SenCept: A Domain-specific Textual Commonsense Concept Acquisition System

Rushdi Shams[1], M.S.A. Shahnawaz Chowdhury[2], and S.M. Abu Saleh Shawon[3]
Computational Linguistics Lab
Department of Computer Science
University of Western Ontario, London, Ontario, N6A 5B7, Canada[1]

Department of Computer Science and Engineering
Khulna University of Engineering & Technology (KUET)
Khulna-9203, Bangladesh[2,3]
rshams@csd.uwo.ca[1], sajib014@gmail.com[2], shawon16@gmail.com[3]

***Abstract*— In this paper, we report the development and the performance of SenCept that acquires textual commonsense concepts to offer better contextualization for the domain DC electrical circuits. It uses a commonsense knowledge-base built upon a linguistic relations framework comprising Clause Level Relations, Sentential Roles, and Rhetorical Relations of a domain-specific corpus. SenCept selects representative commonsense knowledge using several parameters like knowledge weight and average commonsensical distances among knowledge. To extract commonsense concepts for any given sentence, the system uses the latter and the mean of distances among normalized weights of the representative sentences. The system is tested with a set of 100 random domain-specific sentences that are also given to five human subjects. Results show that SenCept achieves a precision and recall of 71.43 and 51.77 percent, respectively with an F-score of 60.03 percent.**

Keywords- Commonsense knowledge, commonsense concept, knowledge acquisition, knowledge engineering.

I. INTRODUCTION

Shams *et al.* reported SenCept that acquires commonsense concepts from domain-specific texts using a corpus [1]. They reported that the system, having a Common Concept Rate (CCR) of 43 percent, produces concepts that are originated from commonsense rather than having domain knowledge apriori. However, the commonsense knowledge-base they used was not methodologically sound as its development did not consider the linguistic relations in texts. This paper is a follow-up that reports the performance of the system after re-adjusting the commonsense knowledge-base using linguistic relations in the texts.

For a sentence like *If you throw a ball in the air, it will come down to earth* it is almost certain that a baby boy will ask- *why don't the planets return to the earth then?* Even if he is not aware of the Law of Gravity by the age of ten, he will not ask this question. As he turns to a young man who eventually comes across the Law of Gravity, he finds the answer of this question. The baby boy neither has knowledge nor commonsense. The boy at ten in this scenario has the commonsense but does not have the knowledge. Finally, in his youth, he has both knowledge and commonsense. Although they have subtle differences between them, commonsense is a type of knowledge. Knowledge varies in human but commonsense should not and it should be present commonly in us- it is what makes the identification of exact commonsense a difficult task.

Identification and extraction of commonsense knowledge has been the center of attraction in natural language understanding [1] and personalized learning [2] over the last three decades. Textual commonsense knowledge is important to understand the context and discourse of the text [3]. For example, from the sentence *"The sum of current flowing into a junction is equal to the sum of current out of the junction"*, a reader can contextualize more precisely if he is aware of the concepts like *current, electron flow, junction, branch* and *Kirchhoff's law* and recent research findings showed that such contextualization promotes personalized learning [4]. These concepts are called the textual commonsense concepts that are derived from the commonsense knowledge associated with the sentence.

In this paper, we propose a system named SenCept that acquires domain-specific commonsense concepts from the commonsense knowledge associated with text of the domain DC Electrical Circuits. We use a commonsense knowledge-base, developed by five human subjects from the linguistic relations, namely clause level, sentence level and rhetorical level, of the text of a DC electrical circuit corpus [18]. Every knowledge in this knowledge-base has been weighted and we calculate the average distance among them. This denotes the statistical distance of one knowledge from the others due to the variation of commonsense present in them. Moreover, for any given sentence, whose textual commonsense concepts are to be acquired, we calculate its normalized weight [9]. Then, we select the relevant commonsense knowledge using statistical analysis on the normalized sentence weight and average distance of knowledge, and select the proper nouns in them. We tested SenCept with a sample of 100 random domain-

specific sentences that are also given to five human subjects. Results show that SenCept achieves a precision and recall of 71.43 and 51.77 percent, respectively with an F-score of 60.03 percent.

The rest of the paper is organized as follows. Section II describes related research work, their contribution and outcome. The working principle of SenCept is described in Section III. In Section IV, we show the performance analysis of SenCept. Section V concludes the paper.

II. Related Work

The acquisition of domain-specific textual commonsense concepts depends on the identification of commonsense knowledge associated with domain-specific text. Commonsense knowledge-base, like Cyc [5] and Open Mind Commonsense (OMCS) [6], has numbers of commonsensical assertions but experience difficulty in computer applications. Moreover, domain-specific commonsense knowledge is rarely available as most of the existing knowledge-base are developed by accumulating commonsense of generic people. Recently, trends of using corpora for developing such knowledge-base have started. Corpora-based knowledge-base, like ConceptNet [7], are aiding projects that involve computer applications. These knowledgebases are more reliable in the sense that they are based on representatively collected text from textbooks, newspapers and web documents. Besides, all commonsense knowledge associated with text are not required to acquire textual commonsense concepts- many commonsense knowledge are related specifically to the domain and have low impact on text. Therefore, commonsense concepts are not simply the concepts present in the commonsense knowledge- they need to be identified by using either statistics [1] [2] or fuzzy rules [8].

We developed our commonsense knowledge-base by involving human subjects and a domain-specific corpus proposed in [18]. Zhu *et al.* [1] proposed an experiment that analyzes the difficulty human faces to acquire commonsense knowledge from web corpora. They asked three human subjects to produce commonsensical assertions of 157 sentences from web corpora. In doing so, the human subjects categorized the acquisition with two difficulty levels- *easy* and *difficult* and three richness levels- *sparse, mediate and rich*. The experiment used κ co-efficient and Kendall's τ values to measure concordance strength among nominal variables [10] [11] that are particularly important to acquire domain-independent knowledge. There are several look-alike commonsense concept tools like Cog-Learn [12] and Cognitor [13]. These tools identify commonsense knowledge from text using textual concept maps [14]. SenCept does not draw concept maps for every commonsensical assertions but the corpus we used has been validated with multi-layer concept maps [15]. Therefore, SenCept uses techniques like Cog-Learn and Cognitor but its working principle is less complex than them. Suanmali *et al.* proposed feature based sentence and knowledge extraction technique in [16], where representativeness of sentences and knowledge depended on fused features like normalization, term frequencies and number of proper nouns. We did not fuse these features rather we used them independently as [9] showed that independent features perform better on domain-specific information retrieval. Cao *et al.* [2] weighed adjectives to extract concepts from commonsense knowledge. As the domain of our corpus contains fewer adjectives but handful of proper nouns [17], we extracted concepts from commonsense knowledge based on the relevant terms.

III. Working Principles of SenCept

A. Development of CorParse: A Corpus Parsing Tool

The linguistic information analysis has been limited to some tools unable to operate on any corpus of interest as they have been built upon a specific corpus [18]. Therefore, we developed a Java-based parser named CorParse, a core component of SenCept, which works on SAX and a relational database, to extract the sentences from XML-based corpora.

CorParse operates on any XML corpus in some predefined steps. On its GUI, the user manually defines the number of tags the corpus has. It then matches the tags and the hierarchy of tags defined by the user with the actual corpus file. The parser first checks the syntactic formation of the XML data, making sure that the start tags have corresponding end tags and that there are no overlapping elements. It also validates the structure and contents of the corpus against the specified Document Type Definition (DTD) or the XML Schema. Finally, the parsing output provides access to the content of the XML document via the APIs. In addition, the corpus in concern can be represented graphically- both in tabular format and in tree format. Tabular format is particularly useful when developing and annotating a new corpus where information needs to be in spreadsheet and tree format is very useful for conceptualization of the domain and for corpus analysis. The other functionalities of CorParse are illustrated in Figure 1.

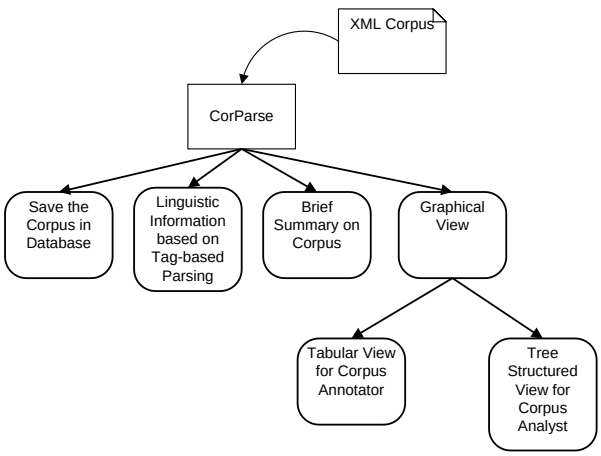

Figure 1. Functioanlity of the CorParse

B. Annotation of Linguistic Relations

Large amounts of text in a corpus of a specific domain will usually contain many types of linguistic relations e.g., clause level relations, sentential roles, and rhetorical relations. Identifying such relations is important for information retrieval systems and for the design and development of a commonsense

knowledge-base. Due to the lack of a uniform framework that can adequately represent all these relations in such domain-specific corpora, we developed a three-level framework that describes the linguistic relations found in this corpus. We annotated the sentences of the corpus with the three different types of linguistic relations that improved the identification of commonsense knowledge by the five human subjects.

Analyzing the whole corpus, we developed 19 Clause Level Relations (CLRs) and included them into the framework. We took a sentence, broken it down into clauses, named them as A1, A2 and so on. We then analyzed the framework prescribed by [19] and tried to fit the clauses according to the suggested relations. Sometimes it is a straightforward match among clauses but most of the time we encountered relations that are not found in the framework. According to the meaning and semantics conveyed by the clauses and their interrelations, we developed new relations and assigned them among clauses. In contrast, it is possible to have more than one relation between two clauses.

We developed seven sentential roles that exist among sentences in the corpus. Though we analyzed the approach of Grice [20], our purpose of using these maxims to outline the sentential roles is to identify the similarities and differences between conversations and written forms. We found that written form of instructional text supports a set of conversational maxims. This finding implies that instructional text from a domain is almost identical to the conversations take place for that domain.

We developed the rhetorical relations existing in the corpus by examining the presentational pattern, subject-matter behavior, behavior between nuclear and satellite, and relative position in the context [21].

Analyzing the corpus, we developed 24 rhetorical relations for the framework. These relations can be used to fit in any text from the domain and of instructional nature.

The framework is shown in Table I. The relations in the framework are detailed in Appendix.

TABLE I. FRAMEWORK FOR LINGUISTIC RELATIONS IN THE CORPUS

Clause Level Relations	Cause, Enablement, Entailment, Prevention, Conjunctive, Disjunctive, Synonymy, Restatement, Evidence, Example, Elaboration, Explanation, Interpretation, Contrast, Spatial, Description, Definition, Property, List
Sentential Role	Quality, Exploitation, Manner, Relevance, Quantity, Implicit Promise, Topic Shifter
Rhetorical Relations	Anti-thesis, Background, Concession, Enablement, Evidence, Justify, Motivation, Restatement, Summary, Circumstances, Condition, Elaboration, Interpretation, Means, Otherwise, Purpose, Solutionhood, Evaluation, Conjunction, Contrast, Disjunction, Joint, List, Sequence

C. Development of a Commonsense Knowledge-base

We provided the annotated sentences to five human subjects who were varied by their ages and their prior knowledge of the domain of the corpus. Each of the human subjects produced numbers of commonsense knowledge for every sentence in the corpus based on the linguistic relations in it. The knowledge-base contains the union (as in the set theory) of commonsense knowledge produced by the human subjects. A partial knowledge-base is listed in Table I.

TABLE II. COMMONSENSE KNOWLEDGE ASSOCIATED WITH SENTENCES IN THE CORPUS (PARTIAL)

Sentences in the corpus	Commonsense Knowledge
Kirchhoff's 2nd Law is based on the principle of conservation of energy.	- Kirchhoff has given more than one law - Energy can be conserved - Laws can be based upon principle.
The Potential Dividers are used to find the EMF of a cell.	- EMF of a cell can be calculated - Potential dividers divide the potential - Potential divider is used to calculate EMF - Cell has EMF
There are different rules for series and parallel circuits.	- There may be two types of circuits. - There are some rules for circuits. - Series circuit is a type of circuit. - Parallel circuit is a type of circuit. - Series and parallel circuit rules are different.

By using eq. (1) and eq. (2), we measured the commonsense knowledge weight, which is equal to the probability of the commonsense to be representative multiplied by summation of term frequencies (tf_{kc}). This probability is a ratio of number of terms (t_{nc}) and number of words (w_{nc}) in the commonsense in order to get the effect of the commonsense on $\sum tf_c$.

$$K_i = P(a\ commonsense\ to\ be\ representative) \times \sum term\ frequency\ in\ commonsense$$

$$= \frac{t_{nc}}{w_{nc}} \times \sum_{k=1}^{t_{nc}} tf_{kc} \quad (1)$$

$$P(a\ commonsense\ to\ be\ representative) = \frac{Number\ of\ terms\ in\ a\ commonsense,\ t_{nc}}{Number\ of\ words\ in\ a\ commonsense,\ w_{nc}} \quad (2)$$

TABLE III. TERM FREQUENCIES IN THE CORPUS (PARTIAL)

Term (noun)	Term Frequency	Term (noun)	Term Frequency
Current	31.85	Electricity	3.53
Charge	49.11	Ohm	10.17
Circuit	100.00	Unit	26.99
Voltage	73.00	Series	22.12
Power	18.58	Law	20.79
resistance	75.22	Wire	29.64
Energy	49.11	Battery	20.79

To measure tf_c, we used terms present in the corpus. For example, a list of Term Frequencies [9] is given in Table II.

When considering domain-specific knowledge, it is usual that some commonsense will be present in more than one knowledge. They are specific to the domain rather than to a particular sentence and have low impact on knowledge acquisition. Therefore, we filtered out these commonsense by normalizing the weight of the knowledge in our knowledge-base. Prior to normalization, we calculated the distance d_{ij} among knowledge with weights K_i and K_j using eq. (3), where $i = 0\ to\ m$; m being the number of knowledge and $i \neq j$.

$$d_{ij} = K_i - K_j \quad (3)$$

This is the statistical distance of a knowledge from every other knowledge due to the variation of commonsense present in them. Thereafter, we find the mean of d_{ij} for every knowledge using eq. (4)-

$$m_i = \frac{\sum d_{ij}}{n} \quad (4)$$

, where n denotes total number of values found by eq. (3).

Lastly, we normalized the weight of the knowledge in our knowledge-base using eq. (5).

$$M = \frac{\sum_{i=0}^{n} m_i}{n} \quad (5)$$

, where n denotes total number of means found by eq. (4).

D. Commonsense Concept Acquisition

SenCept acquires commonsense concepts from a given input sentence by following five steps. First, it calculates the weight of a given input sentence S using eq. (6) [9]-

$$SW = \frac{Sentence\ Weight,\ S_i}{Maximum\ Sentence\ Weight,\ \max(S_i)} \quad (6)$$

Where, sentence weight, $S_i = \frac{t_n}{w_n} \times \sum \mathrm{tf_k}$, t_n = number of terms in the sentence, w_n = number of words in the sentence and tf_k = term frequency.

It then subtracts the normalized sentence weight (SW) from means (m_i) of every knowledge weights and finds their mean (m_{sc}). This is the average statistical distance between the sentence and the knowledge. Now, SenCept finds out five knowledge from the knowledge-base that have weight around m_{sc}, both in positive and negative directions. From these knowledge, it then selects the knowledge that lie between the range of M and m_{sc}. Lastly, it extracts the proper nouns of the selected knowledge, which are the commonsense concepts for the given input sentence, S.

For example, if we consider the input sentence *The current is the same through all the components in series circuit*, then,

Number of terms (t_n) = 8

Number of words (w_n) = 12

Sentence weight, $S_i = \frac{t_n}{w_n} \times \sum \mathrm{tf_k} = \frac{8}{12} \times 159.7344 = 106.48967$

Then, S_i is normalized with eq. (6)-

$$SW = \frac{Sentence\ Weight,\ S_i}{Maximum\ Sentence\ Weight,\ \max(S_i)} = \frac{106.48967}{161.8205}$$

$$= 0.65807$$

Mean of distances between SW and m_i, m_{sc}= 0.4183

From eq. (5), we get the normalized distances among knowledge, $M = 0.2353$

Then, SenCept finds five commonsense knowledge with weights around the value of m_{sc} (0.4183). The five commonsense knowledge for the given sentence are listed in Table III.

TABLE IV. COMMONSENSE KNOWLEDGE WITH WEIGHT AROUND m_{sc}

Commonsense Knowledge	Knowledge Weight
In parallel circuit, high resistance means low current will flow to maintain the same voltage	0.4036
Voltage law is a statement of charge conservation	0.4390
Current can flow in a circuit	0.4449
The sum of the voltage gains and drops around any closed circuit is zero	0.3774
The amount of current is same in everywhere of a series circuit	0.3732

Lastly, SenCept selects the knowledge from the entries of Table III that lie between the range of m_{sc} (0.4183) and M (0.2353). In this case, the selected knowledge are- *The amount of current is same in everywhere of a series circuit (0.3732)*, *The sum of the voltage gains and drops around any closed circuit is zero (0.3774)*, and *In parallel circuit, high resistance means low current will flow to maintain the same voltage (0.4036).* It then tags the knowledge with Part of Speech (POS) tags and extracts the proper nouns- which are the commonsense concepts for the sentence. In this case, the commonsense concepts for the sentence *The current is the same through all the components in series circuit* are *parallel circuits*, *resistance*, *current*, *voltage*, *closed circuit*, and *series circuit*.

IV. RESULTS AND DISCUSSIONS

To evaluate the performance of the system, we provided 100 random sentences from the domain that are not included in the corpus to SenCept and five human subjects to find out commonsense concepts from them and kept the record. A partial record of the concepts generated by SenCept and human subjects is shown in Table V.

TABLE V. COMMONSENSE CONCEPTS ACQUIRED BY SENCEPT AND HUMAN SUBJECTS (PARTIAL)

Input Sentence	**Commonsense Concepts by SenCept**	**Commonsense Concepts by Human Subjects**
Current flows through conductor.	Current, conductor, electron, potential energy, power supply	Current, conductor, electron, material
The supply voltage is divided between the components.	Voltage, current, electrons, resistance, electricity, Kirchhoff's law	Voltage, component, voltage law, supply voltage, current
The current in a parallel circuit depends on the resistance of branch.	Parallel circuit, series circuit, current, voltage, resistance, charge	Parallel circuit, series circuit, current, voltage, resistance, circuit, branch,
The voltage for each component depends on the resistance.	Resistance, current, branch	Resistance, current, voltage, component, charge, device end
Lamps are connected in parallel for some reason	Resistance, component, device, charge	Lamp, parallel circuit, connection, parallel connection
The sum of the voltage gains and drops around any closed circuit is zero.	Voltage, closed circuit, current, resistance, branch	Voltage, closed circuit, current, series circuit, parallel circuit, voltage gain

Then, taking the commonsense concepts by human subjects as our golden standard, we measured the precision, the recall, and their weighted harmonic mean called the F-Score. From the evaluation, we found that SenCept achieved 71.43 percent of precision, 51.77 percent of recall, and an F-Score of 60.03 percent (as shown in Table VI).

TABLE VI. PERFORMANCE OF SENCEPT

True Positives	False Positives	False Negatives	Precision	Recall	F-Score
190	76	177	71.43	51.77	60.03

Its high precision indicates that it finds more true positives than false positives. However, the number of false negatives is high due to the strict evaluation procedures that we followed. Sometimes, SenCept produces concepts like *current* from a sentence where human subjects use *electron flow*- eventually they are synonymous but we did not consider them as the correct acquisitions due to our strict evaluation. False negatives and false positives of the system are also affected by some hierarchical concepts. For example, human subjects sometimes prefer choosing more precise concepts like *series circuit* or *parallel circuit* where SenCept produces concepts like *circuit*. Our evaluation also shows that total number of concepts produced by SenCept is 266 compared to 367 concepts produced by human subjects. This indicates that SenCept works on the concepts that are originated from the commonsense knowledge; we observed that numbers of concepts produced by the human subjects are originated from their prior knowledge on the domain.

We also used the Hooper [22], Rolling [23] and Cosine measures of annotator agreements between the concepts generated by SenCept and humans. The Hooper's annotator agreement is as follows-

$$H = \frac{TP}{TP + FP + FN}$$

Rolling's annotator agreement is defined as follows-

$$R = \frac{2TP}{2TP + FP + FN}$$

And Cosine measure of annotator agreement is defined as-

$$C = \frac{TP}{\sqrt{(TP + FP) \times (TP + FN)}}$$

Where the symbols TP, FP and FN denote true positives, false positives and false negatives respectively. The agreement measure in Table VII shows that SenCept agrees with the gold

standard with $H = 42.89$, $R = 60.03$ and $C = 60.81$ which can be considered to be decent agreement scores.

TABLE VII. SENCEPT'S AGREEMENT WITH THE GOLD STANDARD

Annotator	Hooper	Roling	Cosine
SenCept	42.89	60.03	60.81

V. CONCLUSIONS

In this paper, we presented a domain-specific commonsense concept acquisition system named SenCept that uses commonsense knowledge associated with domain-specific text. We developed a commonsense knowledge-base that contains commonsense knowledge associated with text of a domain-specific corpus. The knowledge-base has several parameters like weight of knowledge, their relative distance with each other due to variation in commonsense, mean of their relative distances and normalized mean to reduce effects of unnecessary commonsense. For any given sentence, the system normalizes its weight using its probability of representativeness and compares the weight with the normalized mean of the knowledge. The system thus selects commonsense knowledge that are closely associated with the sentence and finds out the concepts. Performance results showed that concepts produced by SenCept are originated from textual commonsense in contrast to human analysis that produces concepts from domain knowledge.

REFERENCES

[1] Y. Zhu, L. Zang, Y. Cao, D. Wang, and C. Cao, "A Manual Experiment on Commonsense Knowledge Acquisition from Web Corpora", *Seventh International Conference on Machine Learning and Cybernetics*, Kunming, China, 2008.

[2] Y. Cao, C. Cao, L. Zang, Y. Zhu, S. Wang, and D. Wang, "Acquiring Commonsense Knowledge about Properties of Concepts from Text", *5th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD 2008)*, Shandong, China, 2008.

[3] M. Vallez and R. Pedraza-Jimenez, "Natural Language Processing in Textual Information Retrieval and Related Topics". *Hipertext.net*, ISSN 1695-5498, no. 5, 2007.

[4] A. Carvalho, J. Anacleto, and S. Zem-Mascarenhas, "Planning Learning Activities Pedagogically Suitable by Using Common Sense Knowledge", *16th International Conference on Computing (CIC 2007)*, Mexico, 2007, pp. 1-6.

[5] D. Lenat, "CYC: A Large-Scale Investment in Knowledge Infrastructure", *Communications of the ACM*, vol. 38, no. 11, 1995, pp. 33-38.

[6] D.G. Stork, "The Open Mind Initiative", *IEEE Expert Systems and Their Applications*, 1999, pp. 16-20.

[7] C. Havasi, R. Speer, and J. Alonso, "ConceptNet 3: A Flexible, Multilingual Semantic Network for Common Sense Knowledge", *Proceedings of Recent Advances in Natural Language Processing*, 2007.

[8] L. Zadeh, "Knowledge Representation in Fuzzy Logic", *IEEE Transactions on Knowledge and Data Engineering*, vol. 1, no. 1, 1989.

[9] R. Shams, M.M.A Hashem, A. Hossain, S. R. Akter, and M. Gope, "Corpus-based Text Summarization using Statistical and Linguistic Methods", *3rd IEEE International Conference on Computer and Communication Engineering (ICCCE'10)*, Malaysia, 2010, pp. 115-120.

[10] J. Cohen, "A Coefficient of Agreement for Nominal Scales", *Educational and Psychological Measurement*, vol. 20, pp. 37-46.

[11] J. Wang and X. Liang, "Categorical Data Analysis (in Chinese)", Eastern China Normal University Press, Shanghai, China, 2005.

[12] J. Anacleto, A. Carlos, A. de Carvalho, and M. Godoi. "Using Common Sense Knowledge to Support Learning Objects Edition and Discovery for Reuse", *Proceedings of WebMedia*, Gramado, 2007, pp. 290-297.

[13] J. Anacleto, A. Fabiano P. de Carvalho, A. Ferreira, E. Pereira, and A. Carlos, "Common Sense-based Applications to Advance Personalized Learning", *2008 International Conference on System, Man and Cybernetics (IEEE SMC 2008)*, 2008, pp. 1-10.

[14] J. D. Novak, "A Theory of Education", Cornell University Press, New York, 1977.

[15] R. Shams and A. Elsayed, "Development of a Conceptual Structure for a Domain-Specific Corpus", *3rd International Conference on Concept Mapping 2008 (CMC 2008)*, Estonia and Finland, 2008.

[16] L. Suanmali, N. Salim, and M. Binwahlan, "Feature-Based Sentence Extraction Using Fuzzy Inference rules", *2009 International Conference on Signal Processing Systems*, Singapore, 2009.

[17] R. Shams and A. Elsayed, "A Corpus-based Evaluation of Lexical Components of a Domain-specific Text to Knowledge Mapping Prototype", *11th IEEE International Conference on Computer and Information Technology (ICCIT 2008)*, Bangladesh, 2008, pp. 242-247.

[18] R. Shams, A. Elsayed, and Q. M. Akter, "A Corpus-based Evaluation of a Domain-specific Text to Knowledge Mapping Prototype", *Journal of Computers*, Academy Publisher, ISSN: 1796-203X, vol. 5, no. 1, 2010, pp. 69-80.

[19] K. Barker, T. Copeck, S. Szpakowicz, and S. Delisle, "Systematic construction of a versatile case system", Natural Language Engineering, volume 3, number 4, pp. 279-315, 1997.

[20] P. H. Grice, *Studies in the Way of Words*, Harvard University Press, Cambridge, MA, 1988.

[21] L. Kosseim and G. Lapalme, "Choosing Rhetorical Structures to Plan Instructional Texts", Computational Intelligence, volume 16, number 3, pp. 408-445, 2000.

[22] R. S. Hooper, Indexer Consistency Test: Origins, Measurements, Results and Utilizations, IBM, Bethesda, 1965.

[23] L. Rolling, Indexing Consistency, Quality and Efficiency, Information Processing and Management, 17:69-76, 1981.

APPENDIX

A. Clause Level Relations

Analyzing the whole corpus, we developed 19 CLRs and included them into the framework. We took a sentence, broken it down into clauses, named them as A1, A2 and so on. We then analyzed the framework prescribed by [12] and tried to fit the clauses according to the suggested relations. Sometimes it is a straightforward match among clauses but most of the time we encountered relations that are not found in the framework. According to the meaning and semantics conveyed by the clauses and their interrelations, we developed new relations and assigned them among clauses. Sometimes, it is possible to have more than one relation between two clauses. In that case, we took all of the clause level relations into account.

Cause: A1 makes A2 to occur or exist. In another way, A1 is sufficient enough to cause A2 and the occurrence or existence of A1 is required. For example, *Obviously as you go around the circuit [the potential difference will drop to zero]$_{A2}$ since [one side of the power source is positive and the other negative]$_{A1}$*. This particular CLR is important for knowledge modelling as the ontology of the TKM prototype depends on the qualitative layer of the linguistic side. Therefore, the understanding of

causal relations among clauses is necessary to model the knowledge in the text.

Enablement: A1 makes A2 possible. In other word, A1 is necessary to enable A2 but it is not sufficient and the existence of A1 is not necessary. For example, *[The wire is connected in this way]$_{A1}$ [so a current can flow through it]$_{A2}$*. Sometimes, the TKM prototype during the primary evaluation showed inconsistency in figuring out the proper enabler. This clause level relation is useful to adjust the qualitative layer of the prototype for proper knowledge representation.

Entailment: If A1 exists or occurs, then A2 must also exist or occur. In addition, A1 is not known to exist or occur but the occurrence of A2 is obvious. For Example, *[The more components there are in a series circuit]$_{A1}$, [the greater the circuit's resistance]$_{A2}$*. There is a subtle difference between enablement and entailment. With the raw text by which the prototype was tested first, it was not possible to distinguish between these two lexical semantics.

Prevention: A1 is meant to keep A2 from occurring or existing. A1 is sufficient to prevent A2; no other clause is required for keeping A2 to occur. For example, *[If resistances are joined in parallel]$_{A1}$ [then values cannot be simply added together]$_{A2}$*. This CLR is also helpful for developing causal relationship among subjects and objects.

Conjunctive: A conjunction relationship exists between acts or states about which no more can be said than that they both occur or exist. For instance, *[There are different rules for series and parallel circuits]$_{A1}$ and [you must know these rules]$_{A2}$*. The prototype performed poor due to its inaccuracy to point out conjunctive clauses and compound sentences. The prototype works well with simple sentences only. Therefore, whenever it found conjunctive clauses, although it is not a compound sentence, it ignored parsing and mapping it.

Disjunctive: A disjunction relationship exists between acts or states about which no more can be said than that one or both. For instance, *[In common applications such as determining the direction of force on a current carrying wire, treating current as positive charge motion]$_{A1}$ or [negative charge motion gives identical results]$_{A2}$*.

Synonymy: A1 and A2 connected by verbs that represent the same thing. For example, *[Kirchhoff's 1st Law]$_{A1}$ can be remembered as [the rule that uses nodes to study the flow of current around a circuit]A2*. With the help of identifying clauses that provide synonymy, the prototype now can have the flexibility of representing same knowledge for both of the clauses.

Restatement: A1 and A2 connected by a conjunction where A2 conveys the same meaning of A1 in a different way. For example, *[Kirchhoff's 1st Law states that the current flowing into a junction in a circuit (or node) must equal the current flowing out of the junction]$_{A1}$ – [a direct consequence of the conservation of charge]$_{A2}$*. The prototype in its early trial of test, represented different knowledge for restated clauses- which is unacceptable.

Evidence: A2 is the evidence of A1, supporting the validity of A1. For instance, *[An electrical cell is made from materials]$_{A1}$ [(metal or chemicals, for example)]$_{A2}$*. This CLR evidence and the following named example are necessary to figure out clauses that are associated with these two relations with subtle difference.

Example: A2 is the example of A1. A2 must be an instance from the discourse. *[Having connected our circuit]$_{A1}$, [we can use it to light an electric lamp, to run an electric motor, to heat an electric element and so on]$_{A2}$*.

Elaboration: A2 holds additional or specific information of A1. *[If one lamp in a series circuit breaks or fails, all the others will go out with it]$_{A1}$ – [for this reason, lamps are always connected in parallel]$_{A2}$*

Explanation: A2 holds the factual explanation of A1. This factual explanation is not to convince the reader. *If resistances are joined in parallel [then values cannot be simply added together]$_{A1}$ – [values need to be treated differently]$_{A2}$*

Interpretation: Similar to elaboration but A2 may include other topics to elaborate A1. *[Manual-ranging meters have several different selector positions for each basic quantity]$_{A1}$: [several for voltage, several for current, and several for resistance]$_{A2}$*

Contrast: A2 provides contrast with A1 and they are joined with some definite terms. *[Electrons move about randomly due to thermal energy]$_{A1}$ [but on average, there is zero net current within the metal]$_{A2}$*

Spatial: A1 and A2 can be symmetric or asymmetric based on the verb by which they are related. *[One simple DC circuit consists of a voltage source (battery or voltaic cell)]$_{A1}$ connected to [a resistor]$_{A2}$*

Description: A2 describes A1. A1 and A2 are most often connected by a hyphen (–). *[Some digital multimeters are auto ranging]$_{A1}$ – [an auto ranging meter has only a few selector switch (dial) positions]$_{A2}$*

Definition: A2 defines A1. A1 and A2 are most often connected by a hyphen (–) or by a colon (:). *[Kirchhoff's 1st Law]$_{A1}$ states that [the current flowing into a junction in a circuit (or node) must equal the current flowing out of the junction]$_{A2}$*

Property: A2 holds the property of A1 and they are related with a verb. *This is very useful because it means that [we can switch the lamp]$_{A1}$ [on and off]$_{A2}$*

List: A2 contains number of instances from the discourse which are the categories or derived from A1. *[Digital multimeters have numerical displays, like digital clocks]$_{A1}$, [for indicating the quantity of voltage, current, or resistance]$_{A2}$*

B. Sentential Roles

Quality: If the writer puts a sentence with a technical proof followed by it, then the sentence will be adjudged as quality. In any other case like putting a sentence without proof will be adjudged as exploitation.

Exploitation: If the writer puts a sentence without backup sentences but that sentence conveys a message which does or does not affect any sentence followed by it, then the sentence will be categorized as exploitation.

Manner: The writer avoids obscurity, and ambiguity in this type of sentence and the writer states briefly and in order.

Relevance: Almost every sentence in the instructional text for a particular domain maintains relevance with each other.

Quantity: The writer makes the sentence informative as much as possible but does not provide extra information on the main discourse.

Implicit Promise: The writer changes the topic but promises implicitly he/she will come back to the main topic. For this relation, mainly overlapped text is considered.

Topic Shifter: The writer explicitly changes the topic. The text is never overlapped and there is not even any implicit indication of coming back from the writer.

The roles marked in the following sentences are reflecting the type of sentential roles one sentence has with its neighbouring ones.

[Kirchhoff's 2nd Law is based on the principle of conservation of energy]$_{QUALITY}$. *[No energy can be lost from or gained by the circuit, so the net voltage change must be 0]*$_{QUANTITY,RELEVANCE}$. *[Kirchhoff's 2nd Law can be remembered as the rule that uses loops to study the flow of current around a circuit]*$_{MANNER,QUALITY}$. *[At any junction in a circuit, the sum of the currents arriving at the junction = the sum of the currents leaving the junction]*$_{RELEVANCE,QUANTITY}$. *[In any loop (path) around a circuit, the sum of the emfs = the sum of the pds]*$_{QUANTITY,EXPLOITATION}$.

One sentence can be related with its following and preceding sentences with more than one relation. We compromised all the relations a sentence can have with its neighbouring ones rather than taking just one.

C. Rhetorical Relations

Anti-thesis: Comparison between two or more things but the writer cannot decide which one is good or which one is bad. When modelling the actual domain with the knowledge model, a knowledge representation tool may model the text by representing one thing as good and the other as bad. If the ontology of the tool is developed in way that if it can categorize a sentence as anti-thesis, and the sentence is reflecting in the same way, then proper representation will take place.

Background: Background information is general information of any sort that is likely to help the reader to understand the next part. Identification of background can augment the knowledgebase of any NLP systems.

Concession: Writer knows something is good/ bad, he/ she puts his/ her remark in a way that does not question its validity. This relation exists very less in the corpus but still it exists among text of the domain.

Enablement: Writer discusses some important topics (parts of a multi-meter) and then provides the user a reference (the voltage can be measured by it; the current can measure by it). Similar to enablement relation, but this time this relation is not explicit like the CLR.

Evidence: Writer provides example to support a statement. We found that the original rhetorical structure theory needs to be modified in this case as the evidence needs to be explicit. Implicit evidence can sometimes become example.

Justify: It suggests what the basis is of the writer's right to speak this item. The justification of invoking an event or a phenomenon can be backed up with an evidence or example. TKM prototype could not model justifying sentences.

Motivation: This relation works with enablement relation. There is a slight difference in terms of convey of meanings between motivation and enablement. Enablement leads the reader to come to a decision and motivation does not take the reader to a decision.

Restatement: The same statement previously stated in a text is repeated in a completely different way but the meaning of both the statements are identical. This relation is important when we are not thinking to model the domain in clause level. Restatement in rhetorical level needs proper inference mechanism in the ontology.

Summary: Summarizes the whole tale in a sentence or two. Mostly concludes a paragraph. When any NLP system identifies text as summary, it should recheck its knowledgebase and it should try to compare the knowledge extracted from summarized text and from the context so far.

The subject-matter relations outlined in the corpus are as follows. These relations are also mononuclear relations.

Circumstances: The difference between background and circumstances is in circumstances both the nucleus and the satellite focus on the same subject. The need of including this relation during design of ontology for NLP systems is circumstances provide more detailed information on context than background and can be helpful when comparing context with the summary.

Condition: States a condition on the previous statement. This rhetorical relation is strictly depending upon the statement prior to it. Qualitative layer of any knowledge representation tool should recognize and model such domain-specific text.

Elaboration: If one text elaborates the previously stated statement- no inclusion of any other topic in this part rather than the explanation of the previous statement.

Interpretation: Almost like elaboration. But it may include other topics to elaborate the previously stated statement.

Means: This relation reflects the feature of a method and mostly describes the characteristics of a material.

Otherwise: Otherwise can also be used to describe patterns of the form: If A then B otherwise C.

Purpose: To do one thing, the writer states to do another thing which is actually the purpose. Any backbone of NLP system should be constructed in a way so that the purpose of one task can be delivered when representing knowledge.

Solutionhood: Stating a problem in doing something, the writer proposes the solution- picking up the solution is very

important and carefulness is required so that NLP systems can recognize that the solution is not universal, it is delivered by the writer only.

Evaluation: Evaluation supports the writer's perspective indirectly. Evaluation is also crucial in knowledge representation. We found text on the corpus that the TKM prototype could not model because it seemed the support is universal. Therefore, any other writer differing to support on the same thing would be ignored- which is not acceptable.

The multinuclear relations existing in the corpus are given below.

Conjunction: Connects two sentences with a conjunction. Similar to a CLR but the context is higher than conjunctive relation. In this case, it is more appropriate to state that one rhetorical relation is in conjunction with the other.

Contrast: This relation has been called Neutral Contrast to reflect the balance of nuclearity, unlike Concession or Antithesis.

Disjunction: Two independent sentences but a clear difference is drawn with a word like but or unfortunately- an antonym to conjunction relation.

Joint: Joint represents the lack of a rhetorical relation between the nuclei- one of the difficult rhetorical relations to model.

List: Provides the first of a larger set of background facts, in a list. Unfortunately, due to the lack of this rhetorical relation the TKM prototype could never model such text. List is one of the common relations found in the text. So, the prototype requires to understand the relation if it really wants to model the domain.

Sequence: Sequence includes both presentational sequence, e.g., "Secondly," and also subject matter sequence, e.g., "After that," as in this case. Subject matter sequence is crucial as this represents the same text- which a knowledge representation tool must recognize.

The text is also analyzed for content organization structure according to its relative position in the paragraph. This analysis revealed the following schemes for the text in the corpus- Introduction, Background, Methods and materials, Results, Observations, Priming, Exposition, Description, and Conclusion.