Identifying Technology Terms in Document Text

Master's Thesis

Presented to

The Faculty of the Graduate School of Arts and Sciences

Brandeis University

Department of Computer Science

James Pustejovsky, Advisor

In Partial Fulfillment

Of the Requirements for

Master's Degree

by

Olga Cherenina

February 2013

ABSTRACT

Identifying Technology Terms in Document Text

A thesis presented to the Department of Computer Science

Graduate School of Arts and Sciences

Brandeis University

Waltham, Massachusetts

By Olga Cherenina

Finding technologies in the text of patents or other documents such as medical articles is a subtask of building a technology ontology. Building such a technology ontology was proposed by Brandeis scholars as part of the project aimed at patent classification based on a certain notion of availability of technologies relevant to the patent. Technology ontology represents a database of technologies evaluated by their availability within certain time frame, that is their maturity. Technology terms identification in the text of documents is an initial step necessary for building an ontology. The terms found in the text of the patent will reflect the notion of a technology and constitute the basis for technology maturity identification.

Here, we explore the efficiency of using natural language processing techniques to help identify technologies in patent text. We attempt at creating and using a matcher that uses lexical and syntactic features to look for technology terms. We address the problem of determining the concept of a technology which is important for the task and use an annotation for the evaluation of the matcher. Finally, we analyze the results and propose improvements to the system.

Table of contents

1. Introduction

1.1 Emergence

The attempt to identify technology terms in the text of documents is closely related to the increase in research of technological and scientific emergence. Before we talk about the goal of our project we would like to discuss the phenomenon of emergence and the two main ambiguities in the concepts that underlie our research, the ambiguity of the notion of emergence itself and the precision of the definition of 'technology'.

The question of emergence has been in the center of philosophical, scientific and artistic interest since the time of Aristotle. Emergence can generally be looked upon as arising of new systems and patterns as a result of previously existing systems interaction. The concept of emergence is highly ambiguous and requires a definition depending on the field of study of the phenomenon.

Automated detecting of emergence of new technologies has been one of the areas of research and development in many scientific fields. New domains in science and technology are formed, and changes in existing domains are produced, by human-directed, open-ended extensions and transformations of these networks. Individuals and groups draw on a variety of conceptual, material, and sociocultural resources in efforts to transform existing domains within science and/or technology

against a backdrop of a structured field of intentions. This intention-driven human activity causes networks to form or change. (Pickering, 1993)

In Computational Linguistics, detecting of technology emergence is attempted by means of processing a huge amount of linguistic data pertaining to the field where emergence may be or have taken place.

One of the studies of emergence is presented by BAE Systems. They develop ARBITER (Abductive Reasoning Based on Indicators and Topics of EmeRgence), a multidisciplinary study and development effort to analyze full-text and metadata for indicators of emergent technologies and scientific fields. (Brock, 2012) The goal of the study is to create a pragmatic theory of technological and scientific emergence based on the actant network theory and the notion of the robustness of actant networks. Thus, the ambiguity questions, the definition of emergence and technology, are addressed by this approach through defining every domain of science and technology as an actant network consisting of numerous elements such as individuals, institutions, instruments, practices, terminology, materials, funders, meetings, government organizations, laws, journals, patents, publications, and so on. (Brock, 2012)

Emergence is defined by the ARBITER study in terms of change or development of a network. Any change in a network is considered to be a novelty in the domain, an evolution of the existing network that can potentially develop into a new field of study. Therefore a novelty has antecedents within some existent networks. This argument agrees with the definition of emergence as development of

new systems as a result of existing systems interaction. The concept of emergence, which is usually associated with new domains that attract increasing interest, is treated as pertaining to an existent actant network by the ARBITER approach. They define emergence as *growth in the robustness of an actant network*. The robustness of an actant network in its turn can be roughly defined as the ability of the network to function despite the failure or removal of crucial actants (e.g. star researchers or major funders), grow and have significant effect on science and society. The robustness of a network is associated with such measures as extent, diversity and traffic (information flow).

Defining emergence as an increase of robustness means that the study looks at the concept as mostly referring to mature scientific domains. The study also treats emergence as a phase different from the phase of initial discovery. This approach may potentially make emergence identification a more workable task since discovery of new technologies can be very hard to trace.

The goal of the ARBITER team is to develop tools for identifying and examining the development of emergent fields of technology with the help of certain characteristics of actant networks that they consider important for the study of emergence. They plan on providing the users with emerging technologies and their characteristics by means of assessing the related document groups (RDGs) associated with actant networks within a very large corpus of publications and patents, in multiple languages, from a range of years. The flexibility of such an approach consists in the ability of the user to look at separate indicators of a

network in the process of its examination. Some important aspects of the novelty of their approach are a high speed of document processing, a large amount of documents and a diversity of the document sources.

This approach to technological emergence identification strongly relies on the hypothesis that for any technological field there exists a specific terminology that can be used for emergent technology identification. Major outputs of science and technology are terminological: publications, patents, and, more generally, the generation and communication of information. Within actant networks perhaps the primary type of traffic - the most common form of relationship - is the production and exchange of terminological information. Indeed, for human actors within such a network, their understanding of the technology or field is generated by their consumption of this linguistic traffic. (Latour, 2005)

There also exist discoverable patterns of terms that specifically characterize – that is, identify and bound – the field, technology, or sub-variant thereof. Following discovery of a characteristic terminology, additional measures and characterization of it are possible: novelty, growth, persistence, diversification, consolidation, and connectivity. These features of the characteristic terminology constitute measures of the existence and robustness vector of the actant network constitutive of the particular technology or field. (Brock, 2012)

The assumption of terminology existence and terminology traffic is central both for ARBITER and for our project of technology term identification. Our

research relies on technological terminology exchange through document text and the informativity of the language used in scientific publications.

1.2 Building a technology ontology as part of Patent Classification project

Our research on finding technology terms in the text of documents is a part of the Brandeis Genre Classification project, namely the Classification of Patents as one of the prevalent document types in the dataset. The relevance of genre classification research to the central question of emergence can be expressed in the following way:

The kind of community of actors that is indicative of an emergent concept is reflected in the literature in many ways. One way is the overall genre mix of publications on the concept. That is, the distribution over genres is different for an emergent concept as compared to a random set of documents or publications on a concept that is not emergent. (Verhagen, 2012, working draft)

This hypothesis elaborates on the main axiom of the ARBITER system stated in section 1.1. It basically claims that not only there exists special terminology that measures the existence of the actant network of a technology, but also such terminology may be classified to serve as an indicator of a specific document genre and vice versa – that documents of a specific genre can be used to extract information on emerging technologies.

One of the tasks of Genre Classification project is thus classifying patents. Patents are one of the dominant types of documents in the data set used by

Brandeis. Their classification appears to be closely correlated with the question of emergence identification because the notion of technology maturity is taken as one of the fundamental measures for patent classification by the Brandeis team. They hypothesize that maturity level of technologies referenced in a patent is an indicator of emergence in an RDG.

The approach proposes to classify patents based on the patent type and the maturity of technologies referenced in the patent. The type is derived from the often quoted classification by the U.S. Patent and Trademark Office (PTO). Type detection uses simple heuristics and extracts information from specific sections of the document.

The second dimention is concentrated around technologies referenced in the patent and their position in the life cycle, that is their maturity. The Brandeis team makes an attempt to give each technology an availability score by processing textual data. Technology maturity scores are assigned on three levels: technology level, patent level and RDG level. The patent level score is an aggregate score based on all technologies referenced in the patent and their maturity scores. The maturity score of the RDG is the average of scores of all patents in the RDG.

The system of patent classification designed by the Brandeis team consists of three major components: pattern generator, technology ontology builder and the runtime classification system (see figure 1 taken from (Verhagen, 2012, working draft) for a detailed description).

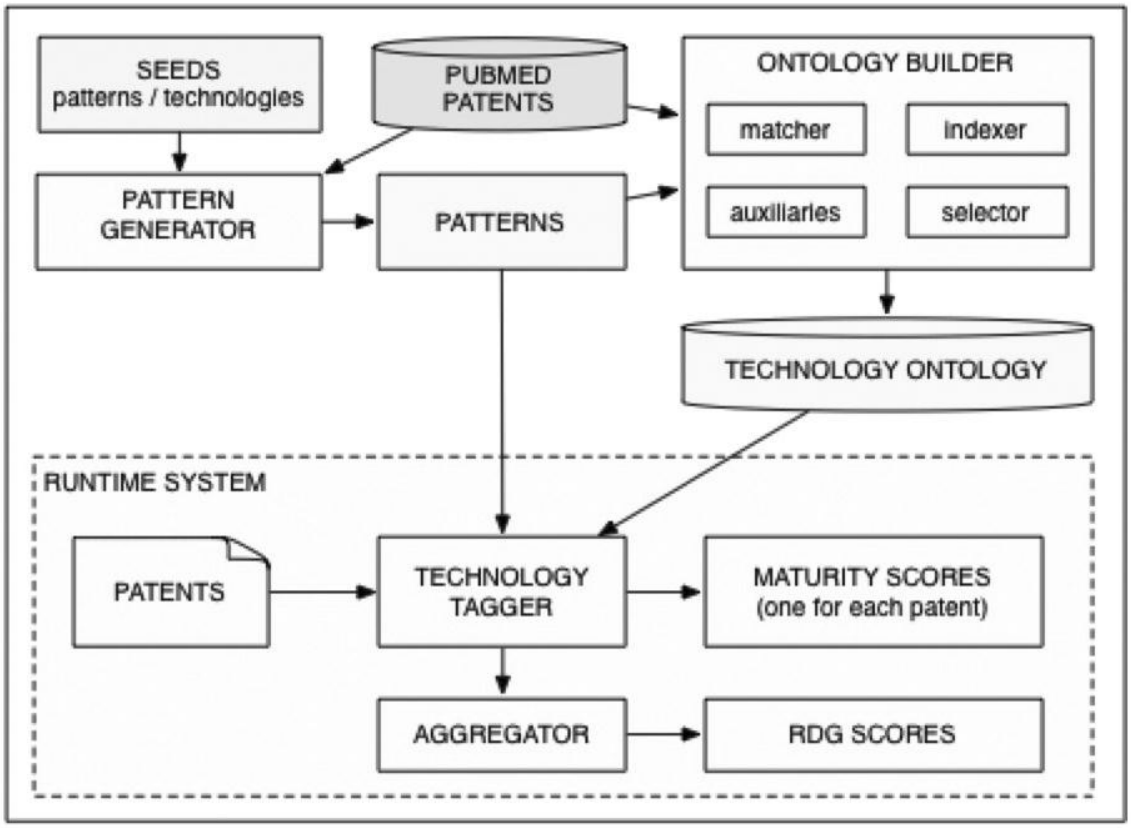


Figure 1.1: The System for Generating Maturity Scores

1.3 Pattern generation for building the technology ontology

Our experiments are related to the pattern generation part of the system. They are in fact initial attempts of using terminology and other linguistic information extracted from document text to generate contexts that may signify the presence of technology terms in a certain position in the sentence. Finding technology terms is necessary for building the ontology of these technologies. On other hand, the patterns themselves can serve as a source of information on the availability of the technology that it describes:

(1) TERM became widely available in YEAR TERM,

(2) an experimental approach to/for

(3) We used TERM to create X

(examples taken from (Verhagen, 2012))

The first example directly states that some technology became available in a certain year. The second one marks the development stage of some technology. The last one can potentially mean that a technology has become available in a certain year, the year being extracted from the title section of the document where this pattern occurred. It should be mentioned that patterns may generate false positives and therefore some filtering of the results needs to be done.

We also need to point out that at this stage of the patent classification project we do not deal with extracting any maturity information from the patterns that we attempt to generate and focus ourselves on developing a set of patterns and extracting technology terms from the text of documents with the help of these patterns. In other words, patterns that we aim to generate are going to be used simply to determine whether something is a technology or not and will be evaluated accordingly. We also do not concern ourselves with determining a set of technologies for each patent in the data set, but rather try to identify technologies that are mentioned in an RDG. Note that a set of patterns used for technology extraction may be different for different RDGs, for example contexts in which technology terms occur in Pubmed data may be different from those of the patents.

The goal of our project is thus to collect a set of technologies needed to constitute a frame for the technology ontology. Building the ontology will require additional evidence that would help identify whether the terms matched by the patterns are indeed

technologies. The ontology itself is in essence a matrix where pieces of evidence are mapped to years that constitute the life cycle of technologies (more on the proposed structure of the ontology in (Verhagen, 2012)).

The pattern generation process as part of building the technology ontology has three inputs: seed patterns, seed technologies and the Pubmed and LexisNexis source data. Pattern generation can be approached in two ways. One approach is to start with seed technologies and use bootstrapping to find contexts that often occur with technology terms. Another possible way is to start with seed patterns, see how well they perform and use lexical substitution and other linguistic information to generate more patterns. It is interesting that the processes of pattern generation and technology extraction are interrelated. We can both apply patterns to look for technologies and use technology terms we find to generate more patterns.

1.4 Initial experiments

For our project, we started with generating patterns by means of heuristics and lexical substitution. As our initial experiments on pattern generation, we applied a set of manually found seed patterns using lexical substitution to refine the pattern set.

The initial set of patterns is chosen based on a web research aimed at finding contexts where biomedical technologies often occur (Peter Anick, personal communication). We used simple lexical substitution to build on the list of contexts that were found to often appear with biomedical technology terms in google search. We used thesaurus look up to manually find synonyms and generate new patters.

For example, the dictionary gives the following synonyms of “innovation” which were used in new patterns: introduction, novelty, progress, advancement etc.

The set of patterns that came out of these initial experiments consisted of about 15 patterns. We applied these patterns to Pubmed data to find out if they could produce any useful matches. The pattern set was only useful to identify sentences that potentially contained technologies in them.

1.5 Project goal

We did not make any attempts to formally evaluate the results of our initial experiments. Their purpose was merely getting a sense of whether or not using heuristics and linguistic information can be indeed helpful for pattern generation. As our main experiment that we would like to describe, we decided to try the approach of seeding manually annotated technologies to generate patterns. We hypothesize that linguistic information can be useful in determining whether something that occurs in a certain context is a technology and that manual annotation of technology terms is a reliable way of evaluation of pattern performance. Our goal is to find out whether language patterns can be useful at all for finding technologies in the text of documents and what methods of pattern selection give better results. We ran some experiments on pattern selection and evaluated the pattern sets we found based on a gold standard annotation of technology terms. We describe our pattern generation approach in detail in the following two chapters.

2. Data

The data used for pattern generation is a random subset of 500 patents taken from the US Patent Office with XML annotations added to the documents by LexisNexis. The experiments were run on fragments of these patents that were extracted by the document structure parser. The fragments represent a limited set of sections most of which are abstracts.

The data that we used for pattern generation was preprocessed by Peter Anick as part of his work on an ML approach to technology identification. (Anick, personal communication) Patent text was tagged using the Stanford POS Tagger and chunked based on a small set of chunking rules. Each Noun Phrase was extracted from the text and considered as an NP in context. That means that the data was formatted to contain each NP and its context on a separate line. The sixteen features that substitute the context of an NP are listed in table 2.1. The tags that mark a particular NP in context are found by analyzing the structure of this NP and the structure of the surrounding context. The tags are chosen to contain mainly syntactic information that is potentially important for finding similar contexts for technologies. For instance, the tag 'prev_N' (previous noun) is meant to give information about such phrases as for example 'process of cloning' or 'process used for cloning' where 'cloning' would be considered an NP and 'process' would be the 'prev_N' feature that captures the information that a technology term is potentially

often preceded by ‘process of ‘, ‘process for’. Similarly, the tag ‘prev_V’, as in ‘curve-negotiating storage and retrieval vehicle’, is intended to find the verb that dominates the NP and excludes cases when a noun that is outside the given NP follows the verb.

Table 2.1. NP features.

prev_n2	Bigram that precedes the NP
prev_n3	Trigram that precedes the NP
next_n2	Bigram that follows NP
next_n3	Trigram that follows NP
section_loc	Location of the NP in a section of a document (patent). This tag also contains the information on whether the NP is located in the first sentence of the section or later.
last_word	Last word of the NP
next2_tags	Next two words’ POS tags
tag_sig	Succession of POS tags for this NP
initial_V	Verb in the initial position in the NP
prev_J	Adjective immediately preceding the NP
of_head	Noun that is the head of the object of the NP
initial_J	Adjective that starts the phrase
following_prep	Preposition immediately following the phrase
prev_N	First noun to the left of the chunk within 3 words.
prev_V	Closest verb to the left, including the particle if there is one.

Bellow we give examples of what an NP in context looks like in the input file. The NPs are followed by the sentences they were extracted from.

(1)

US20090032458A1.xml_1914 2009 biogas producing facility next_n3=is_intended_for next_n2=is_intended prev_n3=on_whether_the prev_n2=whether_the section_loc=DESC_later last_word=facility next2_tags=VBZ_VBN tag_sig=NNPS_VBG_NN prev_V=depending on

If necessary, the second separators, increase the dry matter content to in the order of 10-15% depending on whether the biogas producing facility is intended for livestock dung with a low dry matter content, or for livestock dung with high dry matter content.

(2)

US20090084970A1.xml_1614 2009 cooling panel energizing unit next_n3=for_energizing_the next_n2=for_energizing prev_n3=function_as_a prev_n2=as_a section_loc=DESC_later last_word=unit next2_tags=IN_VBG tag_sig=JJ_NN_VBG_NN initial_J=cooling following_prep=for prev_V=function as chunk_lead_J=cooling

Therefore, the temperature controller, the temperature sensor 138, and the fan of the temperature regulation control means do not function as a cooling panel energizing unit for energizing the cooling panel.

(3)

US6147681A.xml_603 2000 use of a temperature sensor next_n3=for_measuring_a next_n2=for_measuring prev_n3=relates_to_the prev_n2=to_the section_loc=SUMMARY_later last_word=sensor next2_tags=IN_VBG tag_sig=NN_IN_DT_NN_NN of_head=use following_prep=for prev_V=relates

The present invention also relates to the use of a temperature sensor for measuring a temperature adjacent the writing surface.

(4)

US20090032458A1.xml_1788 2009 reactor next_n3=may_be_maintained next_n2=may_be prev_n3=produced_in_the prev_n2=in_the section_loc=DESC_later last_word=reactor next2_tags=MD_VB tag_sig=NN prev_V=produced in

When the biological liquid has reached the decided temperature, communication with the biogas produced in the reactor may be maintained at least for at predetermined period, such as for the entire hydrolysis period.

In these four examples each NP in Context constitutes a single line of a text file where elements are tab separated. The first and second elements of each line are the document index and year of the document correspondingly; the third one is the NP itself; the rest of the elements are the context of this NP.

Examples given above show that if the data was processed for ML and all the NPs that they contain were considered technologies certain elements of their context might be useful for their identification. One example is the succession of tags (tag_sig) that contains a NOUN followed by a Verb in the ING-form which is followed by another Noun. This feature is illustrated by examples 1 and 2. Another useful feature could be having the word 'facility' or 'unit' as the last word of an NP that represents a technology term (examples 1 and 2). It might also turn out to be useful to have the last word of the NP ending in –or, such as 'sensor' in example 3 and 'reactor' in example 4. Seeing such verbs as 'function as' and 'produced in' in the context of the NP may be an indicator of a sentence talking about a technology as well (examples 2 and 4 correspondingly).

We used a set of 211614 of such NPs in Context produced by chunking a set of 500 patents as our training data. We tried to derive a set of patterns helpful in finding technologies using the context of these NPs as features for term occurrence in sentences.

3. Pattern generation

We look at patterns generation as a process of applying a set of manually found seed technologies and using feature frequency counting and lexical substitution to refine the pattern set. In this chapter we will describe the process in detail. It included the following steps:

1. using manually annotated technology terms to generate contexts for such terms;

2. selecting patterns using frequency counts and simple heuristics;

3.1 Using annotated technologies as seed data

As the first step of pattern generation we used a list of manually annotated terms to look for technologies. The seed file represented a list of terms classified by annotators as technologies or non-technologies. The consistency of this list in naming technologies is crucial for our task since this list will then be used as a benchmark for generation of technology term context patterns. It is therefore important to work out a clear and reliable definition of technology to use as the base for annotation.

Defining technology is a difficult task in itself. BAE system's understanding of emergence of technology as occurring in an actant network cannot be turned into a technical definition of technology. In some sense patterns themselves can be

considered operational definition of technologies as ”terms used in certain contexts”. (Verhagen, 2012, working draft) However, since we start with using seed technologies and not patterns, we need to develop a formal definition of technology that will serve as a basis for technology terms identification.

In our research we define technology loosely as “something that can be used in scientific research as a resource that helps further the goal of the research”. This could refer to ‘hard’ technologies like control rods, micro-processors and other physical technologies; medications; drugs and treatments (5-fluorouracil, adjuvant chemotherapy); and medical technologies (MIRNA target site, lambda DNA replication, hairpin probe). It also includes methods, processes, theories and algorithms (random matrix theory, support vector machines). Natural kinds (iron, carbon dioxide) and common artifacts like chair are not technologies. (Verhagen, 2012)

Technology term annotation is a tedious process for several reasons. Many annotation units can be considered as technologies or non-technologies depending on the context. However the list of terms for annotation included the terms taken out of the sentences where they occurred. Our technology definition is therefore is looked at separately from the context and the location of the term in the text of the document. In addition, most of the terms are not easily assigned to either of the two categories of our annotation model by the annotators due to either a lack of domain-specific knowledge or ambiguity in the way the term is used. The annotators relied on online

research as a helpful but time-consuming means of clarifying the meaning of such terms.

We extracted all nouns from a set of 1235 terms annotated as technologies. The nouns were used as textual clues to define whether a noun phrase in the training file is a technology term. We considered a noun phrase in the training file as a technology if it contained at least one of the nouns from our noun set. We thought that this approach was applicable for the kind of experiments we were conducting. It would help us derive a bigger list of potential technology terms. The NPs containing nouns from the set we used might contain technologies as their parts or could potentially be technologies that were not classified as such by the annotators. We found that in many cases the nouns from our noun set can be the heads of a number of technology terms or noun phrases containing such terms verses just one possible term. Examples of such head-nouns are 'apparatus', 'gear', 'system', 'network', 'process', 'transmission', 'use', 'method', 'unit', 'invention'. We believed that identification of frequently happening technology contexts with the help of such nouns is an achievable task. However, the evaluation of the results showed that this was not a reliable approach to using the annotation (we will elaborate on this later on).

After we found all NPs that contained any of the seed nouns in them in the training file, we were able to use all possible contexts in which these NPs were used in the data set in our pattern generation process. The context components were

combined in all possible combinations of one and two elements to serve as a base set of potential patterns.

3.2 Pattern selection

The pattern selection process is the experimental part of the project. Since there were no previous experiment done on this by other researches we eventually had to rely on heuristics and the results of our initial experiment to select the pattern set that we wanted to present for evaluation. We considered the following methods of pattern selection:

- feature frequency counts, including frequencies of feature combinations;

- point wise mutual information;

- simple heuristic knowledge of feature informativity.

3.2.1 Experimenting with frequencies

We initially used all features in our training file to generate all possible combinations of one, two and three elements and intended to use these combinations for deriving an experimental set of patterns. However, we noticed that one-element combinations often contain all useful information included in two-element ones and using two-element patterns alongside single-element ones was not justified.

(1) last_word=method tag_sig=NN with a frequency of 597;

(2) prev_n2=in_the last_word=system with a frequency of 45;

(3) last_word=gear prev_J=first with a frequency of 25.

Note that even though example (1) might have a good chance of being a technology context, the second part of the feature combination is not informative enough to consider this feature separately from the one element feature 'last_word=method' since all phrases in the training file are NPs. This is also true for examples (2) and (3) where 'prev_n2=in_the' and 'prev_J=first' may occur in the context of any NP and therefore do not help identify technologies, but features 'last_word=system' and 'last_word=gear' appear to be informative enough when used in single-element patterns.

Three element patterns turned out to be even less informative and we decided to leave them out in the process of selection. We thought that having more elements in our patterns might narrow our search scope and result in missing technology terms in the text. Nevertheless, we used some of the two-element combinations in our evaluation to get a better idea of whether or not they might be helpful for our purpose.

Additionally, we filtered some of the unwanted feature tags from our feature set. These included 'section_loc' tag, which was not applicable to the task of finding a set of technologies for all kinds of patents and document sections. Other examples are redundant feature sets: 'prev_n3' and 'prev_n2', 'next_n3' and 'next_n2', 'next_n2' and 'next2_tags'.

We analyzed the data and concluded that raw frequency counts do not give a clear picture of feature-technology correlations. The most frequent features included features: ('tag_sig=NN', 'following_prep=in'), ('tag_sig=NN', 'following_prep=for'), ('next2_tags=IN_VBG', 'following_prep=for') etc. Such patterns are widely used in sentence constructions in the context of any NPs and do not necessarily indicate technology terms.

We thought that normalizing frequency counts could give us a clearer picture of feature-technology correlations. We normalized the frequency counts by dividing the frequency of a feature's occurrence with a technology NP by the frequency of its occurrence with any NPs. This did not result in any significant difference either. However, the normalized counts were similar to the raw counts in the sense that they also contained mostly widely used context features.

Since the frequency counts showed us that the correlation between the features and technology terms is very low, we decided to try using simple heuristics for feature selection. We chose the most obvious patterns and aimed at a higher precision at the expense of recall.

3.2.2 Manual feature selection

Using our observations on the data we manually selected a set of 116 patterns 30 of which are two element combinations and the rest are single feature patterns. Note that all of those patterns were chosen among the ones that have the highest frequency scores of occurrence in the contexts of technology terms and that low frequency patterns were not taken into account. Many of these patterns are

close to the initial set of patterns that we used to experiment on unprocessed Pubmed data. However, the new set gives more precise syntactic information due to the data format. Some of the one-element patterns are a part of the two element ones included in the pattern set. Even though they are redundant, we included both types to see how well they perform in comparison.

In this section we would like to highlight some pattern groups among the patterns included in the experimental pattern set and explain the motivation behind choosing these patterns over the others.

The first group is a large group of patterns with the tag 'last_word'. As we have already mentioned, we thought that this tag would be helpful for finding compound technology NPs as we found that out of 1219 annotated technology terms 50 terms end in 'device', 38 terms end in 'system', 30 terms end in 'apparatus'. There are other examples of words that frequently happen to be the last noun of technology terms and we included them into our pattern set as well.

We also thought that it might be useful to look for the same type of nouns outside the NPs as technologies may occur in some other syntactic constructions but the meaning of the words itself seems to indicate that the sentence is talking about some sort of device or technology. Hence patterns like ('prev_N=system', 'following_prep=of'), 'prev_N=apparatus', 'prev_n2=method_for', 'prev_n2=medium_in' etc.

The next group we would like to discuss is verb patterns. We assumed these would produce good matches for the reasons we discussed in chapter 1 – that

technology terms often occur in certain linguistic patterns that describe their application. To this group refer: ‘prev_V=are_for’, ‘prev_V=based_on’, ‘prev_V=displayed_on’, ‘prev_V=formed_in’, ‘prev_V=passed_through’, ‘prev_V=produced_by’, ‘prev_V=programmed’, ‘prev_V=synchronized’, ‘prev_V=transmitted’, ‘prev_V=use’, ‘prev_V=used_as’, ‘prev_V=used_for’, ‘prev_V=used_in’ etc. Other patterns are not verb patterns but that were also chosen based on this intuition are for example: ‘next_n2=under_test’, ‘next_n2=for_detecting’, ‘next_n2=for_manufacturing’, ‘next_n2=for_producing’, ‘next_n3=for_manufacturing’_a etc.

We also included some patterns which have high frequencies of occurrence with technologies and do not seem to be particularly frequent in the contexts of any NPs. Examples of these are: ‘prev_n3=invention_,_the’, (‘of_head=operating’, ‘last_word=transmission’), ‘following_prep=vs.’ etc.

In the next chapter we will discuss the evaluation process and analyze the results of matching the patterns for finding technologies.

4. Evaluation and result analysis

4.1 Evaluation of selected pattern set

We used annotation and a gold standard corpus of phrases to evaluate the performance of the selected patterns. The gold corpus we used consisted of 1444 NPs from which 326 were classified as technologies according to an annotated list of terms. We received both the gold corpus and the annotation file from the members of the Brandeis team. The gold corpus file contained a list of terms with a collection of features selected for these terms from different sentences. This format is slightly different from the development file format that contained only features taken from a single sentence for each NP.

We measured the performance of our pattern set by calculating precision and recall on the gold standard data. Each phrase in the test file that matched at least one pattern in our pattern set was taken for a technology match (a hit). We then used the annotated list of technologies to see which hits were correct and which ones were incorrect. The counts of correct and incorrect hits were then used to calculate precision and recall for the whole test set.

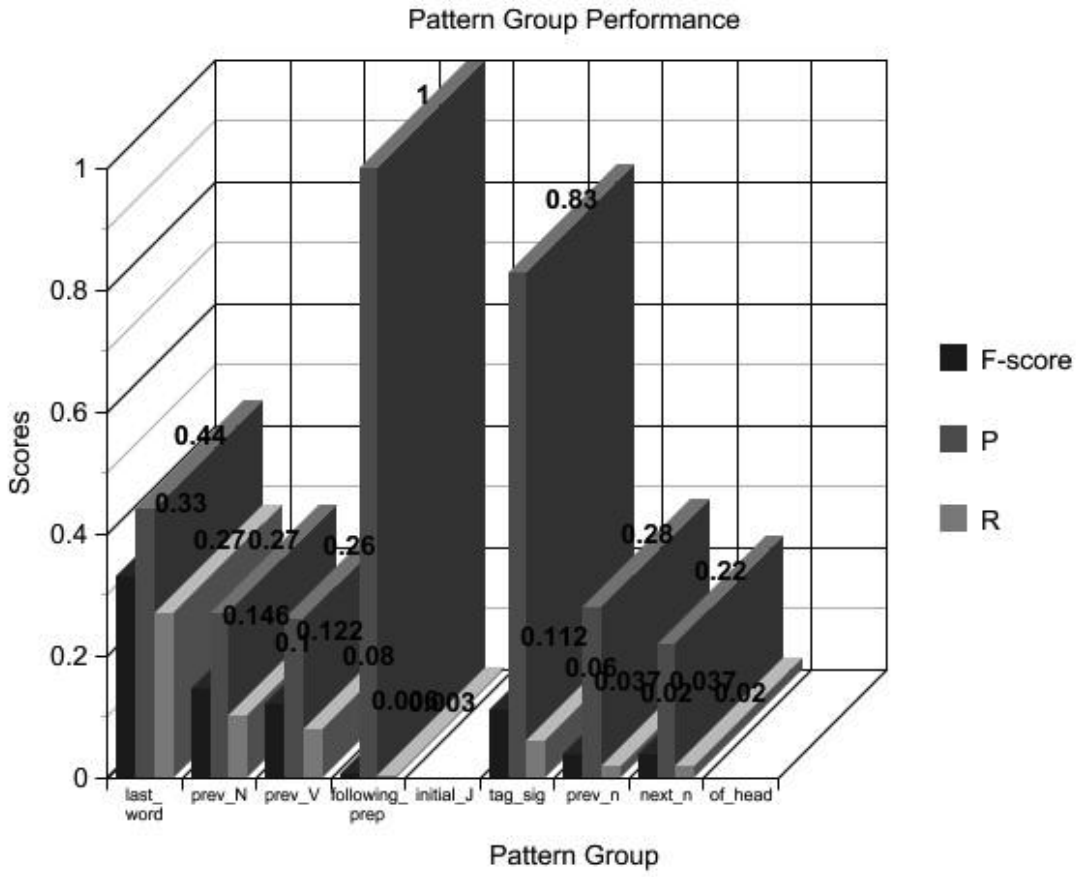
The results of the evaluation are displayed in table 4.1. We also made an attempt of classifying patterns into groups based on their feature labels (see Appendix) and calculated precision, recall and F-measure for each pattern group

separately (image 4.1). These measures allow us to compare the performance of different pattern groups and find out which groups performed better.

Table 4.1. Evaluation scores

	Actual correct	All hits	True positives	Precision	Recall
Manually selected set	326	357	112	0.31	0.34
Random	326	718	163	0.28	0.5

Image 4.1. Evaluation of pattern groups.



The group of patterns that showed the best performance is the 'last_word' feature group. In fact 88 out of 112 correct hits are captures by these patterns. Table 4.2 shows some examples of correct matches for 'last_word' patterns.

Table 4.2. 'Last_word' pattern group correct matches.

Pattern	*Matches*
*last_word=method[a-z]** *tag_sig=NN_NN_NN*	vision imaging devices and methods
last_word=unit tag_sig=NN_NN	image processing unit beam parameter monitoring unit
last_word=system[a-z] tag_sig=NN_NN*	CPU system wrap control system vision system electronics system transport system transportation system VR systems energy detector system image processing system computer vision system object identification system airport instrument landing system lateral-guidance transportation system visibility navigation system vision systems

*last_word=system[a-z]**	CPU system wrap control system vision system electro-optic vision system electro-optic system electronics system electro-optic vision systems lateral guidance transportation systems virtual reality systems transport system communications system global navigational satellite system transportation system energy detector system augmented reality vision systems vision systems image processing system optical combiner of the HUD system lateral guidance transportation system multiprocessing system computer vision system geographic information systems magnetic tape drive recording system GPS satellite navigation system object identification system airport instrument landing system Pentium Pro single or multiprocessor system lateral-guidance transportation system global positioning system visibility navigation system

*last_word=source[a-z]**	EUV source EUV or soft x-ray generating source EUV radiation source gas discharge based photon sources x-ray lithography and microscopy source gas discharge EUV photon source plasma pinch EUV sources intense soft x-ray source pulsed plasma source Xzenon z-pinch source EUV radiation sources monitoring pulse energies and\/or another parameter of an extreme ultraviolet radiation source extreme ultraviolet radiation source soft x-ray radiation sources homogeneous cylindrical plasma source voltage source EUV generating source

Note that the pattern 'system[a-z]*' has 32 correct matches vs. 16 correct hits for pattern (last_word=system[a-z]*, tag_sig=NN_NN). Phrases like 'electro-optic system', 'airport instrument landing system', 'multiprocessing system' did not match the 'NN_NN' pattern due to the presence of adjective or gerund modifiers in them.

Table 4.3. 'Last_word' pattern group false positives.

Pattern	*False positives*
*last_word=source[a-z]**	metal source source output power of gas discharge based photon sources output of the EUV source emission properties of the source radio frequency preionized xenon z-pinch source susceptibility of the metal source many components of the exemplary EUV source output power of the source improved EUV photon source output location of the EUV source emission characteristic of the source quotient of the increase of output and the initial output of each source experimental setup the EUV source location of the EUV source wide variety of EUV sources dissolving the anion source and the metal source
*last_word=system[a-z]**	system main CPU system requirements of CPU system present system other lateral-guidance transportation systems block diagram of an embodiment of an electro-optic system developing field of computer assisted vision systems

	exemplary applications of system real time position and attitude of the vision useful vision system least a portion of the image processing various applications and advantages of electro-optic system other vision systems block diagram of an electro-optic system advanced systems field of view of the system
last_word=application	advanced image augmentation application siting applications
last_word=detector	detector instability of the detector increasing the long term stability of the detector detailed schematic of a detector detailed view of a modified embodiment of the back portion 22b of a vacuum tight detector spectral region of radiation reaching the detector embodiment of a detector

In general the ratio of true positives to false positives appears to be very low. One example is pattern 'system[a-z]*' which has 20 false positives vs. 32 true positives (see table 4.3). The numbers show that the pattern overgenerates to a large extent. In this sense, a more specific pattern such as ('last_word=system[a-z]*', 'tag_sig=NN_NN') may help to improve on the precision score.

Among 'prev_V' patterns the ones based on the word 'use' turned out to perform better than the other patterns in the group (examples are given in table

4.4). The rest of the group produced very few correct matches. The ‘use’ patterns overgenerated a lot as well, producing 32 false negatives alongside 18 correct matches.

Table 4.4. ‘use’ as a previous verb in patterns.

Pattern	*Correct matches*
prev_V=used as	energy monitor detectors initiator composition photon detectors
prev_V=used for	light generation
prev_V=used in	electro-optic system shell synthesis or organic media of similar polarity
prev_V=using	mirror system of contactless energy transmission pulsed plasma source
*prev_V=use[a-z]**	CPU system graphics processor multilayer mirror electro-optic system shell formation anions shell synthesis or organic media of similar polarity energy monitor detectors metal salts lanthanide metal combinations initiator composition photon detectors light generation

The ‘prev_N’ pattern set produced a large portion of correct matches (34 correct hits), but generated 94 false negatives which is almost three times more (examples are given in table 4.5). The patterns ‘prev_n2’, ‘prev_n3’, ‘next_n2’ and ‘next_n3’ that use previous and next n-gram information matched a very small number of technology terms.

Table 4.5. ‘prev_N’ pattern group correct matches.

Pattern	*Correct matches*
prev_N=system[a-z] following_prep=for*	magnetic tape drive
prev_N=data	magnetic tape graphics processor image processing unit
prev_N=invention	luminescent nanoparticles polymeric peroxycarbonates shell material vision system shell salt or oxide SI dead region diode or a PTSI photodiode backhitchless writing of synchronized data image processing system shell synthesis
*prev_N=method[a-z]**	site plan mapping process lanthanide-doped nanoparticles of this type locating and characterizing potential explosive sites
prev_N=process	nanoparticles individual nanoparticles and oswald ripening

	preparation of these nanoparticles preparation of the luminescent nanoparticle cores image-delivery mechanism
*prev_N=system[a-z]**	magnetic tape drive wrap control system contactless energy transmission delivering real scene information Pentium Pro single or multiprocessor system
prev_N=use	EUV radiation source EUV photon source emitting
prev_N=information	image processing unit GPS satellite navigation system
prev_N=device	intense soft x-ray source

The results of the evaluation show very low precision and recall. We ran a random technology classification and evaluated it to have a baseline. Our experiment pattern set resulted in a precision that was slightly higher than the random experiment precision. The recall was even lower than the baseline value. This may be explained by the fact that we chose the most obvious patterns for our experiment at the cost of the actual frequency of features. This helped us improve the precision a little in comparison with random selection. However, we didn't consider many of the high frequency features in cases when they were not obvious clues to the presence of technology terms next to them. Therefore, we didn't seem to have gathered enough information to cover most technology term contexts.

As an attempt to interpret the results of the evaluation, we calculated pointwise mutual information (PMI) between each feature and technology terms. PMI is a measure that would show us the probability of each feature occurrence with technology terms compared to its occurrence independently of the presence of technology terms. We found that PMI values were negative for most of the features. In fact, the only positive PMI value in the whole set of features was 'tag_sig=NN' with the PMI value of 2.0. We concluded that the features selected for the patterns are anti-correlated with the technology terms instead of being positively correlated with them. A possible explanation for this fact might be that we did not use the annotation on the training data as a direct way of finding technologies. Instead, we used a list of nouns that we derived from the annotation file.

We thought that since the features were not correlated with technology terms, the evaluation strategy we chose was not suitable for the data we were working with. As we have described before, we considered an NP a technology if it matched at least one of the patterns in our pattern set. It became obvious that such decisions are not justified considering the anti-correlation between the features and technology terms.

We also thought that trying to elaborate on this evaluation method by accumulating feature matches and setting a threshold for identifying something as a technology based on the number of features matched is not the best way to approach the problem. Our recall score was lower than the baseline while precision improved compared to the baseline. This means that we missed a lot of

technologies. Setting the threshold even higher would give us even worse results on recall, even though it might improve the precision score.

These suppositions were an incentive for us to run a few experiments using the PMI scores of the feature set to select features for training an SVM. Using an SVM would allow for better flexibility in feature combination since an SVM weights the features in the training process. We discuss this experiment in the next section of this chapter.

4.2 Experimenting with PMI values

We used an SVM of the LIBSVM package (Chang, 2011) to run the experiments on PMI scores of the features. Since we thought that our use of annotation for extracting nouns from the list of technologies might not be the best approach judging by the scores of our pattern set, we ran the experiments on two different training sets: one based on the noun extraction for technology identification and another one that used annotated technology terms as whole units for finding technologies in the training data.

The first feature set we used was a subset of patterns derived from the training data based on the list of nouns that we extracted from technology term annotation. We selected 200 patterns with the lowest PMI scores and balanced the feature set with 70 patterns with the highest PMI scores to have some sort of relatively correlated features for our data. We also filtered this feature set using a LIBLINEAR package filter tool. Prefiltering of the features did not influence the evaluation results.

Table 4.6. Experiments with PMI scores.

Feature set	***Cross validation on training data Accuracy***	***Cross validation on test data Accuracy***	***Evaluation on test data Accuracy***
1. 270 PMI scores	94%	75%	61%
2. Top 250 PMI scores	94%	75%	61%
3. Top 150 and bottom 150 PMI scores	90%	—	Split training: 90% Original test: 61%

The second feature set was based on PMI scores recalculated by matching the annotated technologies on the training data without splitting them into a set of nouns. We found that PMI measures calculated in this way are more strongly correlated with technology terms. The set included 250 features with the highest PMI scores.

Finally, we ran cross validation and evaluation on a set consisting of 150 features with the highest PMI scores and 150 having the lowest PMI scores. We

thought that a better balances set of features might show better results, but the experiment showed no difference from the experiment on 250 features.

The results of the experiments are listed in table 7. As one can see, our results improved greatly for cross validation. Note that even though the performance of the feature set improved there is a huge gap between the cross validation accuracy and the performance of the classifier on the test data.

We also tried splitting the training data into half and using one half as a training set and the other half as a test set. We ran this experiment on feature set 2. This experiment resulted in an accuracy of 90% (see "split training" in table 4.6). Since our PMI scores for this experiment were calculated based on the whole training set we also ran a test to make sure that the results of the evaluation were not influenced be the presence of the test part of the data in the PMI score calculations. We recalculated PMI scores on 100,000 phrases and trained the classifier on a feature set consisting of 150 features with the lowest PMI scores and 150 features with the highest PMI scores from the recalculated values. The accuracy score stayed the same for this experiment (90%) thus showing that our previous experiment was valid.

We believe that the difference between the classifier performance on training and test data sets can be explained in two possible ways. It might be that the technology terms in the training and test sets differ significantly, which may be the result of having document sets with different contents or years of publication.

A more likely explanation, considering the big accuracy score difference, would be the possibility of inconsistencies in the annotation of the training and test data caused by the ambiguity inherent in defining what a technology is. We suspect that the two annotations schemes used somewhat different criteria for what constitutes a technology. Having more data annotated in the exact same way would probably help improve on the results as mixing even similar annotation schemes can be detrimental.

5. Future work

We conclude that identifying technologies in the text with the help of context matching is an extremely difficult task. This may be due to a frequent usage of different application descriptions in the text of patents in general to describe both technologies and other objects or processes that are not classified as such based on our definition of technology. In this chapter we suggest that some changes in the present system can help improve the results. We propose two main directions of algorithm improvement: improving on the feature set and using anti-correlated features to get the results.

Improving the feature set can be achieved in several possible ways. One possible improvement is revisiting the annotation schema. Note that some of the false positives do not look like obvious non-technologies. Consider the following examples:

radio frequency preionized xenon z-pinch source

advanced image augmentation application

polymeric peroxycarbonates comprising repeating units of the formula

sensitizer transferring the energy

curve-negotiating storage and retrieval vehicle

other lateral-guidance transportation systems

We believe that some of our system hits could be classified as false positives because of the fact that in many cases technology and non-technology terms have the same or similar linguistic patterns and the distinction between them can be made only based on the annotators' knowledge of the terms' meaning. Many of the false positives selected by our approach indeed seem to be able to serve as technologies for somewhat different tasks. For instance, if we were to include any reference to technologies into the gold standard list we might make use of such false positives as 'present system' and 'present invention'.

Another step that could be done is making the patterns more precise and possibly using negative features to filter out unwanted matches. Many of the false positives are NPs that include potential technology terms in them but are not classified as technologies as a whole. Using the feature of having the preposition 'of' as negative along with having the 'last_word' positive feature in the following NPs would not allow for classifying them as technologies:

block diagram showing an image processing system of the present invention

output power of gas discharge based photon sources

output of the EUV source

emission properties of the source

susceptibility of the metal source

many components of the exemplary EUV source

We could also develop the idea of using measures of association between the features and technologies and experiment with some other classification options. Our experiments with PMI values have shown good results on a five fold cross validation. Considering that those were conducted in a very short term we can assume that it is possible to tune the classifier for an even better performance by balancing the features.

References

1. Brock, D., Babko-Malaya, O., Pustejovsky, J., Thomas, P., Stromsten, S. and Barlos, F. 2012. Applied Actant-Network Theory: Toward the Automated Detection of Technoscientific Emergence from Full-text Publications and Patents.

2. Verhagen, M., Anick, P. and Pustejovsky, J. 2012. Multilingual Patent Classification.

3. Pickering, A. 1993. The Mangle of Practice: Agency and Emergence in the Sociology of Science. *American Journal of Sociology.* 99: 559–89.

4. Latour, B. 2005. Reassembling the social: An introduction to actor-network theory. Oxford, UK: Oxford University Press. Law, J. 1987. Technology and heterogeneous engineering: The case of the Portuguese expansion. In W.

5. Chang, C. C. and Lin, C. J. 2011. LIBSVM : a library for support vector machines. ACM Transactions on Intelligent Systems and Technology, 2:27:1--27:27. *Software available at* http://www.csie.ntu.edu.tw/~cjlin/libsvm

Appendix: Classification of patterns by tag.

Pattern Group	***Patterns***
last_word	[u'last_word=method[a-z]*', u'tag_sig=NN_NN_NN ']
	[u'of_head=process', u'last_word=invention[a-z]*']
	[u'last_word=method[a-z]*', u'tag_sig=NN_NN ']
	[u'prev_V=employed', u'last_word=invention']
	[u'prev_V=is', u'last_word=invention[a-z]*']
	[u'prev_V=improve[a-z]*', u'last_word=apparatus[a-z]*']
	[u'prev_V=operating', u'last_word=gear']
	[u'of_head=operating', u'last_word=transmission']
	[u'initial_J=present', u'last_word=invention[a-z]*']
	[u'of_head=process', u'last_word=invention']
	[u'last_word=invention[a-z]*', u'tag_sig=NN_IN_DT_NN ']
	[u'of_head=method[a-z]*', u'last_word=system[a-z]*']
	[u'last_word=unit[a-z]*', u'tag_sig=NN_NN ']
	[u'initial_J=active', u'last_word=apparatus[a-z]*']
	[u'last_word=gear', u'tag_sig=JJ_NN ']
	[u'last_word=system[a-z]*', u'tag_sig=NN_NN ']
	[u'last_word=unit[a-z]*', u'tag_sig=NNP_NN ']
	[u'last_word=system[a-z]*', u'tag_sig=NN_IN_VBG_DT_JJ_JJ_NN_NN ']
	[u'last_word=apparatus[a-z]*']
	[u'last_word=application']

	[u'last_word=beam'] [u'last_word=computer'] [u'last_word=detector'] [u'last_word=device[a-z]*'] [u'last_word=engine'] [u'last_word=error'] [u'last_word=frame'] [u'last_word=function'] [u'last_word=gear'] [u'last_word=instrument'] [u'last_word=invention'] [u'last_word=metadata'] [u'last_word=method[a-z]*'] [u'last_word=network[a-z]*'] [u'last_word=operation[a-z]*'] [u'last_word=pattern'] [u'last_word=processing'] [u'last_word=processors'] [u'last_word=program'] [u'last_word=research'] [u'last_word=signal'] [u'last_word=source'] [u'last_word=system[a-z]*']

	[u'last_word=template'] [u'last_word=training'] [u'last_word=use[a-z]*']
prev_N	[u'prev_N=system[a-z]*', u'following_prep=for'] [u'prev_N=system', u'following_prep=of'] [u'prev_V=provide[a-z]*', u'prev_N=invention[a-z]*'] [u'prev_N=gear[a-z]*', u'of_head=operating'] [u'prev_N=system', u'of_head=method'] [u'prev_N=apparatus'] [u'prev_N=data'] [u'prev_N=device'] [u'prev_N=evaluation'] [u'prev_N=information'] [u'prev_N=invention'] [u'prev_N=material'] [u'prev_N=method[a-z]*'] [u'prev_N=operation'] [u'prev_N=process'] [u'prev_N=system[a-z]*'] [u'prev_N=use']
prev_V	[u'prev_V=employ[a-z]*', u'of_head=process'] [u'prev_V=provide[a-z]*', u'prev_N=invention[a-z]*'] [u'prev_V=employed', u'last_word=invention']

	[u'prev_V=is', u'last_word=invention'] [u'prev_V=improve[a-z]*', u'last_word=apparatus[a-z]*'] [u'prev_V=operating', u'last_word=gear'] [u'prev_V=allows'] [u'prev_V=are for'] [u'prev_V=based on'] [u'prev_V=displayed on'] [u'prev_V=formed in'] [u'prev_V=passed through'] [u'prev_V=produced by'] [u'prev_V=programmed'] [u'prev_V=synchronized'] [u'prev_V=transmitted'] [u'prev_V=use'] [u'prev_V=used as'] [u'prev_V=used for'] [u'prev_V=used in'] [u'prev_V=using']
Following_prep	[u'prev_N=system[a-z]*', u'following_prep=for'] [u'prev_N=system', u'following_prep=of'] [u'following_prep=for', u'of_head=method[a-z]*'] [u'following_prep=of', u'of_head=method[a-z]*'] [u'following_prep=in', u'of_head=process']

	[u'following_prep=for', u'of_head=process'] [u'following_prep=vs.']
initial_J	[u'initial_J=electronic', u'of_head=method[a-z]*'] [u'initial_J=present', u'last_word=invention[a-z]*'] [u'initial_J=active', u'last_word=apparatus[a-z]*']
tag_sig	[u'last_word=method[a-z]*', u'tag_sig=NN_NN_NN '] [u'last_word=method[a-z]*', u'tag_sig=NN_NN '] [u'of_head=process', u'tag_sig=NN_IN_DT_NN '] [u'last_word=invention', u'tag_sig=NN_IN_DT_NN '] [u'last_word=unit', u'tag_sig=NN_NN '] [u'last_word=gear', u'tag_sig=JJ_NN '] [u'last_word=system', u'tag_sig=NN_NN '] [u'last_word=unit', u'tag_sig=NNP_NN '] [u'last_word=system', u'tag_sig=NN_IN_VBG_DT_JJ_JJ_NN_NN ']
prev_n	[u'prev_n2=embodiment_,'] [u'prev_n2=invention_provides'] [u'prev_n2=medium_in'] [u'prev_n2=method_for'] [u'prev_n2=provided_with'] [u'prev_n3=according_to_the'] [u'prev_n3=applied_to_the'] [u'prev_n3=invention_,_a']

	[u'prev_n3=invention_,_the'] [u'prev_n3=invention_provides_a'] [u'prev_n3=provided_by_the'] [u'prev_n3=provided_on_the'] [u'prev_n3=the_invention_provides']
next_n	[u'next_n2=created_by'] [u'next_n2=for_detecting'] [u'next_n2=for_manufacturing'] [u'next_n2=for_producing'] [u'next_n2=for_the'] [u'next_n2=formed_from'] [u'next_n2=is_applied'] [u'next_n2=is_generated'] [u'next_n2=is_positioned'] [u'next_n2=is_used'] [u'next_n2=under_test'] [u'next_n2=used_for'] [u'next_n2=used_in'] [u'next_n3=for_manufacturing_a'] [u'next_n3=may_be_used'] [u'next_n3=produced_by_the']

of_head	[u'prev_V=employ[a-z]*', u'of_head=process'] [u'of_head=process', u'last_word=invention[a-z]*'] [u'prev_N=gear[a-z]*', u'of_head=operating'] [u'prev_N=system', u'of_head=method'] [u'of_head=process', u'tag_sig=NN_IN_DT_NN '] [u'initial_J=electronic', u'of_head=method[a-z]*'] [u'of_head=operating', u'last_word=transmission'] [u'of_head=process', u'last_word=invention'] [u'following_prep=for', u'of_head=method[a-z]*'] [u'following_prep=of', u'of_head=method[a-z]*'] [u'of_head=method', u'last_word=system'] [u'following_prep=in', u'of_head=process'] [u'following_prep=for', u'of_head=process']