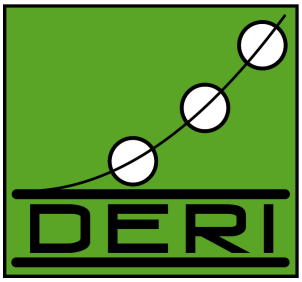
DERI – DIGITAL ENTERPRISE RESEARCH INSTITUTE
UNLP – UNIT FOR NATURAL LANGUAGE PROCESSING



TOWARDS TECHNOLOGY STRUCTURE MINING FROM TEXT BY LINGUISTICS ANALYSIS

Behrang QasemiZadeh

DERI TECHNICAL REPORT 2010-02-15
OCTOBER 2010

UNLP – UNIT FOR NATURAL LANGUAGE PROCESSING
DERI – DIGITAL ENTERPRISE RESEARCH INSTITUTE

DERI Galway
IDA Business Park
Galway, Ireland
www.deri.ie

DERI Technical Report
DERI Technical Report 2010-02-15, October 2010

Towards Technology Structure Mining from Text by Linguistics Analysis

Behrang QasemiZadeh[1]

Abstract. This report 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 research challenges and our research plan.

[1]Unit for Natural Language Processing, Digital Enterprise Research Institute, National University of Ireland Galway, Ireland
E-mail: behrang.qasemizadeh@deri.org.

Acknowledgements: The work presented in this report has been funded in part by Science Foundation Ireland under Grant No. SFI/08/CE/I1380 (Lion-2).

Contents

1 Introduction **1**
1.1 Research Challenges 1
1.2 Document Outline 2

2 Related Work **3**

3 Task Definition **4**

4 Proposed Methodology **6**

5 Data Analysis and Dataset Development **6**
5.1 Text Processing 9
5.2 Indexing and Storage 10
5.3 Concept Identification 10
5.4 Sentence Selection 11
5.5 Manual Verification of Analysis, Annotation and Grouping of Relations 11

6 Conclusion **15**
6.1 Research Agenda 15

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. In knowledge-based economy, organizational capability is not defined by what they know or what they can buy; however how well they learn and adapt. [2].

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. [3]

A long discussed topic in MoT is Technology-structure relationships [4]. One empirical research aspect of technology-structure relationship deals with *interdependence of technologies* i.e. how technologies are related to each other. In this report, I 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 from scientific literatures. Considering technology as applied science there is no doubt about the importance of science for technological innovation. [5] And therefore, scientific publications can be considered as a primary source of information about technological advances, and trends.

We have named the proposed task "Technology Structure Mining". The proposed task involves several established research challenges in Information Extraction and Natural Language Processing such as Named Entity Recognition [6], Semantic Role Identification [7], Relation Extraction [8],[9], and in a broader sense, Natural Language Understanding and Semantic Computing.

Figure 1 illustrates an example of the expected output for the proposed task in the domain of Human Language Technology (HLT). The figure has been generated semi-automatically, and by careful study of publications in the domain of HLT. The given example in the figure suggests that Information Extraction, Semantic Role Labeling, Entity Extraction, and Component Technology are amongst the identified technologies in the domain of HLT. In addition, it further suggests relations between technology concepts e.g. Semantic Role Identification *plays role in* Information Extraction.

The research in this area can result in methodologies for smoothing the progress of knowledge acquisition from natural language text; the acquired knowledge then models the domain's semantics in terms of the technologies that are involved in the domain. This further results in tools for contextual information retrieval, and respectively assists the process of technology management by providing means that supports higher level queries about technology concepts.

1.1 Research Challenges

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. [10],[11]. In other words such research is more task-driven rather than

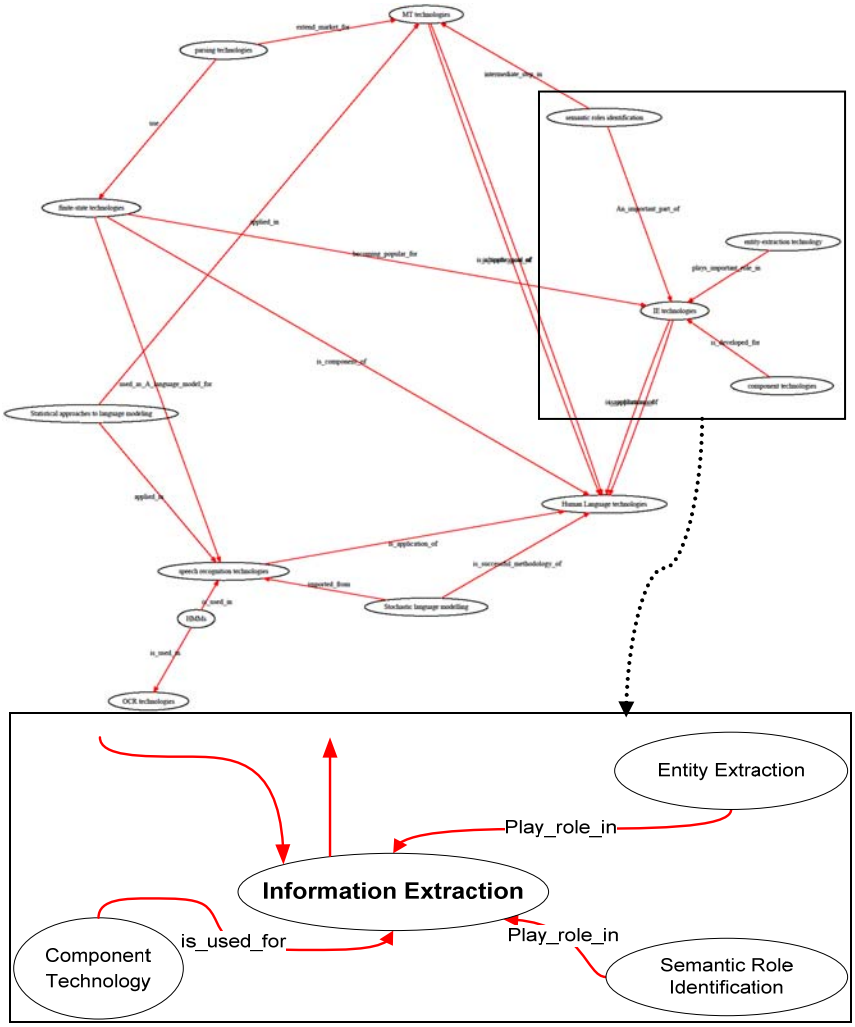


Figure 1: In the above figure, ellipses show technologies and each labeled edge shows a relationship between pairs of technologies. The represented figure above has been generated from a part of publications in the ACL anthology reference corpus in the domain of Human Language Technology. The graph illustrates the goal of our proposed research where concepts are related to each other by help of natural language processing techniques for relation extraction.

fact-driven. We address and target these issues in our research. We are particularly interested in the use and evaluation of generic natural language processing tools in a domain specific task.

1.2 Document Outline

The rest of this document is organized as follows. Section 2 introduces related work. A formal definition for the proposed task which further explains the research goals through examples is given in section 3. The applied methodology for approaching the task is explained in section 4. In section 5, we report experimental results which leads to developing a dataset for development and

evaluation purposes. Finally we conclude in section 6 and give the direction for future work.

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. [12], assisting Intellectual Property Management [13], and technology road-mapping [14]. 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)[15] and Open (Domain) Information Extraction (OIE)[16]. OL tries to extract *related* concepts and relations from a given corpus automatically. In [15], 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 [17]. 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.

In [18], 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.

Khoo and Na [8] 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. 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, 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:

1. identification of technology terms at the lexical level
2. mapping the lexical representation of technologies into a termino-conceptual level
3. extracting relations between pairs of termino-conceptual technologies at the lexical level (i.e. at sentence surface structure)
4. 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." [19]

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

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

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.

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 semantically interpret this input data. Any semantic interpretation by 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. 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 involves an eco-system that comprises concepts, relations and linking/connecting concepts to each other through these relations, in addition to the user's understanding 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 example 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."[20]

In the above sentence, the author(s) addresses the interaction between two technologies and provides information about an interdependence. Our definition 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.

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. Linguistic analysis of a digital natural language text requires clearly defined characters, words, and sentences in a document. This step cope with converting a raw text file into a well defined sequence of linguistically meaningful units.

2. *Indexing and storage*: provides a suitable machine readable representation of extracted text. Figure 3 shows our proposed index scheme. In fact, the proposed scheme offers lexical objects with linguistic annotations as the units of indexing. In other words, the provided indexing and storage scheme provides an information space where the features are lexical units that are lemmatized, and part of speech tagged.

3. *Concept Identification*: marks technologies and their definitions in a (semi)-automatic manner. A range of concept identification methods may be used at this stage. Currently, we have employed a method that is based on queries from the indexing scheme; the concepts are identified by help of querying and filtering that are enriched by linguistic features.

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. *Post-processor*: provides a suitable representation of the extracted information e.g. a visualization for the proposed definition of Technology Structure Graph such as figure 1, or/and converting Technology Structure Graph to further standard representation such as RDF, and linking the results into the Linked Open Data cloud [1].

5 Data Analysis and Dataset Development

As mentioned earlier, one of the main challenges to pursuing the proposed tasks is the lack of linguistic resources for evaluation and development. 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. 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.

[1] `http://www.linkeddata.org`

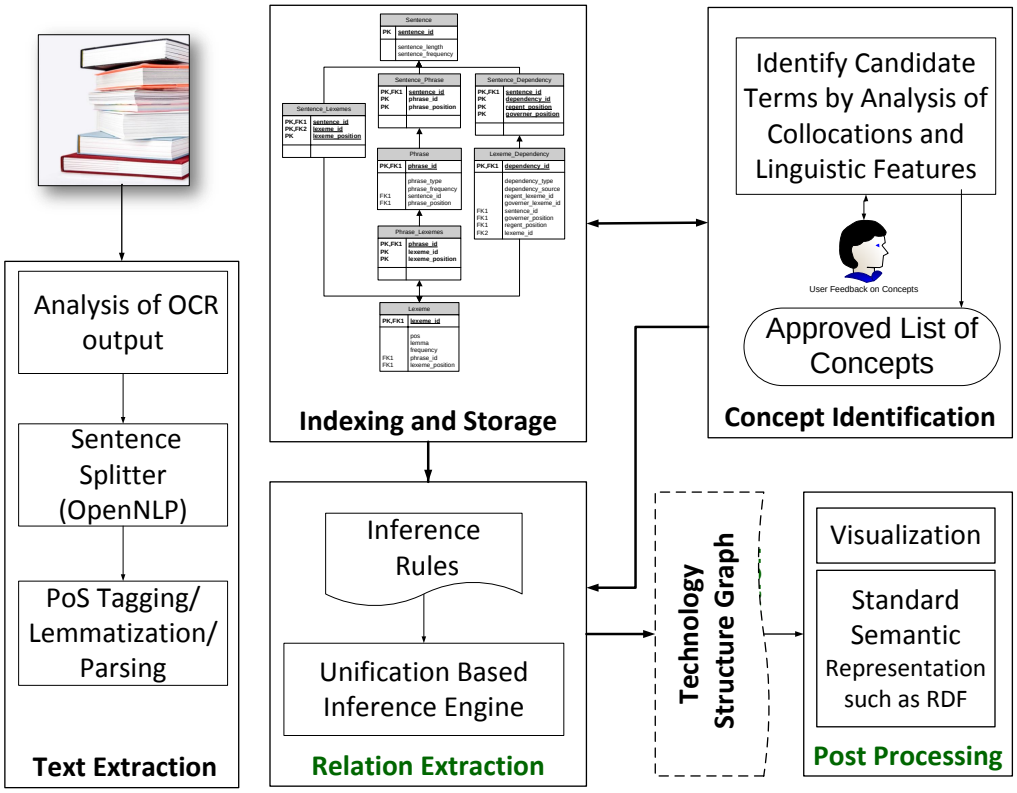


Figure 2: Schematic overview of the Proposed Methodology

The dataset comprises of sentences with at least two technology terms and their interdependencies. The sentences are extracted from the ACL Anthology Reference Corpus [21](ACL ARC) i.e. a corpus of scholarly publications about Computational Linguistics consisting of 10,921 articles which can be downloaded from [22]. The ACL ARC is represented in three different formats: source PDF files of articles, plain text, and an 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 further divided into different sections in directories labeled with a single letter, with 11 sections in total.

The dataset development essentially comprised 4 steps, similar to proposed methodology in section 4:

1. Text Processing
2. Indexing and Storage
3. Concept (technology) Identification
4. Manual Annotation and Compilation of dataset.

We studied the selected sentences manually, verified the processes, and annotated the sentences with the lexical/semantic relations between pairs of technologies.

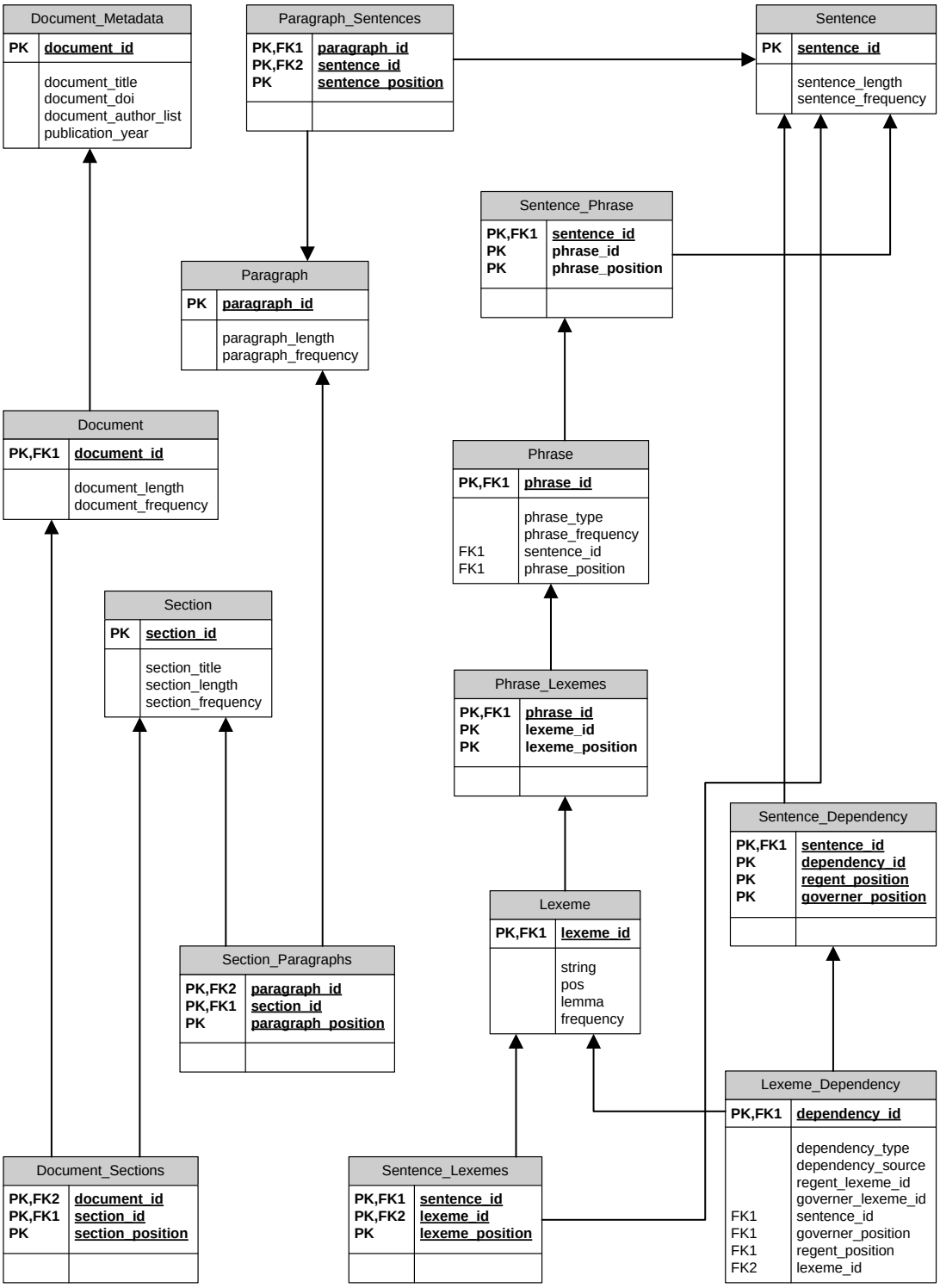


Figure 3: Data-Entity Relationship Diagram of the Indexing System

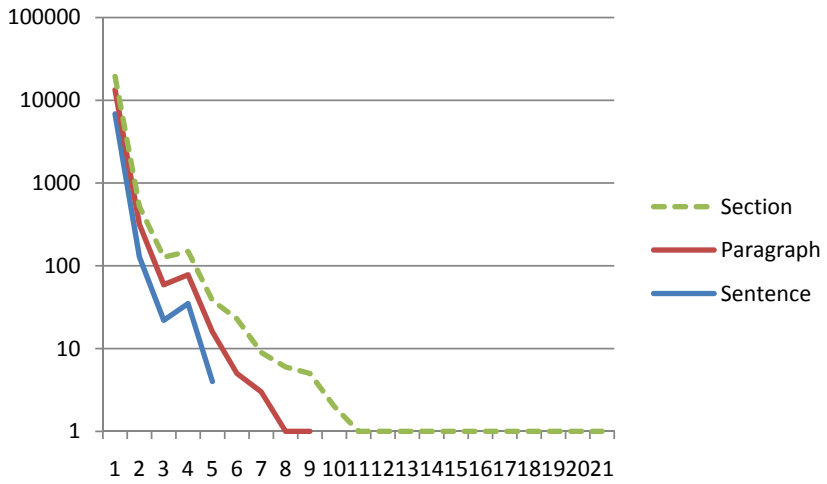


Figure 4: Distribution of co-occurrences of technology terms: The analysis shows that the co-occurrences of two technology terms tend to be at the boundary of sentences; The above diagram shows that if two technologies appeared together in a text boundary then it is most probable that these two terms 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.

We followed an iterative process for the dataset development. In the first step, the main issue is to find the optimum boundary size of text for dataset development e.g. should we focus at paragraph level or sentence level. To answer this question, in the first step we chose 1,424 random papers from the corpus and performed the following analysis. The selected papers consist of 45,031 paragraphs, 168,028 sentences, 4,524,062 tokens, and 124,525 types[1]. We studied 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 4). This means that if two technologies 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 the 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 [23]. In the remainder of this section we describe each step of the analysis in detail.

5.1 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

[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

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 performed 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 [24]. Since OpenNLP tools are trained on scientific publications, they tended to perform better when compared to other available tools. Then, We used the Stanford Part of Speech (POS) tagger [25] for tagging and lemmatization. The generated files are available for download[2]. The indexed sentences were also processed with open source dependency parsers: Malt Parser[26], BioLG [27], and Stanford Dependency Parser [28].

5.2 Indexing and Storage

The next step of the process involved indexing and storage of the corpus. We have used a data model -available at the URL in the footnote- that lets us dynamically generate a lexicon out of the POS 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 as part of speech. We have used the model to retrieve data from the corpus with queries similar to the Corpus Query Language[29] but at uniquely indexed text segments. Improved performance, reduced 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 **??**.

5.3 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 end with a token that has "technology" as its lemma. In addition, we applied 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[30]) 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 the evaluation of P, V, and the function β in Definition 1 in section 3. As an example, at the end of this step, P includes these strings: *information retrieval technology,information retrieval technologies,information retrieval,IR technology, IR*, while V has a member $\langle \text{IR}, \mathsf{TECHNOLOGY} \rangle$ and function β maps all the given values above for P to

[2] `http://nlp.deri.ie/behrang/sepid_arc.html`

$\langle \text{IR}, \textsf{TECHNOLOGY} \rangle$ in V. This processing step has been carried out on the sub-corpus of 1,424 random papers described above.

5.4 Sentence Selection

After choosing the technology classes and defining P, V and β for the corpus, we identified sentences that contain more than one string term from P. In this step, we extracted the sentences for each section of the ACL ARC; e.g. we were able to extract text from 2,435 papers out of section C (failing on 432 papers; either because of 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 1 and Table 2 show summarized statistics of the performed processes. Table 1 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 that failed to segment and index *(UNSUC-ARTICLE#)*. Table 2[2] shows statistics for the successfully indexed documents. This includes the numbers of tokens, types, identical sentences *(SENT)*, identical sentences with a minimum of 1 technology term *(SST1)* and identical sentences with more than one technology term *(SST2)* for each section of the corpus.

Table 1: Statistics For Text Processing Step

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**

5.5 Manual Verification of Analysis, Annotation and Grouping of Relations

In the final step of dataset development, we chose and annotated sentences from section C of the corpus. This section of the corpus comprises papers from different conferences from the years 1965 to 2004. Among the 230,936 sentences in this section of the corpus, only 2,012 sentences contain a technology term, and amongst these sentences only 482 have two or more lexical chains that signal

[2]The total numbers of articles proposed here are not identical to the numbers proposed in [21] due to corruptions in the source XML files; we have excluded these files from the corpus

Table 2: 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

appearance of technologies of different classes in the sentence. We manually read the extracted sentences and annotated them with the following information:

1. Whether the text processing step has been performed correctly: this comprised 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, this gives the linguistic context for the relation as described below.

4. Lexical Relation: whether a sentence implies a relation and how that is is expressed.

5. Grouping of Lexical Relations into Semantic Relations: 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 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))." [31]

   The above sentence suggests a relation as follows:

   $\langle\langle \text{NLP}, \mathsf{technology}\rangle, \mathsf{hasSubClass}, \langle \text{EB-NLP}, \mathsf{technology}\rangle\rangle$

   Noun-Compound is the only context that 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."[32]

   the above sentence suggests a relation as follows:

   $\langle\langle \text{MT}, \mathsf{technology}\rangle, \text{hasComponent}, \langle \text{NLP}, \mathsf{technology}\rangle\rangle$

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

   > "lexical Knowledge acquisition *plays an important role* in Corpus-Based NLP."[33]

   However, extracting relations of this type may not be as straight-forward because other relations e.g. noun-compounds may occur at the same time. For example, relations in the above sentence are as follows:
   $\langle\langle \text{lexical-KA}, \mathsf{technology}\rangle, \mathsf{isSubClassOf}, \langle \text{KA}, \mathsf{technology}\rangle\rangle$
   $\langle\langle \text{CB-NLP}, \mathsf{technology}\rangle, \mathsf{isSubClassOf}, \langle \text{NLP}, \mathsf{technology}\rangle\rangle$
   $\langle\langle \text{lexical-KA}, \mathsf{technology}\rangle, \mathsf{playsRoleIn}, \langle \text{CB-NLP}, \mathsf{technology}\rangle\rangle$

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

   > "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)."[34]

   This suggests the relation: $\langle\langle \text{POS-tagging}, \mathsf{technology}\rangle, \text{isProblemExampleOf}, \langle \text{NLP}, \mathsf{technology}\rangle\rangle$

5. *Other*: this category refers to relations that do not fit into any of the first three above categories and/or are too complicated to be automatically inferred via structure, e.g.:

   > "finite-state rules are represented Using regular expressions and they are transformed into finite-state automata by a rule compiler."[35]

   This conveys a relation between *Finite Automata* and *Compiler*. Consider another example sentence:

   > "In translation memory or Example-Based machine translation systems, one of the decisive tasks is to retrieve from the database ,the example that best approaches the input sentence." [36]

This expresses a relation between *Database Technology* and *Machine Translation Technology*. However, we believe that the expressed relations in these sentences are too complex: automatic extraction and expression of such relations by *TSG* may be far from reality. It is worthwhile to mention that we have identified some of the relations expressed by sentence structure that are difficult to extract automatically. For example, the temporal relation between the time of introducing "translation by Analogy" and "Example-Based NLP" expressed in the above sentence, and the temporal relation conveyed by the sentence given previously as an example of a noun-compound relation. We have grouped these relations under the *Other* category.

These different contexts have been studied in previous research e.g. [37, 38, 9, 39] and [40]. However, the authors are unaware of any reported research on the analysis of the distribution of these contexts, nor any corpus that provides linguistic context annotations 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 does not allow the measurement of the recall for text extraction at the 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 gives precision of 81% at the 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 (summarized in table 3): 37 *Noun-Compounds*, 26 *Prepositional*, 59 *Verb-based*, and 79 *Structural* relations. 55 sentences are annotated with relations of the type of *Other*. Other sentences are 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." [41]

Table 3: Frequency of Relation Contexts in the Dataset of 482 Sentences

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

We finally mapped the lexical relations into the termino-conceptual relations manually (Defining $\omega : S \rightarrow \Sigma$ in Definition 1 in section 3). 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.

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

6 Conclusion

We introduce the task of *Technology Structure Mining* as an example of a broader task of extracting concepts and the relationships between them for a given text corpus. We propose 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 482 sentences from section C of the ACL ARC corpus. Each sentence is annotated with at least two technology terms and their interdependencies. We have also annotated the sentences with a linguistic context category that relations may be inferred from. Moreover, sentences are accompanied by other miscellaneous annotations such as the modality of the relations, and the position of the sentence in the article.

6.1 Research Agenda

The main focus of the future work is on relation extraction between technology concepts. This mainly comprises of developing models for automatic construction of Σ, S, ω, and α from the definition 1. We aim to employ the developed dataset for studying different aspect of the use of machine learning techniques for accomplishing the task. We are specifically interested in developing models for automatic classification of natural language sentences for the task of relation extraction.

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 definition 1 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 another part 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. This may future requires the manual correction/annotation of part of speech tags and dependency parses for the selected sentences in the developed dataset. Moreover, this will enable us to study the performance of generic parsers on our dataset.

References

[1] Afie M. Badawy. Technology management simply defined: A tweet plus two characters. *J. Eng. Technol. Manag.*, 26:219–224, 2009.

[2] John Bessant, David Francis, Sandie Meredith, Raphael Kaplinsky, and Steve Brown. Developing manufacturing agility in smes. *IJMTM*, 2(1-7):730–756, 2000.

[3] Diana Maynard, Milena Yankova, Ros Kourakis, and Antonis Kokossis. Ontology-based information extraction for market monitoring and technology watch. In *End User Apects of the Semantic Web*, 2005.

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

[5] Marc De Vries and Guest Article. Guest article technology education: Beyond the "technology is applied science" paradigm. *Applied Science Paradigm. Journal of Technology Education*, 8:7–15, 1996.

[6] David Nadeau and Satoshi Sekine. A survey of named entity recognition and classification. *Linguisticae Investigationes*, 30:3–26, 2007.

[7] Daniel Gildea and Daniel Jurafsky. Automatic labeling of semantic roles. *Computational Linguistics*, 28(3):245–288, 2002.

[8] Christopher S. G. Khoo and Jin-Cheon Na. Semantic relations in information science. *Annual Review of Information Science and Technology*, 40(1):157–228, 2006.

[9] Dmitry Zelenko, Chinatsu Aone, and Anthony Richardella. Kernel methods for relation extraction. *J. Mach. Learn. Res.*, 3:1083–1106, 2003.

[10] Rebecca Hwa. *Learning probabilistic lexicalized grammars for natural language processing*. PhD thesis, Harvard University, Cambridge, MA, USA, 2001. Adviser-Shieber, Stuart.

[11] Chengzhi Zhang. Extracting chinese-english bilingual core terminology from parallel classified corpora in special domain. In *WI-IAT '09*, pages 271–274, Washington, DC, USA, 2009. IEEE Computer Society.

[12] Yuen-Hsien Tseng, Chi-Jen Lin, and Yu-I Lin. Text mining techniques for patent analysis. *Information Processing & Management*, 43(5):1216 – 1247, 2007. Patent Processing.

[13] Nelleke Oostdijk, Suzan Verberne, and Cornelis Koster. Constructing a broad-coverage lexicon for text mining in the patent domain. In *LREC'10*, Valletta, Malta, may 2010.

[14] Byungun Yoon, Robert Phaal, and David Probert. Structuring technological information for technology roadmapping: data mining approach. In *AIKED'08*, pages 417–422, Stevens Point, Wisconsin, USA, 2008. World Scientific and Engineering Academy and Society (WSEAS).

[15] Philipp Cimiano, Paul Buitelaar, and Johanna Völker. Ontology construction. In Nitin Indurkhya and Fred J. Damerau, editors, *Handbook of Natural Language Processing, Second Edition*, pages 577–605. 2010.

[16] Michele Banko, Michael J. Cafarella, Stephen Soderland, Matthew Broadhead, and Oren Etzioni. Open information extraction from the web. *IJCAI*, pages 2670–2676, 2007.

[17] Alexander Yates, Michael Cafarella, Michele Banko, Oren Etzioni, Matthew Broadhead, and Stephen Soderland. Textrunner: open information extraction on the web. In *NAACL '07*, pages 25–26, Morristown, NJ, USA, 2007. Association for Computational Linguistics.

[18] Jerry R. Hobbs and Ellen Riloff. Information extraction. In Nitin Indurkhya and Fred J. Damerau, editors, *Handbook of Natural Language Processing, Second Edition*. CRC Press, Taylor and Francis Group, Boca Raton, FL, 2010. ISBN 978-1420085921.

[19] Masaru Tomita, Marion Kee, Hiroaki Saito, Teruko Mitamura, and Hideto Tomabechi. 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*, pages 94–114, 1988.

[20] Koji Kakigahara and Teruaki Aizawa. Completion of japanese sentences by inferring function words from content words. In *Proceedings of the 12th conference on Computational linguistics*, pages 291–296, Morristown, NJ, USA, 1988.

[21] Steven Bird, Robert Dale, Bonnie Dorr, Bryan Gibson, Mark Joseph, Min-Yen Kan, Dongwon Lee, Brett Powley, Dragomir Radev, and Yee Fan Tan. The acl anthology reference corpus: A reference dataset for bibliographic research in computational linguistics. In *LREC'08*, Marrakech, Morocco, May 2008.

[22] Acl anthology reference corpus (acl arc). http://acl-arc.comp.nus.edu.sg/.

[23] Tom M. Mitchell, Justin Betteridge, Andrew Carlson, Estevam Hruschka, and Richard Wang. Populating the semantic web by macro-reading internet text. In *ISWC '09: Proceedings of the 8th International Semantic Web Conference*, pages 998–1002, Berlin, Heidelberg, 2009. Springer-Verlag.

[24] The opennlp project. http://opennlp.sourceforge.net/.

[25] Stanford log-linear part-of-speech tagger. http://nlp.stanford.edu/software/tagger.shtml/.

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

[27] Sampo Pyysalo, Tapio Salakoski, 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.

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

[29] Using corpus query language for complex searches. http://www.fi.muni.cz/ thomas/corpora/CQL/.

[30] Mitchell P. Marcus, Beatrice Santorini, and Mary A. Marcinkiewicz. Building a large annotated corpus of english: The penn treebank. *Computational Linguistics*, 19, 1994.

[31] Takehito Utsuro, Kiyotaka Uchimoto, Mitsutaka Matsumoto, and Makoto Nagao. Thesaurus-based efficient example retrieval by generating retrieval queries from similarities. 1994.

[32] Yutaka Sasaki and Yoshihiro Matsuo. Learning semantic-level information extraction rules by type-oriented ilp. In *Proceedings of the 18th conference on Computational linguistics*, pages 698–704, Morristown, NJ, USA, 2000. Association for Computational Linguistics.

[33] Anoop Sarkar and Woottiporn Tripasai. Learning verb argument structure from minimally annotated corpora. In *Proceedings of the 19th international conference on Computational linguistics*, pages 1–7, Morristown, NJ, USA, 2002. Association for Computational Linguistics.

[34] Dekai Wu, Grace Ngai, and Marine Carpuat. Why nitpicking works: evidence for occam's razor in error correctors. In *COLING '04: Proceedings of the 20th international conference on Computational Linguistics*, page 404, Morristown, NJ, USA, 2004. Association for Computational Linguistics.

[35] Kimmo Koskenniemi, Pasi Tapanainen, and Atro Voutilainen. Compiling and using finite-state syntactic rules. In *COLING*, pages 156–162, 1992.

[36] Emmanuel Planas Cyber and Emmanuel Planas. Multi-level similar segment matching algorithm for translation memories and example-based machine translation. In *COLING*, 2000.

[37] Marti A. Hearst. Automatic acquisition of hyponyms from large text corpora. In *In Proceedings of the 14th International Conference on Computational Linguistics*, pages 539–545, 1992.

[38] Dan I. Moldovan and Roxana Girju. An interactive tool for the rapid development of knowledge bases. *International Journal on Artificial Intelligence Tools*, 10(1-2):65–86, 2001.

[39] Peyman Sazedj and Helena Sofia Pinto. Mining the web through verbs: A case study. In Enrico Franconi, Michael Kifer, and Wolfgang May, editors, *ESWC*, volume 4519 of *Lecture Notes in Computer Science*, pages 488–502. Springer, 2007.

[40] Vivek Srikumar, Roi Reichart, Mark Sammons, Ari Rappoport, and Dan Roth. Extraction of entailed semantic relations through syntax-based comma resolution. In *ACL*, pages 1030–1038, 2008.

[41] Takeshi Masuyama, Satoshi Sekine, and Hiroshi Nakagawa. Automatic construction of japanese katakana variant list from large corpus. In *Proceedings of Coling 2004*, pages 1214–1219, Geneva, Switzerland, Aug 23–Aug 27 2004. COLING.