2009

Information Extraction from Interviews to Obtain Tacit Knowledge: A Text Mining Application

Ramesh Sharda
Oklahoma State University, ramesh.sharda@okstate.edu

Michael Henry
Oklahoma State University, michael.henry@okstate.edu

Information Extraction from Interviews to Obtain Tacit Knowledge: A Text Mining Application

ABSTRACT

One of the most challenging knowledge management tasks is to obtain, summarize, and present tacit knowledge. It is important to develop approaches for creating insightful summaries from the knowledge obtained. In this paper, we present different information extraction methods for summarizing interview transcripts. Manual, semi-automatic, and automatic text analysis are evaluated to transform tacit knowledge into explicit form and to substantially reduce the time required to perform this transformation. These approaches are described in the context of a real application.

Keywords (Required)

Information extraction, tacit knowledge, knowledge summary, text mining

1. INTRODUCTION

Knowledge worker has become an essential component of organizational functionality. In fact, the most important and useful knowledge in an organization are the skills and expertise that a worker possesses (Koenig et al., 2004). Therefore, it is extremely important that systems are in place to gather the organizational "know-how" that an employee has obtained during their time with an organization. Unfortunately, the ability to capture worker knowledge is extremely difficult (Leonard and Sensiper, 1998). The tacit knowledge that humans possess is as valuable as it is elusive. Two reasons for this elusiveness: people are not fully aware of the working knowledge they possess and there is a lack of incentive on the individual level to make it explicit (Stenmark, 1999). Tacit knowledge "incorporates so much accrued and embedded learning that its rules may be impossible to separate from how an individual acts" (Davenport and Prusak, 1998). Additionally, the lack of incentive to make knowledge explicit is simply because knowledge resides within us and only presents itself through our actions. Therefore, we have no reason to document and share this information since we only draw upon it when we need it (Stenmark, 1999).

These issues present an interesting problem when attempting to obtain and codify tacit knowledge. Therefore, the objective of our study is to evaluate methods that can transform tacit knowledge into explicit form. Tacit knowledge is gathered through structured interviews. Then it is presented as a knowledge summary which contains punch-line statements, bullet point summary sentences, and keywords. We apply text mining techniques to generate these summaries in a semi-automatic and fully automatic mode. We compare efficiency of text mining approaches with a manual process. The remainder of this paper is organized as follows. Section 2 provides the definition of text mining and knowledge, as well as an explanation of knowledge summaries. Section 3 provides details for our case study. Section 4 details the manual, semi-automatic, and fully automatic knowledge summary creation process. Finally, section 5 concludes with a summary of experimental results and future research tasks.

2. BRIEF LITERATURE REVIEW

2.1 Text Mining

Text mining generally refers to the process of extracting interesting and non-trivial patterns of knowledge from unstructured text (Tan, 1999). As textual data continues to grow, the importance of text mining becomes more evident. There are two types of text mining systems: information extraction and text understanding systems. Information extraction systems are useful when only a fraction of the text is significant, information is mapped into a predefined target representation, and the writer's meaning in writing the text are of no interest (Appelt et al., 1993). In contrast, text understanding systems are relevant when the aim is to make sense of the entire text, the ability to understand the full complexities of the language are important, and there is an interest in recognizing the meaning of the writer (Appelt et al., 1993). Our use of text mining is the information extraction approach to obtain textual data containing tacit knowledge. We will explore use of text understanding techniques in the future.

2.2 Knowledge

In the business realm, knowledge is information that is relevant, actionable, and based partially on experience (Leonard and Sensiper, 1998). It is common to define knowledge of two types: tacit and explicit knowledge. Tacit knowledge is defined as the cognitive skills that are learned from experience and resides in the unconscious and semiconscious. Explicit

knowledge on the other hand, is codified and structured, and is accessible to people from sources other than the individuals originating it (Leonard and Sensiper, 1998). In our study, the methods are aimed at taking tacit knowledge in the form of interviews and transforming it into explicit form. We present the explicit form of knowledge as a knowledge summary.

3. CASE STUDY IMPLEMENTATION

The Defense Ammunition Center (DAC) is a division of the U.S. Army Joint Munitions Command (JMC). Their employees include Quality Assurance Specialists and trailer tracking personnel. Quality Assurance Specialists were ammunition experts deployed to the Middle East to inspect foreign ammunition and trailer tracking personnel monitored ammunition shipments throughout the United States. The tacit knowledge from workers returning from the field is gathered through interviews. Generation of knowledge summaries can be performed through a manual process. However, to speed up the process, we explored use of text mining for generating such summaries. This paper presents and evaluates three approaches: manual, semi-automatic, and automatic.

4. KNOWLEDGE SUMMARY CREATION

4.1 Manual Process

We adopt the IDEF0 (National Institute of Standards and Technology, 1993) and IDEF3 (Mayer et al., 1992) modeling approach to describe our process. The context diagram in Figure 1 shows four different components, the input, the controls, the mechanisms, and the output. Process details outlining the steps to create knowledge summaries are displayed in Figure 2.

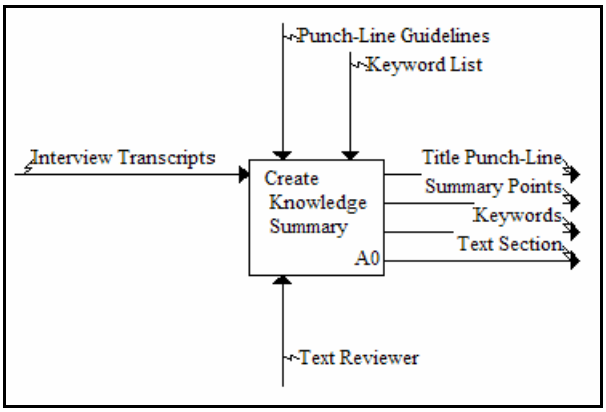


Figure 1. Context diagram (IDEF0) for Obtaining Knowledge Summaries

The input is the interview transcript, which consists of a dialogue between the interviewer and the interviewee. The controls are the guidelines for developing punch lines and the keyword list. A punch line is defined as a single sentence that actively portrays what the knowledge summary contains. The keyword list is a group of words that accurately defines the topic of the knowledge summary. The mechanism is the text reviewer, who is an individual that has a good understanding of any language and is able to accurately summarize sections of text. The output is the punch-line, summary points, keywords, and the sections of text used in the summary.

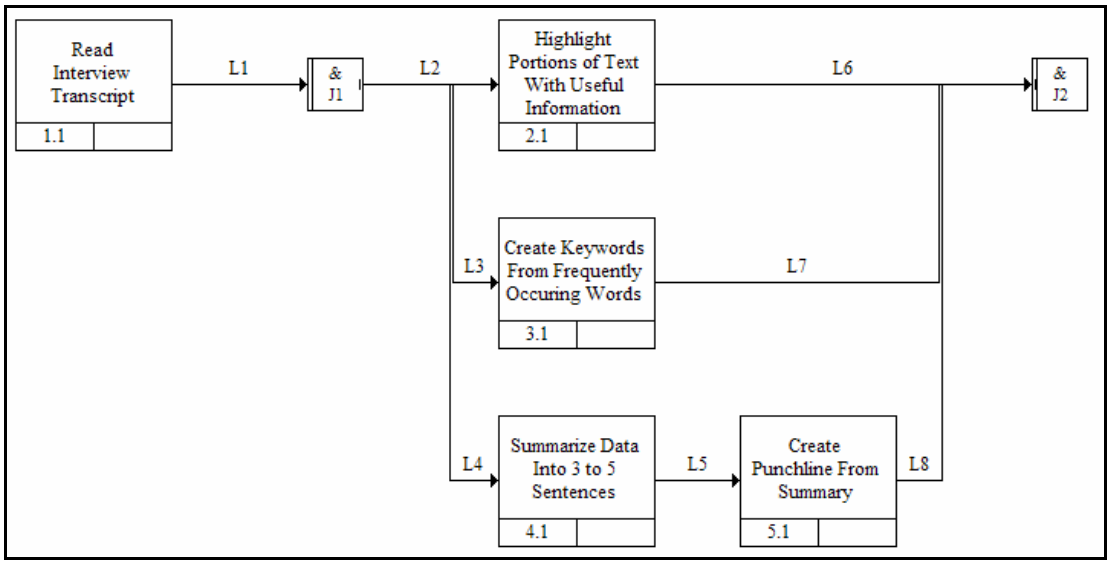


Figure 2. Process diagram (IDEF3) for Obtaining Knowledge Summaries

The creation of knowledge summaries begins with at least two reviewers reading an interview transcript and following a list of asynchronous tasks. For this example, we present the tasks in the order of highlighting text, creating keywords, summary sentences, and a punch-line statement. The reviewer highlights portions of text that contain useful information and adds keywords that accurately define each section. If keywords that describe the text are not available, the reviewer would create a new keyword based upon the frequency of words that occur throughout the transcript. The reviewer needs to create three to five summary points that accurately define the lessons learned and from these points, a punch-line statement is created. An example of a knowledge summary is displayed in Table 1.

Punch-Line	**It is essential to bring extra supplies and reference materials to deployment site.**
Keywords	**Kuwait, supplies**
Summary	**1. Bring extra batteries and Gatorade.** **2. Essential materials, such as MRE's and water, will be supplied.**

Table 1. Example of a Knowledge Summary

Manual extraction has the following advantages and disadvantages. Advantages include high retention as the reviewer receives a good understanding of the interview transcript, and high accuracy as the reviewer cross-checks summaries that were created. Disadvantages stem from high page length and high reading difficulty where each transcript contains 20 to 30 pages of domain specific abbreviations and vocabulary. Conflicts also occur when reviewers create variations of the same knowledge summary and keyword saturation develops as the keyword list escalates into a larger number of unmanageable keyword terms. Finally, the manual process is time consuming as each interviewee occasionally covers many topics when asked a specific question. This requires the reviewer to spend time re-reading and linking together fragmented paragraphs to obtain useful summary information. This leads us to explore using a computer assisted approach to creating such summaries.

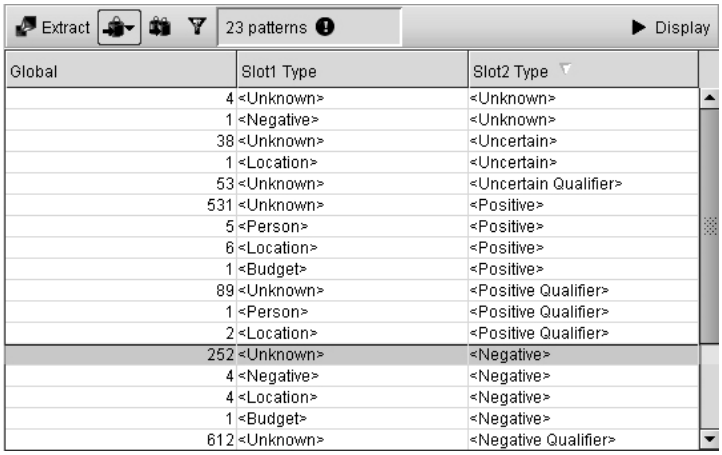
4.2 Semi-Automatic Summaries

Text mining techniques of text link analysis and categorization can be used to assist the reviewers in generating a knowledge summary. Text link analysis can obtain how a person feels about a certain subject and categorization can group different text sources together under similar concepts. Both techniques allow for the quick acquisition of important sections of text and keywords. To assist in the implementation of these methods, the text mining capabilities of SPSS Clementine were used.

4.2.1 Text Link Analysis

During text extraction, text is processed into single words and word phrases, which are labeled as terms. Relevant terms are then grouped together under a lead term called a concept. Pattern matching algorithms are applied to the extracted concepts in order to identify relationships across all documents (SPSS, Crowsey et al., 2007). Text link analysis can be grouped under the text mining topic of 'sentiment classification'. A further discussion of this topic can be found in (Pang et al., 2002).

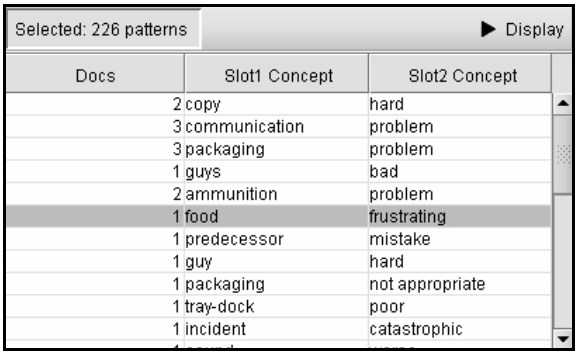
Figure 3 displays text link analysis patterns uncovered for a group of interview transcripts. The Global column shows the number of concepts discovered which pertain to a particular pattern. Slot1 Type contains a description of the terms found and Slot2 Type defines whether the pattern refers to a positive, negative, uncertain, or unknown opinion type.



Global	Slot1 Type	Slot2 Type
4	<Unknown>	<Unknown>
1	<Negative>	<Unknown>
38	<Unknown>	<Uncertain>
1	<Location>	<Uncertain>
53	<Unknown>	<Uncertain Qualifier>
531	<Unknown>	<Positive>
5	<Person>	<Positive>
6	<Location>	<Positive>
1	<Budget>	<Positive>
89	<Unknown>	<Positive Qualifier>
1	<Person>	<Positive Qualifier>
2	<Location>	<Positive Qualifier>
252	<Unknown>	<Negative>
4	<Negative>	<Negative>
4	<Location>	<Negative>
1	<Budget>	<Negative>
612	<Unknown>	<Negative Qualifier>

Figure 3. Identification of opinion types

Figure 4 shows the concepts found within a pattern where [Global] = 252, [Slot1 Type] = <Unknown>, and [Slot2 Type] = <Negative>. The Docs column represents the number of documents that contain a concept and an opinion. Slot1 Concept refers to a particular concept and Slot2 Concept refers to an opinion term. From this information, we can derive how an interviewee feels about a particular subject. In addition, other interviewees that have the same opinion about the same subject are presented. This allows for faster creation of knowledge summaries as the reviewer is no longer required to read the entire transcript. The reviewer can also add additional points to an existing knowledge summary from the different interview transcripts where a similar opinion is presented.



Docs	Slot1 Concept	Slot2 Concept
2	copy	hard
3	communication	problem
3	packaging	problem
1	guys	bad
2	ammunition	problem
1	food	frustrating
1	predecessor	mistake
1	guy	hard
1	packaging	not appropriate
1	tray-dock	poor
1	incident	catastrophic

Figure 4. A detailed view of concept-opinion link

Figure 5 displays the text area where [Slot1 Concept] = food and [Slot2 Concept] = frustrating. From this textual link, we find the story of an individual who was deployed to the Middle East and discovered essential food items were not always available. With the text highlighted, the reviewer can focus on a single section of text rather than having to evaluate the entire transcript. This alleviates the difficulties associated with the manual technique as high transcript length, difficult reading comprehension, and fragmented interviewee responses are less of an issue.

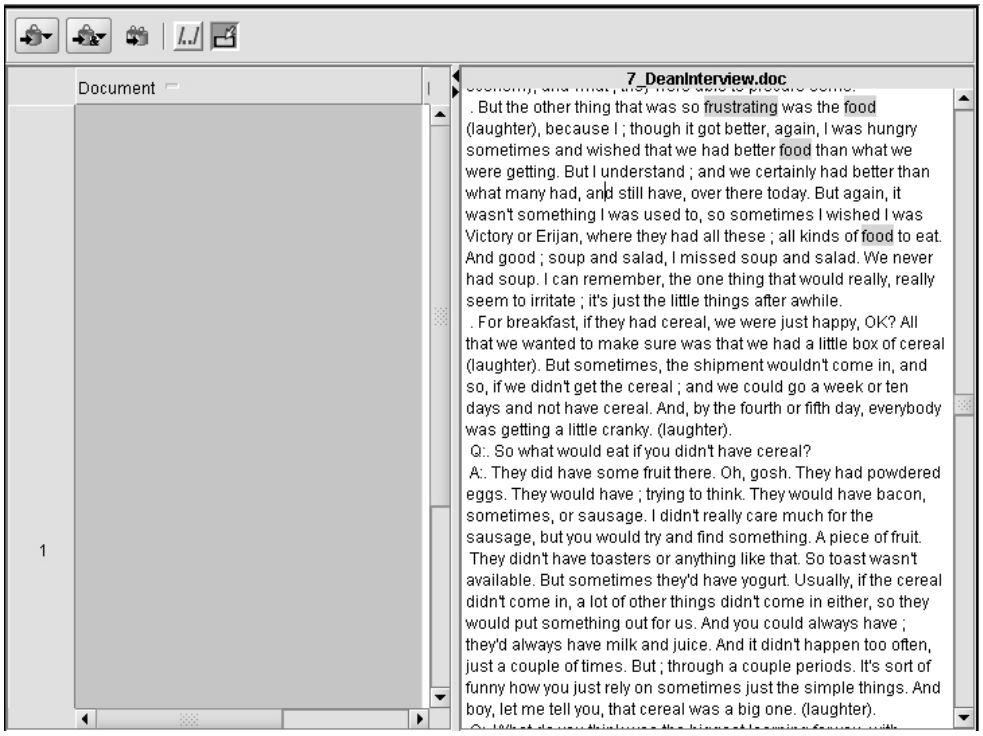
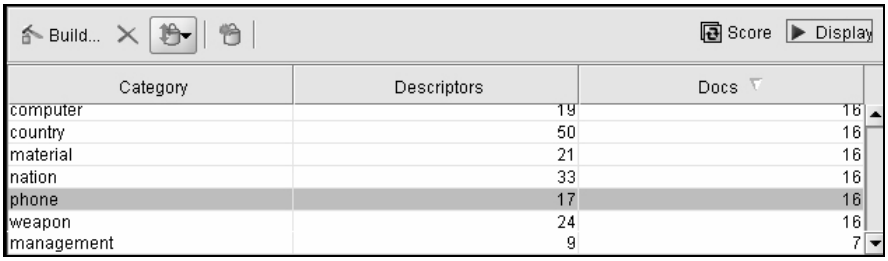


Figure 5. Text section identified from [food] concept and [frustrating] opinion link

4.2.2 Categorization

Categories are a set of descriptors which contain concepts, types, and rules. Once concepts, types, and rules have been defined, they are grouped together to form categories using concept-grouping techniques. These techniques include concept derivation, concept inclusion, semantic networks, and co-occurrence rules. Concept derivation creates categories by taking a concept and finding other concepts that are related to it (SPSS, Bontcheva et al., 2002, Turenne and Rousselot, 1998). Concept inclusion, on the other hand, groups concepts which include other concepts (Vogel and Powers, 2000).

Semantic networks create categories by grouping concepts together using word relationships found through online lexical databases, such as WordNet (SPSS, Miller, Mack and Hehenberger, 2002). Co-occurrence rules group concepts together that are strongly related within a set of documents (SPSS, Papernick and Hauptmann, 2005). An example of categories created is displayed in Figure 6. The Descriptors column displays the number of concepts, types, and rules that are contained within each category. The Docs column is the number of documents that pertain to each category.



Category	Descriptors	Docs
computer	19	16
country	50	16
material	21	16
nation	33	16
phone	17	16
weapon	24	16
management	9	7

Figure 6. Example of categories derived from interview transcripts

Figure 7 displays the text which pertains to the phone category within a single transcript. With the categorization method, a reviewer can instantly see text data that is related to the phone category across multiple documents. This allows for knowledge summaries to quickly be created and similarities between summaries to easily be discovered. In addition, the reviewer can form keywords from the concepts found within categories. For example, the summary that is created from the text in Figure 7 could have the keywords ‘satellite phone’ and ‘cell phone’.

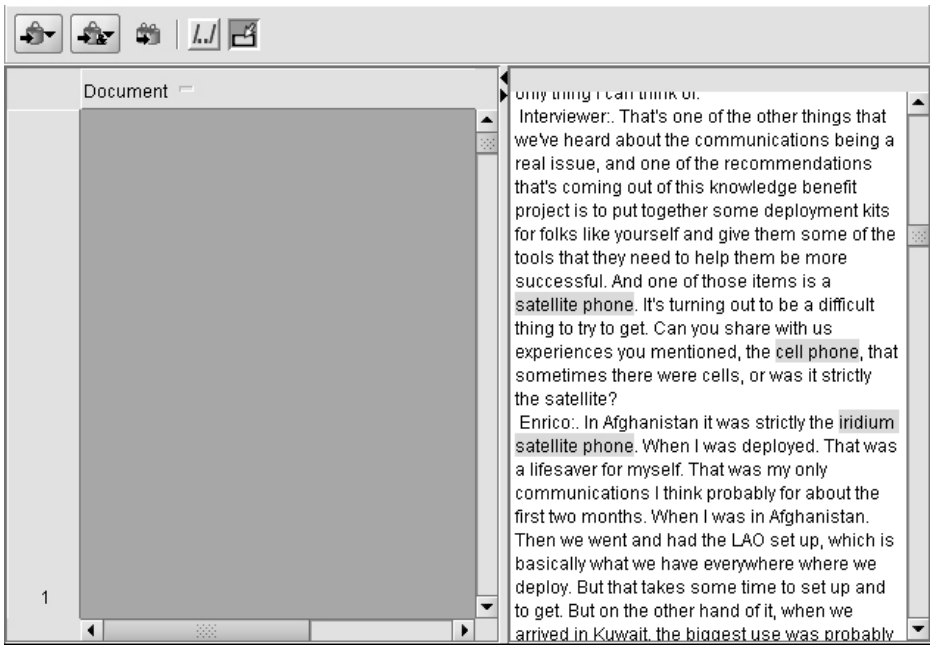


Figure 7. Screenshot of Text Section Identified

To gauge the advantages that categorization provides, a list of phone related knowledge summaries from the manual and semi-automatic processes are displayed in Table 3. Through categorization, reviewers were able to create six additional text summaries that were not created through the manual process. Reviewers may have missed these important summaries as they were located across multiple interview transcripts. While a reviewer may be considering the phone category when reviewing one transcript, they may lose focus on this category when studying different transcripts. With categorization, the reviewer can examine all transcripts at once while focusing on a single category.

MANUAL SUMMARY	SEMI-AUTOMATIC SUMMARY
DNS telephone limitations	**DNS telephone limitations**
Units strictly used the Iridium satellite phone which worked on DNS	**Units strictly used the Iridium satellite phone which worked on DNS**
Missed by Reviewers	**Purchase of phone cards needed in order to communicate**
"	**Satellite phones can only be used for a few minutes**
"	**E-mail is main source of communication**
"	**Specific type of pre-paid calling card must be used**
"	**Ensuring phone communication is available is a necessity**
"	**Satellite and cell phones will have spotty or no reception**

Table 3. List of Knowledge Summaries Created Through Manual and Semi-Automatic Process

Semi-automatic extraction thus improves speed as text sections containing knowledge are quickly identified, diversity as text sections across several documents can be indentified at the same time, and better insight as identification of new text sections is possible. A disadvantage is that the reviewer has to sort through identified text sections and still manually create the summaries.

4.3 Automatic Summaries

To address the disadvantages inherent in both manual and semi-automatic processing, a combination of readability statistics and word weighing algorithms were used to automatically obtain relevant sections of text, identify keywords, create summary points, and punch line titles. The application begins by processing an entire transcript by segmenting it into equal sentence sections. On these sections, computations for readability statistics and the mean and standard deviation of those statistics are used to determine the upper and lower bounds of retained sentences and delete the rest. Once a section has been retained, keywords, key summary points, and a punch line title are created. The process of our application can be seen in Figure 8.

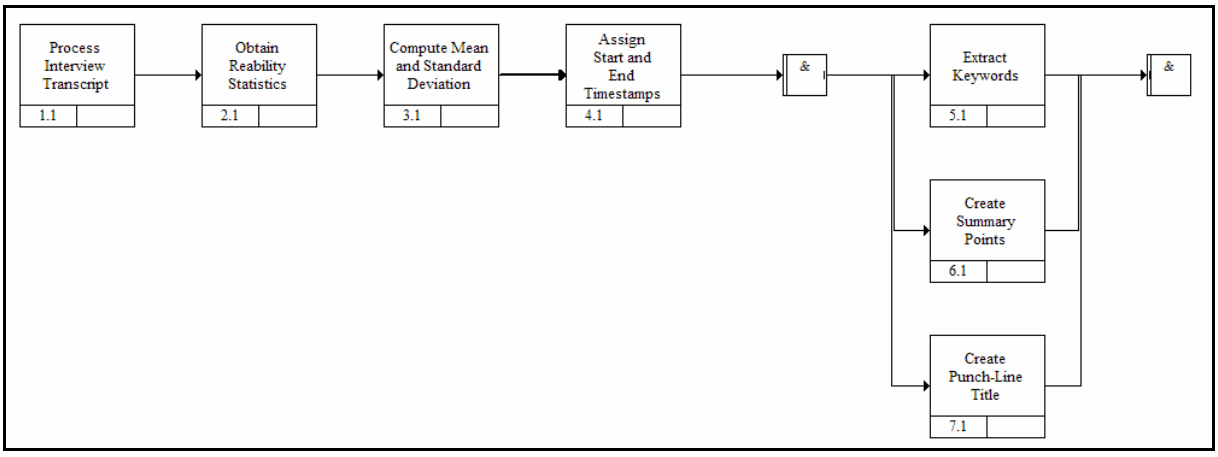


Figure 8: Automatic Summary Process

4.3.1 Obtaining Text Sections

The process begins with a reviewer submitting sections of text that contain knowledge. These sections are used as 'training data' to determine the criteria that the automatic application should use when it processes transcripts. The average number of sentences within the training data is used to obtain text sections from an unprocessed transcript. We label the sections obtained as the 'test data'. In addition to the word, character, and sentence count, readability statistics of Flesch Reading Ease (Flesch, 1948), the Flesh-Kincaid Grade Level (Kincaid, 1975), and the Coleman-Liau Index (Coleman and Liau, 1975) were calculated. A more detailed discussion of these concepts can be found at (Klare, 1974). The readability statistics are calculated on the training data and the mean and standard deviation are calculated from the readability statistics. Sections from the test data that are within one standard deviation of the mean for each readability statistic calculated from the training data, are identified as containing potential tacit knowledge. The accuracy of the application is also determined by comparing time stamps between the training data and the test data. Time stamps, in the form of minutes and seconds, are available in the original transcript and correspond to when an interview question has been asked and answered. If the test data time stamps fall within the start and end times of the training data time stamps, then the application is considered to have obtained text sections that contain tacit knowledge.

4.3.2 Identifying Keywords

The keyphrase extraction engine "XtraK4Me" was used to automatically obtain keywords from text. A component developed by Schutz (Schutz, 2008b), XtraK4Me processes textual documents into linguistic information such as part-of-speech tags, morphological base-forms, and noun chunks. With this information, XtraK4Me applies statistical algorithms to obtain important keywords. An overview of the underlying technology can be found at (Schutz, 2008a)

4.3.3 Creating Summary Points and Punch Line Title

To automatically identify key summary points, the auto summarize feature in Microsoft Word was used. The auto summarize feature determines key text points by analyzing the document and assigning a score to each sentence. Sentences that contain the most frequently used words in the document are given a higher score. The results returned are based upon a threshold percentage that is set by the user (Microsoft, 2009). For the creation of summary points, we define a 35% threshold. This threshold percentage was obtained through experimentation, as lower thresholds did not reveal useful

summaries and higher thresholds returned too many sentences from the original text. Additionally, the punch line title was created using a threshold of 5%.

5. CASE STUDY RESULTS

Through the manual process, 148 text sections were identified from a collection of DAC interview transcripts. We analyzed these through the automatic process. Using the Microsoft Word object library (Wyatt, 2009), we calculated the readability statistics for each text section. The mean and standard deviation were calculated for the training data and is presented in Table 4. The average number of sentences calculated from the training data was twenty sentences. The NLP sentence breaking function of Microsoft Word was used to obtain these sentence segments (CodeProject, 2006). Using the data from Table 4, the following SQL Query obtained sentences that contained tacit knowledge: SELECT * FROM AutoTextSection WHERE (wordCount BETWEEN 115 AND 453) AND (characterCount BETWEEN 480 AND 1876) AND (FReadingEase BETWEEN 71 AND 89) AND (FKincaidGrade BETWEEN 4 AND 8) AND (ColemanIndex BETWEEN 5 AND 7).

	Word Count	Character Count	Flesch Reading Ease	Flesh Kincaid Grade Level	Coleman Liau Index
Mean	**284**	**1178**	**80**	**6**	**6**
Standard Deviation	**169**	**698**	**9**	**2**	**1**
Within one SD of mean	**[115, 453]**	**[480, 1876]**	**[71, 89]**	**[4, 8]**	**[5,7]**

Table 4. Mean and Standard Deviation Readability Scores from the Control Group Text Sections

The text selection step (4.3.1) resulted in four text sections being identified as relevant. To determine the accuracy of our test data, Table 5 displays the timecodes for eight knowledge summaries obtained through the manual and automatic extraction process. Out of the eight knowledge summaries, the automatic method obtained sections relating to four of the existing knowledge summaries. Although this only points to a 50% success in automatic identification, as a first test of a commonly available tool, it is encouraging.

ID	Manual Text Section Timecode	Auto Text Section Timecode
1	**8:59 to 12:13**	**7:10 to 9:02, 9:02 to 10:51, 10:51 to 12:46**
2	**12:21 to 14:17**	**Missed by Text Mining**
3	**16:48 to 19:39**	**“**
4	**19:41 to 22:20**	**19:11 to 21:10**
5	**22:31 to 23:19**	**21:17 to 23:17**
6	**23:23 to 25:00**	**23:23 to 25:47**
7	**25:07 to 27:15**	**Missed by Text Mining**
8	**27:20 to 28:11**	**“**

Table 5. Time codes for manual and auto text sections

The four knowledge summaries identified by the application are used in extraction of keywords, summary points, and a punch line title (section 4.3.2). Table 6 displays the keywords extracted from the test data. Words marked in bold are keywords that matched between the different extraction methods. Out of the four knowledge summaries, automatic keyword extraction obtained similar keywords in three of the knowledge summaries. It is worthwhile to note that even though the automatic process did not obtain keywords for knowledge summary 5 and 6, new keywords such as ‘operations analyst’, ‘emergency’, and ‘message’ were identified that accurately describe the section of text. For this reason, we italicize these keywords to show that they still hold value despite not being obtained in the manual process.

ID	Manual	Automatic
1	Transportation **message**, overdues	**Messages**, time, yeah, people, minutes
4	**Emergency**, panic button	**Emergency**, message, truck
5	Split codes, SOP	Time, *operations analyst*, *emergency*
6	Fast paced environment	*Message*, people

Table 6: Keywords created through manual and automatic extraction

The results of the summary extraction process (section 4.3.3) are displayed in Table 7. The sections marked in bold are the sections of text that match between the two extraction methods. The automatic process was able to obtain similar data for all four existing knowledge summaries. Though the automatic summary was not able to obtain all of the key points defined through the manual process, it was able to capture the main topic points found within each text section.

ID	Manual Summary	Automatic Summary
1	**1.Messages arrive every 15 minutes.** 2.Work with messages as soon as they come in. **3. If message appears later that requires immediate attention, change focus to that task before completing previous task.**	**Q: Because, if I understand you correctly, you're getting messages every 15 minutes.** **Q: Do you look through those messages that you get at 9:15 and if there's one that's more important, do you stop what you're doing and focus on that**
4	**1. When accidents occur, such as hitting a deer, the drivers may not hit the panic button to indicate an emergency.** 2. Emergencies may come about one to two times a week.	**Q: Well, let's – all right.** **Have you ever gotten a message – I'm just curious about your perception – a message – you might receive a message that it should have been emergency?** **We've heard from others where somebody might've hit a deer where they should've actually pressed the emergency panic button and reported it through that.**
5	**1. Reference manuals with SOP and split codes available.** 2. Custom notes from individual employees on SOP are available upon request.	**A: There's a – there's book up there with split codes, and usually there's a book with the SOP in there, and I'll refer to that, refer to notes that I've had since I've been trained, and I'll try to listen and write down things that I don't know.**
6	**1. Work is demanding and may be frustrating at times.** 2. System currently being used is outdated and involves a lot of paperwork in a fast paced environment. **3. Will need to ask fellow employees for assistance if needed.**	**Q: Why is the how frustrating?** **A: And most of the time I reach a point, I stop, and I ask for help. I have no problem asking for help at all on anything.**

Table 7. Summary points created through manual and automatic extraction

Table 8 displays the results from obtaining punch line summaries. Recall that these were also obtained using the same process as in 4.3.3, but with a lower threshold. Though the sentences are not the same between the two methods, it is interesting to note that the automatic sentences directly refer to the topic of the knowledge summary. For example, in knowledge summary 1, the test data refers to messages that need to be prioritized, summary 2 refers to how emergencies are handled when they are not reported by the drivers, and summary 6 states that the installed legacy system currently prints so much paper that the worker gets lost in the rapid delivery of messages every 15 minutes. Summary 5 did not refer to the SOP's available, however, as Table 5 shows, an additional one minute and fourteen seconds were available in the section that was produced by the automatic extraction method. Reviewing this additional time, the title that was produced accurately defines the conversation as the text relates to drivers who did not send an alarm message when an accident occurred.

ID	Manual Title	Automatic Title
1	Prioritize work based upon importance of transportation message.	Because, if I understand you correctly, you're getting messages every 15 minutes.
4	Emergency may not be reported by driver depending on circumstances of the accident.	It'd be handled just like an emergency.
5	Standard and customized SOP's are available to the employee.	Especially – yeah, especially when drivers don't hit the panic button.
6	Daily activities of tracker can be demanding and frustrating.	Less paper and more of the screen – screen time looking at you.

Table 8. Punch-Line title created through manual and automatic extraction

CONCLUSION

These results illustrate the promise of using automated methods in summarizing texts of interview transcripts. The success in text selection, key word generation, and summary text and punch line identification while using a combination of off the shelf tools is quite encouraging.

Manual extraction has the benefits of accuracy and high retention; however, it is inherently slow due to the density of the text and the massive amounts of information that must be translated. The semi-automatic process was able to improve tacit knowledge acquisition by allowing the reviewer to see multiple ranges of text that had been pre-defined by text link analysis and categorization. We also saw that tacit knowledge unavailable through the manual process was discovered through the semi-automatic method. Yet, the semi-automatic process requires the reviewer to filter through multiple text sections. In addition, the implementation of text link analysis and categorization on a large number of documents remains to be evaluated.

We highlighted methods that are currently in development to automate the process of obtaining tacit knowledge. Although the current algorithm requires improvements, it is encouraging to see that implementations of the automatic process obtained text sections, keywords, summary points, and punch line titles that were discovered by the reviewers. The benefit of the automated method is that hundreds of transcripts can be processed in the time it takes a reviewer to complete a single interview transcript. Therefore, future implementations of the automated system will need to be tested on a higher number of transcripts. Also, other advanced text mining techniques applied at each of these steps or other summarization algorithms can also be explored. This project is very much a work in progress at this time.

ACKNOWLEDGEMENT

We wish to acknowledge other colleagues who have worked on knowledge transfer aspects of this project: David Biros, Joyce Lucca, Ridhima Nerlekar, and Spruthi Parupalli.

REFERENCES

1. Appelt, D. E., Hobbs, J. R., Bear, J., Israel, D. & Tyson, M. (1993) FASTUS: A Finite-state Processor for Information Extraction from Real-world Text. In: *Proceedings IJCAI-93*, pp., Chambery, France.
2. Bontcheva, K., Dimitrov, M., Maynard, D., Tablan, V. & Cunningham, H. (2002) Shallow Methods for Named Entity Coreference Resolution. In: *In Proceedings of TRAITEMENT AUTOMATIQUE DES LANGUES NATURELLES (TALN)*, pp. 24-32.
3. CodeProject (2006) Sentence Breaker using Microsoft Word. Vol. 2009, pp. CodeProject.
4. Coleman, M. & Liau, T. L. (1975) *Journal of Applied Psychology,* **60,** 283-284.
5. Crowsey, M. J., Ramstad, A. R., Gutierrez, D. H., Paladino, G. W. & White, K. P. (2007) An Evaluation of Unstructured Text Mining Software. In: *Systems and Information Engineering Design Symposium, 2007. SIEDS 2007. IEEE*, pp. 1-6.
6. Davenport, T. H. & Prusak, L. (1998) *Working knowledge : how organizations manage what they know,* Harvard Business School Press, Boston, Mass.
7. Flesch, R. (1948) *Journal of Applied Psychology,* **32,** 221-233.
8. Kincaid, J. P. (1975) Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel. pp. 49. National Technical Information Service, Springfield, Virginia 22151 (AD-A006 655/5GA, MF.
9. Klare, G. R. (1974) *Reading Research Quarterly,* **10,** 62-102.
10. Koenig, M. E. D., Srikantaiah, T. & American Society for Information Science and Technology. (2004) *Knowledge management lessons learned : what works and what doesn't,* Published for the American Society for Information Science and Technology by Information Today, Medford, N.J.
11. Leonard, D. & Sensiper, S. (1998) *California Management Review,* **40,** 112.
12. Mack, R. & Hehenberger, M. (2002) *Drug Discovery Today,* **7,** S89-S98.
13. Mayer, R. J., Cullinane, T. P., DeWette, P. S., Knappenberger, W. B., Perakath, B. & Knowledge Based Systems Inc College Station, T. X. (1992) *Information Integration for Concurrent Engineering (IICE) IDEF3 Process Description Capture Method Report,* Defense Technical Information Center, Ft. Belvoir.
14. Microsoft (2009) About Automatically Summarizing a Document. Vol. 2009, pp. Microsoft.
15. Miller, G. A. WordNet - Princeton University Cognitive Science Laboratory. pp. Princeton University.
16. National Institute of Standards and Technology (1993) Integration definition for function modeling (IDEFO). pp. US Department of Commerce, Technology Administration, Federal Information Processing Standards Publication, Report Number FIPS PUB 183, Gaithersburg, MD.
17. Pang, B., Lee, L. & Vaithyanathan, S. (2002) Thumbs up?: sentiment classification using machine learning techniques. In: *Proceedings of the ACL-02 conference on Empirical methods in natural language processing - Volume 10*, pp. Association for Computational Linguistics.
18. Papernick, N. & Hauptmann, A. G. (2005) Summarization of Broadcast News Video through Link Analysis of Named Entities. In: Proceedings of the AAAI workshop on link analysis, pp., Pittsburgh, PA.
19. Schutz, A. (2008a) Extraction of Keyphrases for Metadata. Vol. 2009, pp. SmILE.
20. Schutz, A. T. (2008b) Keyphrase Extraction from Single Documents in the Open Domain Exploiting Linguistic and Statistical Methods. In: Applied Science, Vol. Masters of Applied Science, pp. 131. National University of Ireland, Galway.
21. SPSS Text Mining for Clementine 5.0 User Manual. pp. Integral Solutions Limited, Chicago.
22. Stenmark, D. (1999) Using Intranet Agents to Capture Tacit Knowledge. In: Proceedings of WebNet World Conference on the WWW and Internet 1999, pp. 1000-1005. Chesapeake, VA: AACE.
23. Tan, A.-H. (1999) Text mining: The state of the art and the challenges. In: Pacific Asia Conference on Knoweldge Discovery and Data Mining PAKDD'99 Workshop on Knowledge Discovery from Advanted Databases, pp.
24. Turenne, N. & Rousselot, F. (1998) Application of clustering in a system of query reformulation. In: Presentation of Saros, ERIC-LIIA report, pp. University of Strasbourg, France.
25. Vogel, C. & Powers, J. (2000) Quality Metrics: How to Ensure Quality Taxonomies, Information Today, Medford, NJ.
26. Wyatt, A. (2009) Only Showing Readability Statistics. Vol. 2009, pp. WordTips.