Towards Technology Structure Mining from Scientific Literature

Behrang QasemiZadeh

Unit for Natural Language Processing, DERI
National University of Ireland, Galway
behrang.qasemizadeh@deri.org

Abstract. This paper introduces the task of Technology-Structure Mining to support Management of Technology. We propose a linguistic based approach for identification of Technology Interdependence through extraction of technology concepts and relations between them. In addition, we introduce Technology Structure Graph for the task formalization. While the major challenge in technology structure mining is the lack of a benchmark dataset for evaluation and development purposes, we describes steps that we have taken towards providing such a benchmark. The proposed approach is initially evaluated and applied in the domain of Human Language Technology and primarily results are demonstrated. We further explain plans and research challenges for evaluation of the proposed task.

1 Introduction

We are drowning in the sea of data and effective intelligent-contextual information retrieval systems have turned out to be strategic tools in different disciplines, among them interdisciplinary field of Management of Technology [1](MoT). The role technology plays in shaping our lives, and its critical role in an increasingly competitive knowledge based economy is a matter of fact. Technology is developed and propagates globally with a surprising velocity, and managing the accelerated rate of technology development becomes a universal challenge. MoT tries to bring efficiency in technology organization mainly through the process of Technology Watch. Technology Watch in general is the process of extracting tactical information about technology. However, the manual process of extracting such information is tedious and time consuming considering the gigantic amount of information. [2]

A long discussed topic in MoT is Technology-structure relationships [3]. One empirical research aspect of technology-structure relationship deals with *interdependence of technologies* i.e. how technologies are related to each other. We propose a linguistic based approach to facilitate the process of extracting information about technologies by proposing a methodology for extracting information about interdependencies of technologies -e.g. how technologies are built on top of each other. We have named the proposed task "Technology Structure Mining".

The proposed research involves several established research challenges in Information Extraction and Natural Language Processing such as Named Entity

P.F. Patel-Schneider et al.(Eds.): ISWC 2010, Part II, LNCS 6497, pp. 305–312, 2010.

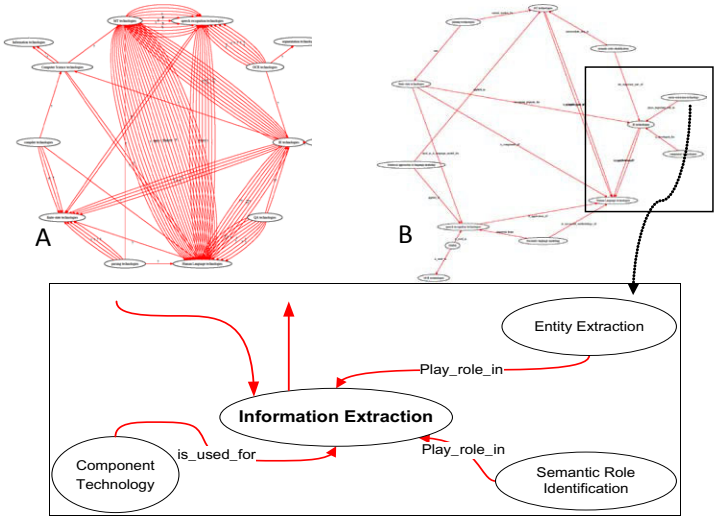

Fig. 1. In the above figure, ellipses show technologies and each labeled edge shows a relationship between pairs of technologies. The represented figure above has been generated from a part of publications in the ACL anthology reference corpus. Graph A illustrates state of the art in most text mining/ontology learning systems where co-occurrences of terms are usually considered as a measure for relating concepts. Graph B illustrates the goal of our proposed research where concepts are related to each other by help of natural language processing techniques for relation extraction. Graph A is generated automatically by help of our proposed method, while graph B is extracted and annotated by a careful study of graph A.

Recognition [4], Semantic Role Identification [5], and Relation Extraction [6],[7]. Considering technology as applied science, then scientific publications can be considered as a primary source of information about technologies and emerging technological trends. Figure 1 illustrates an example of the result of the proposed task after analysis of publications in the domain of Human Language Technology from ACL Anthology Reference Corpus (ACL ARC)[8] and offers a graphical representation of the outcome of analysis.

Evaluation and Understanding of the outcome of any task like the one proposed here remains a research challenge. In addition, 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. [9],[10]. In other words such research is more task-driven rather than fact-driven. We address and target these issues in our proposed research.

The rest of the paper is organized as follows. The next section briefly introduces related work. In section 3, we propose a formal definition for the proposed task and explain our goals through some examples. The applied methodology for approaching the task is briefly explained in section 4. In section 5, we report

statistical information of our analysis on our reference corpus. Finally we conclude and give the direction of our future work in section 6.

2 Related Work

There has been number of research directions for supporting MoT and the task of Technology Watch. Most of the reported research is focusing on the task of patent mining e.g. [11], assisting Intellectual Property Management [12], and technology road-mapping [13]. However, as to the knowledge of the author there is no research reported on mining information specifically from scientific publications for the task of technology interdependency mining.

We classify the task of Technology Structure Mining as an activity situated between two emerging research areas: Ontology Learning (OL)[14] and Open (Domain) Information Extraction (OIE)[15]. OL tries to extract *related* concepts and relations from a given corpus automatically. In [14], 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. However, OIE is an extraction paradigm that extracts a large set of relational tuples from a given corpus without requiring any human input e.g. TextRunner System [16]. As defined, OIE gets a corpus as an input and it generates a list of relational tuples as output. Although it is claimed that the sole input to an OIE system is a corpus, these systems still use self-supervised learners that rely on a classifier that needs to be trained prior to full scalable applications. Evaluation of both OL and OIE remains to be a research challenge and unclear.

Finally, we 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 enables Bio-Text miners to build new semantic layers on top of already existing semantic resources (ontologies).

3 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.

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. In a similar way, a semantic relation between pairs of technologies can be conveyed by different lexical representation e.g. lexical relations such as *used in*, *applied in*, and *employed by* are expressing the same conceptual relation *DEPEND ON*.

We name the result of the above 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.,* $\langle NLP, \mathsf{TECHNOLOGY} \rangle$ *or* $\langle Lexicon, \mathsf{RESOURCE} \rangle$ *or* $\langle Quality, \mathsf{PROPERTY} \rangle$*. 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.,* $\alpha : V \times V \rightarrow \Sigma \times M$*.* 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.,* $\beta : P \rightarrow V$*.*
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." [17]

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

$V = \{\langle \text{NLU}, \mathsf{TECHNOLOGY} \rangle, \langle \text{SR}, \mathsf{TECHNOLOGY} \rangle\}$
$P = \{\text{natural language understanding}, \text{speech recognition}\}$
$\Sigma = \{\text{MERGE}\}$
$S = \{integrate\ with\}$
$\beta = \text{natural language understanding} \mapsto \langle \text{NLU}, \mathsf{TECHNOLOGY} \rangle$

speech recognition $\mapsto \langle \text{SR}, \text{TECHNOLOGY} \rangle$
$\omega =$ integrate with $\mapsto \text{MERGE}$
$\alpha = \langle \langle \text{SR}, \text{Technology} \rangle, \langle \text{NLU}, \text{Technology} \rangle \rangle \mapsto \langle \text{MERGE}, \Diamond \rangle$

In our proposed definition, we have considered the computational cost and complexity of the processes that are involved in the automatic generation of structured representation from natural language text. Therefore, in the proposed definition above the expressiveness of the model is not the only concern but also the practical computational aspect of converting natural language text into a structured model like the one we have proposed here.

As a step towards the proposed research goals in this paper, we have used the provided baseline in *Definition 1* for annotating a development dataset comprising of 486 sentences from the domain of Human Language Technology. Further information about the annotated corpus can be found in [18].

4 Proposed Methodology

Figure 2 presents a schematic view of the proposed methodology. The proposed method comprises of 5 major steps. (1) Text extraction deals with identification and extraction of text from scientific publications, (2) Indexing and storage

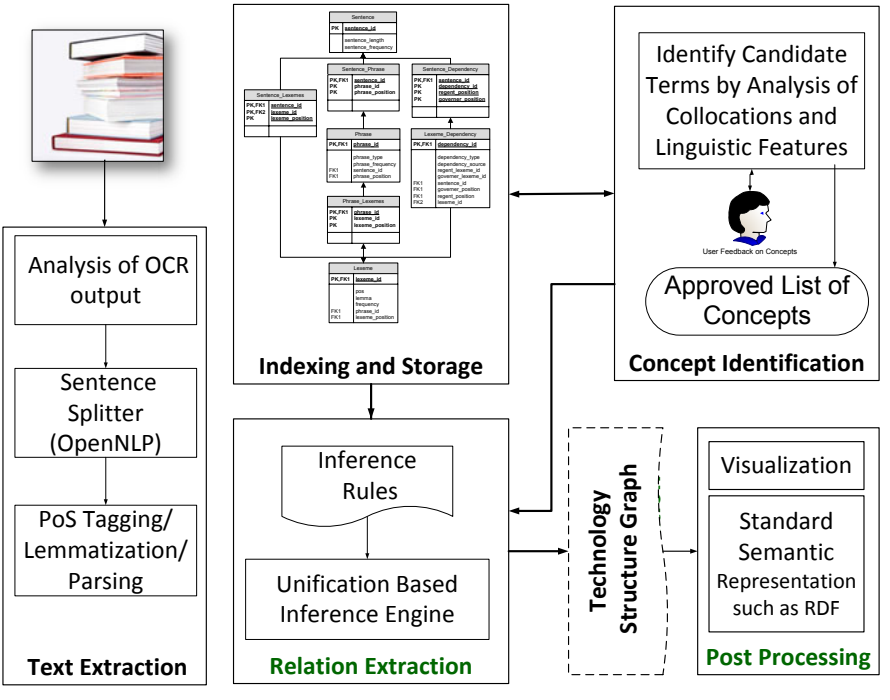

Fig. 2. Schematic of the Proposed Methodology

provides a suitable machine readable representation of extracted text (More information about the index scheme can be found in [18]) (3) Concept Identification marks technologies and their definitions in a semi-automatic manner (4) Parsing and Relation Extraction (RE) currently provides deep syntactic analysis of the stored sentences and extract relations between previously identified concepts by help of a unification based pattern matching over the syntactic annotations of the text (5) finally Post-processor provides a suitable representation of the extracted information e.g. a visualization for the proposed definition of Technology Structure Graph, or/and converting Technology Structure Graph to further standard representation such as RDF, and llinking the results into the Linked Open Data cloud[1].

5 Experimental Results

We have evaluated the proposed methodology on the C section of ACL Anthology Reference Corpus (ACL ARC)[8], which comprises of 2,435 articles from conferences in the domain of Human Language Technology. In the first step, we have been able to extract and index text from 2,003 articles. We fail to extract the text from the remaining 432 papers either because of deficiency in our heuristics for text extraction, or errors in the source XML files. The extracted text comprises of 6,168,312 tokens, 172,077 lexemes, and 230,936 sentences.

As figure 2 suggests, we then applied a set of heuristics to extract technology terms from the corpus. As a result, 147 different technology terms are extracted and suggested to the domain expert; this step finally results in 43 different technology classes where each technology class has different lexical variations. The corpus is then annotated with technology classes automatically. Figure 3 shows an example of the distribution of 4 technology classes over a time line of 25 years.

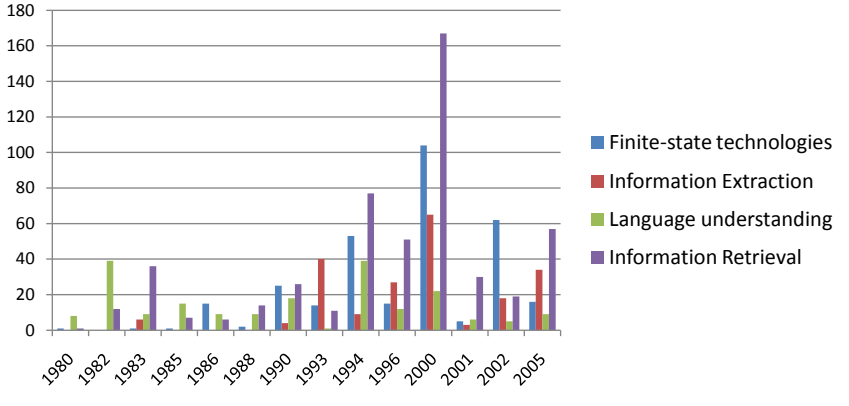

Fig. 3. Frequency of Four Technology Classes ordered in a time line of 25 years

[1] www.linkeddata.org

While we have been able to identify some of the relations automatically (Verb-Based Relations) between technology classes, the rest of the relations were extracted and tagged manually by an expert of the domain. This results in a development dataset for the proposed task of Technology Structure Mining. Details about dataset development can be found in [18]. The presented graph in Figure 1 has been generated with the help of the current post-processing module on the basis of automatically and semi-automatically extracted facts.

6 Conclusion and Future Work

In this paper we introduced the task of technology structure mining and proposed a formal definition for the task. Our efforts have resulted in the generation of a data set comprising of 486 sentences for training and evaluation purposes. Our future work will focus on step 4 and 5 of the method proposed in section 4. While current systems only extract verb-based relations, our experiment on the corpus of 486 sentences shows that only 10% of relations are conveyed by verbs. Therefore, extending the functionality of the relation extraction module beyond verb-based relations, e.g. relations expressed by apposition, is one of the goals of our future work.

We consider the mapping of extracted information to standard semantics and linking the information into the Linked Open Data cloud as an important step in our future research work. This comprises of mapping Σ, and V from the proposed definition1 in section 3 into already published ontologies or the ontologies that are going to be developed as part of our future work.

Methodologies for the evaluation of the proposed task is the other important focus of our future research. Each step of the proposed task is subject to error and each of the proposed processes is facing accumulated errors from the previous processes. We especially would be interested to investigate the role of the quality of each of the processes in the overall result, e.g., how errors at parsing natural language sentences effects the relation extraction step, and what is the impact of this error in the overall quality of the output of the system. We consider development of the dataset as an important step towards this goal.

Acknowledgements. This research is supported by Science Foundation Ireland grant SFI/08/CE/I1380(Lion-2). The author wishes to express sincere gratitude to Dr Paul Buitelaar for his supervision.

References

1. Badawy, A.M.: Technology management simply defined: A tweet plus two characters. J. Eng. Technol. Manag. 26, 219–224 (2009)
2. Maynard, D., Yankova, M., Kourakis, R., Kokossis, A.: Ontology-based information extraction for market monitoring and technology watch. In: End User Apects of the Semantic Web (2005)
3. Fry, L.W.: Technology-structure research: three critical issues. Academy of Management Journal 25, 532–552 (1982)

4. Nadeau, D., Sekine, S.: A survey of named entity recognition and classification. Linguisticae Investigationes 30, 3–26 (2007)
5. Gildea, D., Jurafsky, D.: Automatic labeling of semantic roles. Computational Linguistics 28(3), 245–288 (2002)
6. Khoo, C.S.G., Na, J.-C.: Semantic relations in information science. Annual Review of Information Science and Technology 40(1), 157–228 (2006)
7. Zelenko, D., Aone, C., Richardella, A.: Kernel methods for relation extraction. J. Mach. Learn. Res. 3, 1083–1106 (2003)
8. Bird, S., Dale, R., Dorr, B., Gibson, B., Joseph, M., Kan, M.-Y., Lee, D., Powley, B., Radev, D., Tan, Y.F.: The acl anthology reference corpus: A reference dataset for bibliographic research in computational linguistics. In: LREC 2008, Marrakech, Morocco (May 2008)
9. Hwa, R.: Learning probabilistic lexicalized grammars for natural language processing. PhD thesis, Harvard University, Cambridge, MA, USA. Adviser-Shieber, Stuart (2001)
10. Zhang, C.: Extracting chinese-english bilingual core terminology from parallel classified corpora in special domain. In: WI-IAT 2009, Washington, DC, USA, pp. 271–274. IEEE Computer Society, Los Alamitos (2009)
11. Tseng, Y.-H., Lin, C.-J., Lin, Y.-I.: Text mining techniques for patent analysis. Information Processing & Management 43(5), 1216–1247 (2007); Patent Processing
12. Oostdijk, N., Verberne, S., Koster, C.: Constructing a broad-coverage lexicon for text mining in the patent domain. In: LREC 2010, Valletta, Malta (May 2010)
13. Yoon, B., Phaal, R., Probert, D.: Structuring technological information for technology roadmapping: data mining approach. In: AIKED 2008, Stevens Point, Wisconsin, USA, pp. 417–422. World Scientific and Engineering Academy and Society, WSEAS (2008)
14. Cimiano, P., Buitelaar, P., Völker, J.: Ontology construction. In: Indurkhya, N., Damerau, F.J. (eds.) Handbook of Natural Language Processing, 2nd edn., pp. 577–605 (2010)
15. Banko, M., Cafarella, M.J., Soderland, S., Broadhead, M., Etzioni, O.: Open information extraction from the web. In: IJCAI, pp. 2670–2676 (2007)
16. Yates, A., Cafarella, M., Banko, M., Etzioni, O., Broadhead, M., Soderland, S.: Textrunner: open information extraction on the web. In: NAACL 2007, pp. 25–26. ACL, Morristown (2007)
17. Tomita, M., Kee, M., Saito, H., Mitamura, T., Tomabechi, H.: The universal parser compiler and its application to a speech translation system. In: Proceedings of the 2nd Inter. Conf. on Theoretical and Methodological Issues in Machine Translation of Natural Languages, pp. 94–114 (1988)
18. QasemiZadeh, B., Buitelaar, P., Monaghan, F.: Developing a dataset for technology structure mining. In: Proc. of IEEE International Conference on Semantic Computing (2010)