Segregating Discourse Segments from Engineering Documents for Knowledge Acquisition

Madhusudanan N., B. Gurumoorthy, Amaresh Chakrabarti

Segregating discourse segments from engineering documents for knowledge acquisition

Madhusudanan N[1] , Gurumoorthy B[2], Amaresh Chakrabarti[3],

Virtual Reality Laboratory, Centre for Product Design and Manufacturing,
Indian Institute of Science, Bangalore – 560 012

[1] madhu@cpdm.iisc.ernet.in, [2]bgm@cpdm.iisc.ernet.in, [3]ac123@cpdm.iisc.ernet.in

Abstract. The broader goal of the research being described here is to automatically acquire diagnostic knowledge from documents in the domain of manual and mechanical assembly of aircraft structures. These documents are treated as a discourse used by experts to communicate with others. It therefore becomes possible to use discourse analysis to enable machine understanding of the text. The research challenge addressed in the paper is to identify documents or sections of documents that are potential sources of knowledge. In a subsequent step, domain knowledge will be extracted from these segments. The segmentation task requires partitioning the document into relevant segments and understanding the context of each segment. In discourse analysis, the division of a discourse into various segments is achieved through certain indicative clauses called cue phrases that indicate changes in the discourse context. However, in formal documents such language may not be used. Hence the use of a domain specific ontology and an assembly process model is proposed to segregate chunks of the text based on a local context. Elements of the ontology/model, and their related terms serve as indicators of current context for a segment and changes in context between segments. Local contexts are aggregated for increasingly larger segments to identify if the document (or portions of it) pertains to the topic of interest, namely, assembly. Knowledge acquired through such processes enables acquisition and reuse of knowledge during any part of the lifecycle of a product.

Keywords: Knowledge acquisition, mechanical assembly, discourse analysis, segmentation.

1 Introduction

In the process of realizing industrial scale products, assembly is a critical and integrative step. If potential assembly issues can be detected during the planning stages, expensive repetitions in assembly planning can be reduced. In order to do so, knowledge of assembly issues is necessary during the planning stage. Knowledge based systems have been explored as a means of providing such knowledge [1].

Knowledge based systems have been in use in a variety of applications for quite some time now.

The need for using knowledge entities in PLM systems has also been stressed in literature [2, 3]. The ability to manage knowledge generated during a product's lifecycle is key towards automation of design process [4]. Literature indicates the ability [5], importance [6] as well as the limited capability of PLM systems to manage and re-use of knowledge across life-cycles [7].

The acquisition of knowledge for such systems however, remains a bottleneck [8]. Automation of such knowledge acquisition is a larger goal of this research. Specifically, the work reported in this paper is intended to serve as the first step in automatically acquiring diagnostic knowledge pertaining to assembly from documents.

1.1 Background

The research reported here is part of a research work aimed at building a diagnostic system for mechanical assembly. In particular the focus is on the manual assembly of aircraft structures. Aircraft assembly is largely a manual process. The planning of such large scale part-assembly processes is a complex task. After an assembly plan is drawn up, in case there are issues while performing the actual assembly, the assembly plan might have to be revised, and each such revision adds to both cost and time delays. If assembly planners possess prior knowledge of such issues in advance, expensive and time-consuming iterations in the planning- assembly- replanning loop can be reduced. Sources of such knowledge are assembly experts, and documented collections of such issues. We chose documents as the source of knowledge for this research, since they in turn reflect the knowledge of experts who prepared them.

1.2 Documents as a knowledge source

In professional organizations, documents can be considered authoritative sources of knowledge, since they are usually prepared by multiple experts and undergo many reviews and revisions. They represent a repository of the experiences of multiple personnel. Documents that would be useful for our purpose are incident reports, standards manuals and best practices. Documents are also a step closer to being machine processible than knowledge that comes directly from experts.

2 Document Segmentation and Classification

Towards acquiring the necessary knowledge from documents, the first step is to identify whether a given document or some sections of the document belong to the relevant domain of interest – in this case, aircraft assembly. For example, a document pertaining to issues in assembly of aft-fuselage is relevant, whereas a document about annual sales of a toy is not. Methods for classifying text are available, notably from the domain of pattern classification and machine learning.

However, such methods typically require training data sets to be available for them to work effectively. However, due to reasons that concern the activities downstream in the knowledge acquisition process (elaborated later in the paper) we chose not to use these methods.
It may not be useful at all times to perform this classification only at a document level – sometimes only some parts of a document may be related to assembly (For example, in a document that contains feedback about workplace difficulties from an organization's employees, only the feedback from shop-floor employees is of interest). The challenge here is to filter such relevant and coherent chunks of text. Relevant chunks of text are those that semantically relate to the domain of aircraft assembly. By coherent chunks, we mean that these are collections of continuous and meaningful parts of a discourse. These pieces of text then serve as input for acquisition of diagnostic knowledge. We concentrate only on the sections of a document, rather than the entire document here. To summarize, the objectives of this paper are,

- To identify coherent sections of a given document
- To classify whether such coherent sections of the document pertain to the domain of aircraft assembly (and it's related domains)

2.1 Current methods

A number of methods are available to segment given data into meaningful chunks. As mentioned hitherto, machine learning based methods are quite useful [9]. However, such methods usually require large amounts of training data to be available, with the data being manually labeled *a priori*. There are mathematical methods combined with semantics available for text categorization as a standalone application [9]. Also dedicated efforts have been made to link the referred entity to its counterpart in a knowledge base, based on the topic of relevance [10].
The collection of words in a document can be used to determine the topic of discussion in the document, this being termed as a 'bag of words' approach in literature [11]. On a similar note one method uses word-sequences as a means of classification [11]. Document clustering is a popular application of techniques that can work without training data, as opposed to classification methods [12]. There is existing literature about the use of phrases and their semantic relationship, as well as the use of ontology for clustering [13]. Clustering documents based on a graph-based technique by detecting frequent sub-graphs of related terms is another method found in literature [14]. Another method uses sampling to discriminate segments of documents [9]. In this, a probabilistic method called Generalized Mallows Model (GMM) is used to model the topics of a text, and is used for segmentation. As regards to current PLM systems, there exists a piece of work to model and elicit information about key relationships and stakeholders by looking at emails [15].
Another relevant research is based on multi-paragraph segmentation using TextTiling algorithm [16], which divides a given text into predetermined blocks of equal size, and then looks at the semantic relatedness of words between these blocks. Related blocks are chunked together if they are closer than a specified threshold. This method is tested against the method proposed in this paper.

However, use of such methods may not help in the subsequent steps of knowledge acquisition, which demands an understanding of the document.

3 Discourse

Discourses are a common form of communication using natural language. They are considered useful to analyze and track the semantic content of a natural language exchange. Discourse analysis has been the focus of study for quite some time now, and there are different theories and approaches to doing so, see for instance [17]. A discourse can be considered to have a hierarchical structure [18] of segments, each of which is a sequence of clauses. The discourse itself may proceed in various ways, with interruptions, digressions and itemizations amongst the different segments.

3.1 Cue phrases

One of the means of distinguishing the boundaries between discourse segments is the use of cue phrases, also known as discourse markers [19]. Cue phrases such as "after that" and "by the way" signal the transition from one segment to the other. The type of deviation in the discourse context is associated with the type of cue phrase used.
Since discourse analysis helps to track how the previous sentence in a text influences the understanding of the current sentence [18], it is useful to consider documents as discourses, in which one or more authors try to communicate with the reader. The documents that are intended to be used here are those mentioned in the first section. However, technical documents are usually written in a formal manner, and do not resemble other forms of discourse such as conversations. The presence of discourse markers such as cue phrases is not guaranteed in this case.

4 Proposed Method

4.1 Assumptions

Before discussing the proposed method it is appropriate to state the assumptions that have been made here,

- A document is treated as a one-way discourse between the author and the reader;
- The knowledge represented in documents is correct and valid knowledge;
- Available semantic resources such as dictionaries and lexica are sufficient to cover the range of terminology used in technical documents.

4.2 Comparative studies

An intuitive means of classifying a document or parts of it is to look at the words used and their frequency. In a preliminary exercise, this approach was tried on documents and the results of such a classification were not always indicative of the content at the sentence level.

As mentioned in Section 2.1, TextTiling is another useful way of segmenting sections from a given text. An implementation of the TextTiling algorithm available as part of the NLTK-tokenizer [20] module was tested on a test document [case study]. Extracts of the text as segmented by the researchers and by the tiling algorithm used are presented in the box below.

The document was 4303 words in length, and was a case study of a wing manufacture [21]. Only a small portion of the entire document was considered. The result was compared with the segmentation of the same portion of the document obtained manually by eight test subjects, including the researcher. During the course of using the algorithm, two parameters needed to be adjusted to get a reasonable number of segments. The parameters that were varied were the block length and the block size. The combination which resulted in maximum number of segments was finally considered. The final number of segments using TextTiling was 39.

Figure 1 shows a comparison of how the TextTiling implementation performed against the manual segmentation.

(Drawing from NASA CR-4735.) Cost is the main barrier to use of carbon fiber composites in aircraft. They can cost from $60 to 400 per pound, compared to $0.33 for steel and $1.00 for aluminum.4 The main cost element is the fiber. In automobiles and recreational boats, glass fibers are used with epoxies and other polymers. These support much lower stresses but are sufficiently strong and stiff for those applications. They cost much less and are quite economical for those products. Another component of the cost is the molds. These are usually made of Invar or another material with very low thermal expansion coefficient in order that the curing process does not introduce size or shape variations. A third significant cost component is layup, by which is meant placing uncured composite materials onto the mold. This can be done by NC machines if the shape is flat or nearly flat, such as a wing skin, but is mainly done manually. A fourth cost is rework and repair. Composite parts are made in layers, and a major potential failure mode is delamination, or interior separation of the layers due to such causes as gas bubbles or insufficient bonding. Ultrasonic inspection is used to find such flaws, and increasingly they can be repaired even in thermosets. The process is still very costly, however, and the prospect of generating a flawed large assembly that becomes expensive scrap is a deterrent. This is ironic inasmuch as the ability to produce a large assembly all at once is one of the most attractive features of composite construction. Even though large subassemblies can be made all at once in an oven, final assembly still requires drilling holes and installing fasteners. This is just as critical and costly as in metal structures. Furthermore, the structural engineers worry every time a hole is drilled and fibers are cut. In some cases, the parts can be glued together. The parts and subassemblies at this stage are remarkably rigid. If they do not fit properly, it is not feasible to use the fasteners to draw them together. While solid or liquid shims are the only recourse, they reduce the strength of the structure and dilute many of the advantages of the method. Only small errors can be corrected this way. Therefore, part and subassembly size and shape accuracy are essential, and great effort is expended on molds and process control to achieve the necessary accuracy.

Cost is the main barrier to use of carbon fiber composites in aircraft. They can cost from $60 to 400 per pound, compared to $0.33 for ste
for aluminum.4 The main cost element is the fiber. In automobiles and recreational boats, glass fibers are used with epoxies and other polymers. These support much lower stresses but are sufficiently strong and stiff for those applications. They cost much less and are quite economical for those products.

Another component of the cost is the molds. These are usually made of Invar or another material with very low thermal expansion coefficient in order that the curing process does not introduce size or shape variations.

A third significant cost component is layup, by which is meant placing uncured composite materials onto the mold. This can be done by NC machines if the shape is flat or nearly flat, such as a wing skin, but is mainly done manually.

A fourth cost is rework and repair. Composite parts are made in layers, and a major potential failure mode is delamination, or interior separation of the layers due to such causes as gas bubbles or insufficient bonding. Ultrasonic inspection is used to find such flaws, and increasingly they can be repaired even in thermosets. The process is still very costly, however, and the prospect of generating a flawed large assembly that becomes expensive scrap is a deterrent. This is ironic inasmuch as the ability to produce a large assembly all at once is one of the most attractive features of composite construction. Even though large subassemblies can be made all at once in an oven, final assembly still requires drilling holes and installing fasteners. This is just as critical and costly as in metal structures. Furthermore, the structural engineers worry every time a hole is drilled and fibers are cut. In some cases, the parts can be glued together.

The parts and subassemblies at this stage are remarkably rigid. If they do not fit properly, it is not feasible to use the fasteners to draw them together. While solid or liquid shims are the only recourse, they reduce the strength of the structure and dilute many of the advantages of the method. Only small errors can be corrected this way. Therefore, part and subassembly size and shape accuracy are essential, and great effort is expended on molds and process control to achieve the necessary accuracy.

Fig. 1 An extract of the text showing segmentation by the researchers (left)and the tiling algorithm (right). Each change of color in a column indicates a shift in segment

Some observations on the results are as follows.

- In the graph, the red blocks on the second row indicate that 50% or more subjects have indicated a discourse segment i.e. where a shift in focus occurs, similar to that indicated in [Hearst]. This is compared against the segmentation provided by TextTiling, which matches up most of the segments as provided by the manual segmentation too. However TextTiling, by default looks at paragraph breaks as a shift in focus. On such instance in the test document, there was an itemization in the document, which was not perceived as a shift by all but one of the subjects. But tiling treated this as four segments as they appeared on different paragraphs.
- For the converse case, where there are multiple segments within a paragraph, tiling had only one exception (due to formatting issues in the input) and performed as expected. Other than these, the segments given by tiling matched with 3 subjects on 4 instances, with 2 subjects on 3 instances, with 1 subject on 4 instances, and with no subjects on 1 instance. Hence the performance for tiling was satisfactory in this case.

4.3 Discourse context for segmentation

As seen in the previous subsections, methods such as looking at the frequency of occurrence of words are not useful, since they do not concentrate on the semantic content of the discourse. The semantic content is important from the point of view of future activities in the research, such as identifying the entities in the domain, and extracting diagnostic knowledge that concern these entities.

Although TextTiling has performed segmentation at the most prominent segment boundaries, it identifies other boundaries that are not identified as so by the test subjects. Also, it is a difficult task to keep varying the parameters, namely the block-length and block-size parameters, for every document.

These parameters are important since the number of segments that are recognized are dependent on them. Moreover, since segmentation is only a preliminary step to enable filtering of relevant text portions. More importantly, we need to understand the content of a document and extract diagnostic knowledge from it. By understanding we mean that one should be able to list the entities and events in the text, and the relations among them. An additional case for using discourse analysis techniques is made by the fact that methods that look at words and their meanings do not address the task of resolving pronouns and anaphora. This is important since pronouns implicitly contain references to other words, and may not be captured by such methods.

In such a situation, discourse context is useful. In a given discourse the current context is defined by the entities that are being talked about, the activities that concern them and the relations amongst these entities. The list of entities is called Discourse Entity (DE) list [18]. In the domain of assembly the two important factors are product information and the process information [22]. These translate respectively to the nouns and verbs of sentences in natural language. Nouns would also cover the peripheral but related terms such as tools and the assembly environment. By treating the document as a discourse, it is also possible to find out which fact entails others by

means of inference. With this explanation the procedure for extracting relevant segments from a document is proposed as follows:

- Given text from a document, tokenize it into sentences;
- Resolve anaphora and pronouns on a per-sentence basis - This gives a DE list for every sentence;
- Segment the sentences which are both contiguous (i.e. within a certain distance d) and share parts of their DE list, within a specified threshold, say N_{common};
- Once the segments are recognized and marked, compare entities in the DE list to determine to how many of them relate to the assembly domain. The basis for comparison here would the set of terms (and their semantic neighbors) from one or more assembly ontologies;
- If, for a given segment, the semantic similarity (as indicated by a certain measure) is greater than a threshold, say D_{sem}, then classify that segment as being related;to assembly. For example, in the WordNet ontology, Jiang-Conrath similarity is one such example of semantic similarity.

4.4 Plan for implementation

For tokenization, any standard tokenizer (e.g. NLTK Punkt tokenizer [20]) that can split the input into sentences is an adequate choice. To resolve anaphora (back references to entities in previous sentences), pronouns and to perform related discourse analyses, methods of representation such as Discourse Representation Structure (DRS) [23] are available. Once the raw, tokenized text is represented in DRS, existing anaphora and pronoun resolution methods can be utilized. From the DE list, for a combination of d (See Section 4.3) and N_{common} (yet to be decided) the related segments can be separated. Alternatively, one could use unsupervised methods of classification such as k-means to automatically infer two groups. Then the DE list for every segment can be compared against one or more assembly (and related) ontologies [24] and classify whether that segment is related to assembly or not, based on the value of D_{sem} (See Section 4.3).

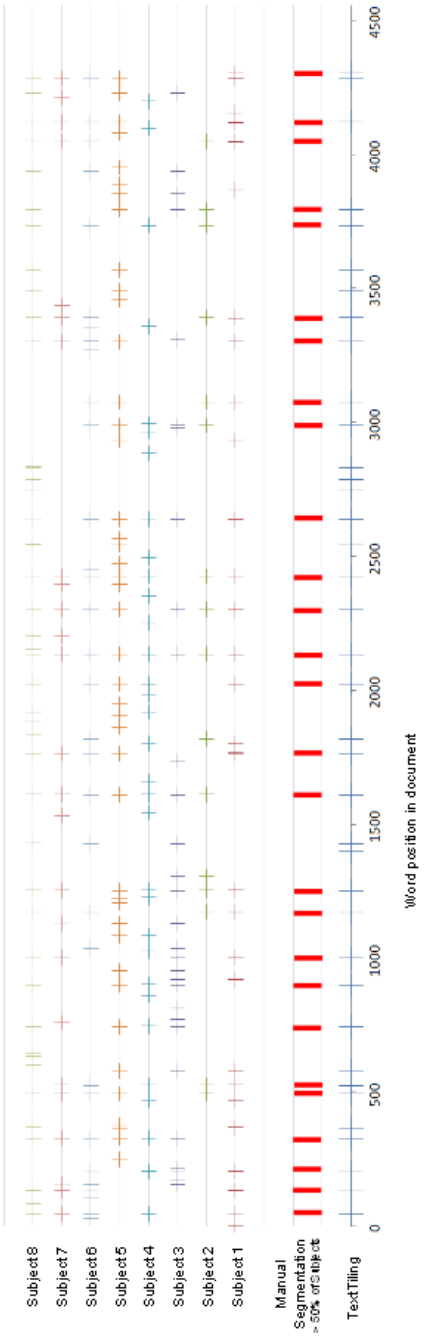

Fig. 2 Comparison of TextTiling vs manual segmentation for eight readers.

5 Conclusions

This paper has discussed the beginnings of a piece of work to acquire diagnostic knowledge for aircraft assembly from documents. In particular a method for segmenting relevant parts of a document that are related to assembly is proposed. Previous methods have been revised, and one method in particular, namely the TextTiling approach, has been tested on a typical aircraft assembly document with segmentation by human subjects as benchmark. As shown in Figure 2, majority of the subjects' segmentation have corresponded to the segments given by TextTiling. However there have been some specific instances where the desired result has not been achieved. The performance of the existing TextTiling method cannot be conclusively ruled out for our purposes - however, a different approach that is more suited to the future needs of the current research has been proposed. TextTiling does not ensure understanding of the text in the document as natural language and there are no measures such as resolution of pronouns and anaphora being employed to acknowledge their role in segmenting coherent sections.
The proposed approach treats documents as a discourse from the experts to the reader. Techniques from discourse analysis such as pronoun and anaphora resolution can be used to recognize and build coherent sections of a document. The discourse entity list can then be collected from such coherent sections and compared to those from domain ontologies to classify whether each segment is related to assembly or not.

6 Future Work

The paper has described a method for using discourse analysis techniques to classify relevant sections of a document. Some potential directions for implementation have also been touched upon. Future work includes implementing the method as a computer based program. This implementation then needs to be comprehensively tested to evaluate its effectiveness and to obtain feedback. The results of the implementation then have to be compared against the manual segmentation data as shown in this paper.

Acknowledgments. The authors wish to thank the members of IDeaS Laboratory at Centre for Product Design and Manufacturing, Indian Institute of Science, who volunteered as subjects for manual segmentation of the assembly text.

References

1. Madhusudanan, N. and Chakrabarti, A. (2013). *Implementation and initial validation of a knowledge acquisition system for mechanical assembly*. In CIRP Design 2012, pages 267–277. Springer.
2. Samira Sadeghi, Frederic Noel, Cedric Masclet, 'Collaborative specification of virtual environments to support PLM activities', PLM11 8th International Conference on Product Lifecycle Management

3. Teng, Fei, Néjib Moalla, and Abdelaziz Bouras. "A PPO Model-based Knowledge Management Approach for PLM Knowledge Acquisition and Integration." *International Conference on Product Lifecycle Management Eindhoven*. 2011.
4. Pugliese, Dante, Giorgio Colombo, and Maurizio Saturno Spurio. "About the integration between KBE and PLM." Advances in Life Cycle Engineering for Sustainable Manufacturing Businesses. Springer London, 2007. 131-136.
5. Briggs, Hugh C. Knowledge management in the engineering design environment. Pasadena, CA: Jet Propulsion Laboratory, National Aeronautics and Space Administration, 2006.
6. Penoyer, J. A., Burnett, G. J. F. D., Fawcett, D. J., & Liou, S. Y. (2000). Knowledge based product life cycle systems: principles of integration of KBE and C3P. Computer-Aided Design, 32(5), 311-320.
7. Brandt, S. C., Morbach, J., Miatidis, M., Theißen, M., Jarke, M., & Marquardt, W. (2008). An ontology-based approach to knowledge management in design processes. Computers & Chemical Engineering, 32(1), 320-342.
8. SE Savory. Some views on the state of the art in artificial intelligence. In Artificial intelligence and expert systems, pages 21–34. John Wiley & Sons, Inc., 1988.
9. Harr Chen. Learning semantic structures from in-domain documents. PhD thesis, Massachusetts Institute of Technology, 2010.
10. Xianpei Han and Le Sun. An entity-topic model for entity linking. In Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pages 105–115. Association for Computational Linguistics, 2012.
11. Yanjun Li, Soon M Chung, and John D Holt. Text document clustering based on frequent word meaning sequences. Data & Knowledge Engineering, 64(1):381–404, 2008.
12. Nicholas O Andrews and Edward A Fox. Recent developments in document clustering. Computer Science, Virginia Tech, Blacksburg, VA, Technical Report TR-07-35, 2007.
13. Hai-Tao Zheng, Bo-Yeong Kang, and Hong-Gee Kim. Exploiting noun phrases and semantic relationships for text document clustering. Information Sciences, 179(13):2249–2262, 2009.
14. M Shahriar Hossain and Rafal A Angryk. Gdclust: A graph-based document clustering technique. In Data Mining Workshops, 2007. ICDM Workshops 2007. Seventh IEEE International Conference on, pages 417–422. IEEE, 2007.
15. Loftus, C., Hicks, B. and McMahon, C., 2009. Capturing key relationships and stakeholders over the product lifecycle: an email based approach. *In*: *6th International Conference on Project LifeCycle Management (PLM 09)*, 2009-07-06 - 2009-07-08, Bath.
16. Marti A Hearst. Multi-paragraph segmentation of expository text. In Proceedings of the 32nd annual meeting on Association for Computational Linguistics, pages 9–16. Association for Computational Linguistics, 1994.
17. Barbara J Grosz and Candace L Sidner. Attention, intentions, and the structure of discourse. Computational linguistics, 12(3):175–204, 1986.
18. James Allen. Natural Language Understanding, 2/e. Pearson, 2011.
19. Bruce Fraser. What are discourse markers? Journal of pragmatics, 31(7):931–952, 1999.
20. Ntlk tokenize package, text tiling module, http://nltk.org/api/nltk.tokenize.html#modulenltk.tokenize.texttiling, October 2013.
21. Case study of aircraft wing manufacture, http://www.oup.com/us/static/companion.websites/9780195157826/chapter 19.pdf, October 2013.
22. N Madhusudanan and Amaresh Chakrabarti. Combining product information and process information to build virtual assembly situations for knowledge acquisition. ASME, 2011.

23. Hans Kamp, Josef Van Genabith, and Uwe Reyle. Discourse representation theory. In Handbook of philosophical logic, pages 125–394. Springer, 2011.
24. N. Lohse, H. Hirani, S. Ratchev, and M. Turitto. An ontology for the definition and validation of assembly processes for evolvable assembly systems. In Assembly and Task Planning: From Nano to Macro Assembly and Manufacturing, 2005. (ISATP 2005). The 6th IEEE International Symposium on, pages 242–247, 2005.