Developing a Dataset for Technology Structure Mining

Behrang QasemiZadeh and Paul Buitelaar

***Abstract*—This paper describes steps that have been taken for constructing a development dataset for the task of Technology Structure Mining. We have defined the proposed task as the process of mapping a scientific corpus into a labeled digraph named Technology Structure Graph as described in the paper; the generated graph expresses the domain semantics in terms of interdependencies between pairs of technologies that are named (introduced) in the target scientific corpus. The dataset comprises of a set of sentences extracted from the ACL Anthology Corpus; each sentence is annotated with at least two technologies in the domain of Human Language Technology and the interdependence between them. The annotations, technology mark-up and their interdependencies, are expressed at two levels: terminological and conceptual. Terminological representation of technologies comprises of variant lexicalization of a technology e.g. at the lexical level Human Language Technology may be signaled by** `HLT`**,** `Human Language Technology`**,** `Natural Language Processing`**, and** `NLP`**; however, at the conceptual level all these terminologies refer to the same concept i.e. *HLT*. We have adopted the same approach for representing Semantic Relations; at the terminological level a semantic relation is a predicate i.e. defined based on the sentence surface structure; however at the conceptual level semantic relations are classified into conceptual relations either taxonomic or non-taxonomic e.g. lexical relations such as** `used_in`**,** `applied_in`**, and** `employed_by` **are classified under a conceptual relation *DEPEND_ON*. The contexts where interdependencies are extracted from are classified into five groups based on the linguistic criteria and syntactic structure that are identified by the annotators. These are Prepositional, Noun Compound, Verb Based, and Structural contexts, as well as Residuals as described in the paper. The dataset initially comprises of 482 sentences. Other annotations along the sentences are the author names, the year of publication, the position of text in the paper (visual position); we hope this effort results in a benchmark that can be used for the technology structure mining task as defined in the paper.**

I. Introduction

Technology Management [1] is a strategic research topic dealing with innovation, efficiency and organization structure management in rapidly changing technology world. Started in the 60s, a long discussed topic in this area is technology-structure relationships[3]. Among the category definitions for empirical technology-structure research is Technology Interdependence. Technology Interdependence potentially can be used for "minding the technology gap" as defined by Bailey et al [2]:

> "We define a technology gap as the space in a work flow between two technologies wherein the output of the first technology is meant to be the input to the second one."

The automatic extraction of such information involves several established research challenges in Information Extraction and Natural Language Processing namely Named Entity Recognition (NER) [4], Semantic Role Identification [5], Relation Extraction (RE) [7], [6]; and in a broader sense, Natural Language Understanding and Semantic Computing with two emerging research application areas: Open (Domain) Information Extraction (OIE) [8], and Ontology Learning (OL) [9]. We classify the task of Technology Structure Mining as an activity situated between OIE and OL.

One of the main challenges to pursuing such tasks is the lack of linguistic resources for evaluation and development. While any task like the one we will introduce here tackles the problem of knowledge acquisition and tries to engineer the bottleneck of knowledge acquisition through automated methodologies and algorithms, the development and evaluation of such methods relies closely on the provided dataset for testing and training e.g. [23], [24]. In addition, understanding and evaluation of the outcome of an IE/OL task is subject to the understanding of domain experts and the sort of information they are looking for; generally speaking, these activities are more task-driven rather than fact-driven. In addition, research studies in these domains usually focus on evaluation of engaged activities such as NER or RE in isolation. There is no report on the impact of the quality of these activities in the overall quality of the task performance.

For the reasons mentioned above, we have developed a dataset that will ideally result in a benchmark to evaluate the proposed task in section 3. The dataset comprises of sentences in the domain of Human Language Technology from the ACL Anthology Reference Corpus (ACL ARC)[13]. The annotations are provided at two layers, lexical and termino-conceptual. At the lexical layer the representation of an identical technology may comprise of lexical variants e.g. Human Language Technology may be signaled by `HLT`, `Human Language Technology`, `Natural Language Processing`, and `NLP`. However, at the conceptual level all these lexical variations refer to the same concept i.e. *HLT*. We have adopted the same approach for representing Semantic Relations; at the lexical level a semantic relation is a predicate i.e. defined based on the sentence surface structure. However at the termino-conceptual level, semantic relations are classified into conceptual relations, either taxonomic or non-taxonomic e.g. lexical relations such as `used_in`, `applied_in`, and `employed_by` are classified under a conceptual relation *DEPEND_ON*. This layered representation will assist us in modularizing the task of Technology Structure Mining into several sub-task, including detecting technologies at the lexical level, mapping the technology lexicalizations to concepts, relation extraction between pairs of technology concepts at the lexical level, and finally mapping the lexical relations to conceptual semantic relations.

The rest of the paper is organized as follows: in the next section we briefly introduce related work. In section 3, we propose a formal task definition. Section 4 describes the methodology for generating the dataset out of the ACL ARC corpus. Section 5 describes manual annotations, and gives examples of sentences in the corpus together with statistics. Finally we conclude and give the direction of our future work in section 6.

II. Related Work

Besides existing research in information extraction from patents e.g. [10], there is not much research reported towards extracting information from scientific publications for mining technology interdependence. Considering technology as applied science then it is not far from reality to consider scientific publications as a primary source of information for the task of technology structure mining. The research in this area can result in methodologies for smoothing the process of domain-semantic modeling in terms of technologies that are involved in a scientific domain. This may result in a strategic tool for intelligent information retrieval as well as assisting the process of technology management.

As stated in [7], the information science research community and the Natural Language Processing (NLP) community [18] have focused on concepts and terms, but “the focus is increasingly shifting to the identification, process and management of relations to achieve greater effectiveness”. However, none of research in these domains explicitly mentions the correlation between concepts and relations, particularly in their task formalization. They either have considered this as an obvious fact, or this has not been the focus of their theoretical foundation. What is required here is a model that can combine and express properties of semantic relations from both the lexical and logical perspectives at a scalable size. We consider our research towards this goal. The most prominent research in recent years have approached the problem from the ontology engineering and population point of view. The main power of this research resides in the use of ontologies as a foundation for expressing domain-semantics. However, just until recently [12] this research lacked the concern about lexical properties of concepts.

In [14], Hobbs and Riloff provide an overview of research in the Information Extraction (IE) domain. With emphasis on diversity in IE tasks, they have identified *named entity recognition*, *relation extraction*, and the task of *event identification* under the IE research topic and provide a classification over the existing approaches from various perspectives and a comparison between finite state based methods versus machine learning approaches. They have discussed the complexity of the tasks of detecting complex words, basic phrases, complex phrases, as well as event detection and assigning them a unique identifier and a semantic type. The importance of real-world knowledge and its encoding into such systems is also emphasized.

In [9], Cimiano et al give a survey of current methods in ontology construction and discuss the relation between ontologies and lexica as well as ontology and natural language. They illustrate different engineering approaches to ontology design and enumerate their excellence and deficiency. Under the topic of ontology learning, authors contemplate controversies in concept identification and relation extraction. They emphasize the distinction between linguistic representation of concepts and the concepts themselves and make a difference between concept hierarchy and relation extraction since they see these as the difference between paradigmatic versus syntagmatic relations. The importance of selectional restriction and choosing the right level of abstraction has been mentioned as other challenges in this field.

Khoo and Na [7] provide a survey on semantic relations. Their survey describes the nature of semantic relations from the perspective of linguistics and psychology, in addition to a detailed discussion of types of semantic relations including lexical-semantic relations, case relations, and relations between larger text segments. They clarify the definition of semantic relation in knowledge structures such as thesauri, and ontologies. Although some semantic relations can be extracted/inferred from syntactic structures, there are other semantic relations that require multi step sequence of reasoning. Their survey enumerates a number of approaches for automatic/semi-automatic extraction of relations and ends up with explaining the application of semantic relations in applications such as question-answering, query-expansion, and text summarization.

Finally, consider much of the work in BioNLP as the closest to the proposed task here. Bio texts are usually written for describing a specific phenomenon e.g. gene expression, protein pathways etc. in a very specific context. Extracting such information, e.g. extracting instances of specific relations or interactions between genes and proteins, from Bio-literature is similar to the task of technology structure mining. However, despite the proposed application here, Bio-Text Mining is well supported by ontologies, and language resources; the context and concepts are usually clearly defined and tools which are tuned for the domain are available. The availability of knowledge resources such as well defined ontologies in this domain lets Bio-Text miners to build new semantic layers on top of already existing semantic resources (ontologies).

III. Task Definition

We identify the task of technology structure extraction to comprise of four major processes: identification of technology terms at the lexical level, mapping the lexical representation of technologies into a termino-conceptual level, extracting relations between pairs of termino-conceptual technologies at the lexical level (i.e. at sentence surface structure), and finally mapping/grouping relations at the lexical level into canonical relation classes at the conceptual level. We name the result of the proposed processes the *Technology Structure Graph* (TSG). Therefore, we define the task of technology structure extraction as the process of mapping a scientific corpus into a TSG graph with the following definition:

Definition 1: A *Technology Structure Graph (TGS)* is a tuple $G = \langle V, P, S, \Sigma, \alpha, \beta, \omega \rangle$ where:

1) V is a set of pairs $\langle W, T \rangle$ where $\langle W, T \rangle$ is a uniquely identifiable terminology from a set of identifiers N and T is the terminology semantic type, e.g., $\langle \text{NLP}, \mathsf{TECHNOLOGY} \rangle$ or $\langle \text{Lexicon}, \mathsf{RESOURCE} \rangle$ or $\langle \text{Quality}, \mathsf{PROPERTY} \rangle$. To support different level of granularity of information abstraction we also consider V can contain pairs $\langle G_i, \mathsf{GRAPH} \rangle$ where G_i has the same definition as G above.
2) P is a set of technology terms at lexical level, uniquely identifiable from a set of identifiers R, e.g., `Natural Language Processing`, `NLP`, `Human Language Technology`.
3) S is a set of lexical relations, uniquely identifiable from a set of identifiers Q, e.g., `used by`, `applied for`, `is example of`.
4) Σ is a set of relations, i.e., the canonical relations vocabulary, e.g., $\{\mathsf{DEPEND\_ON}, \mathsf{KIND\_OF}, \mathsf{HAS\_A}\}$.
5) α is a partial function that maps $\langle W, T \rangle$ to a label of Σ annotated by a symbol from a fixed set M, i.e., $\alpha : N \times N \rightarrow \Sigma \times M$. M can be, e.g., the symbols $\{\Box, \Diamond\}$ from modal logic.
6) β is a function that maps P to a tuple in V i.e., $\beta : R \rightarrow N$.
7) ω is a function that maps S to a term in Σ i.e., $\omega : S \rightarrow \Sigma$.

Considering the following input sentence:

> "There have been a few attempts to integrate a speech recognition device with a natural language understanding system." [16]

with M defined as *possible* and *certain* modalities, i.e., $\{\Box, \Diamond\}$, then the expected output of analysis will be as follows:

$$\begin{aligned}
&V = \{\langle \text{NLU}, \mathsf{TECHNOLOGY} \rangle, \langle \text{SR}, \mathsf{TECHNOLOGY} \rangle\} \\
&P = \{\text{natural language understanding}, \text{speech recognition}\} \\
&\Sigma = \{\text{MERGE}\} \\
&S = \{integrate\ with\} \\
&\beta = \ \text{natural language understanding} \\
&\mapsto \langle \text{NLU}, \mathsf{TECHNOLOGY} \rangle \\
&\qquad\qquad \text{speech recognition} \mapsto \langle \text{SR}, \mathsf{TECHNOLOGY} \rangle \\
&\omega = \ \text{integrate with} \mapsto \text{MERGE} \\
&\alpha = \langle \langle \mathsf{SR}, \mathsf{Technology} \rangle, \langle \mathsf{NLU}, \mathsf{Technology} \rangle \rangle \\
&\qquad\qquad\qquad\qquad\qquad\qquad \mapsto \langle \mathsf{MERGE}, \Diamond \rangle
\end{aligned}$$

The main goal of the introduced task is in giving unstructured data (i.e. natural language text) a machine tractable structure in a way that we can make semantically interpret this input data. Any semantic interpretation in machines is limited to our definition of symbols and their interpretations. In fact, since our knowledge of (natural language) understanding is limited we move towards human understanding of language through an engineering approach, and the proposed definition above can provide us with a base-line to perform and evaluate this task.

As with previous research in this domain, our task definition deals with two major sub-tasks: concept and relation identification/definition; it considers concepts as the building blocks of knowledge and relations as the elements that are connecting these concepts into a structure. However, we emphasize the interaction between concept definition and relation definition. In addition, we make the boundaries in the process more visible so we can divide the task into sub-tasks in a more modular manner enabling us to study their interconnections in a more systematic way. We argue it is not possible to define what we call relations vocabulary Σ without considering the definition of V. The task of semantic interpretation of a natural language text is an eco-system that comprises of concepts, relations, and linking/connecting concepts to each other through these relations, in addition to the understanding of the user of the system of the provided symbols in V, and Σ. The other research challenge resides in mapping lexically introduced "concepts and relations" to a canonical termino-conceptual format. As stated in the given definition, we only focus on binary relations; the proposed model only concentrates on the relation between two technologies and we are aware of the limitations of the proposed model e.g. in modeling and representing the following sentence:

> "This method eliminates possible errors at the interface between speech Recognition and machine translation(component technologies of an AUTOMATIC Telephone Interpretation system) and selects the most appropriate candidate from a lattice of typical phrases output by the speech Recognition system."[17]

In the above sentence, the author(s) addresses the interaction between two technologies and provides information about an interdependence. Our defenition does not support representation of such information.

As mentioned, *Definition 1* provides us with a base-line to approach the task of Technology Structure Mining; our first attempt towards this goal starts with developing a dataset for further experiments as described in the next section.

IV. Dataset Development

As mentioned above the dataset comprises of sentences with at least two technology terms and their interdependencies. The sentences are extracted from the ACL Anthology Reference Corpus (ACL ARC) i.e. a corpus of scholarly publications about Computational Linguistics consisting of 10,921 articles and can be downloaded from [15]. The ACL ARC is represented in three different formats: source PDF files of articles, plain text, and a XML version of the articles i.e the OCR output of PDF files with additional information of visual features of the text e.g. font face, font size, the position of text etc. The corpus is furthermore divided into different sections in directories labeled with a single letter, 11 sections in total.

Dataset development essentially has comprised of 4 steps: Text Processing, Indexing and Storage, Concept (technology) Identification, and Compilation of dataset (Figure 2). Then we have studied the selected sentences manually, verified the processes, and annotated the sentences with the lexical/semantic relations between pairs of technologies. In the remainder of this section we give a description of each step of the task with results on the corpus.

We have gone through an iterative process for the dataset development. In the first step, the main question to answer was finding the optimum boundary size of text for dataset development e.g. should we focus at paragraph level or sentence level. To answer the question, in the first step we have chosen 1424 random papers from the corpus and performed following analysis. Selected papers consist of 45,031 paragraphs, 168,028 sentences, 4,524,062 tokens, and 124,525 types [1]. We studies the distribution of terms that can be considered as a representation of a technology in the domain. Our experiment showed that the co-occurrences of pairs of technologies tend to happen at sentence level(Figure 1). This means that if two technology occur within a text segment then it is more likely that this happens within a sentence. In addition, studying the relations at a greater boundary such as paragraph level imposes computational costs that may not be desirable considering the size of corpus, the cost of annotating a dataset, and the current state of technologies such as anaphora resolution. This has been also discussed from another perspective in [19]. In the remainder of this section we describe each step of the analysis in detail.

A. Text Processing

The ACL ARC corpus does not provide text sections and segments. The first stage of our process therefore involved text sectioning, and structuring. The text sectioning step involved converting provided XML files in ACL ARC into a more structured XML document where different sections of a paper such as titles, abstract, references etc. were identified using a set of heuristics. The heuristic rules are based on provided visual information in the source XML files such as font face, font size, position of text segments, and their frequency distribution. As for any other text sectioning task, this step involves noise and error in the output. In the next step, we did text segmentation including the detection of boundaries of paragraphs, sentences, and tokens. We have also performed part-of-speech tagging, and lemmatization. For detecting paragraph boundaries we have used a set of heuristics. However sentence segmentation, and tokenization has been carried out with OpenNLP [20]. Since OpenNLP tools are trained on scientific publications, they tends to have better quality compared to other available tools. Then, We used the Stanford Part of Speech tagger [21] for tagging and lemmatization. The generated files can

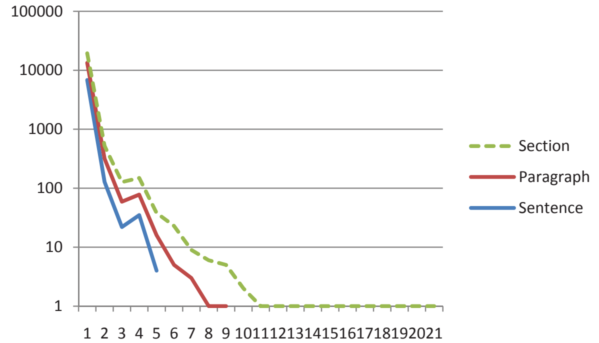

Fig. 1. Distribution of co-occurrences of technology terms: The analysis shows that the co-occurrences of two technology terms tend to be at the boundary of sentences; The above diagram shows that if two technologies appeared together in a text boundary then it is most probable that these two terms are situated within a sentence. Here, the vertical axis shows the number of technology terms and the horizontal axis shows the number of terms (in logarithmic scale) in sentence, paragraph and sections segments e.g. the diagram shows that we have 10,000 sections, paragraphs, and sentences with one technology term while there are no paragraphs or sentences with more than 10 technology terms within their boundaries.

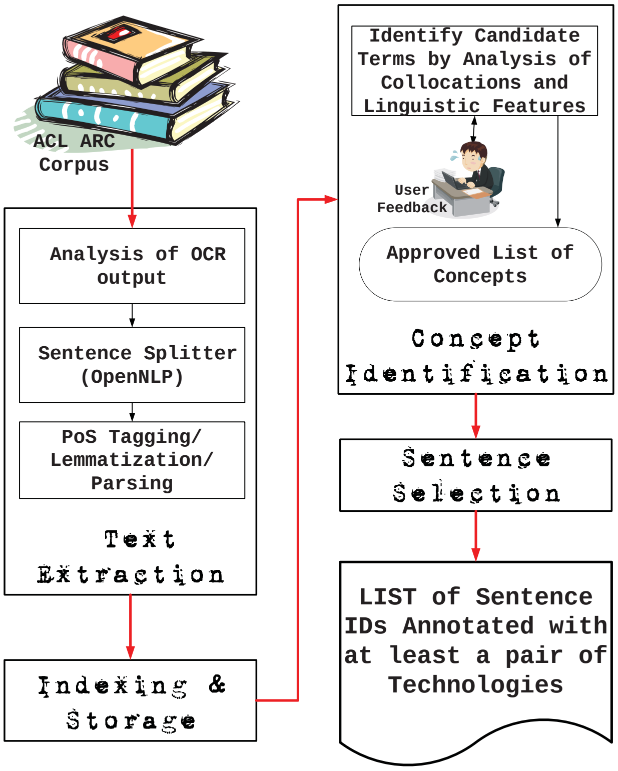
ACL ARC Corpus
Analysis of OCR output
Sentence Splitter (OpenNLP)
PoS Tagging/ Lemmatization/ Parsing
Text Extraction
Indexing & Storage
Identify Candidate Terms by Analysis of Collocations and Linguistic Features
User Feedback
Approved List of Concepts
Concept Identification
Sentence Selection
LIST of Sentence IDs Annotated with at least a pair of Technologies

Fig. 2. Dataset Development: Steps that have been taken for selecting sentences

be downloaded from http://nlp.deri.ie/behrang/sepid_arc.html. The indexed sentences were also processed with open source

[1] The numbers proposed here are subject to the errors that are imposed by text processing/extraction process and may not be identical using different approaches for text extraction

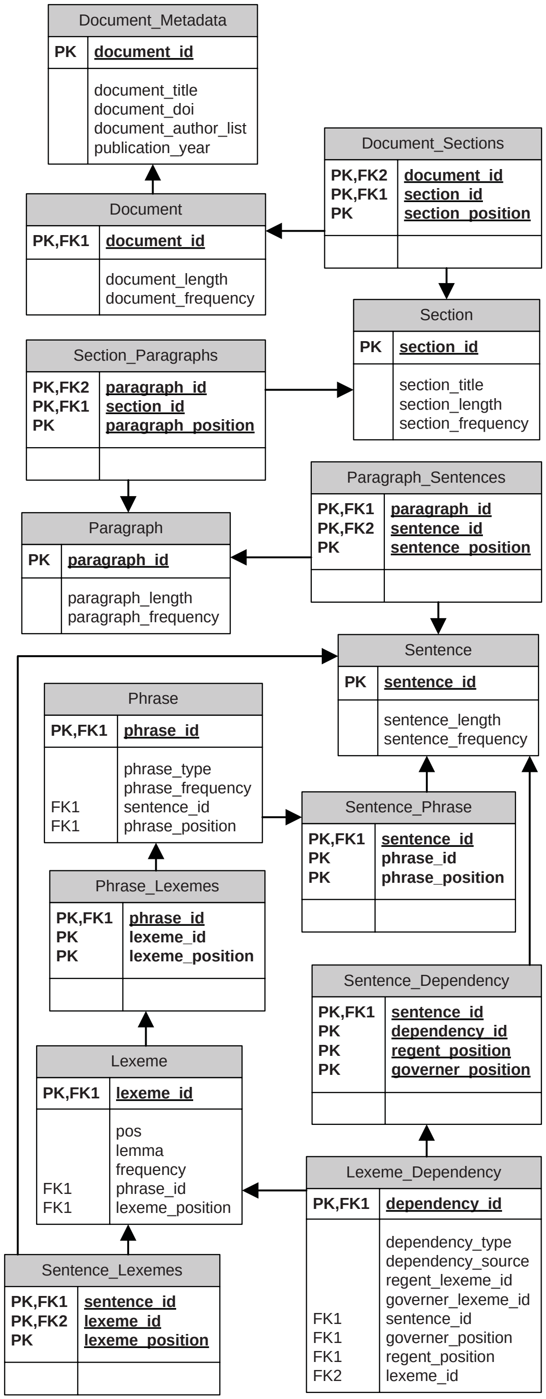

Fig. 3. Entity-Relationship diagram of the indexing system

dependency parsers: Malt Parser[26], BioLG [25], and Stanford Dependency Parser [27].

B. Indexing and Storage

The next step of the process involves indexing and storage of the corpus in the proposed data model in Figure 3. The proposed model let us dynamically generate a lexicon out of the part of speech tagged and lemmatized tokens in the corpus, along with the frequency of words. This also enables us to keep track of the position of words, sentences, paragraphs, and sections within a document. For example, we can easily identify all the sentences, paragraphs, and sections that have the word *technology* with a specific linguistic annotation such part of speech. We have used the model to retrieve data from the corpus with queries similar to the Corpus Query Language[22] but at uniquely indexed text segments. Performance, reducing processing time, ability for concurrent parsing of sentences, as well as flexibility in modification of metadata have been among the other reasons for using the proposed model in Figure 3.

C. Concept Identification

The concept identification (technology term recognition) process starts with selecting all the phrases in the corpus with the word "technology/ies". In fact we queried the corpus for the chain of tokens/lexemes that are ended with a token that has "technology" as its lemma, in addition to applying a set of filters which have been defined based on part of speech, and the position of the tokens. For example if we found a lexeme chain starting with a *verb in gerund or present participle form* (i.e. VBG part of speech in Penn Style Treebank[28]) then the chain would be accepted only if a determiner appeared before the token with VBG part of speech. In the next step, the extracted technology terms were manually refined. Among the 147 extracted lexeme chains, 31 terms were rejected manually(this includes meaningless terms in addition to very specific terms such as "Japaneses sentence parsing technology"). Then, we manually grouped the remaining terms into 43 different classes, each class refers to a specific technology in the domain of Human Language Technology e.g. finite-state, segmentation, parsing,entity-extraction, etc. As a matter of fact, this processing step comprises of defining P, V, and the function β in the Definition 1 in section III. As an example, at the end of this step, P includes these strings: *information retrieval technology,information retrieval technologies,information retrieval,IR technology, IR*; and V has a member $\langle \mathrm{IR}, \mathsf{TECHNOLOGY} \rangle$, and function β maps all the given value above for P to $\langle \mathrm{IR}, \mathsf{TECHNOLOGY} \rangle$ in V. This process step has been carried out on the sub-corpus of 1424 random papers described above.

D. Sentence Selection

After choosing the technology classes and defining P, V, β for the corpus, we identified sentences that contain more than one string term from P. In this step, we have extracted the sentences for each section of the ACL ARC; e.g. we were able to extract text from 2435 papers out of section C (failing on 432 papers; either because of the errors in the source XML files or deficiency in our heuristics for corpus processing). This step has been carried out on all sections of the corpus. Table I and Table II shows summarized statistics of the performed processes. Table I shows the overall number of articles that have been extracted from the XML source files *(ARTICLES#)*, the number of documents successfully segmented and indexed *(SUC-ARTICLE#)*, and the number of documents failed to segment and index *(UNSUC-ARTICLE#)*. Table II shows statistics for the successfully indexed documents; this includes the number of tokens, types, the number of identical sentences *(SENT)*, the number of identical sentences with minimum 1 technology term *(SST1)*, and the number of identical sentences with more than one technology term *(SST2)* for each section of corpus.

TABLE I
STATISTICS FOR TEXT PROCESSING STEP[2]

Section	ARTICLES#	SUC-ARTICLE#	UNSUC-ARTICLE#
A	404	265	139
C	2,435	2,003	432
E	846	463	383
H	897	828	69
I	146	113	33
J	922	114	808
M	180	168	12
N	371	365	6
P	2028	1873	155
T	120	81	39
W	2281	2121	160
Total	**10,630**	**8,394**	**2,236**

TABLE II
STATISTICS FOR EXTRACTED TEXT FROM ACL-ARC SECTIONS

Section	Token#	Type#	SENT#	SST1#	SST2#
A	955761	40938	35439	2012	134
C	6168312	172077	230936	7514	482
E	1901481	61854	67588	1646	81
H	2107057	56470	78797	4777	330
I	358358	20299	14258	721	52
J	612692	23702	22061	496	25
M	400398	20807	14903	592	52
N	1164215	38772	44103	2349	180
P	7446189	152890	272706	8833	603
T	122969	10882	4693	65	1
W	8169591	167107	300612	na	na

E. Manual Verification of Analysis, Annotation and Grouping of Relations

In the final step of dataset development, we chose and annotated sentences from the C section of the corpus. This section of the corpus comprises of papers from different conferences from the year 1965 to 2004. Among the 230,936 sentences in this section of the corpus, only 2012 sentences contain a technology term, and amongst these sentences only 482 have two or more lexical chains that signal appearance of technologies of different classes in the sentence. We manually read the extracted sentences and annotated them with the following informations:

1) Check whether the text processing step has been performed correctly; this comprises of checking the sectioning/segmentation of the source XML files, sentence splitting, and tokenization.
2) Technology Mark-up: whether the applied approach for detecting the technologies has been successful
3) Type of Relation: whether the sentence implies/expresses a relation between marked-up technologies, moreover what is the linguistic context for the relation as described below
4) Lexical Relation: if a sentence implies a relation, how is it expressed
5) Grouping Lexical Relations into Semantic Relations: This step comprises of classification of detected lexical relations into semantic relations

As mentioned earlier, we have identified and classified 5 different types of contexts for relation extraction as follows:

1) *Noun-Compound*: This context refers to a relation that can be inferred from the combination of nouns in a compound e.g.

   > "Since a model of machine translation(MT) called translation by Analogy was first proposed in Nagao (1984) , much work has been undertaken in *Example-Based NLP*(e.g. Sato and Nagao (1990) and Kurohashi and Nagao (1993))." [33]

   The above sentence suggests a relation as follows:

   $$\begin{array}{l}\langle\langle \mathrm{NLP}, \mathsf{TECHNOLOGY}\rangle, \\ \mathsf{HAS}-\mathsf{SUB}-\mathsf{CLASS} \\ \langle \mathrm{EB\text{-}NLP}, \mathsf{TECHNOLOGY}\rangle\rangle\end{array}$$

   Noun-Compound is the only context provides termino-conceptual relations directly.
2) *Prepositional*: This class of relations can be inferred from prepositional attachment, e.g.

   > "*NLP components of a machine translation* system are used to automatically generate semantic representations of text corpus that can be given directly to an ILP system."[34]

   the above sentence suggests a relation as follows:

   $$\begin{array}{l}\langle\langle \mathrm{MT}, \mathsf{TECHNOLOGY}\rangle, \\ \text{has-component-of}, \langle \mathrm{NLP}, \mathsf{TECHNOLOGY}\rangle\rangle\end{array}$$

3) *Verb-based*: This refers to contexts where two technology terms are directly/indirectly related to each other by a verb:

   > "lexical Knowledge acquisition *plays an important role* in Corpus-Based NLP."[35]

   however, extracting relations of this type may not be as straight-forward as expressed by, because other relations e.g. noun-compounds may occur at the same time. For example, relations in the above sentence are as follows:

[2] The total numbers of articles proposed here are not identical to the numbers proposed in [13] due to corruptions in the source XML files; we have excluded these files from the corpus

a) $\langle\langle \text{LEXICAL-KA}, \mathsf{TECHNOLOGY}\rangle,$
$\mathsf{IS-SUB-CLASS-OF},$
$\langle \text{KA}, \mathsf{TECHNOLOGY}\rangle\rangle$

b) $\langle\langle \text{CP-NLP}, \mathsf{TECHNOLOGY}\rangle,$
$\mathsf{IS-SUB-CLASS-OF}, \langle \text{NLP}, \mathsf{TECHNOLOGY}\rangle\rangle$

c) $\langle\langle \text{LEXICAL-KA}, \mathsf{TECHNOLOGY}\rangle,$
play-role-in, $\langle \text{CP-NLP}, \mathsf{TECHNOLOGY}\rangle\rangle$

4) *Structural*: this context refers to relations that can be inferred based on the structure of a sentence:

> "Transformation-Based learning has been used to tackle *a wide range of* NLP problems , *ranging from* part-of speech tagging(Brill , 1995) to parsing(Brill , 1996) *to* segmentation and message understanding(Day et al. , 1997)."[36]

suggesting the relation:
$\langle\langle \text{POS-TAGGING}, \mathsf{TECHNOLOGY}\rangle,$
is-problem-example-of, $\langle \text{NLP}, \mathsf{TECHNOLOGY}\rangle\rangle$

5) *Residuals*: this category refers to relations that cannot be fitted in any of the first three above categories and/or are too complicated to be inferred automatically e.g.:

> "finite-state rules are represented Using regular expressions and they are transformed into finite-state automata by a rule compiler."[37]

conveys a relation between *Finite Automata* and *Compiler*, or the following sentence:

> "In translation memory(TM) or Example-Based machine translation(EBMT) systems, one of the decisive tasks is to retrieve from the database ,the example that best approaches the input sentence." [38]

express a relation between *Database Technology* and *Machine Translation Technology*. However, we believe that the expressed realtions in these sentences are too complex, and automatic extraction of such relations and expressing them by *TSG* may be far from reality. Worthwhile to mention that we have identified some of the relations expressed by sentence structure that are also difficult to be extracted automatically e.g. the temporal relation conveyed by the sentence given above as an example of noun-compound relation; the above sentence expressing a temporal relation between the time of introducing "translation by Analogy" and "Example-Based NLP". We have also grouped this relations under residuals category.

These different contexts have been studied in previous research e.g. [30], [29], [6], [32] and [31]. However, as to the knowledge of the authors no research has been reported on the analysis of the distribution of these contexts, nor exists a corpus that provides an annotation about the linguistic contexts for relation extraction.

Among the 482 annotated sentences, the text extraction process has been carried out correctly for 425 sentences, and it fails for 57 cases. This gives the precision of 89% for this process step. Unfortunately, our approach won't let us measure the recall for text extraction at sentence level. However, Table 1 may be used for measuring recall at the document level. The process of concept identification (technology recognition) has been done correctly for 385 sentences; this will gives the precision of 81% at sentence level. However, among the total number of 982 instances of technologies, 78 cases were marked up incorrectly; this will give the precision of 92% for technology recognition ignoring the text segmentation error. [3]

Among the 482 sentences, 201 sentences are annotated with at least one relation context: 37 *Noun-Compounds*, 26 *Prepositional* , 59 *Verb-based*, and 79 *Structural* relations. 55 sentences are annotated with relations of the type of *Residual*. Other sentences do not accompanied by a relation since they do not express any relation between the marked-up technologies e.g.

> "the result could be helpful to solve the variant problems of information retrieval , information extraction , question answering , and so on." [39]

Table III summarizes the frequency of relation contexts for the dataset.

TABLE III
FREQUENCY OF RELATION CONTEXTS IN THE DATASET OF 482 SENTENCES

Context	Frequency
Noun-compound	37
Verb-based	26
Prepositional	59
Structural	79
Residual	55

We finally mapped the lexical relations into the termino-conceptual relations manually (Defining $\omega : S \rightarrow \Sigma$ in Definition I in section III). For example, the lexical relations, S, such as `incorporate`, `is_combined_with`, and `integrate_with` are mapped into the termino-conceptual relation MERGE in Σ.

V. Conclusion and Future Work

We introduce the task of *Technology Structure Mining* as an example of a broader task of extracting concepts and relationships between them for a given text corpus. We proposed a "Technology Structure Graph" for formalizing the task. The major challenge is the lack of a benchmark dataset for evaluation and development purposes. The paper reports steps taken for constructing such a dataset which comprises of 482 sentences from the C section of the ACL ARC corpus. Each sentence is annotated with at least two technology terms and their interdependencies. We have also annotated

[3] We have defined precision as the number of correct annotations divided by the total number of annotations

the sentences with a linguistic context category that relation may be inferred from. Moreover, sentences are accompanied by other miscellaneous annotations such as modality of the relations, and the position of sentence in the article.

Future work will be the manual correction/annotation of part of speech tags and dependency parses for the selected sentences. This will enable us to study the performance of generic parsers on our dataset. Since the proposed task consists of several steps including text sectioning and segmentation, part of speech tagging etc. and as each of these processes is subject to error, there may be the danger of accumulated errors. The annotated dataset will enable us to study this in details.

Acknowledgment

This research is funded by Science Foundation Ireland under grant number SFI/08/CE/I1380(Lon-2).

References

[1] Afie M. Badawy, *Technology management simply defined: A tweet plus two characters*, J. Eng. Technol. Manag. Pages 219-224, ISSN 0923-4748, 2009.

[2] Diane Bailey and Paul M. Leonardi and Jan Chong, *Minding the Gaps: Understanding Technology Interdependence and Coordination in Knowledge Work*, Forthcoming Organization Science http://ssrn.com/paper=1334107, 2009.

[3] Louis W. Fry, *Technology-structure research: three critical issues*, Academy of Management Journal Volume 25, Pages 532-52, 1982.

[4] David Nadeau and Satoshi Sekine, *A survey of named entity recognition and classification*, ALinguisticae Investigationes Volume 30, Pages 3-26, 2007.

[5] Daniel Gildea and Daniel Jurafsky, *Automatic Labeling of Semantic Roles*, Computational Linguistics Volume 28, Pages 245-288, 2002.

[6] Dmitry Zelenko and Chinatsu Aone and Anthony Richardella, *Kernel methods for relation extraction*, J. Mach. Learn. Res. Volume 3, Pages 1083-1106, 2003.

[7] Christopher S. G. Khoo and Jin-Cheon Na, *Semantic relations in information science*, Annual Review of Information Science and Technology Volume 40, Pages 157-228, 2006.

[8] Michele Banko and Michael J. Cafarella and Stephen Soderland and Matthew Broadhead and Oren Etzioni, *Open Information Extraction from the Web*, IJCAI Pages 2670-2676, 2007.

[9] Philipp Cimiano and Paul Buitelaar and Johanna Volker, *Ontology Construction*, Handbook of Natural Language Processing, Second Edition Pages 577-605, 2010.

[10] Yuen-Hsien Tseng and Chi-Jen Lin and Yu-I Lin, *Text mining techniques for patent analysis*, Information Processing & Management Volume 43, Pages 1216-1247, 2007.

[11] Claire Cardie, *Empirical Methods in Information Extraction*, AI Magazine Volume 18, Pages 65-80, 1997.

[12] Paul Buitelaar and Philipp Cimiano and Peter Haase and Michael Sintek, *Towards Linguistically Grounded Ontologies*, 6th Annual European Semantic Web Conference (ESWC2009) Pages 111-125, 2009.

[13] Steven Bird and Robert Dale and Bonnie Dorr and Bryan Gibson and Mark Joseph and Min-Yen Kan and Dongwon Lee and Brett Powley and Dragomir Radev and Yee Fan Tan, *The ACL Anthology Reference Corpus: A Reference Dataset for Bibliographic Research in Computational Linguistics*, Proceedings of the Sixth International Language Resources and Evaluation (LREC'08) 2008.

[14] Jerry R. Hobbs and Ellen Riloff, *Information Extraction*, Handbook of Natural Language Processing, Second Edition Pages 511-533, 2010.

[15] *ACL Anthology Reference Corpus (ACL ARC)*, http://acl-arc.comp.nus.edu.sg/.

[16] Masaru Tomita and Marion Kee and Hiroaki Saito and Teruko Mitamura and Hideto Tomabechi, *The Universal Parser Compiler and Its Application to a Speech Translation System*, Proceedings of the 2nd International Conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages Pages 94-114, 1988.

[17] Koji Kakigahara and Teruaki Aizawa, *Completion of Japanese sentences by inferring function words from content words*, Proceedings of the 12th conference on Computational linguistics Pages 291-296, 1988.

[18] Iris Hendrickx and Su Nam Kim and Zornitsa Kozareva and Preslav Nakov and Diarmuid Ó Séaghdha and Sebastian Padó and Marco Pennacchiotti and Lorenza Romano and Stan Szpakowicz, *SemEval-2010 task 8: multi-way classification of semantic relations between pairs of nominals*, DEW '09: Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions Pages 94-99, 2009.

[19] Tom M. Mitchell and Justin Betteridge and Andrew Carlson and Estevam Hruschka and Richard Wang, *Populating the Semantic Web by Macro-reading Internet Text*,ISWC '09 Pages 998-1002, 2009.

[20] *The OpenNLP project.*, http://opennlp.sourceforge.net/.

[21] *Stanford Log-linear Part-Of-Speech Tagger*, http://nlp.stanford.edu/software/tagger.shtml/.

[22] *Using Corpus Query Language for complex searches*, http://www.fi.muni.cz/∼thomas/corpora/CQL/.

[23] Rebecca Hwa, *Learning probabilistic lexicalized grammars for natural language processing*,PhD Thesis, Harvard University, Adviser-Shieber, Stuart 2001.

[24] Chengzhi Zhang, *Extracting Chinese-English Bilingual Core Terminology from Parallel Classified Corpora in Special Domain*,WI-IAT '09: Proceedings of the 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology Pages 271-274, 2009.

[25] Sampo Pyysalo and Tapio Salakoski and Sophie Aubin and Adeline Nazarenko, *Lexical Adaptation of Link Grammar to the Biomedical Sublanguage: a Comparative Evaluation of Three Approaches*,CoRR , abs/cs/0606119, 2006.

[26] Joakim Nivre and Johan Hall and Sandra Kbler and Erwin Marsi, *Maltparser: A language-independent system for data-driven dependency parsing*,In Proc. of the Fourth Workshop on Treebanks and Linguistic Theories , Pages 13-95, 2005.

[27] Marie-Catherine de Marneffe and Bill MacCartney and Christopher D. Manning, *Generating Typed Dependency Parses from Phrase Structure Parses*,Proceedings of the IEEE / ACL 2006 Workshop on Spoken Language Technology , 2006.

[28] Mitchell P. Marcus and Beatrice Santorini and Mary A. Marcinkiewicz, *Building a Large Annotated Corpus of English: The Penn Treebank*,Computational Linguistics Volume 19, 1994.

[29] Dan I. Moldovan and Roxana Girju, *An Interactive Tool for the Rapid Development of Knowledge Bases*,International Journal on Artificial Intelligence Tools Volume 10, Pages 65-86, 2001.

[30] Marti A. Hearstu, *Automatic Acquisition of Hyponyms from Large Text Corpora*,In Proceedings of the 14th International Conference on Computational Linguistics Pages 539-545, 1992.

[31] Vivek Srikumar and Roi Reichart and Mark Sammons and Ari Rappoport and Dan Roth, *Extraction of Entailed Semantic Relations Through Syntax-Based Comma Resolution*,Proceedings of ACL-08 , Pages 1030-1038, 2008.

[32] Peyman Sazedj and Helena Sofia Pinto, *Mining the Web Through Verbs: A Case Study*,ESWC Pages 488-502, 2007.

[33] Takehito Utsuro and Kiyotaka Uchimoto and Mitsutaka Matsumoto and Makoto Nagao, *Thesaurus-based Efficient Example Retrieval by Generating Retrieval Queries from Similarities*,ESWC , 1994.

[34] Yutaka Sasaki and Yoshihiro Matsuo, *Learning semantic-level information extraction rules by type-oriented ILP*,Proceedings of the 18th conference on Computational linguistics , Pages 698-704 , 2000.

[35] Anoop Sarkar and Woottiporn Tripasai, *Learning verb argument structure from minimally annotated corpora*,Proceedings of the 19th international conference on Computational linguistics , Pages 1-7 , 2002.

[36] Dekai Wu and Grace Ngai and Marine Carpuat, *Why nitpicking works: evidence for Occam's Razor in error correctors*,COLING '04 , Pages 404-410 , 2004.

[37] Kimmo Koskenniemi and Pasi Tapanainen and Atro Voutilainen, *Compiling and Using Finite-State Syntactic Rules*,COLING, , Pages 156-162 , 1992.

[38] Emmanuel Planas Cyber and Emmanuel Planas, *Multi-level Similar Segment Matching Algorithm for Translation Memories and Example-Based Machine Translation*, COLING2000, , 2000.

[39] Takeshi Masuyama and Satoshi Sekine and Hiroshi Nakagawa, *Takeshi Masuyama and Satoshi Sekine and Hiroshi Nakagawa*,Proceedings of Coling 2004, ,Pages 1214-1219 , 2004.