On Breadth Alone: Improving the Precision of Terminology Extraction Systems on Patent Corpora

Sean Nordquist
New York University
nordquist@nyu.edu

Adam Meyers
New York University
meyers@cs.nyu.edu

Abstract

Automatic Terminology Extraction (ATE) methods are a class of linguistic, statistical, machine learning or hybrid techniques for identifying terminology in a set of documents. Most modern ATE methods use a statistical measure of how important or characteristic a potential term is to a foreground corpus by using a second background corpus as a baseline. While many variables with ATE methods have been carefully evaluated and tuned in the literature, the effects of choosing a particular background corpus over another are not obvious. In this paper, we propose a methodology that allows us to adjust the relative breadth of the foreground and background corpora in patent documents by taking advantage of the Cooperative Patent Classification (CPC) scheme. Our results show that for every foreground corpus, the broadest background corpus gave the worst performance, in the worst case that difference is 17%. Similarly, the least broad background corpus gave suboptimal performance in all three experiments. We also demonstrate qualitative differences between background corpora – narrower background corpora tend towards more technical output. We expect our results to generalize to terminology extraction for other legal and technical documents and, generally, to the foreground/background approach to ATE.

1 Introduction

Terminology extraction is the process by which specialized and domain-specific words and phrases are extracted from a set of documents. These techniques are actively used and researched to identify trends in technical documents, create domain glossaries, and improve the readability of technical documents, among their many uses. Automatic Terminology Extraction (ATE) methods are the class of linguistic, statistical, machine learning, or hybrid techniques designed to identify terminology in a specialized set of documents. ATE now covers a broad class of methods that are in real use today and continues to receive research attention.

Most modern ATE methods take advantage of a statistical measure of how important or characteristic a potential term is to a foreground corpus by using a second background corpus as a baseline. Systems that use these statistics rely on an assumption that the foreground corpus is specialized and the background corpus is less specialized. The statistical methods can then use relative frequencies in the less specialized corpus and compare them to the specialized corpus – if a term is significantly more common in the specialized than the unspecialized, we may have identified a domain-specific term. Techniques that use this statistical strategy work well. While many variables with ATE methods have been carefully evaluated and tuned in the literature, the effects that come from choosing a particular background corpus over another are not obvious. More specifically, what would happen if one were to use a more broad background corpus that contained a wider variety of subject matter?

This paper presents an experiment carried out with Termolator (Meyers et al., 2018), a high-performing open-source ATE system. The system allows for the specification of a foreground corpus consisting of the target topic area and a background corpus that can be customized. We explore the results from running this test on three distinct patent topic areas, using the Cooperative Patent Classification (CPC) scheme to curate five background corpora for each foreground. Our results show that the choice of background corpus has a significant effect on the precision of the words extracted.

For every foreground corpus, the broadest background corpus gave the worst performance, in the worst case that difference is 17%. Similarly, the least broad background corpus gave suboptimal performance in all three experiments. Indeed, the ideal background corpus seemed to occupy some middle position – broader than the foreground cor-

Proceedings of the Natural Legal Language Processing Workshop 2022, pages 1 - 11

pus, but not too general either. For example, we found that highest results (72% precision) for a foreground of semiconductor (H01L 21) patents was derived from a background of patents related to electricity (H), which is more general than "electric solid state devices" (HOIL) and more specific than patents in general or than a combination of patents and non-patents.

We perform a qualitative analysis of words extracted to see how different background breadths affect the words extracted. For example, when the top 100 term candidates from the data input patent foreground corpus were analyzed, the most general background corpus produced a set of terminology that, while technical, was less characteristic of data inputs than the all patent background corpus (e.g., the most general corpus: *fingerprint sensor*, *social media* vs. a patent corpus: *focal vergence*, *selectable interaction element*).

We expect that our results will generalize to terminology extraction for other legal and technical documents and, generally, to the foreground/background approach to ATE.

2 Related Work

The definition of 'terminology' in the context of ATE systems is still a point of discussion in modern literature (Rigouts Terryn et al., 2020). In this study we use the word terminology to describe specialized language that is domain specific. Notionally, we distinguish a word or phrase as terminology if it is sufficiently specialized that a typical naive adult would not be expected to know the meaning of the term (Meyers et al., 2018).

ATE methods are generally split into 3 different categories: linguistic, statistical, and hybrid. Linguistic methods use linguistic features such as parts of speech patterns and chunking to extract term candidates. Statistical methods usually use a statistical measure of how characteristic a term is to a foreground corpus by comparing it to a baseline background corpus. Hybrid methods combine the linguistic and statistical methods, usually by using linguistic methods to identify term candidates and the statistical methods to rank the candidates.

The statistics in the hybrid methods work by comparing a foreground corpus from which terminology should be extracted, with a background corpus which serves as a baseline to identify terms characteristic of the foreground. The use of a foreground and a background corpus (or sometimes an analysis and reference corpus, respectively) has existed for a long time (e.g. (Kageura and Umino, 1996; Tomokiyo and Hurst, 2003; Drouin, 2003)). The intuition is that by using and combining these statistics, one can rank the words and phrases which are most likely to be specialized language from the foreground higher. A variety of statistics have been used in the literature (e.g. TF-IDF, KL divergence, etc.) (Kosa et al., 2020).

The assumption behind using a foreground and background corpus is that the foreground is sufficiently specialized and the background corpus is sufficiently general that the way they use potential terms will be different. This assumption is powerful and effective and has led some research to stick to a single general background corpus (Drouin, 2003) and some research to allow varying background corpora (Meyers et al., 2018).

By taking advantage of both linguistic and statistical techniques, hybrid methods have proven to be some of the most effective in ATE for the last decade (Macken et al., 2013; Rigouts Terryn et al., 2020). While most systems now fall into the hybrid category, there is growing interest in machine learning methods for ATE with a variety of methodologies (Kucza et al., 2018; Hätty and Schulte im Walde, 2018). In this paper we use an open-source hybrid method called Termolator that combines chunking and statistical ranking of term candidates using two corpora: the foreground corpus and the background corpus (Meyers et al., 2018).

Termolator is a flexible hybrid ATE system that allows us to vary the background corpus for a given foreground corpus. We are assuming that Termolator is representative of other hybrid systems which use a foreground and background corpus in the same way. We believe this is a valid assumption because such ATE systems are based on the idea that comparing the distribution of terms candidates across two different corpora helps identify them. Terms that appear frequently in foreground documents but not background documents are more likely to be terms and vice versa. We do not make any assumptions about the relative performance of Termolator and other comparable systems.

In this work, we focus on patents, a technical document in the legal domain, and the relationship between foreground and background corpora. We examine how the choice of background corpora might affect the performance of existing systems and the output of those systems.

Drouin et al. (2020) discussed how the distance between foreground and background corpora affects terms in unspecialized corpora. However, their paper focuses on design choices to optimize ATE for unspecialized corpora, like news articles.

3 Experimental Setup

3.1 Data Set

Patents will be the main document of study. We used the United States Patent Office Bulk Storage System to download all patent grants from the years 2016 to 2022. This will be the set of patents we sample from to construct our corpora. We also combine the Open American National Corpus (OANC) (Ide and Suderman, 2006) with a sample of patents to construct our general corpus.

3.2 Foreground Corpora

To better understand the generalizability of our results across other patent subject areas, we conduct experiments using corpora in three different subject areas. Each foreground corpus corresponds to a CPC classification code that corresponds to a particular "group" in the CPC scheme. The corpus is created by sampling 5,000 patents from each of these "groups." We chose to sample from the "group" rather than the "subgroup" because in most cases subgroups did not have enough patents in that time period for the experiment.

Table 1 shows the patent CPC codes from which we will sample documents for our foreground corpora. We select these three topic areas because they provide a good range of technical topics and types of terminology to test across.

3.3 Breadth of Corpora

We define the breadth of a patent corpus as how much variety in subject matter there is in the corpus. Reducing the problem of breadth to similarity opens us to a significant amount of existing research in computational linguistics on the problem.

Understanding how semantically similar two sets of words, documents, or corpora are is an important problem in natural language processing. Saying two texts are similar relies on an explicit normative definition of what makes them similar (Bär et al., 2011). Without a taxonomy that all speakers of every language agree on, little can be done to create a universal concept of similarity. A specialist, for example, has a richer and deeper ontology than a layman that will change the relative similarities of words and concepts. The precise layout of that ontology is based on circumstances such as what was being researched at the time and the interests of the people involved. Even word embeddings – our best attempt at making the problem numeric – do not assign a transparent measure of magnitude to semantic similarity (Faruqui et al., 2016).

Reconciling all the potential taxonomies that exist or that could exist is beyond this paper. We need not, however, look at precisely how much broader a corpus is than another, just the fact that it is broader. If we examine breadth as a measure that monotonically increases with the addition of dissimilar documents, we can define an ordinal notion of breadth that would serve our purpose. In other words, we need not look at precisely how much more broad a corpus is than another, just the fact that it is more broad. In effect, we create a rank-ordering of our patent corpora that will correspond to five different breadths (Stevens, 1946).

3.4 Background Corpora

To create our background corpora, we use the Cooperative Patent Classification (CPC) scheme. CPC is a classification system that classifies all US patent grants. The CPC scheme defines a hierarchy that organizes patents into sections, classes, subclasses, groups, and subgroups (Table 2) (USPTO, 2016–2022). As one moves down the hierarchy, one describes an increasingly specific set of patents. The CPC scheme thus describes a tree of classifications with the patents themselves at the leaves. Patents are always assigned a 'main' category which we will focus on. Each patent's main classification is a code in the format of "H01L 21/02".

We create a total of five background corpora of increasing breadth: one CPC level removed, two CPC levels removed, three CPC levels removed, a corpus sampled from all patent topics, and a general corpus composed of the OANC mixed with a sample of patents, which we will refer to as OANC+ [1]. To illustrate the curation process, we use a "F03G 7" foreground corpus as an example. The first background corpus is sampled from "F03G" – one level above in the hierarchy. The second background corpus is sampled from "F03". The third background corpus is sampled from "F". Finally, we create a general patent corpus, by sampling from all CPC classification codes, which we

[1]We have released a version of OANC+ to the public at the following link: https://drive.google.com/file/d/1VNFzZb6DyrNozBxiBcf07C83A13PM0RS/view

H01L 21	"Semiconductors"	Processes or apparatus adapted for the manufacture or treatment of semiconductor or solid state devices or of parts thereof
A61B 17	"Surgical Instruments"	Surgical instruments, devices or methods
G06F 3	"Data Input"	Input arrangements for transferring data to be processed into a form capable of being handled by the computer; Output arrangements for transferring data from processing unit to output unit

Table 1: The 3 patent classes that make up our 3 different foreground corpora and our labels for them.

H01L 21/02	Section H/Class 01/Subclass L/Group 21/Subgroup 02
H01L 21	Section H/Class 01/Subclass L/Group 21
H01L	Section H/Class 01/Subclass L
H01	Section H/Class 01
H	Section H

Table 2: Cooperative Patent Classification hierarchy breakdown for a CPC code 'H01L 21/02'.

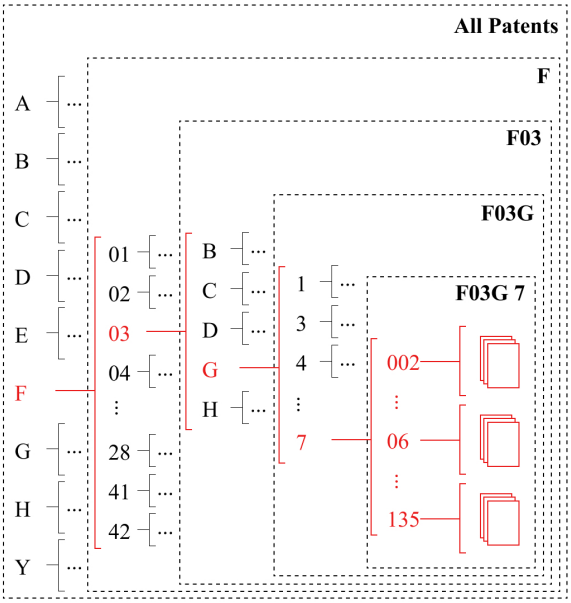

Figure 1: Diagram of the four different sampling levels for a foreground corpus 'F03G 7'. Each increasingly broad background corpus is constructed by sampling one level higher in the CPC hierarchy.

refer to as the 'All Patent' corpus. We repeat this process for each of the foreground corpora by simply moving up the hierarchy in each case. We see this sampling process visually in Figure 1. At each level we sample from a progressively wider range of patents.

Finally, we also create a 'general' background corpus. Ideally, the general background corpus will consist of a variety of different types of documents (court decisions, scholarly papers, patents, news articles, etc.) that will serve as our broadest possible background corpus. In addition, because we are attempting to create a broader corpus – not just a general contrasting corpus – we will also include a sample of patents to make it broader according to our definition. We use 5,000 total documents from both the OANC and a general sampling of patents to create our OANC+ background corpus (87.5% OANC documents and 12.5% patents). Now we have five total background corpora of five distinct breadths, each of which will be run against our foreground corpora using Termolator.

3.5 Annotation and Evaluation

For the purpose of annotation, we follow the convention given in Meyers et al. (2018). We define a valid term as a word or multi-word nominal expression that is specific to some technical field. A valid term should be definable within the field and reused. We do not consider term-like phrases to be valid terms unless they are reused verbatim either in the same or other documents. Next, we also require that valid terms be sufficiently specialized to a field's technical language. For a term to be considered specialized, a naive adult should not be expected to know the meaning of the term. We adopt the same intuitive model as Meyers et al. (2018), asking would Homer Simpson – an animated television character who is a caricature of a naive adult – know this term?

Following the evaluation strategy from Meyers

et al. (2018), we randomly sample 20 terms from each fifth of the output for a total of 100 terms. Then, we manually annotate each term as valid terminology or not. From these annotated terms, we can calculate a precision score that corresponds to that run of Termolator.

To calculate recall, one would need to annotate every document in the foreground corpus. For that reason, calculating recall is a labor-intensive and time consuming process when working with large corpora. This task is uniquely difficult because the experiment uses three 5,000 document-wide foreground corpora and therefore would require the annotation of 15,000 patents. In addition, the annotation of these three particular patent corpora do not serve a larger purpose at the moment. We make a preliminary effort nonetheless to examine a potential proxy for recall obtained by annotating a small subsample of documents in 4.3.

We also examine the words themselves. Specifically, we want to look at how the words change as the background corpora change. First, we examine how the outputs change by determining agreement between the output's top 100 words. We also perform a qualitative investigation of the terms extracted with each background corpus. We do this by looking at where the outputs disagree and examining those differences.

4 Results

4.1 Precision Scores

Table 3 presents the precision scores across experiments for all three foreground corpora. The scores that correspond to the best-performing background corpora for each analysis corpus are in bold.

Generally, we see the hybrid ATE method used by Termolator works better on some patent topics than others. For all three foreground corpora, we tend to achieve the lowest precision with the most general background corpus consisting mostly of non-patent documents. Interestingly, the highest precision is achieved at neither the most specific corpus nor the most general corpus, which suggests that the breadth of the background corpus is a tunable parameter for hybrid ATE methods.

What happens to the precision scores? Examining the first foreground corpus consisting of semiconductor patents and its respective background corpora, we notice a clear break that occurs between 'H01' and 'H' on the CPC hierarchy. Between this break, precision jumps a full 11% from 61% to 72%. Precision falls marginally to 70% in the 'All Patents' corpus and falls all the way to 45% on the general corpus.

The second foreground corpus with surgical instrument patents is similar with a break occurring in the exact same place jumping 6% from 72% to 78%. Yet again the general corpus performed considerably worse than all other background corpus, achieving a precision of only 50%.

The third foreground corpus consisting of data input patents has a slightly different pattern. There is a break that occurs between 'G' and 'All Patents' of a considerable 11%. However the general corpus only performs marginally worse than the other narrower patent categories, namely 'G' and 'G06'.

Why do some background corpora perform better than others? For the semiconductor patents, the best performance (72%) was achieved when the foreground corpus was compared to a background corpus consisting of patents about electricity and electrical devices. Using a background corpus that consisted of only semiconductor related patents resulted in worse performance (64%). This is likely because the patents about semiconductors provide a background corpus that is too similar to the foreground corpus, as a result candidate terms which are terminology are ranked lower than they should be because they occur and co-occur too frequently in the background corpus.

A similar rationale could be applied to the surgical cutting instruments patents. The background corpus about surgical instruments performed much worse (70%) than the background corpus that consisted of patents for human necessities.

The data input patents, on the other hand, did not perform very well at all at the level where the other two foreground corpora performed the best. In fact, the second-worst performance was at that level (50%). Instead, the best performance by far was at the level of all patents (61%). This result may be because the data input patents appear in general to use less specialized language than the other two patent categories.

The general background corpus resulted in the worst performance in all three cases. This result indicates that the wide ranging classes of documents of various technical and non-technical types do not establish as good of a frequency and co-occurrence baseline as documents of the same type.

		H01L	H01	H	All Patents	OANC+
Semiconductors	H01L 21	0.63	0.61	**0.72**	0.70	0.45
		A61B	A61	A	All Patents	OANC+
Surgical Instruments	A61B 17	0.70	0.72	**0.78**	0.77	0.50
		G06F	G06	G	All Patents	OANC+
Data Input	G06F 3	0.55	0.50	0.50	**0.61**	0.47

Table 3: Precision scores of Termolator after being run on three distinct foreground corpora and their corresponding five background corpora of increasing breadth.

What do these results mean? This analysis reveals that there is not a set distance at which all background corpora can be placed optimally when extracting terminology from patents. In fact, it appears the optimal choice is dependent on the foreground corpus. Moreover, the results taken in full suggest that for each foreground corpus there exists some 'optimal' background corpus that can be used to optimize for precision. At this point, the breadth of the optimal background corpus seems to be a variable that needs to be tuned for.

Generally, however, we are able to give some specific prescriptions. Our results suggest that it is important to choose a background corpus that is composed of the same types of documents as your foreground corpus if enough of them exist. What this means in general is if one is running an ATE system on a set of scholarly papers about sorting algorithms, using news articles as a background corpus would likely not result in the best precision; rather, one would prefer to use a set of scholarly documents from all of computer science or perhaps scholarly documents from a range of disciplines as the background corpus.

4.2 Word Analysis

Conducting any rigorous analysis of the qualities of these words is challenging and outside the scope of this paper; instead, we will focus on a qualitative analysis of observations from the words using the intuitive model we described in the annotation step. Each run (we discussed 15 runs above) of Termolator produces 5,000 output words. To narrow our investigation, we will only be looking at the top 100 words from each run.

We begin by examining how the top terms vary across the runs. A matrix is used to show the number of words each run, using each background corpus, agrees on. Next, because each output is from the same foreground corpus, many of the words across the top 100 term outputs will be shared,

	G06F	G06	G	All Patents	OANC+
G06F	100	91	86	85	75
G06		100	88	85	73
G			100	91	79
All Patents				100	82
OANC+					100

Table 4: Number of terms shared in the output of the run with each background corpus with the Data Input 'G06F 3' foreground corpus.

however, we are most interested in what one background corpus picked up but another background corpus did not. For that reason, we will be looking at the term candidates the runs did not agree on. In other words, the term candidates that were extracted using one background corpus, but not the other, and vice versa. We will start our discussion with the patent category G06F 3.

Table 4 shows the share of the top 100 terms that are the same between each pair of background corpora used with patent class G06F 3. We notice that corpora that are further away from each other in the CPC hierarchy have fewer words in common. This difference is explained by the difference in the contents of the background corpora. This confirms that our notion of ordinal breadth of the background corpora has a significant effect on the top terms extracted. Specifically, the greatest disagreement occurs between the second most specific corpus (G06) and the most general corpus (general) with only 73% agreement. Whereas, the greatest agreement occurred between corpora that are adjacent in the hierarchy (G06F and G06; G and All Patents).

Table 5 shows the term candidates extracted using the All Patent background corpus but not the OANC+ background corpus in the left column and the vice versa in the right column. Term candidates

All Patent But Not OANC+	OANC+ But Not All Patent
EXTENSION APP	FINGERPRINT SENSOR
DATA PROCESSING ENGINE	TARGET VOLUME
VEHICLE DATA PARAMETER	SOCIAL MEDIA
SELECTABLE INTERACTION ELEMENT	HEAD NODE
SELECTABLE INTERACTION	VIEW ANGLE
MULTI-FUNCTIONAL INPUT BUTTON	SURROUND VIEW
HIGHLIGHT MESSAGE	PHY
GRAPHICAL ASSET	DISPLAY VIEW
FOCAL VERGENCE	DETECTOR ELEMENT
ENVIRONMENT CONTENT	VIBRATION DEVICE
CLIP AREA	UNIT MEMORY
USER INPUT ATTACHMENT	SUBARRAY
UNIT TOUCH	SERVICE REQUEST
TOUCH SENSOR SURFACE	SELECTION INDICATOR
TOUCH NODE	PRESENTATION DEVICE
PROCESS MANAGEMENT SERVICE	OPERATION REGION
POSITION POINTER	MULTI-FUNCTIONAL
PORTABLE MEDIA DEVICE	INPUT METHOD EDITOR
...	...

Table 5: Potential terms that were extracted using the All Patent background but not OANC (left column) and the OANC but not the All Patent background (right column) with the 'G06F 3' foreground corpus.

All Patent But Not OANC+	OANC+ But Not All Patent
REMOVAL MAP	HEATER ELEMENT
Q-CARBON	SHIELD PLATE
PROTECTOR LAYER	LIQUID LEVEL
N-TYPE GALLIUM OXIDE SUBSTRATE	FLUID MIXTURE
LIQUID NOZZLE	DIW
GROUND SECTION	DEVICE PACKAGE
FRONT OPENING UNIVERSAL POD	CARRIER STRUCTURE
CERAMIC POROUS BODY	SIDEWALL STRUCTURE
VERTICAL SEMICONDUCTOR FIN	CONDUCTIVE POWDER
THERMAL CENTER	BIAS GENERATOR
SURFACE WF	TUNNEL FET
POLYOLEFIN SHEET	STRESS LAYER
OPTICAL MATERIAL LAYER	EPITAXIAL FIN
MEOL LAYER	CARRIER WAFER
III-V COMPOUND LAYER	CARBON PRECURSOR
HOLDING ARM	C1-C10
...	...

Table 6: Potential terms that were extracted using the All Patent background but not OANC (left column) and the OANC+ but not the All Patent background (right column) with the ' H01L 21' foreground corpus.

	H01L	H01	H	All Patents	OANC+
H01L	100	85	79	76	69
H01		100	79	78	72
H			100	88	81
All Patents				100	84
OANC+					100

Table 7: Number of terms shared in output of the run with each background corpus with the Semiconductor 'H01L 21' foreground corpus.

extracted using the general background corpus are more likely to be well-formed words or phrases that are not terms in our sense of the word (*fingerprint sensor, social media, multifunctional, etc.*). Generally, the terms extracted using the OANC+ background corpus appear to be less specialized and more accessible to a naive adult.

In contrast, term candidates extracted using the base background corpus have on average greater length and apparently more specialized subject matter (*data processing engine, portable media device, focal vergence, etc.*). Even the simpler terms candidates extracted using the base background corpus (*clip area, graphical asset, touch node, etc.*) refer to specialized subject matter. Nonetheless, there are exceptions. For instance, *PHY* is short-hand for the physical layer in the *Open Systems Interconnection* model which is quite a bit more specialized than the other terms in the column.

We shift our analysis to the patent class H01L 21 in Table 7. Again, agreement appears to be decreasing in distance in the CPC hierarchy. The lowest agreement occurs between the least broad (H01L) and the most broad (OANC+) background corpora with 69% agreement. This result lines up with our expectations.

As seen in Table 6, there is not as clear of a separation between the types of words extracted using the All Patents background corpus and the OANC+ background corpus as there were in the previous patent class tested. Both sets of words appear to contain term candidates that a naive adult would not be expected to know (*optical material layer, MEOL layer, front opening universal pod, etc.* vs. *bias generator, epitaxial fin, carrier wafer, etc.*). This is likely due to the nature of the terminology in patents about semiconductors. Namely, it is, on average, a more specialized subject matter than data input patents and requires the description of concepts that are more advanced concepts in physics and chemistry.

Nonetheless, there do appear to be more basic term candidates extracted using the OANC+ background corpus than the All Patents background corpus (*heater element, shield plate, liquid level, fluid mixture, device package*). There are exceptions, however (*carrier wafer, epitaxial fin, carbon precursor*).

We also performed the same analysis for the surgical instrument patents with results similar to the semiconductor patents included in Appendix A.

4.3 Preliminary Recall Scores

One possible solution to calculating recall on such a large corpus is randomly sampling documents to annotate. For this sample, one would want to ensure that their sampling is representative of the 5,000 documents. Take the data input patents foreground corpus for example. We obtained the foreground by selecting 5,000 patents that shared the G06F 3 group level, meaning that there are even more granular classification of patents under the G06F 3 level (over 200 subgroups). To properly represent these subgroups, one should collect a number of patents from each subgroup proportional to how the subgroups are represented in the foreground corpus. Therefore, even with sampling, recall proves to be an expensive metric to calculate.

Nonetheless, in an attempt to find a proxy for recall for one of our experiments, we manually annotated 10 patents that were randomly sampled from the data input foreground corpus. We then compared the correct terms found in these patents to the top 5,000 terms extracted using each background corpus to calculate a total of five recall scores. These results are shown in Table 8.

We observed that a significant portion of the correct terms in the patents are either specific to the document or a small sub-field and therefore appear with low frequency in the overall foreground corpus. One of the reasons for this is, although we sampled from patents in the same group, they still varied in subgroup so there was greater diversity in the subject matter than there would be at the subgroup level.

Moreover, the design of ATE systems is based on the distribution of terms across a large set of documents. Based on this distribution, a ranked list of

	G06F	G06	G	All Patents	OANC+
Data Input	0.061	0.049	0.061	**0.074**	**0.074**

Table 8: Recall scores obtained from a sample of 10 documents from running Termolator on one foreground corpus and its corresponding five background corpora of increasing breadth.

terms is produced. Terms that occur in many foreground documents are more likely to be detected than terms that occur in only a few documents. Zipf's Law tells us that it is likely that most of the terms will be relatively rare, but the "important" terms are likely to occur in many documents (the TF in TF-IDF stands for term frequency). Thus, if we look at individual documents, the recall of an ATE system designed to extract terms from a large corpus should be relatively low. However, if we could somehow manually examine a set of 5,000 documents and only pay attention to terms with a high frequency (100 times in the corpus, rather than five times or less), we might expect a system to achieve a higher recall, but only for these high-frequency words.

Low recall scores are also a consequence of the cut-off chosen and the construction of the task. The task is to extract the top 5,000 terms from the documents with high precision. Naturally in a set of documents as technical as patents there are significantly more terms than documents, resulting in lower recall. Adjusting the cut-off to, for example, 10,000 terms would result in higher recall and lower precision on those terms. We believe determining how to best choose this cut-off with different background corpora is worth investigating.

This is a preliminary investigation into recall. We believe more work should be done to investigate how recall changes as the breadth of the background corpus changes.

5 Future Work

In our experiment, we used a general corpus that was composed of a mixture of OANC and a subset of general patents. We made this choice because our focus was making broader corpora not contrasting corpora. Nonetheless, the effect of using a truly general corpus would be an important baseline to compare in future research.

We limited our evaluation in this paper to precision and a qualitative analysis of the words themselves. We believe it would be relevant to devise a methodology that would allow us to further investigate the differences in the words extracted using the different background corpora.

A relevant extension would be to perform similar experiments using other document types. For instance, a natural extension would be to perform a similar set of experiments on medical scholarly text from PubMed or Wikipedia articles and examine if the trends we observed with patents remain true for other kinds of technical documents.

6 Conclusion

In this paper we investigated how varying the breadth of the background corpus affects hybrid ATE systems. After creating five background corpora for each foreground corpora using the CPC hierarchy, we ran three experiments on three different patent groups. We examined both the precision scores and the output words themselves. In this analysis, we were unable to find a single "best choice" for all patent classes. We found that for all three patent groups neither the narrowest nor the most broad background corpus achieved the best precision; rather, it was always a background corpus that consisted of patents that performed best. In addition, we found that the words we extracted varied with the background corpus we chose. For one patent class there was a clear separation between less specialized terms for the general corpus and the more specialized terms from the all patent corpus. This separation was not clear for the other two patent classes.

We showed that the choice of background corpus has a significant effect on the precision of the output of an ATE system. We found that optimizing for precision in all three cases meant choosing a patent only corpus. We also studied the words we extracted by comparing differences across runs. We found that the breadth of the corpora had a significant effect on the words extracted. Moreover, we informally analyzed how the words from the general background corpus differed from the patent background corpus, concluding that the term candidates were on average less specialized with the general corpus.

References

Daniel Bär, Torsten Zesch, and Iryna Gurevych. 2011. A reflective view on text similarity. In *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011*, pages 515–520, Hissar, Bulgaria. Association for Computational Linguistics.

Patrick Drouin. 2003. Term extraction using non-technical corpora as a point of leverage. *Terminology*, 9.

Patrick Drouin, Jean-Benoît Morel, and Marie-Claude L' Homme. 2020. Automatic term extraction from newspaper corpora: Making the most of specificity and common features. In *Proceedings of the 6th International Workshop on Computational Terminology*, pages 1–7, Marseille, France. European Language Resources Association.

Manaal Faruqui, Yulia Tsvetkov, Pushpendre Rastogi, and Chris Dyer. 2016. Problems with evaluation of word embeddings using word similarity tasks. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pages 30–35, Berlin, Germany. Association for Computational Linguistics.

Anna Hätty and Sabine Schulte im Walde. 2018. Fine-grained termhood prediction for German compound terms using neural networks. In *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, pages 62–73, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Nancy Ide and Keith Suderman. 2006. Integrating linguistic resources: The American national corpus model. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy. European Language Resources Association (ELRA).

Kyo Kageura and Bin Umino. 1996. Methods of automatic term recognition: A review. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication, Volume 3, Issue 2*.

Victoria Kosa, David Chaves-Fraga, Hennadii Dobrovolskyi, and Vadim Ermolayev. 2020. Optimized term extraction method based on computing merged partial c-values. In *Information and Communication Technologies in Education, Research, and Industrial Applications*, pages 24–49, Cham. Springer International Publishing.

Maren Kucza, Jan Niehues, Thomas Zenkel, Alex Waibel, and Sebastian Stüker. 2018. Term extraction via neural sequence labeling a comparative evaluation of strategies using recurrent neural networks. In *Interspeech 2018*, pages 2072–2076.

Lieve Macken, Els Lefever, and Véronique Hoste. 2013. Texsis: Bilingual terminology extraction from parallel corpora using chunk-based alignment. *Terminology*, 19.

Adam L. Meyers, Yifan He, Zachary Glass, John Ortega, Shasha Liao, Angus Grieve-Smith, Ralph Grishman, and Olga Babko-Malaya. 2018. The termolator: Terminology recognition based on chunking, statistical and search-based scores. *Frontiers in Research Metrics and Analytics*, 3.

Ayla Rigouts Terryn, Veronique Hoste, Patrick Drouin, and Els Lefever. 2020. TermEval 2020: Shared task on automatic term extraction using the annotated corpora for term extraction research (ACTER) dataset. In *Proceedings of the 6th International Workshop on Computational Terminology*, pages 85–94, Marseille, France. European Language Resources Association.

S. S. Stevens. 1946. On the theory of scales of measurement. *Science*, 103(2684):677–680.

Takashi Tomokiyo and Matthew Hurst. 2003. A language model approach to keyphrase extraction. In *Proceedings of the ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, pages 33–40, Sapporo, Japan. Association for Computational Linguistics.

USPTO. 2016–2022. Patent grant full text data (no images). *Bulk Data Storage System*.

A Word Analysis Tables for Surgical Instrument Patents

	A61B	A61	A	All Patents	OANC+
A61B	100	83	78	73	64
A61		100	88	80	72
A			100	78	72
All Patents				100	84
OANC+					100

Table 9: Number of terms shared in the output of the run with each background corpus with the Surgical Instruments 'A61B 17' foreground corpus.

All Patent But Not OANC+	OANC+ But Not All Patent
ROBOTIC DEBRIDEMENT APPARATUS	DISTAL BODY
TARGET VESSEL	DISTAL CROWN
SUPPORT CATHETER	CAMMING
CUTTING BLOCK	ATTACHMENT SIDE
CAMMING	TARGET VESSEL
ATTACHMENT SIDE	INTERSPINOUS PROCESS SPACING DEVICE
TUBULAR ELEMENT	FORMING POCKET ARRANGEMENT
DISTAL CROWN	DILATOR TUBE
CUTTING ASSEMBLY	COMPRESSIBLE ADJUNCT
CLAMP PAD	TUBULAR ELEMENT
PENETRATOR	SUPPORT CATHETER
OCCLUSION DEVICE	SCALPET ARRAY
INVENTIVE CONCEPT	SACROILIAC JOINT
INTERSPINOUS PROCESS	REMOVING DEVICE
FORMING SURFACE	MONOMER LIQUID
ENDOSCOPIC INSTRUMENT	CUTTING ASSEMBLY
DISTAL BASKET	BIOCOMPATIBLE LAYER
DILATOR TUBE	TISSUE THICKNESS COMPENSATOR
COMPRESSIBLE ADJUNCT	THROMBUS EXTRACTION DEVICE
SACROILIAC JOINT	THICKNESS COMPENSATOR
PIEZOELECTRIC ELEMENT	SURGICAL INSTRUMENT GUIDE
MICROBUBBLE	SHOCK WAVES
INTERSPINOUS PROCESS SPACING DEVICE	SCALPET DEVICE
I-BEAM	RETRACTION ELEMENT
FORMING POCKET ARRANGEMENT	RECEIVER MEMBER
BIOCOMPATIBLE LAYER	PERIANAL SUPPORT MEMBER
BASEPLATE	PERIANAL SUPPORT
TISSUE THICKNESS COMPENSATOR	PENETRATOR
...	...

Table 10: Potential terms that were extracted using the All Patent background but not OANC (left column) and the OANC but not the All Patent background (right column) with the 'A61B 17' foreground corpus.