SEGREGATING DISCOURSE SEGMENTS FROM ENGINEERING DOCUMENTS

Madhusudanan N , Gurumoorthy B, Amaresh Charkrabarti

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

madhu@cpdm.iisc.ernet.in, bgm@cpdm.iisc.ernet.in, ac123@cpdm.iisc.ernet.in

Abstract. The automation of knowledge acquisition for knowledge based systems is a research challenge. The broader goal of the research being described here is to acquire diagnostic knowledge from documents in the domain of manual and mechanical assembly of aircraft structures. Natural language understanding techniques are useful tools for this, and in particular, discourse analysis is a means of understanding a discourse. We consider that a document is a discourse used by experts to communicate with others. The research challenge addressed in the paper is to identify documents or sections of documents that are potential sources of knowledge. From such segments, we later plan to extract the required knowledge. The segmentation task requires partitioning the document segment-wise and understanding the context of each segment. In discourse analysis, the division of a discourse into various segments is made by means of 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 an 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 are would 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, which is assembly.

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. This research uses knowledge based systems as a means of providing such knowledge. 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 [Sadeghi, Teng]. The acquisition of knowledge for such systems however,

remains a bottleneck [Savory]. 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 from documents.

1.1 Background

The research reported here is part of a larger effort to build a diagnostic system for mechanical assembly. In particular we focus 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 many such revisions might prove expensive. If the assembly planners possess prior knowledge of such issues in advance, iterations in the planning- assembly- replanning loop can be reduced. Sources of such knowledge are assembly experts, and documented collections of such issues. We choose documents as the source of knowledge for this research, since they would, 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 are the result of collecting the experiences of multiple personnel and aggregating them. Examples of documents that would be useful for our purpose are incident reports, standards manuals, best practices etc. Documents are also a step closer to being machine processible than the knowledge that comes directly from experts.

2. Document Segmentation

Towards acquiring the necessary knowledge from documents, the first step is to identify whether a given document belongs to the relevant domain of interest – in this case, aircraft assembly. Many methods of classifying are available in the present day, 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. Also, 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 is not enough to say if whether an entire document pertains to aircraft assembly (or related domains). Only some portions of a document may be relevant. 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 [Chen]. 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 [Chen]. Also dedicated efforts have been made to link the topic of relevance when an entity is being referred to [Han].

The collection of words in a document can be used to determine the topic of discussion in a document, this being termed as a bag of words approach in literature [Li]. On a similar note one method uses word sequences as a means of classification [Li]. Document clustering is a popular application of techniques that can work without training data, as opposed to classification methods [Andrews]. There is existing literature about the use of phrases and their semantic relationship, as well as the use of ontology for clustering [Zheng]. Clustering documents based on a graph-based technique by detecting frequent sub-graphs of related terms is another method found in literature [Hossain]. Another method uses sampling to discriminate segments of documents [Chen]. 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 [Loftus, Hicks, McMahon].

Another relevant research is the multi-paragraph segmentation using the TextTiling algorithm [Hearst], 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 are closer than a specified threshold. This method is tested against the proposed method in this paper.

The use of such methods may not aid us in the future steps of our knowledge acquisition, which demands 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, e.g. [Grosz]. A discourse can be considered to have a hierarchical structure [Allen] of segments, each of which is a sequence of clauses. The discourse itself may proceed in various ways, with interruptions, digressions, itemizations etc 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 [Fraser]. 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 [Allen], 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 are being made here,

- A document is treated as a one-way discourse between the author and the reader
- The knowledge represented in documents are correct and valid knowledge
- Available semantic resources such as dictionaries and lexica are sufficient to cover the range of language 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 and their frequency. In a preliminary exercise, this approach was tried on a document 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 ways of segmenting sections from a given text. An implementation of the TextTiling algorithm available as part of the NLTK-tokenizer [nltk] module was tested on a test document [case study]. An extract of the text as segmented by the researchers and tiling algorithm is presented in the box below.
The document was 4303 words in length, and was a case study of a wing manufacture [case_study URL]. Only a small portion of the entire document was considered. These were tested against copies of these documents which were manually segmented 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 blocksize.

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.

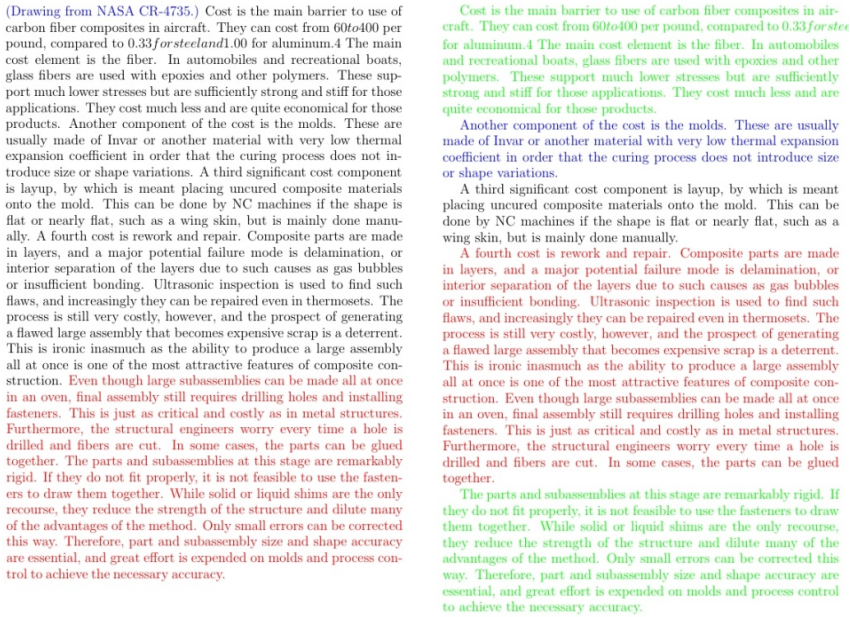

Figure 1 An extract of the text showing segmentation by the researchers (left) and the tiling algorithm (right).

Some of the observations 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 Text-Tiling, 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, that was not perceived as a shift by all but one of the subjects. But tiling treated this as four segments as they appear 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.

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 the 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 known segment boundaries, there are still other boundaries which have not been classified by the test subjects. Also, it becomes a difficult task to keep varying the parameters, namely the block length and blocksize parameters, for every document we encounter.
These parameters are important since the number of segments that are recognized are dependent on them. Moreover, from the perspective of the larger goal of this research, segmentation is not the only objective to be achieved - it 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 between 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 this 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 [3]. In the domain of assembly two important factors are the product information and the process information [Madhu]. These translate 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 out relevant segments from the document can be listed 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 specified distance d) and share parts of their DE list, within a specified threshold, say N_{common}
- Once the segments are recognized and marked, look at the DE list and compare them 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 measure) is greater than a threshold, say D_{sem}, then classify that segment as being related to assembly

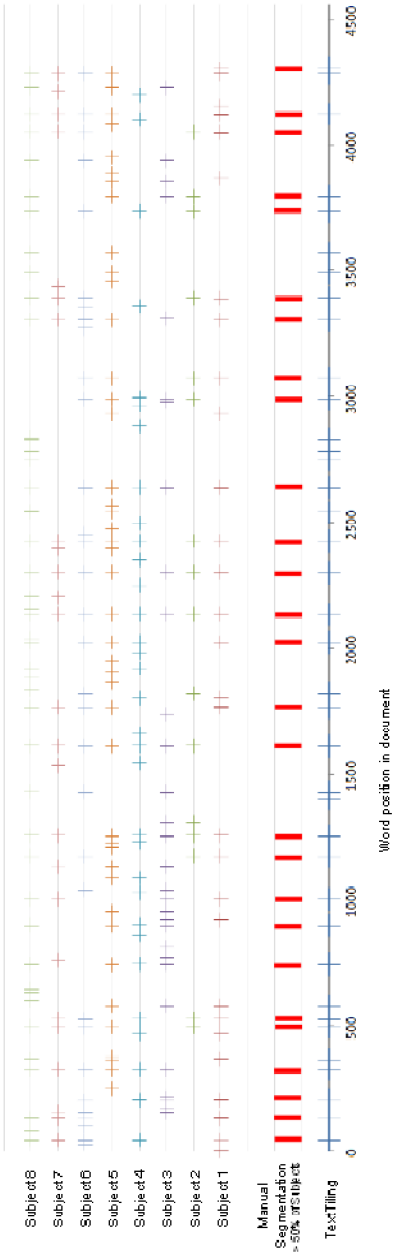

Figure 2 Comparison of TextTiling vs manual segmentation for eight readers.

4.4. Plan for implementation

For the tokenization, any standard tokenizer that can split the input into sentences is a useful choice. To resolve anaphora, pronouns and to perform related discourse analyses, methods of representation such as Discourse Representation Structure (DRS) [Kamp] 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 and N_{common} (yet to be decided) the related segments can 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 [Lohse] and classify whether that segment is related to assembly or not, based on the value of D_{sem}.

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 of segmenting out relevant parts of a document that are related to assembly is proposed. A set of previous methods have been referred to, and one method in particular, namely the TextTiling approach, has been tested on a typical aircraft assembly document. As shown in Figure 2, a 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. Text-Tiling does not ensure the 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 of using discourse analysis techniques to classify relevant sections of a document. Some potential directions for implementation have also been touched upon. The future work of this paper is to implement 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 as shown in this paper.

Acknowledgements. 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. Case study of aircraft wing manufacture, http://www.oup.com/us/static/companion.websites/9780195157826/chapter 19.pdf, October 2013.
2. Ntlk tokenize package, text tiling module, http://nltk.org/api/nltk.tokenize.html#modulenltk. tokenize.texttiling, October 2013. 3. James Allen. Natural Language Understanding, 2/e. Pearson, 2011.
3. James Allen. Natural Language Understanding, 2/e. Pearson, 2011.
4. Nicholas O Andrews and Edward A Fox. Recent developments in document clustering. Computer Science, Virginia Tech, Blacksburg, VA, Technical Report TR-07-35, 2007.
5. Kerem C¸ elik and Tunga Gungor. A comprehensive analysis of using semantic information in text categorization. In Innovations in Intelligent Systems and Applications (INISTA), 2013 IEEE International Symposium on, pages 1–5. IEEE, 2013.
6. Harr Chen. Learning semantic structures from in-domain documents. PhD thesis, Massachusetts Institute of Technology, 2010.
7. Bruce Fraser. What are discourse markers? Journal of pragmatics, 31(7):931–952, 1999.
8. Barbara J Grosz and Candace L Sidner. Attention, intentions, and the structure of discourse. Computational linguistics, 12(3):175–204, 1986.
9. 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.
10. 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.
11. 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.
12. Hans Kamp, Josef Van Genabith, and Uwe Reyle. Discourse representation theory. In Handbook of philosophical logic, pages 125–394. Springer, 2011.
13. 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.
14. 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.
15. N Madhusudanan and Amaresh Chakrabarti. Combining product information and process information to build virtual assembly situations for knowledge acquisition. ASME, 2011.
16. 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.
17. 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.

18. Samira Sadeghi, Frederic Noel, Cedric Masclet, 'Collaborative specification of virtual environments to support PLM activities', PLM11 8th International Conference on Product Lifecycle Management
19. 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.
20. 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.