Jargon-Term Extraction by Chunking

Adam Meyers[†], Zachary Glass[†], Angus Grieve-Smith[†], Yifan He[†],
Shasha Liao[‡] and Ralph Grishman[†]
New York University[†], Google[‡]
meyers/angus/yhe/grishman@cs.nyu.edu, zglass@alumni.princeton.edu

Abstract

NLP definitions of *Terminology* are usually application-dependent. IR terms are noun sequences that characterize topics. Terms can also be arguments for relations like abbreviation, definition or IS-A. In contrast, this paper explores techniques for extracting terms fitting a broader definition: noun sequences specific to topics and not well-known to naive adults. We describe a chunking-based approach, an evaluation, and applications to non-topic-specific relation extraction.

1 Introduction

Webster's II New College Dictionary (Houghton Mifflin Company, 2001, p.1138) defines terminology as: *The vocabulary of technical terms and usages appropriate to a particular field, subject, science, or art.* Systems for automatically extracting instances of terminology (terms) usually assume narrow operational definitions that are compatible with particular tasks. Terminology, in the context of Information Retrieval (IR) (Jacquemin and Bourigault, 2003) refers to keyword search terms (*microarray*, *potato*, *genetic algorithm*), single or multi-word (mostly nominal) expressions collectively representing topics of documents that contain them. These same terms are also used for creating domain-specific thesauri and ontologies (Velardi et al., 2001). We will refer to these types of terms as *topic-terms* and this type of terminology *topic-terminology*. In other work, types of terminology (genes, chemical names, biological processes, etc.) are defined relative to a specific field like Chemistry or Biology (Kim et al., 2003; Corbett et al., 2007; Bada et al., 2010). These classes are used for narrow tasks, e.g., Information Extraction (IE) slot filling tasks within a particular genre of interest (Giuliano et al., 2006; Bundschus et al., 2008; BioCreAtIvE, 2006). Other projects are limited to Information Extraction tasks that may not be terminology-specific, but have terms as arguments, e.g., (Schwartz and Hearst, 2003; Jin et al., 2013) detect abbreviation and definition relations respectively and the arguments are terms. In contrast to this previous work, we have built a system that extracts a larger set of terminology, which we call *jargon-terminology*. Jargon-terms may include *ultracentrifuge*, which is unlikely to be a topic-term of a current biology article, but will not include *potato*, a non-technical word that could be a valid topic-term. We aim to find all the jargon-terms found in a text, not just the ones that fill slots for specific relations. As we show, jargon-terminology closely matches the notional (e.g., Webster's) definition of terminology. Furthermore, the important nominals in technical documents tend to be jargon-terms, making them likely arguments of a wide variety of possible IE relations (concepts or objects that are invented, two nominals that are in contrast, one object that is "better than" another, etc.). Specifically, the identification of jargon-terms lays the ground for IE tasks that are not genre or task dependent. Our approach which finds all instances of terms (tokens) in text is conducive to these tasks. In contrast, topic-term detection techniques find smaller sets of terms (types), each term occurring multiple times and the set of terms collectively represents a topic, in a similar way that a set of documents can represent a topic.

Proceedings of SADAATL 2014, pages 11–20,
Dublin, Ireland, August 24, 2014.

This paper describes a system for extracting jargon-terms in technical documents (patents and journal articles); the evaluation of this system using manually annotated documents; and a set of information extraction (IE) relations which take jargon-terms as arguments. We incorporate previous work in terminology extraction, assuming that terminology is restricted to noun groups (minus some left modifiers) (Justeson and Katz, 1995); [1] and we use both topic-term extraction techniques (Navigli and Velardi, 2004) and relation-based extraction techniques (Jin et al., 2013) in components of our system. Rather than looking at the distribution of noun groups as a whole for determining term-hood, we refine the classes used by the noun group chunker itself, placing limitations on the candidate noun groups proposed and then filtering the output by setting thresholds on the number and quality of the "jargon-like" components of the phrase. The resulting system admits not only topic-terms, but also other non-topic instances of terminology. Using the more inclusive set of jargon-terms (rather than just topic-terms) as arguments of the IE relations in section 6, we are able to detect a larger and more informative set of relation. Furthermore, these relations are salient for a wide variety of genres (unlike those in (BioCreAtIvE, 2006)) – a genre-neutral definition of terminology makes this possible. For example, the CONTRAST relation between the two bold face terms in **necrotrophic effector system**$_{A1}$ *that is an exciting contrast to the* **biotrophic effector models**$_{A2}$. would be applicable in most academic genres. Our jargon-terms also contrast with the tactic of filling terminology slots in relations with any noun-group (Justeson and Katz, 1995), as such a strategy overgenerates, lowering precision.

2 Topic-term Extraction

Topic-term extractors (Velardi et al., 2001; Tomokiyo and Hurst, 2003) collect candidate terms (N-grams, noun groups, words) that are more representative of a foreground corpus (documents about a specific topic) than they are of a background corpus (documents about a wide range of topics), using statistical measures such as $\frac{Term\ Frequency}{Inverse\ Document\ Frequency}$ (TFIDF), or a variation thereof. Due to the metrics used and cutoffs assumed, the list of terms selected is usually no more than a few hundred distinct terms, even for a large set of foreground documents and tend to be especially salient to that topic. The terms can be phrases that lay people would not know (e.g., *microarray*, *genetic algorithm*) or common topics for that document set (e.g., *potato*, *computer*). Such systems rank all candidate terms, using cutoffs (minimum scores or percentages of the list) to separate out the highest-ranked terms as output. Thus sets of topic-terms, derived this way, are dependent on the foreground and background assumed, and the publication dates. So a precise definition would include such information, e.g., *topic-terms(biomedical-patents, random-patents, 1990–1999)* would refer to those topic-terms that differentiate a foreground of biomedical patents from the 1990s from a background of diverse patents from the same epoch. Narrower topics are possible (e.g., comparing DNA-microarray patents to the same background); or broader ones (e.g., if a diverse corpus including news articles, fiction and travel writing are the background set instead of patents, then patent terms such as *national stage application* may be highly ranked in the output). Thus topic-terms generated by these methods model a relationally based definition and are relative to the chosen foregrounds, backgrounds and dates.

Topic-terms can include words/phrases like *potato, wheat, rat, monkey*, which may be common subjects of some set of biomedical documents, but are not specific to a technical field. In contrast, jargon-terms would include words (like *ultracentrifuge, theorem, graduated cylinder*) that are specific to technical language, but don't tend to be topics of any current document of interest. Jargon-terms, like topic-terms, can be defined relative to a particular foreground (which can also be represented as a set of documents), but there is the implicit assumption that they all share the same background set: non-genre-specific language (or simply a very diverse set of documents). It is also possible to refer to terminology in general as the union of jargon-terms with respect to the set of specialized knowledge areas as foregrounds and all sharing the same background of non-genre-specific language. Jargon-terms, like topic-terms, are also time dependent, since some terms will eventually be absorbed into the common lexicon, e.g., *computer*. However, we can make the simplifying assumption that we are talking about jargon in the present

[1] We restrict our scope to nominal terminology, but acknowledge the importance of non-nominal terminology, e.g., event verb terms (*calcify, coactivate*) which are crucial to IE.

time. Furthermore, jargon-term status is somewhat less time sensitive than topic-term status because terminology is absorbed very sparingly (and very slowly) into the popular lexicon, whereas topics go in and out of fashion quickly within a literature that is meant for an expert audience. Ignoring the *potato* type cases, topic-terms are a proper subset of jargon-terms and, thus, the set of jargon-terms is larger than the set of topic-terms. Finally, topic terms are ranked with respect to how well they can serve as keywords, i.e., how specific they are to a particular document set, whereas +/-jargon-term is a binary distinction.

We built a topic term extractor that combines several metrics together in an ensemble including: TFIDF, KL Divergence (Cover and Thomas, 1991; Hisamitsu et al., 1999) and a combination of Domain Relevance and Document Consensus (DRDC) based on (Navigli and Velardi, 2004). Furthermore, we filtered the output by requiring that each term would be recognized as a term by the jargon-term chunker described below in section 3. We manually scored the top 100 terms generated for two classes of biology patents (US patent classes 435 and 436) and achieved accuracies of 85% and 76% respectively. We also manually evaluated the top 100 terms taken from biology articles, yielding an accuracy of about 88%. As discussed, we use the output of this system for our jargon-term extraction system.

3 Jargon-term Extraction by Chunking

(Justeson and Katz, 1995) uses manual rules to detect noun groups (sequences of nouns and adjectives ending in a noun) with the goal of detecting instances of topic-terms. They filter out those noun groups that occur only once in the document on the theory that the multiply used noun groups are more likely to be topics. They manually score their output from two computer science articles and one biotechnology article, with 146, 350 and 834 instances of terms and achieve accuracies of 96%, 86% and 77%. (Frantzi et al., 2000) uses linguistic rules similar to noun chunking to detect candidate terms; filters the results using a stop list and other linguistic constraints; uses statistical filters to determine whether substrings are likely to be terms as well; and uses statistical filters based on neighboring words (context). (Frantzi et al., 2000) ranks their terms by scores and achieve about 75% accuracy for the top 40 terms – their system is tested on medical records (quite a different corpus form ours). Our system identifies all instances of terminology (not just topic terms) and identifies many more instances per document (919, 1131 and 2166) than (Justeson and Katz, 1995) or (Frank, 2000). As we aim to find all instances of jargon-terms, we evaluate for both precision and recall rather than just accuracy (section 5). Two of the documents that we test on are patents, which have a very different word distribution than articles. In fact, due to both the amount of repetition in patents and the presence of multiple types of terminology (legal terms as well as topic-related terms), it is hard to imagine that eliminating terms occurring below a frequency threshold (as with (Justeson and Katz, 1995)) would be an effective method of filtering. Furthermore, (Frank, 2000) used a very different corpus than we did and they focused on a slightly different problem (e.g., we did not attempt to find the highest-ranked terms and we did not attempt to find both long terms and substrings which were terms). Thus while it is appropriate to compare our methodology, it is difficult to compare our results.

We have implemented a hand-crafted term extractor, which we will call a jargon-term chunker because it functions in much the same way as a noun group chunker. It uses a deterministic finite state machine, based on parts of speech (POS) and a fine-tuned set of lexical categories. We observed that jargon-terms are typically noun groups, minus some left modifiers, and normally include words that are not in standard vocabulary or belong to certain other classes of words (e.g., nominalizations). While topic-term techniques factor the distribution of whole term sequences into the choice of topic-terms, our method focuses on the distribution of words within topic-term sequences. The primary function of POS classification is to cluster words distributionally in a language. A POS tag reflects the syntactic distribution of the word in the sense that words with the same POS should be able to replace each other in sentences. Morphologically, POSs are subject to the same morphological variation (prefixes, suffixes, tense, gender, number, etc.). For example, the English word *duck* belongs to the POS *noun* because it tends to occur: after a determiner, after an adjective, and ending a unit that can be the subject of a verb: nouns are substitutable for each other. Furthermore, it has a plural form resulting from an -s or -es suffix, etc. Similarly, we hold that the presence of particular classes of words within a noun group affects its potential to function

as a jargon-term. As will become evident, we can use topic-term-like metrics to identify some of these word classes. Furthermore, given our previous assertion that topic-terms are a subset of jargon-terms, we assume that the most saliently ranked topic-terms are also jargon-terms and words that are commonly parts of topic-terms tend to be parts of jargon-terms. There are also "morphological properties" that are indicative of subsets of jargon-terms: allCap acronyms, chemical formulas, etc.

Our system classifies each word using POS tags, manually created dictionaries and the output of our own topic-term system. These classifications are achieved in four stages. In the first stage we divide the text into smaller segments using coordinate conjunctions (*and*, *or*, *as well as*, . . .) and punctuation (periods, left/right parentheses and brackets, quotation marks, commas, colons, semi-colons). These segments are typically smaller than the level of the sentence, but larger than most noun groups. These segments are good units to process because they are larger than jargon-terms (substrings of noun groups) and smaller than sentences (and thus provide a smaller search space). In the second stage, potential jargon-term (PJs) are generated by processing tokens from left to right and classifying them using a finite state machine (FSM). The third stage filters the PJs generated with a set of manually constructed constraints, yielding a set of jargon-terms. A final filter (stage 4) identifies named entities and separates them out from the true jargon-terms: it turns out that many named entities have similar phrase-internal properties as jargon-terms.

The FSM (that generates PJs) in the second stage includes the following states (Ramshaw and Marcus, 1995): START (S) (marking the beginning of a segment), Begin Term (B-T), Inside Term (I-T), End Term (E-T), and Other (O). A PJ is a sequence consisting of: (a) a single E-T; or (b) exactly one B-T, followed by zero or more instances of I-T, followed by zero or one instances of E-T. Each transition to a new state is conditioned on: (a) the (extended) POS tag of the current word; (b) the extended POS tag of the previous word; and (c) the previous state. The extended POSs are derived from the output of a Penn-Treebank-based POS tagger and refinements based on machine readable dictionaries, including COMLEX Syntax (Macleod et al., 1997), NOMLEX (Macleod et al., 1998), and some manually encoded dictionaries created for this project. Table 1 describes the transitions in the FSM (unspecified entries mean no restriction). ELSE indicates that in all cases other than those listed, the FST goes to state O. Extended POS tags are classified as follows.

Adjectives, words with POS tags JJ, JJR or JJS, are subdivided into:

STAT-ADJ: Words in this class are marked adjective in our POS dictionaries and found as the first word in one of the top ranked topic-terms (for the topic associated with the input document).

TECH-ADJ: If an adjective ends in a suffix indicating (*-ic*, *-cous*, *-xous*, and several others) it is a technical word, but it is not found in our list of exceptions, it is marked TECH-ADJ.

NAT-ADJ: An adjective, usually capitalized, that is the adjectival form of a country, state, city or continent, e.g., *European*, *Indian*, *Peruvian*, . . .

CAP-ADJ: An adjective such that the first letter is capitalized (but is not marked NAT-ADJ).

ADJ: Other adjectives

Nouns are marked NN or NNS by the POS tagger and are the default POS for out of vocabulary (OOV) words. POS tags like NNP, NNPS and FW (proper nouns and foreign nouns) are not reliable for our POS tagger (trained on news) when applied to patents and technical articles. So NOUN is also assumed for these. Subclasses include:

O-NOUN: (Singular or plural) nouns not found in any of our dictionaries (COMLEX plus some person names) or nouns found in lists of specialized vocabulary which currently include chemical names.

PER-NOUN: Nouns beginning with a capital that are in our dictionary of first and last names.

PLUR-NOUN: Nouns with POS NNS nouns that are not marked O-NOUN or PER-NOUN.

C-NOUN: Nouns with POS NN that are not marked O-NOUN or PER-NOUN.

Verbs Only **ING-VERB**s (VBG) and **ED-VERB**s (VBN and VBD) are needed for this task (other verbs trigger state O). Finally, we use the following additional POS tags:

POSS: POS for *'s*, split off from a possessive noun.

PREP: All prepositions (POS IN and TO)

ROM-NUM: Roman numerals (I, II, . . ., MMM)

Previous POS	**Current POS**	**Previous State**	**New State**
	DET, PREP, POSS, VERB		O
O-NOUN, C-NOUN, PLUR-NOUN	ROM-NUM	B-T, I-T	E-T
	PLUR-NOUN	B-T,I-T	I-T
	ADJ, CAP-ADJ	I-T	I-T
	C-NOUN, PER-NOUN, O-NOUN	B-T, I-T	I-T
O-NOUN	CAP-ADJ, TECH-ADJ, STAT-ADJ, NAT-ADJ	B-T, I-T	I-T
	CAP-ADJ, TECH-ADJ, NAT-ADJ, ING-VERB, ED-VERB, STAT-ADJ C-NOUN, O-NOUN, PER-NOUN	E-T, O, S	B-T
TECH-ADJ, NAT-ADJ ADJ, CAP-ADJ	TECH-ADJ, NAT-ADJ ADJ, CAP-ADJ	B-T, I-T	I-T
ELSE			O

Table 1: Transition Table

A potential jargon-term (PJ) is an actual jargon-term unless it is filtered out as follows. First, a jargon term J must meet all of these conditions:

1. J must contain at least one noun.
2. J must be more than one character long, not counting a final period.
3. J must contain at least one word consisting completely of alphabetic characters.
4. J must not end in a common abbreviation from a list (e.g., cf., etc.)
5. J must not contain a word that violates a morphological filter, designed to rule out numeric identifiers (patent numbers), mathematical formulas and other non-words. This rules out tokens beginning with numbers that include letters; tokens including plus signs, ampersands, subscripts, superscripts; tokens containing no alphanumeric characters at all, etc.
6. J must not contain a word that is a member of a list of common patent section headings.

Secondly, a jargon-term J must satisfy at least one of the following additional conditions:

1. J = highly ranked topic-term or a substring of J is a highly ranked topic-term.
2. J contains at least one O-NOUN.
3. J consists of at least 4 words, at least 3 of which are either nominalizations (C-NOUNs found in NOMLEX-PLUS (Meyers et al., 2004; Meyers, 2007)) or TECH-ADJs.
4. J = nominalization at least 11 characters long.
5. J = multi-word ending in a common noun and containing a nominalization.

A final stage aims to distinguish named entities from jargon-terms. It turns out that named entities, like jargon terms, include many out of vocabulary words. Thus we look for NEs among those PJs that remain after stage 3 and contain capitalized words (a single capital letter followed by lowercase letters). These NE filters are based on manually collected lists of named entities and nationality adjectives, as well as common NE endings. Dictionary lookup is used to assign GPE (ACE's Geopolitical Entity) to *New York* or *American*; LOC(ation) to *Aegean Sea* and *Ural Mountains*; and FAC(ility) to *Panama Canal* and *Suez Canal*. Plurals of nationality words, e.g., *Americans* are filtered out as non-terms. PJs are filtered by endings typically associated with non-terms, e.g., *et al* signals PJs as citations to articles and honorifics (Esq, PhD, Jr, Snr) signal PER(son) named entities. Finally, if at least one of the words in a multi-word term is a first or last person name, we can further filter them by endings, where ORGanization endings

include *Agency*, *Association*, *College* and more than 65 others; GPE endings include *Heights, Township, Park*; LOC(ation) endings include *Street*, *Avenue* and *Boulevard*. It turns out that 2 word capitalized structures including at least one person name are usually either ORG or GPE in our patent corpus, and we maintain this ambiguity, but mark them as non-terms.

We have described a first implementation of a jargon-term chunker based on a combination of principles previously implemented in noun group chunking and topic-term extraction systems. The chunker can use essentially the same algorithms as previous noun group chunkers, though in this case we used a manual-rule based FSM. The extended POSs are defined according to conventional POS (representing substitutability, morphology, etc.), statistical topic-term extraction, OOV status (absence from our dictionary) or presence in specialized dictionaries (NOMLEX, dictionary of chemicals, etc.). We use topic-term extraction to identify both particular noun sequences (high-ranked topic-terms) and some of their components (STAT-ADJ), and could extend this strategy to other components, e.g., common head nouns. We approximated the concept of "rare word" by noting which words were not found in our standard dictionary (O-NOUN). As is well-known, "noun" and "adjective" are the first and second most frequent POS for OOV words and both POSs are typically found as part of noun groups. Furthermore, rare instances of O-NOUN (and OOV adjectives) are typically parts of jargon-terms. This approximation is fine-tuned by the addition of word lists (e.g., chemicals). In future work, we can use more distributional information to fine-tune these categories, e.g., we can use topic-term techniques to identify single topic words (nouns and adjectives) and experiment with these additional POS (instead of or in addition to the current POS classes).

4 The Annotator Definition of Jargon-Term

For purposes of annotation, we defined *jargon-term* as a word or multi-word nominal expression that is specific to some technical sublanguage. It need not be a proper noun, but it should be conventionalized in one of the following two ways:

1. The term is defined early (possibly by being abbreviated) in the document and used repeatedly (possibly only in its abbreviated form).
2. The term is special to a particular field or subfield (not necessarily the field of the document being annotated). It is not enough if the document contains a useful description of an object of interest – there must be some conventional, definable term that can be used and reused. Thus multi-word expressions that are defined as jargon terms must be somewhat word-like – mere descriptions that are never reused verbatim are not jargon terms. (Justeson and Katz, 1995) goes further than we do: they require that terms be reused within the document being annotated, whereas we only require that they be reused (e.g., frequent hits in a web search).

Criterion 2 leaves open the question of how specific to a genre an expression must be to be considered a jargon-term. At an intuitive level, we would like to exclude words like *patient*, which occur frequently in medical texts, but are also commonly found in non-expert, everyday language. By contrast, we would like to include words like *tumor* and *chromosome*, which are more intrinsic to technical language insofar as they have specialized definitions and subtypes within medical language. To clarify, we posited that a jargon-term must be sufficiently specialized so that a *typical naive adult* should not be expected to know the meaning of the term. We developed 2 alternative models of a naive adult:

1. *Homer Simpson*, an animated TV character who caricatures the typical naive adult–the annotators invoke the question: *Would Homer Simpson know what this means?*
2. **The Juvenile Fiction sub-corpus of the COCA:** The annotators go to `http://corpus.byu.edu/coca/` and search under FIC:Juvenile – a single occurrence of an expression in this corpus suggests that it is probably not a jargon-term.

In addition, several rules limited the span of terms to include the head and left modifiers that collocate with the heads. Decisions about which modifiers to include in a term were difficult. However, as this

			Strict				Sloppy			
	Doc	Terms	Matches	Pre	Rec	F	Matches	Pre	Rec	F
Annot 1	SRP	1131	798	70.8%	70.6%	70.7%	1041	92.5%	92.0%	92.2%
	SUP	2166	1809	87.5%	83.5%	85.5%	1992	96.3%	92.0%	94.1%
	VVA	919	713	90.9%	77.6%	83.7%	762	97.2%	82.9%	89.5%
Annot 2	SRP	1131	960	98.4%	84.9%	91.1%	968	99.2%	85.6%	91.9%
	SUP	2166	1999	95.5%	92.3%	93.8%	2062	98.5%	95.2%	96.8%
	VVA	919	838	97.4%	91.2%	94.2%	855	99.4%	93.0%	96.1%
Base 1	SRP	1131	602	24.3%	53.2%	33.4%	968	44.2%	96.8%	60.7%
	SUP	2166	1367	36.5%	63.1%	46.2%	1897	50.6%	87.6%	64.2%
	VVA	919	576	28.5%	62.7%	39.2%	887	44.0%	96.5%	60.4%
Base 2:	SRP	1131	66	24.9%	5.8%	9.5%	151	57.0%	13.4%	21.6%
	SUP	2166	771	52.3%	35.6%	42.4%	1007	68.4%	46.5%	55.3%
	VVA	919	270	45.8%	29.4%	35.8%	392	66.5%	42.6%	51.9%
System Without Filter	SRP	1131	932	39.0%	82.4%	53.0%	1121	46.9%	99.1%	63.7%
	SUP	2166	1475	39.7%	68.1%	50.2%	1962	52.8%	90.6%	66.7%
	VVA	919	629	27.8%	68.4%	39.5%	900	39.8%	97.9%	56.6%
System	SRP	1131	669	69.0%	59.2%	63.7%	802	82.8%	70.9%	76.4%
	SUP	2166	1193	64.7%	55.1%	59.5%	1526	82.8%	70.5%	76.1%
	VVA	919	581	62.1%	63.2%	62.7%	722	77.2%	78.6%	77.9%

Table 2: Evaluation of Annotation, Baseline and Complete System Against Adjudicated Data

evaluation task came on the heels of the relation extraction task described in section 6, we based our extent rules on the definitions and the set of problematic examples that were discussed and cataloged during that project. This essentially formed the annotation equivalent of case-law for extents. We will make our annotation specifications available on-line, along with discussions of these cases.

5 Evaluation

For evaluation purposes, we annotated all the instances of jargon-terms in a speech recognition patent (SRP), a sunscreen patent (SUP) and an article about a virus vaccine (VVA). Each document was annotated by 2 people and then adjudicated by Annotator 2 after discussing controversial cases. Table 2 scores the system, annotator 1 and annotator 2, by comparing each against the answer key providing: number of terms in the answer key, number of matches, precision, recall and F-measure. The "strict" scores are based on exact matches between system terms and answer key terms, whereas the "sloppy" scores count as correct instances where part of a system term matches part of an answer key term (span errors). As the SRP document was annotated first, some of specification agreement process took place after annotation and the scores for annotators are somewhat lower than for the other documents. However, Annotator 1's scores for SUP and VVA are good approximations of how well a human being should be expected to perform and the system's scores should be compared to Annotator 1 (i.e., accounting for the adjudicator's bias).

There are 4 system results: two baseline systems and two stages of the system described in section 3. Baseline 1 assumes terms derived by removing determiners from noun groups – we used an MEMM chunker using features from the GENIA corpus (Kim et al., 2003). That system has relatively high recall, but overgenerates, yielding a lower precision and F-measure than our full system – it is also inaccurate at determining the extent of terms. Baseline 2 restricts the noun groups from this same chunker to those with O-NOUN heads. This improves the precision at a high cost to recall. Similarly, we first ran our system without filtering the potential jargon-terms, and then we ran the full system. Clearly our more complex strategy performs better than these baselines and the linguistic filters increase precision more than they reduce recall, resulting in higher F-measures (though low-precision high-recall output may be better for some applications).

6 Relations with Jargon-Terms

(Meyers et al., 2014) describes the annotation of 200 PubMed articles from and 26 patents with several relations, as well as a system for automatically extracting relations. It turned out that the automatic system depended on the creation of a jargon-term extraction system and thus that work was the major motivating factor for the research described here. Choosing topic-terms as potential arguments would have resulted in low recall. In contrast, allowing any noun-group to be an argument would have lowered precision, e.g., *diagram*, *large number*, *accordance* and *first step* are unlikely to be valid arguments of relations. In the example: *The resequencing* **pathogen microarray**$_{A2}$ *in the diagram is a promising new technology.*, we can detect that the authors of the articles view *pathogen microarray* as significant, and not the NG *diagram*. By selecting jargon-terms as potential arguments we are selecting the most probable noun group arguments for our relations. For the current system (which does not use a parser), the system performs best if non-jargon-terms are not considered as potential relation arguments at all. However, one could imagine a wider coverage (and slower) system incorporating a preference for jargon-terms (like a selection restriction) with dependency-based constraints.

We will only describe a few of these relations due to space considerations. Our relations include: (1) **ABBREVIATE**, a relation between two terms that are equivalent. In the normal case, one term is clearly a shorthand version of the other, e.g., "The **D. melanogaster gene Muscle LIM protein at 84B**$_{A1}$ (abbreviated as **Mlp84B**$_{A2}$)". However, in the special case (**ABBREVIATE:ALIAS**) neither term is a shorthand for the other. For example in "**Silver behenate**$_{A1}$, also known as **CH3-(CH2)20-COOAg**$_{A2}$", the chemical name establishes that this substance is a salt, whereas the formula provides the proportions of all its constituent elements; (2) **ORIGINATE**, the relation between an ARG1 (person, organization or document) and an ARG2 (a term), such that the ARG1 is an inventor, discoverer, manufacturer, or distributor of the ARG2 and some of these roles are differentiated as subtypes of the relation. Examples include the following: "**Eagle**$_{A1}$**'s minimum essential media**$_{A2}$ and **DOPG**$_{A2}$ was obtained from **Avanti Polar Lipids**$_{A1}$". (3) **EXEMPLIFY**, an IS-A relation (Hearst, 1992) between terms so that ARG1 is an instance of ARG2, e.g., "**Cytokines**$_{A2}$, for instance **interferon**$_{A1}$"; and "**proteins**$_{A2}$ such as **insulin**$_{A1}$"; (4) **CONTRAST** relations, e.g., "**necrotrophic effector system**$_{A1}$ that is an exciting contrast to the **biotrophic effector models**$_{A2}$"; (5) **BETTER_THAN** relations, e.g., "**Bayesian networks**$_{A1}$ hold a considerable advantage over **pairwise association tests**$_{A2}$"; and (6) **SIGNIFICANT** relations, e.g., "**Anaerobic SBs**$_{A2}$ are an emerging area of research and development" (ARG1, the author of the article, is implicit). These relations are applicable to most technical genres.

7 Concluding Remarks

We have described a method for extracting instances of jargon-terms with an F-measure of between 62% and 77% (strict vs sloppy), about 73% to 84% of human performance. We expect this work to facilitate the extraction of a wide reange of relations from technical documents. Previous work has focused on generating topic-terminology or term types, extracted over sets of documents. In contrast, we describe an effective method of extracting term tokens, which represent a larger percent of the instances of terminology in documents and constitute arguments of many more potential relations. Our work on relation extraction yielded very low recalls until we adopted this methodology. Consequently, we have obtained recall of over 50% for many relations (with precision ranging from 70% for OPINION relations like Significant to 96% for Originate.).

Acknowledgments

Supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center contract number D11PC20154. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/NBC, or the U.S. Government.

References

M. Bada, L. E. Hunter, M. Eckert, and M. Palmer. 2010. An overview of the craft concept annotation guidelines. In *The Linguistic Annotation Workshop, ACL 2010*, pages 207–211.

BioCreAtIvE. 2006. Biocreative ii.

M. Bundschus, M. Dejori, M. Stetter, V Tresp, and H. Kriegel. 2008. Extraction of semantic biomedical relations from text using conditional random fields. *BMC Bioinformatics*, 9.

P. Corbett, C. Batchelor, and S. Teufel. 2007. Annotation of chemical named entities. In *BioNLP 2007*, pages 57–64.

Thomas M. Cover and Joy A. Thomas. 1991. *Elements of Information Theory*. Wiley-Interscience, New York.

A. Frank. 2000. Automatic F-Structure Annotation of Treebank Trees. In *Proceedings of The LFG00 Conference*, Berkeley.

K. Frantzi, S. Ananiadou, and H. Mima. 2000. Automatic recognition of multi-word terms:. the C-value/NC-value method. *International Journal on Digital Libraries*, 3(2):115–130.

C. Giuliano, A. Lavelli, and L. Romano. 2006. Exploiting shallow linguistic information for relation extraction from biomedical literature. In *EACL 2006*, pages 401–408, Trento.

M. Hearst. 1992. Automatic Acquisition of Hyponyms from Large Text Corpora. In *ACL 1992*, pages 539–545.

T. Hisamitsu, Y. Niwa, S. Nishioka, H. Sakurai, O. Imaichi, M. Iwayama, and A. Takano. 1999. Term extraction using a new measure of term representativeness. In *Proceedings of the First NTCIR Workshop on Research in Japanese Text Retrieval and Term Recognition*.

Houghton Mifflin Company. 2001. *Webster's II New College Dictionary*. Houghton Mifflin Company.

C. Jacquemin and D. Bourigault. 2003. Term Extraction and Automatic Indexing. In R. Mitkov, editor, *Handbook of Computational Linguistics*. Oxford University Press, Oxford.

Y. Jin, M. Kan, J. Ng, and X. He. 2013. Mining scientific terms and their definitions: A study of the acl anthology. In *EMNLP-2013*.

J. S. Justeson and S. M. Katz. 1995. Technical terminology: some linguistic properties and an algorithm for identification in text. *Natural Language Engineering*, 1(1):9–27.

J. D. Kim, T. Ohta, Y. Tateisi, and J. I. Tsujii. 2003. Genia corpus–a semantically annotated corpus for bio-textmining. *Bioinformatics*, 19 (suppl 1):i180–i182.

C. Macleod, R. Grishman, and A. Meyers. 1997. COMLEX Syntax. *Computers and the Humanities*, 31:459–481.

C. Macleod, R. Grishman, A. Meyers, L. Barrett, and R. Reeves. 1998. Nomlex: A lexicon of nominalizations. In *Proceedings of Euralex98*.

A. Meyers, R. Reeves, C. Macleod, R. Szekeley, V. Zielinska, and B. Young. 2004. The Cross-Breeding of Dictionaries. In *Proceedings of LREC-2004*, Lisbon, Portugal.

A. Meyers, G. Lee, A. Grieve-Smith, Y. He, and H. Taber. 2014. Annotating Relations in Scientific Articles. In *LREC-2014*.

A. Meyers. 2007. Those Other NomBank Dictionaries – Manual for Dictionaries that Come with NomBank. `http:nlp.cs.nyu.edu/meyers/nombank/nomdicts.pdf`.

R. Navigli and P. Velardi. 2004. Learning Domain Ontologies from Document Warehouses and Dedicated Web Sites. *Computational Linguistics*, 30.

L. A. Ramshaw and M. P. Marcus. 1995. Text Chunking using Transformation-Based Learning. In *ACL Third Workshop on Very Large Corpora*, pages 82–94.

A. Schwartz and M. Hearst. 2003. A simple algorithm for identifying abbreviation definitions in biomedical text. In *Pacific Composium on Biocomputing*.

T. Tomokiyo and M. Hurst. 2003. A language model approach to keyphrase extraction. In *ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*.

P. Velardi, M. Missikoff, and R. Basili. 2001. Identification of relevant terms to support the construction of domain ontologies. In *Workshop on Human Language Technology and Knowledge Management - Volume 2001*, pages 5:1–5:8.