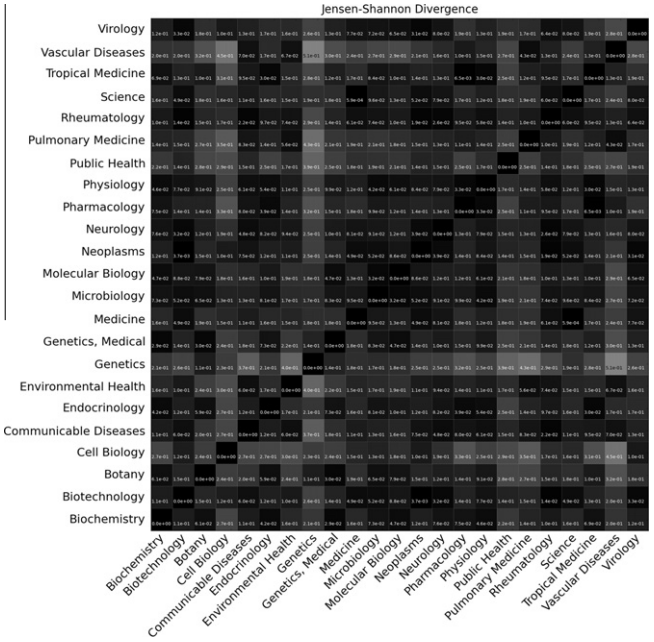

Fig. 9. Heat map of Jensen–Shannon divergence between subdomains for the SCF distributions of *express*.

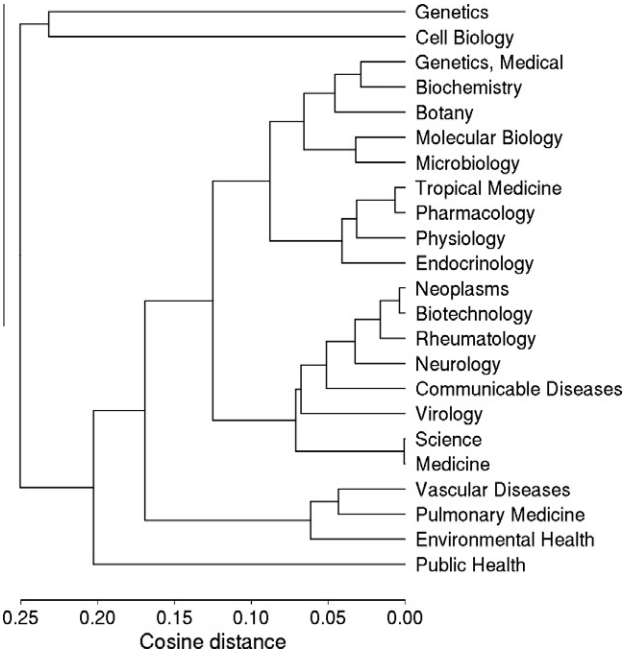

Fig. 10. Hierarchical clustering of subdomains via average-linking for the SCF distributions of *express*.

This method has several drawbacks. First, frame definitions as in Fig. 5 must be manually written and maintained: not only is this difficult work, it also ties the definitions to particular formalisms, such as the POS and grammatical relation inventories. It also precludes the question of whether a different inventory might be

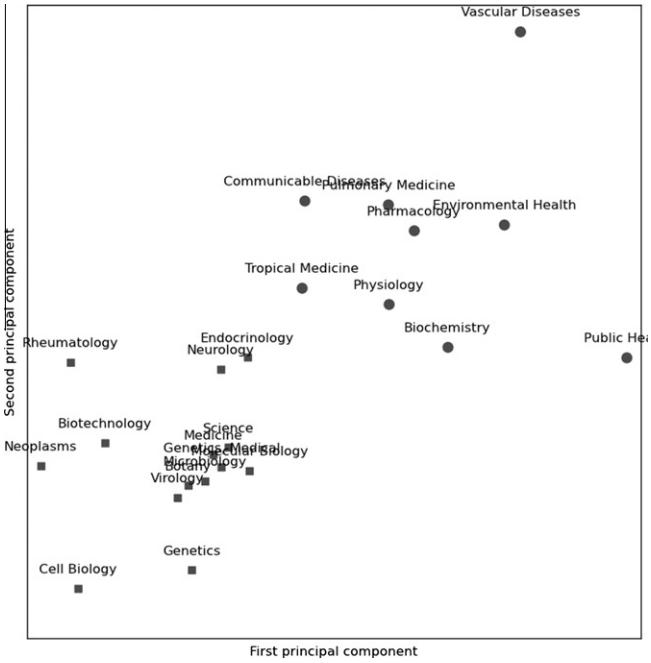

Fig. 11. Two-dimensional PCA reduction with Gap-statistic-optimal clustering for the SCF distributions of *express*.

more suitable for specialized language. Second, the method is sensitive to parsing errors, which are known to increase when dealing with biomedical text [37]. Finally, there has been no evaluation so far of how the method performs in biomedicine.

3.2. Biomedicine-specific SCF resources

3.2.1. Existing lexicons

A small number of verb lexicons already exist for biomedicine. BioFrameNet [38] extends FrameNet with links to biomedical resources (e.g. gene ontologies) for verb frames related to intracellular transport. The UMLS SPECIALIST Lexicon [39] includes coarse verb subcategorization information for some 11,000 verbs, but is manually built from a variety of biomedical and general language dictionaries. BioProp [13] adds PropBank-style annotation to 500 abstracts from the GENIA corpus. PASBio [12] is an inventory of predicate-argument structure frames for 30 verbs, focused on molecular biology. The frames were constructed through expert examination of MEDLINE sentences, using guidelines similar to those of PropBank. The resource most relevant to this study is the BioLexicon [22], which includes semi-automatically acquired verb subcategorization information for 658 verbs.

3.2.2. Acquisition methodology and the Biolexicon system

When producing an SCF lexicon for a specialized domain, there are three typical approaches. First, a manual approach where linguists and domain experts produce a lexicon via introspection and/or annotation of data. Second, an automatic approach where a system designed for general language, such as the Cambridge system, is simply applied to the specialized domain. Third, an automatic approach using a system that utilizes components designed for the specialized domain, or some approximation to it.

We now describe the only existing system specifically for automatic SCF acquisition in biomedicine, that was used to produce the BioLexicon [22] (hereafter referred to as the *BioLexicon system*). Where the Cambridge system uses the unlexicalized general-language RASP parser, the BioLexicon system uses a version of the lexicalized Enju parser [40] that has been trained on the GENIA treebank of molecular biology abstracts as described in [41]. Like the Cambridge system, the BioLexicon system considers a verb's grammatical relations to indicate its frame, but no SCF inventory is assumed in advance; rather, the set of grammatical relations for each verb instance are considered as a potential SCF. These are filtered at a relative frequency threshold of 0.03, i.e. for any given verb, all SCFs with a relative frequency less than 0.03 are discarded. To produce the lexicon, this procedure is run over six million words of MEDLINE E. Coli abstracts and articles, leading to an inventory of 136 SCFs. Further arguments and strongly-selected adjuncts are chosen according to their log-likelihood with respect to the verb.

It is important to note that the BioLexicon system draws on a single subdomain of biomedical literature, and uses manually-annotated training data that would be expensive to produce for new subdomains. Moreover, the parsing model used in SCF discovery is lexicalized and therefore adapted to the subcategorization phenomena present in the training data. While there are immediate benefits to these approaches in terms of accuracy in SCF acquisition

Table 6
Top three SCFs, by subdomain, for *express*.

Subdomain	Top three SCFs					
Genetics	NP	0.484719	NP-PP-PRED	0.088202	NP-PRED-RS	0.077303
Cell Biology	NP	0.436123	NP-PRED-RS	0.256388	NP-PP-PRED	0.183260
Genetics, Medical	NP	0.445434	NP-PRED-RS	0.084633	NP-PP-PRED	0.082405
Biochemistry	NP	0.320611	NP-PRED-RS	0.122137	NP-AS-NP-SC	0.113700
Botany	NP	0.457393	NP-PRED-RS	0.107769	PP-PRED-RS	0.084586
Molecular Biology	NP	0.401806	NP-PP-PRED	0.151806	NP-PRED-RS	0.125282
Microbiology	NP	0.393716	NP-PRED-RS	0.192811	NP-PP-PRED	0.152821
Tropical Medicine	NP	0.362590	NP-AS-NP-SC	0.152518	NP-AS-NP	0.152518
Pharmacology	NP	0.300459	NP-AS-NP-SC	0.181193	NP-AS-NP	0.181193
Physiology	NP	0.320866	NP-AS-NP-SC	0.140748	NP-AS-NP	0.140748
Endocrinology	NP	0.389426	NP-PRED-RS	0.131325	NP-AS-NP-SC	0.117112
Neoplasms	NP	0.439103	NP-PP-PRED	0.200038	NP-PRED-RS	0.171003
Biotechnology	NP	0.416469	NP-PRED-RS	0.182106	NP-PP-PRED	0.165479
Rheumatology	NP	0.435431	NP-PRED-RS	0.136413	NP-PP-PRED	0.132412
Neurology	NP	0.384721	NP-PP-PRED	0.137646	NP-PRED-RS	0.135582
Communicable Diseases	NP	0.336735	NP-AS-NP-SC	0.204082	NP-AS-NP	0.204082
Virology	NP	0.388041	NP-PRED-RS	0.227216	NP-PP-PRED	0.185567
Science	NP	0.392503	NP-PRED-RS	0.172770	NP-PP-PRED	0.138302
Medicine	NP	0.396785	NP-PRED-RS	0.167203	NP-PP-PRED	0.154984
Vascular Diseases	NP-AS-NP-SC	0.281022	NP-AS-NP	0.281022	NP	0.253650
Pulmonary Medicine	NP	0.328225	NP-AS-NP-SC	0.186462	NP-AS-NP	0.186462
Environmental Health	NP	0.281679	NP-AS-NP-SC	0.167877	NP-AS-NP	0.167877
Public Health	NP	0.266667	NP-PP-PRED	0.183333	NP-PP	0.126190

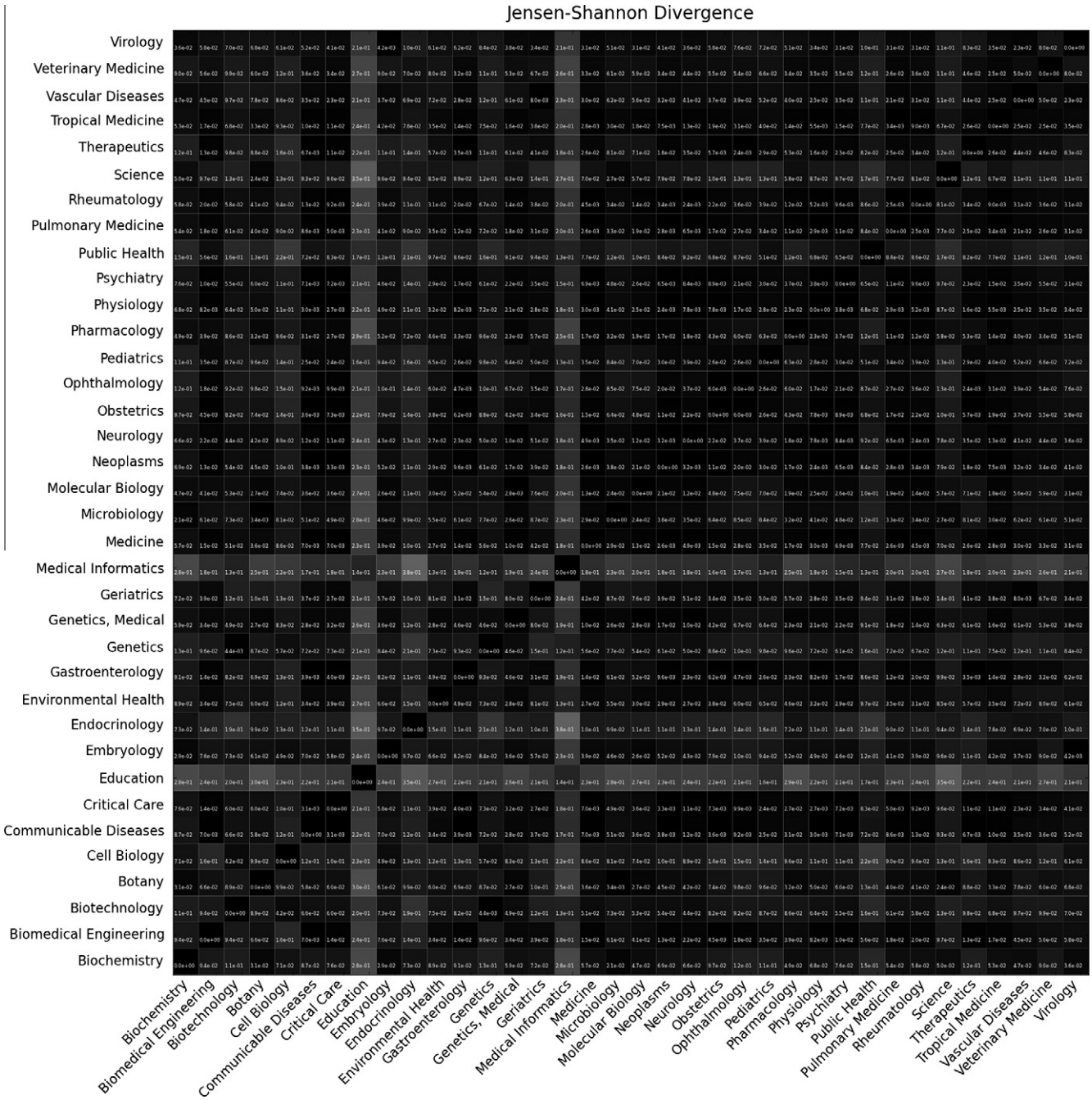

Fig. 12. Heat map of Jensen–Shannon divergence between subdomains for the SCF distributions of *perform*.

within the same subdomain as the training data, the model's reliance on manual annotation is costly, and its preconception of subcategorization may introduce bias against new subdomain behaviors. Finally, since the resources used to build and evaluate the BioLexicon system are drawn from a subset of the biomedical literature, there has been no study of how it performs on a broader range of subdomains.

4. Investigation of subdomain variation

4.1. Motivation

Both approaches we have described, applying a general language system or exploiting domain-adapted resources, potentially suffer from the effects of subdomain variation. The Cambridge and Biolexicon systems exemplify this: the former because its components are trained on general language, and the latter because its parser is tuned on a small subset of biomedical text, and applied to abstracts regarding a single organism. While we presently lack a gold standard for measuring absolute performance of these systems on biomedical text, we can consider the question of how much subdomains of biomedicine vary in SCF behavior. If this variation is high, it implies that even using adapted resources like the BioLexicon system will lead to problems when applied to subdomains that it was not trained on. The infeasibility of creating manual resources for each biomedical subdomain would then require less supervised approaches.

4.2. Subdomain variation methods

This section describes our approach to quantifying differences in verb subcategorization behavior across subdomains of biomedicine. The primary type of data that we investigate is a verb's *SCF distribution*, that is, the probability distribution representing the relative frequency of the verb appearing with a given SCF. Our goal is to discover the presence or absence of significant differences between a verb's SCF distribution in different subdomains. By investigating whether individual verbs exhibit specialized behavior across subdomains, we build up an overall picture of subdomain variation in verb subcategorization.

4.2.1. Data and SCF extraction

To obtain the SCF distributions we use one of the general language systems, namely the Cambridge system, because it is unbiased with respect to a given subdomain of biomedicine. The PubMedCorpus Open Access subset (PMC OA) includes a classification of journals by subdomain. We apply the Cambridge system to the 37 largest subdomains, which produces an SCF distribution for each combination of verb and subdomain.

4.2.2. Measuring divergence

To measure the distance between two SCF distributions we use the Jensen–Shannon divergence (JSD) [42], a finite and symmetric measurement of divergence between probability distributions, defined as:

$$JSD = H(X+Y) - H(X) - H(Y)$$

where H is the Shannon entropy of a distribution

$$-\sum_{x} x \log x$$

JSD values range between 0 (identical distributions) and 1 (disjoint distributions), and is closely related to the familiar, but asymmetric, Kullback–Leibler divergence [43]. We calculate the JSD between a given verb's SCF distributions for each pair of subdomains.

4.2.3. Presentation

We applied this methodology to 30 verbs, and present detailed results for six: *develop*, *express*, *perform*, *predict*, *recognize* and *treat*. These verbs were chosen because they exemplify one or more interesting properties, such as sharp divergence in a single subdomain or a wide variety of behaviors across all subdomains. For a given verb, we only show subdomains in which it occurs a minimum of 200 times. For each of the six verbs we present four different views of the data.

Heat maps present pairwise calculations of a metric between a set of objects: cell $\langle x,y \rangle$ is shaded according to the value of *metric*(x,y). Our heat maps show the JSD values between pairs of subdomains for a given verb: the cells are shaded from white (JSD value of 1, maximum divergence) to black (JSD value of 0, identity). The actual values are inscribed in each cell.

Dendrograms present the results of hierarchical clustering performed directly on the JSD values. The algorithm begins with each instance (in our case, subdomains) as a singleton cluster, and repeatedly joins the two most similar clusters until all the data is clustered together. The order of these merges is recorded as a tree structure that can be visualised as a dendrogram in which the length of a branch represents the distance between its child nodes. Similarity between clusters is calculated using average cosine distance between all members, known as "average linking". The tree leaves represent data instances (subdomains) and the paths between them are proportional to the pairwise distance. This allows visualization of multiple potential clusterings, as well as a more intuitive sense of how distinct the clusters truly are. Rather than choosing a set number of flat clusters, the trees mirror the nested structure of the data.

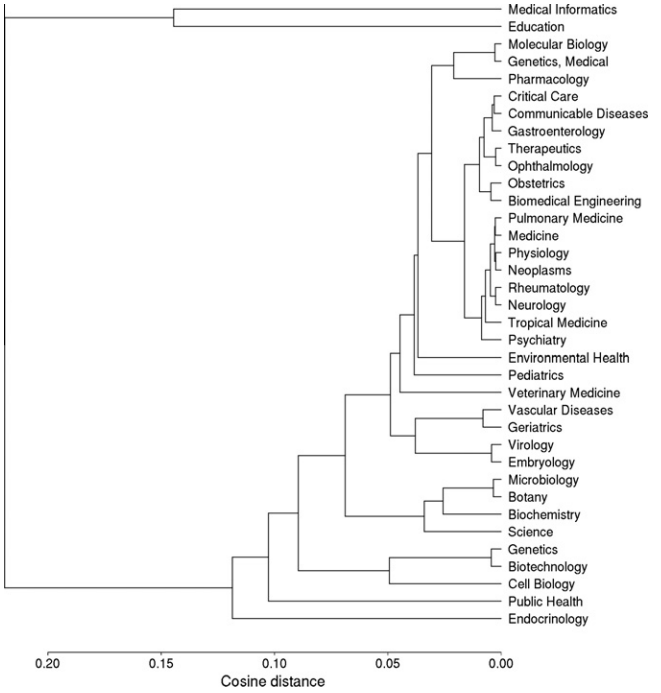

Fig. 13. Hierarchical clustering of subdomains via average-linking for the SCF distributions of *perform*.

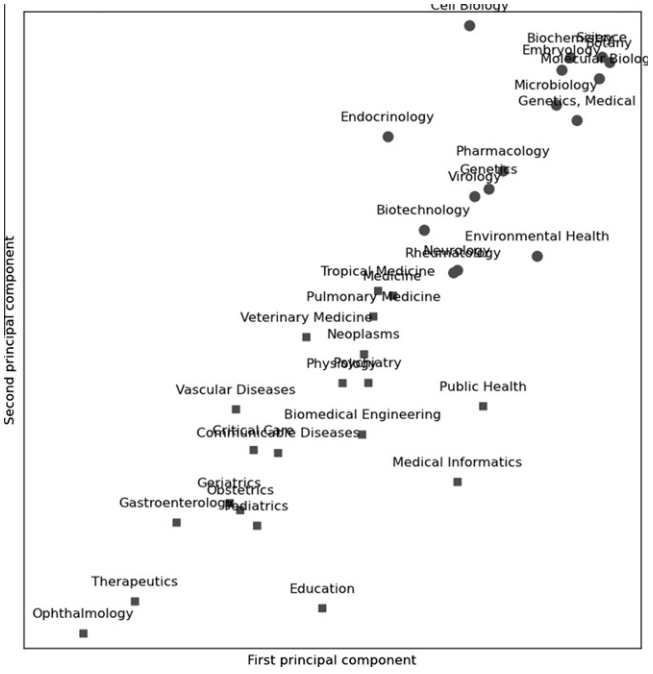

Fig. 14. Two-dimensional PCA reduction with Gap-statistic-optimal clustering for the SCF distributions of *perform*.

Scatter plots project the optimal K-Means clustering onto the first two principal components of the data. The optimal clustering was determined via the Gap Statistic [44], which increases the cluster count and runs K-Means until the improvement in error on the data is within a small range of the improvement on randomly-generated data with similar statistical properties. The principal components are normalised, and points coloured according to

Table 7
Top three SCFs, by subdomain, for *perform*.

Subdomain	Top three SCFs					
Medical Informatics	NP	0.361941	NP-PP-PRED	0.177756	NP-PRED-RS	0.084217
Education	NP	0.442718	NP-PRED-RS	0.116505	INTRANS	0.100971
Molecular Biology	NP	0.248283	NP-ING-SC	0.124142	NP-ING-OC	0.124142
Genetics, Medical	NP	0.262342	NP-ING-SC	0.120675	NP-ING-OC	0.120675
Pharmacology	NP	0.304765	NP-PP-PP PFORM	0.146923	NP-ING-SC	0.102581
Critical Care	NP	0.441001	NP-PP-PP PFORM	0.080182	NP-ING-SC	0.075064
Communicable Diseases	NP	0.431208	NP-ING-SC	0.075201	NP-ING-OC	0.075201
Gastroenterology	NP	0.484485	NP-PP-PP PFORM	0.070187	NP-ING-SC	0.069537
Therapeutics	NP	0.511537	NP-FOR-NP	0.065702	NP-ING-SC	0.057880
Ophthalmology	NP	0.536599	NP-PP-PRED	0.057788	NP-ING-SC	0.055036
Obstetrics	NP	0.454327	NP-PP-PRED	0.079327	NP-ING-SC	0.066106
Biomedical Engineering	NP	0.393035	NP-PP-PRED	0.074627	NP-TO-INF-OC	0.073383
Pulmonary Medicine	NP	0.374464	NP-PP-PP PFORM	0.094271	NP-ING-SC	0.092366
Medicine	NP	0.362900	NP-ING-SC	0.099518	NP-ING-OC	0.099518
Physiology	NP	0.394495	NP-ING-SC	0.083524	NP-ING-OC	0.083524
Neoplasms	NP	0.382559	NP-PP-PP PFORM	0.091148	NP-ING-SC	0.083187
Rheumatology	NP	0.333756	NP-PP-PP PFORM	0.106480	NP-ING-SC	0.089181
Neurology	NP	0.331288	NP-PP-PP PFORM	0.105171	NP-ING-SC	0.088721
Tropical Medicine	NP	0.370042	NP-ING-SC	0.105513	NP-ING-OC	0.105513
Psychiatry	NP	0.381216	NP-PRED-RS	0.092344	NP-PP-PRED	0.092344
Environmental Health	NP	0.300141	NP-PP-PRED	0.103796	NP-ING-SC	0.091065
Pediatrics	NP	0.450953	NP-PRED-RS	0.073572	NP-PP-PRED	0.062320
Veterinary Medicine	NP	0.407389	NP-ING-SC	0.099351	NP-ING-OC	0.099351
Vascular Diseases	NP	0.444747	NP-ING-SC	0.089117	NP-ING-OC	0.089117
Geriatrics	NP	0.457423	NP-PRED-RS	0.080250	NP-TO-INF-VC	0.072671
Virology	NP	0.312346	NP-ING-SC	0.115070	NP-ING-OC	0.115070
Embryology	NP	0.260802	NP-ING-SC	0.135802	NP-ING-OC	0.135802
Microbiology	NP	0.276414	NP-ING-SC	0.126016	NP-ING-OC	0.126016
Botany	NP	0.249518	NP-ING-SC	0.131218	NP-ING-OC	0.131218
Biochemistry	NP	0.264828	NP-ING-SC	0.134100	NP-ING-OC	0.134100
Science	NP	0.255107	NP-ING	0.130580	NP-ING-OC	0.130580
Genetics	NP	0.305055	NP-ING	0.114337	NP-ING-OC	0.114337
Biotechnology	NP	0.337702	NP-ING	0.107471	NP-ING-OC	0.107471
Cell Biology	NP	0.297386	NP-ING	0.153232	NP-ING-OC	0.153232
Public Health	NP	0.338684	NP-PRED-RS	0.097372	NP-TO-INF-VC	0.081143
Endocrinology	NP	0.352185	NP-ING	0.141674	NP-ING-OC	0.141674

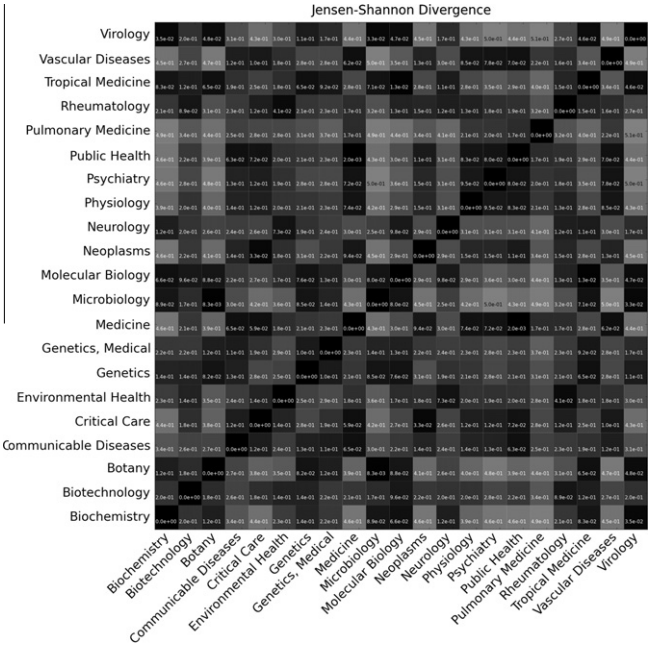

Fig. 15. Heat map of Jensen–Shannon divergence between subdomains for the SCF distributions of *predict*.

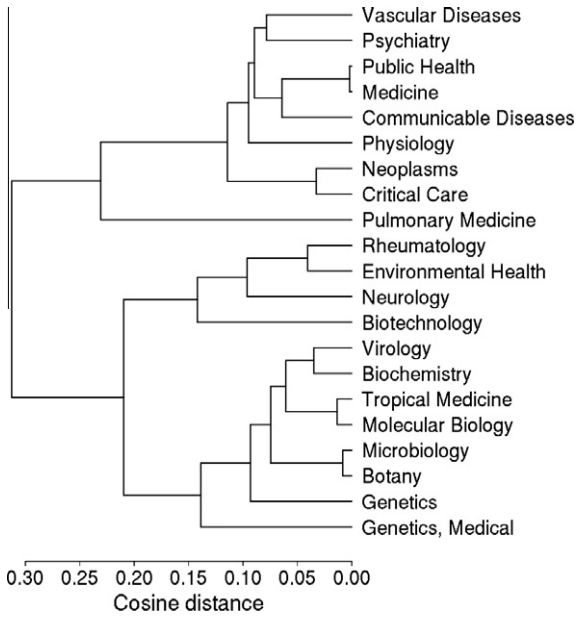

Fig. 16. Hierarchical clustering of subdomains via average-linking for the SCF distributions of *predict*.

cluster membership, with the subdomain written immediately above. The clustering is performed using the full SCF distributions, while the principle component analysis relies on decomposing the distributions into two optimal dimensions.

Top SCF tables show the top three SCFs for each subdomain, along with their relative frequencies. The SCFs are shown in their

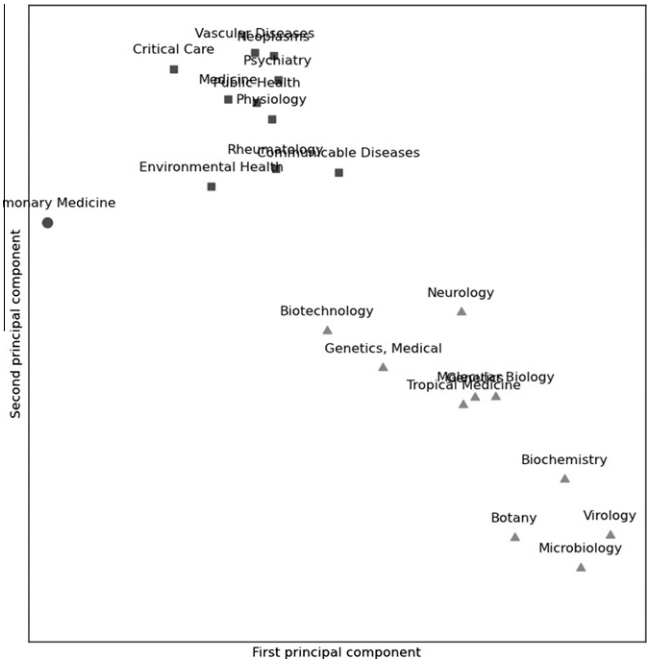

Fig. 17. Two-dimensional PCA reduction with Gap-statistic-optimal clustering for the SCF distributions of *predict*.

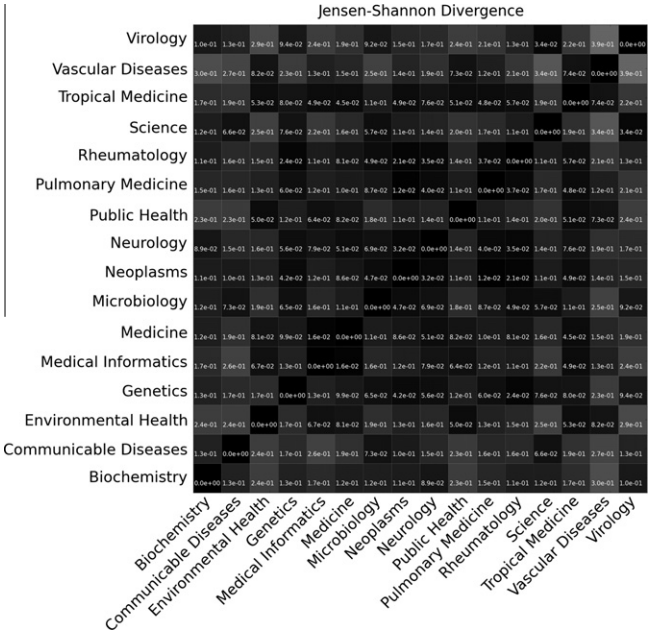

Fig. 18. Heat map of Jensen–Shannon divergence between subdomains for the SCF distributions of *recognize*.

equivalent COMLEX forms, which reflect the complements involved, as described in Section 2.1.

4.3. Discussion

4.3.1. Other views of subdomain variation

In previous studies [8,9] biomedical subdomains have been compared in terms of the frequencies of basic lexical items (verb, noun, adverb and adjective lemmas, part-of-speech tags, etc.) and using topic and selectional preference modeling methods. The results often contrast with those of the current paper, and we briefly review them here for easier comparison.

In [9] it was found that subdomains formed stable clusters in terms of basic lexical behavior, and several recurrent clusters were identified, shown in Table 4. The first cluster includes subdomains dealing primarily with microscopic processes and can be further subdivided into groupings of biochemical (*Biochemistry, Genetics*) and cellular (*Cell Biology*, *Embryology*) study. The second cluster includes subdomains focused on specific anatomical systems (*Endocrinology*, *Pulmonary Medicine*). The third cluster includes subdomains focused on clinical medicine (*Psychiatry*) or specific patient-types (*Geriatrics*, *Pediatrics*). The fourth and final cluster includes subdomains focused on social and ethical aspects of medicine (*Ethics*, *Education*).

Almost all variation was significant at a high (>0.99) level, supporting the intuition that lexical features such as vocabulary are primary aspects of different subdomains. It was also noted that the handful of syntactic features considered, such as average sentence length and grammatical relation types, did not necessarily

Table 8
Top three SCFs, by subdomain, for *predict*.

Subdomain	Top three SCFs					
Vascular Diseases	NP-PP-PRED	0.319039	NP	0.259005	NP-PRED-RS	0.197256
Psychiatry	NP	0.296053	NP-PP-PRED	0.265351	NP-PRED-RS	0.155702
Public Health	NP	0.313056	NP-PP-PRED	0.258160	NP-PRED-RS	0.143917
Medicine	NP	0.333758	NP-PP-PRED	0.249682	NP-PRED-RS	0.152866
Communicable Diseases	NP-PP-PRED	0.272923	NP	0.242837	NP-PRED-RS	0.139685
Physiology	NP	0.297170	NP-PP-PRED	0.266509	NP-PRED-RS	0.127358
Neoplasms	NP-PP-PRED	0.301850	NP	0.252678	NP-PP	0.176241
Critical Care	NP	0.321659	NP-PP-PRED	0.291244	NP-PRED-RS	0.185253
Pulmonary Medicine	NP	0.610138	NP-PP-PRED	0.117051	NP-PRED-RS	0.073733
Rheumatology	NP	0.287570	NP-PP-PRED	0.257885	NP-PRED-RS	0.150278
Environmental Health	NP	0.356804	NP-PP-PRED	0.259309	NP-PRED-RS	0.119838
Neurology	NP	0.239140	NP-PP-PRED	0.174610	HAT-S	0.115141
Biotechnology	NP	0.304348	NP-PP-PRED	0.214393	NP-TOBE	0.143928
Virology	NP	0.176289	NP-TOBE	0.139175	NP-PP-PRED	0.126804
Biochemistry	NP-PP-PRED	0.190345	NP	0.167586	NP-TOBE	0.124138
Tropical Medicine	NP	0.261468	NP-PP-PRED	0.133486	NP-PRED-RS	0.104587
Molecular Biology	NP	0.212812	NP-PP-PRED	0.185082	NP-TOBE	0.105761
Microbiology	NP	0.211287	NP-TOBE	0.165237	NP-TO-INF-OC	0.125508
Botany	NP	0.265457	NP-TOBE	0.139535	NP-TO-INF-OC	0.119682
Genetics	NP	0.258138	NP-PP-PRED	0.137358	NP-TOBE	0.103301
Genetics, Medical	NP	0.277823	NP-PP-PRED	0.187652	NP-TO-INF-OC	0.130788

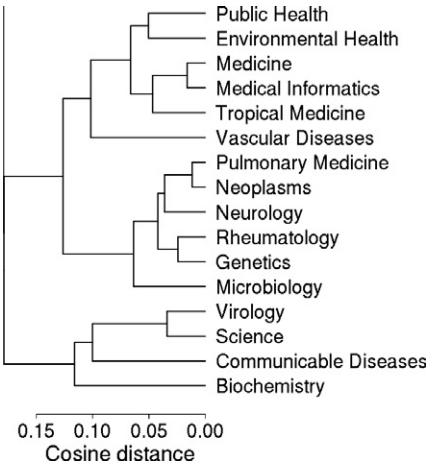

Fig. 19. Hierarchical clustering of subdomains via average-linking for the SCF distributions of *recognize*.

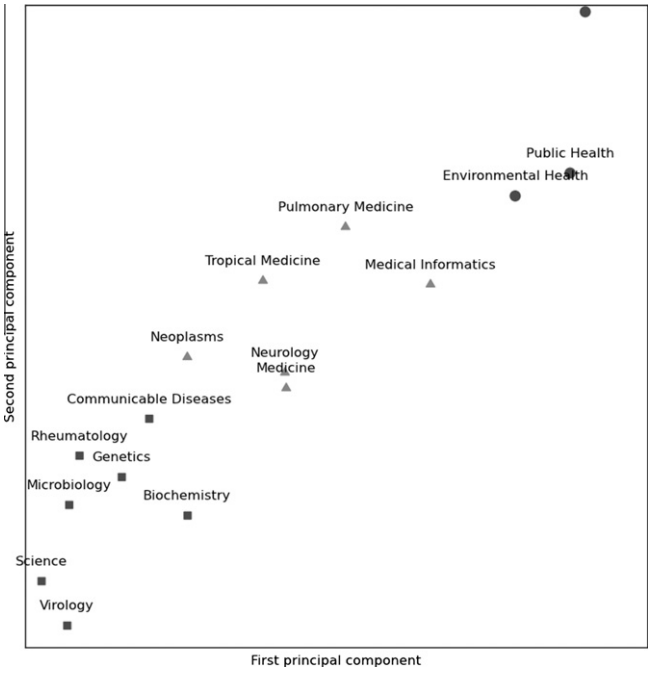

Fig. 20. Two-dimensional PCA reduction with Gap-statistic-optimal clustering for the SCF distributions of *recognize*.

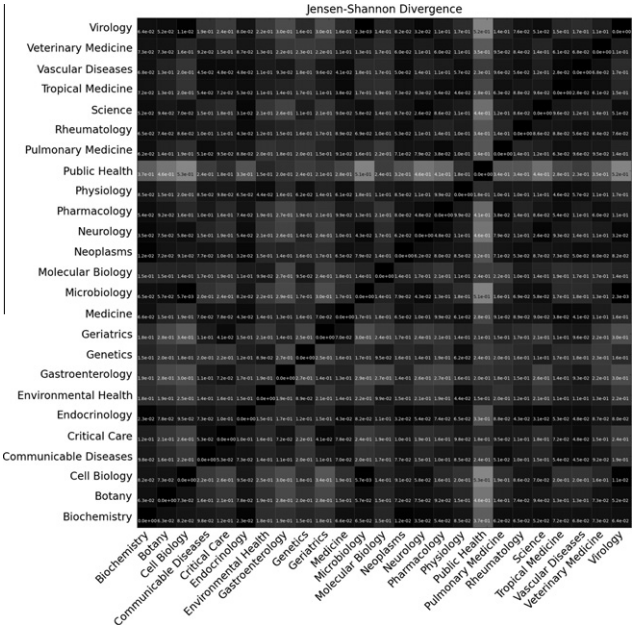

Fig. 21. Heat map of Jensen–Shannon divergence between subdomains for the SCF distributions of *treat*.

align with the more stable lexical clusters. Verbs showed a mixture of syntactic and lexical variation, reflecting their combined semantic and syntactic roles.

4.3.2. Verb subcategorization behavior

We now discuss the results of our study of SCF behavior across subdomains as described in Section 4.2. At a high level, our experiments found large differences in the amount of variation a verb could exhibit between subdomains. For example, the verb *induce* has a maximum JSD of 0.07 (low variation, between Botany and Physiology), while *develop* has a maximum of 0.62 (high variation, between Embryology and Therapeutics). Similarly, some verbs shift behavior in just one or two subdomains (e.g. *activate* in Molecular Biology and Biochemistry) while others are broadly heterogeneous (e.g. *predict*).

In contrast to the lexical results, verb subcategorization tends to show small pockets of specialized behavior, and the distinction between microscopic, systemic, clinical and social subdomains is less consistent. Instead, there are cases where verbs have taken on a specific usage in a single subdomain. The clearest example of this is *develop* (Figs. 6–8 and Table 5), which has a distinct emphasis

Table 9
Top three SCFs, by subdomain, for *recognize*.

Subdomain	Top three SCFs					
Public Health	NP	0.257610	NP-PP-PRED	0.125464	NP-AS-NP	0.096511
Environmental Health	NP	0.302128	HAT-S	0.093617	NP-PP-PRED	0.093617
Medicine	NP	0.413386	NP-PP-PRED	0.118110	NP-PRED-RS	0.100394
Medical Informatics	NP	0.332331	NP-PP-PRED	0.169925	IT-PASS-SFIN	0.075188
Tropical Medicine	NP	0.423986	NP-S	0.108108	IT-PASS-SFIN	0.104730
Vascular Diseases	NP	0.251641	IT-PASS-SFIN	0.157549	NP-AS-NP-SC	0.135667
Pulmonary Medicine	NP	0.362429	IT-PASS-SFIN	0.132827	NP-S	0.121442
Neoplasms	NP	0.447775	NP-PP-PRED	0.117166	NP-AS-NP-SC	0.101726
Neurology	NP	0.396584	NP-PP-PRED	0.146110	NP-PRED-RS	0.104364
Rheumatology	NP	0.505841	NP-PP-PRED	0.156542	NP-PRED-RS	0.096963
Genetics	NP	0.491974	NP-PP-PRED	0.130016	NP-PRED-RS	0.108347
Microbiology	NP	0.505447	NP-PP-PRED	0.159041	NP-PRED-RS	0.100218
Virology	NP	0.525084	NP-PP-PRED	0.158863	NP-PRED-RS	0.107023
Science	NP	0.530660	NP-PP-PRED	0.136792	NP-PRED-RS	0.106132
Communicable Diseases	NP	0.463087	NP-AS-NP-SC	0.194631	NP-AS-NP	0.194631
Biochemistry	NP	0.465596	NP-PP-PRED	0.135321	NP-PRED-RS	0.080275

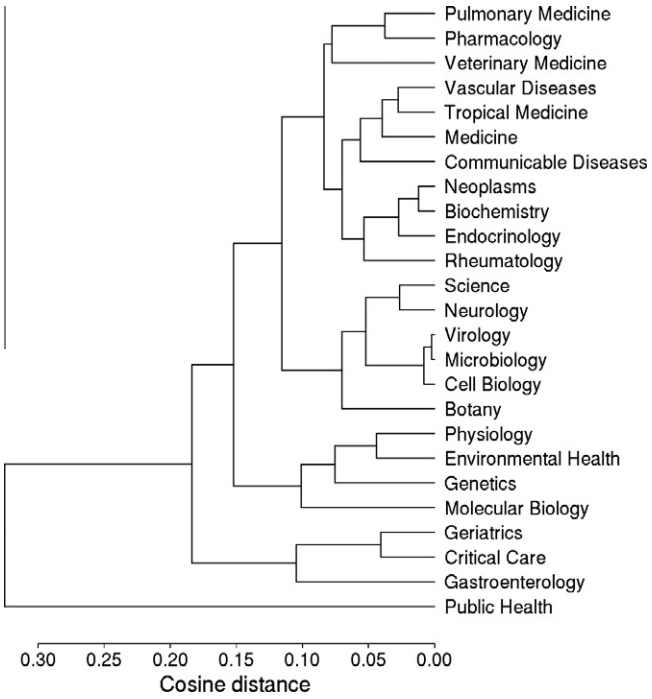

Fig. 22. Hierarchical clustering of subdomains via average-linking for the SCF distributions of *treat*.

on intransitive usage INTRANS in Embryology (“The fetus develops”), compared to its typical transitive usage NP in other subdomains (“The patient developed a tumor”).

A similar example is the verb *express* (Figs. 9–11 and Table 6), which takes NP-AS-NP-SC (“We express it as a ratio”) frequently in most subdomains, but not in Genetics and Cell Biology, where the simple transitive NP is unusually common. Sometimes the reasons for specialized behavior are not so obvious: *perform* (Figs. 12–14 and Table 7) behaves differently in Medical Informatics and Education as compared to other subdomains. Both subdomains show unusually high usage of NP-PRED-RS, and Education is unique in its frequent use of TRANS.

Not all verb behavior follows the pattern of extreme specialization in one or two subdomains: the heatmap for *predict* (Figs. 15–17 and Table 8), for example, is extremely diverse. The corresponding dendrogram shows a clear distinction between system-specific and clinical subdomains in the top half, and the microscopic subdomains in the bottom half. The top SCFs show that the microscopic subdomains use *predict* in conjunction with infinitival forms (e.g. NP-TOBE, “We predicted it to be”). *Recognize* (Figs. 18–20 and Table 9), like *predict*, shows a diverse set of JSD values. It is unclear why some subdomains prefer e.g. THAT-S or NP-AS-NP, except perhaps that diagnosis-oriented subdomains prefer the latter.

Some verbs may have more than one specialized behavior: *treat* (Figs. 21–23 and Table 10) is generally either used in a clinical sense (NP-FOR-NP, “We treat the patient for concussion”) or attributive (NP-AS-NP-SC, “We treat the infection as a separate issue”). The most distinct subdomain, Public Health, appears as an outlier

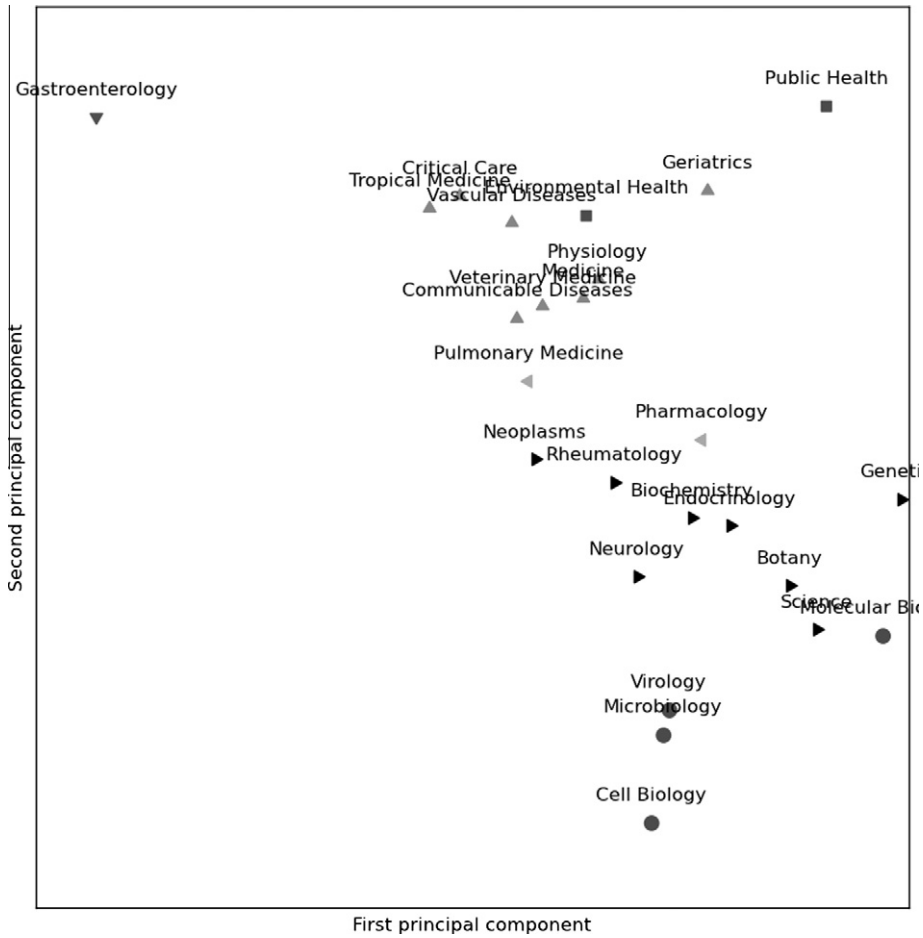

Fig. 23. Two-dimensional PCA reduction with Gap-statistic-optimal clustering for the SCF distributions of *treat*.

Table 10
Top three SCFs, by subdomain, for *treat*.

Subdomain	Top three SCFs					
Pulmonary Medicine	NP	0.337748	NP-PP-PP PFORM	0.167770	NP-NP-PRED	0.129139
Pharmacology	NP	0.274845	NP-NP-PRED	0.184783	NP-NP	0.184783
Veterinary Medicine	NP	0.360000	NP-FOR-NP	0.120000	NP-PP-PP PFORM	0.106667
Vascular Diseases	NP	0.388060	PP	0.099502	PP-PRED-RS	0.099502
Tropical Medicine	NP	0.425547	NP-NP-PRED	0.103035	NP-NP	0.103035
Medicine	NP	0.355288	NP-PP-PP PFORM	0.126160	NP-PRED-RS	0.080705
Communicable Diseases	NP	0.353806	NP-PP-PP PFORM	0.173010	NP-FOR-NP	0.121107
Neoplasms	NP	0.314900	NP-PP-PP PFORM	0.219662	PP-PRED-RS	0.094470
Biochemistry	NP	0.252427	NP-PP-PP PFORM	0.200647	PP-PRED-RS	0.101942
Endocrinology	NP	0.240283	NP-PP-PP PFORM	0.207303	PP-PP	0.089517
Rheumatology	NP	0.283192	NP-PP-PP PFORM	0.203390	PP	0.133475
Science	NP-PP-PP PFORM	0.224299	NP	0.190314	PP-PP	0.115548
Neurology	NP	0.260030	NP-PP-PP PFORM	0.228826	NP-NP-PRED	0.123328
Virology	NP-PP-PP PFORM	0.300000	NP	0.209524	NP-NP-PRED	0.102381
Microbiology	NP-PP-PP PFORM	0.322925	NP	0.201828	PP-PRED-RS	0.105864
Cell Biology	NP-PP-PP PFORM	0.389027	NP	0.182045	PP-PP	0.114713
Botany	NP	0.214421	NP-PP-PP PFORM	0.204934	NP-NP	0.100569
Physiology	NP	0.358191	NP-PP-PP PFORM	0.107579	NP-NP-PRED	0.074572
Environmental Health	NP	0.385877	NP-PP-PP PFORM	0.091298	NP-AS-NP-SC	0.077746
Genetics	NP	0.211664	NP-PP-PP PFORM	0.189040	NP-AS-NP-SC	0.096531
Molecular Biology	NP-PP-PP PFORM	0.281690	NP	0.170775	PP-PP	0.070423
Geriatrics	NP	0.346975	NP-PRED-RS	0.097865	NP-PP-PRED	0.088968
Critical Care	NP	0.413424	NP-PP-PP PFORM	0.108949	PP-PRED-RS	0.090467
Gastroenterology	NP	0.546099	NP-PP-PP PFORM	0.148936	PP-PRED-RS	0.083333
Public Health	NP	0.342735	NP-FOR-NP	0.124786	NP-AS-NP-SC	0.101709

because of its unique combination of both usages. This is an example of a heterogeneous subdomain merging SCF behaviors into a third, unique distribution.

There are several reasons why our results with SCFs differ from the results obtained with lexical features in previous subdomain comparisons [8,9]. One factor is that we considered individual verbs, whereas lexical studies average variation across all lexical items of a given class. This has a smoothing effect on the specialized behavior. Another factor is that distinct senses of a verb, e.g. general and specialized, may create confounding effects when the SCF behavior of the two senses is overlaid in a subdomain. There are two possible reasons for this: that distinct usages exist side-by-side within individual documents, or the subdomains are grouping together documents that are linguistically quite different. Either case implies that flexible, data-driven SCF lexicons are particularly important for the PMC OA.

Our results here show that there is considerable subdomain variation in verb SCFs in biomedicine which should be taken into account in the development and application of SCF systems in this domain. Future work could look at the nature of this variation in more detail, e.g. by broadening the set of verbs considered and averaging the divergence in their SCF distributions to determine whether there is a correlation with the lexical results. This would require a principled way of combining the distributions, beyond simple equal weighting, because the proportion of verbs that change SCF behavior is small and would be overwhelmed by noise.

5. Conclusions and recommendations

Our review of the state of SCF acquisition in biomedical text processing has found very little in the way of direct (i.e. intrinsic) performance evaluation. Basic questions, such as how general language systems perform on biomedicine, and how well a lexicon acquired from one subdomain translates to others, are best answered by a human-annotated gold standard. While gold standards have been produced for syntactic analysis of the biomedical literature (e.g. GENIA [41] and CRAFT [37]), domain-specific lexical resources have been severely limited in scale (PASBio [12] and BioFrameNet [38]) or in scope (BioLexicon [22]). Currently, no gold standard lexical resource exists representative of biomedicine in general, even as research pushes forward with domain-specific approaches. It is crucial that we have a gold standard to guide efforts in domain adaptation, and simply to evaluate the real-world performance of proposed systems.

Although direct evaluation of SCF acquisition is important, it could be supplemented with task-based (i.e. extrinsic) evaluation which uses the output of a system to augment performance on a downstream task that is easier to assess [45]. For example, an unlexicalized parser or relationship extractor could be augmented with SCF, and then re-evaluated to determine improvement. In this setup, the definition of subcategorization and the SCF inventories used by each system would not need to be reconciled: the candidate parses would simply be reranked based on the new probabilities from the lexicon. Decoupling evaluation from a particular definition and inventory would facilitate the development and comparison of new approaches to SCF acquisition.

We found significant variation in SCF behavior between biomedical subdomains, with different properties than in previously studied lexical variation. Most notably, subdomain clusters produced from the subcategorization behavior of individual verbs did not align well with clusters based on simple lemma frequencies [9], and often were not readily interpretable in terms of major subdomain-spanning topics. Some verb behavior occurred in discrete pockets, just one or two subdomains, rather than in one of the major clusters identified in lexical studies. While future work could broaden the scope of these experiments and aim to obtain a more precise idea of the nature of subdomain variation in biomedicine, the results already presented here highlight the need for subdomain-adaptation in SCF acquisition.

Unsupervised approaches to SCF acquisition have a particular advantage in domain adaptation, since they do not rely on manually created resources and because their definitions and inventories emerge from their domain-specific input data. Ideally, such approaches would also involve moving away from features that require manual domain-adaptation for optimal performance (such as parser output), to shallower and more robust features like parts-of-speech or phrase chunking (e.g. [46]). There are a range of semi-supervised methods between these extremes, such as self-training and hybrid graphical modeling [47], which may help yield optimal

performance on SCF acquisition while minimising the need for manual annotation. An interesting area for future work is determining an optimal middle ground.

References

[1] Hunter L, Cohen KB. Biomedical language processing: what's beyond PubMed? Mol Cell 2006;21(5):589–94. http://dx.doi.org/10.1016/j.molcel.2006.02.012.
[2] Harmston N, Filsell W, Stumpf M. What the papers say: text mining for genomics and systems biology. Hum Genom 2010;5:17–29.
[3] Ananiadou S, Thompson P, Nawaz R. Improving search through event-based biomedical text mining. In: Proceedings of the first international workshop on automated motif discovery in cultural heritage and scientific communication texts (AMICUS 2010), CLARIN/DARIAH 2010. Vienna, Austria; 2010.
[4] Rupp C, Thompson P, Black W, McNaught J. A specialised verb lexicon as the basis of fact extraction in the biomedical domain. In: Proceedings of interdisciplinary workshop on verbs: the identification and representation of verb features (Verb 2010). Pisa, Italy; 2010.
[5] Korhonen A. Subcategorization acquisition. Ph.D. thesis, University of Cambridge Computer Laboratory; 2002.
[6] Korhonen A, Krymolowski Y, Briscoe T. A large subcategorization lexicon for natural language processing applications. In: Proceedings of LREC; 2006.
[7] Preiss J, Briscoe T, Korhonen A. A system for large-scale acquisition of verbal, nominal and adjectival subcategorization frames from corpora. In: Proceedings of the 45th annual meeting of the association for computational linguistics. Prague, Czech Republic; 2007.
[8] Verspoor K, Cohen KB, Hunter L. The textual characteristics of traditional and open access scientific journals are similar. BMC Bioinform 2009;10. 183-1.
[9] Lippincott T, Ó Séaghdha D, Korhonen A. Exploring subdomain variation in biomedical language. BMC Bioinform 2011;12(1).
[10] Grishman R, Macleod C, Meyers A. COMLEX syntax: building a computational lexicon. In: Proceedings of COLING. Kyoto; 1994.
[11] NIH. The pubmed central open access subset; 2009. <http://www.pubmedcentral.nih.gov/about/openftlist.html>.
[12] Wattarujeekrit T, Shah P, Collier N. PASBio: predicate-argument structures for event extraction in molecular biology. BMC Bioinform 2004:5.
[13] Tsai RTH, Chou WC, Lin YC, Sung CL, et al. W.K. BIOSMILE: adapting semantic role labeling for biomedical verbs: an exponential model coupled with automatically generated template features. In: Proceedings of the BioNLP'06 workshop on linking natural language processing and biology. Association for Computational Linguistics; 2005. p. 57–64.
[14] Tsai RTH, Dai HJ, Huang CH, Hsu WL. Semi-automatic conversion of BioProp semantic annotation to PASBio annotation. BMC Bioinform 2008;9(Suppl. 12). S18-1.
[15] Grimshaw J. Argument structure. MIT Press; 1990.
[16] Pollard C, Sag I. An information-based syntax and semantics. CSLI lecture notes, vol. 13. Stanford University; 1987.
[17] Merlo P, Ferrer EE. The notion of argument in pp attachment. Comput Linguist 2006:32.
[18] Abend O, Rappoport A. Fully unsupervised core-adjunct argument classification. In: Proceedings of the 48th annual meeting of the association for computational linguistics. Association for Computational Linguistics; 2010. p. 226–36.
[19] Harris Z. Discourse and sublanguage. In: Kittredge R, Lehrberger J, editors. Sublanguage: studies of language in restricted semantic domains. Walter de Gruyter; 1982. p. 231–45.
[20] Sager N. Syntactic formatting of science information. In: Kittredge R, Lehrberger J, editors. Sublanguage: studies of language in restricted semantic domains. Walter de Gruyter; 1982. p. 9–26.
[21] Cohen KB, Hunter L. A critical review of pasbio's argument structures for biomedical verbs. BMC Bioinform 2006;7(Suppl. 3). S5-1.
[22] Thompson P, McNaught J, Montemagni S, Calzolari N, Gratta RD, Lee V, et al. The biolexicon: a large-scale terminological resource for biomedical text mining. BMC Bioinform 2011;12. 397-397.
[23] Klein D, Manning CD. Accurate unlexicalized parsing. In: Proceedings of ACL; 2003. p. 423–30.
[24] Choi JD, Nicolov N. K-best, locally pruned, transition-based dependency parsing using robust risk minimization. In: Collections of recent advances in natural language processing V. John Benjamins; 2009. p. 205–16.
[25] Cohen KB, Palmer M, Hunter L. Nominalization and alternations in biomedical language. PLoS ONE 2008;3(9).
[26] Baker CF, Fillmore CJ, Lowe JB. The berkeley framenet project. In: Proceedings of the 36th annual meeting of the association for computational linguistics and 17th international conference on computational linguistics, vol. 1, ACL '98. Stroudsburg, PA, USA: Association for Computational Linguistics; 1998. p. 86–90. doi:http://dx.doi.org/10.3115/980845.980860.
[27] Kipper-Schuler K. Verbnet: a broad-coverage, comprehensive verb lexicon. Ph.D thesis, University of Pennsylvania; 2005.
[28] Palmer M, Gildea D, Kingsbury P. The proposition bank: an annotated corpus of semantic roles. Comput Linguist 2005;31(1):71–106. http://dx.doi.org/10.1162/0891201053630264. http://www.mitpressjournals.org/doi/abs/10.1162/0891201053630264.
[29] Marcus MP, Santorini B, Marcinkiewicz MA. Building a large annotated corpus of English: the Penn Treebank. Comput Linguist 1993;19(2):313–30.
[30] Im Walde SS. The induction of verb frames and verb classes from corpora. In: Lüdeling A, Kytö M, editors. Corpus linguistics. An international handbook. Berlin: Mouton de Gruyter; 2009. p. 952–71.
[31] O'Donovan R, Burke M, Cahill A, van Genabith J, Way A. Large-scale induction and evaluation of lexical resources from the penn-ii treebank. In: Proceedings of the 42nd annual meeting on association for computational linguistics, ACL '04. Stroudsburg, PA, USA: Association for Computational Linguistics; 2004. doi:http://dx.doi.org/10.3115/1218955.1219002.
[32] Messiant C. A subcategorization acquisition system for French verbs. In: ACL HLT '08 student research workshop; 2008.
[33] Lenci R, Mcgillivray B, Montemagni S, Pirrelli V. Unsupervised acquisition of verb subcategorization frames from shallow-parsed corpora. In: LREC '08; 2008.
[34] Han X, Lv C, Zhao T. Weakly supervised SVM for Chinese–English cross-lingual subcategorization lexicon acquisition. In: The 11th joint conference on information science; 2008.
[35] Uzun E, Klaslan Y, Agun H, Uar E. Web-based acquisition of subcategorization frames for Turkish. In: The eighth international conference on artificial intelligence and soft computing; 2008.
[36] Briscoe E, Carrol J, Watson R. The second release of the RASP system. In: Proceedings of the COLING/ACL 2006 interactive presentation sessions. Sydney, Australia; 2006.
[37] Verspoor KM, Cohen KB, Lanfranchi A, Warner C, Johnson HL, Roeder C, et al. A corpus of full-text journal articles is a robust evaluation tool for revealing differences in performance of biomedical natural language processing tools. BMC Bioinform 2012:13.
[38] Dolbey A, Ellsworth M SJ. BioFrameNet: a domain-specific framenet extension with links to biomedical ontologies. In: Bodenreider O, editor. Proceedings of KR-MED; 2006. p. 87–94.
[39] McCray A, Srinivasan S, Browne A. Lexical methods for managing variation in biomedical terminologies. In: Proceedings of the 18th annual SCAMC. Washington: McGraw Hill; 1994. p. 235–9.
[40] Miyao Y, Tsujii J. Feature forest models for probabilistic HPSG parsing. Comput Linguist 2008;34:35–80.
[41] Ohta T, Tsuruoka Y, Takeuchi J, Kim JD, Miyao Y, Yakushiji A, et al. An intelligent search engine and gui-based efficient medline search tool based on deep syntactic parsing. In: Proceedings of the COLING/ACL on interactive presentation sessions, COLING-ACL '06. Stroudsburg, PA, USA: Association for Computational Linguistics; 2006. p. 17–20. doi: http://dx.doi.org/10.3115/1225403.1225408.
[42] Grosse I, Bernaola-Galván P, Carpena P, Román-Roldán R, Oliver J, Stanley HE. Analysis of symbolic sequences using the Jensen–Shannon divergence. Phys Rev E 2002;65(4). http://dx.doi.org/10.1103/PhysRevE.65.041905. 041905-1.
[43] Cover TM, Thomas JA. Elements of information theory. New York: Wiley; 1991.
[44] Tibshirani R, Walther G, Hastie T. Estimating the number of clusters in a data set via the gap statistic. J R Stat Soc B 2001;63(2):411–23.
[45] Vlachos A. Evaluating unsupervised learning for natural language processing tasks. In: Proceedings of the EMNLP 2011 workshop on unsupervised learning in NLP. Edinburgh, UK; 2011.
[46] Kang N, van Mulligen EM, Kors JA. Comparing and combining chunkers of biomedical text. J Biomed Inform 2011;44(2):354–60. http://dx.doi.org/10.1016/j.jbi.2010.10.005. <http://www.sciencedirect.com/science/article/pii/S1532046410001577>.
[47] Zhu X. Semi-supervised learning literature survey; 2006.