Automatic Extraction of Conceptual Relations from Children's Stories

A Thesis Proposal
Presented to
the Faculty of the College of Computer Studies
De La Salle University Manila

In Partial Fulfillment
of the Requirements for the Degree of
Master of Science in Computer Science

by

SAMSON, Briane Paul V.

Ethel ONG
Adviser

August 4, 2010

Abstract

People use storytelling as a natural and familiar means of conveying information and experience to each other. During this interchange, people understand each other because we rely on a large body of shared common sense knowledge. But computers do not share this knowledge, causing a barrier in human-computer interaction and in applications requiring computers to generate coherent text. To support this task, computers must be provided with a usable knowledge about the basic relationships between concepts that we find everyday in our world.

Picture Books is a story generation system that generates stories for children age 4 to 6. To achieve this, it uses a semantic ontology containing conceptual knowledge about objects, activities and their relationships in a child's daily life. But the task of building this knowledge base is tedious and time consuming, thus limiting the variants of stories and themes that Picture Books is able to generate. This research involves the development of a software tool that will automatically extract concepts and their relations from existing children's stories, and store these in a knowledge base that Picture Books and other NLP applications can utilize to do their tasks.

Keywords: language parsing and understanding, text analysis, semantic networks, natural language processing

Contents

List of Figures

List of Tables

1 Research Description

This chapter discusses an overview of the current state of technology, the research objectives, scope and limitations, and its significance.

1.1 Overview of the Current State of Technology

Natural language processing systems use a set of knowledge base in order to do tasks such as text generation. But simple lexicons and large unstructured corpora may be insufficient as knowledge base of these systems. Storytelling, for instance, is a natural task for humans. Armed with a library of words, their meanings and their relationships, we combine words and events to tell stories about ourselves, our community, and our experiences. Thus computers must be provided with the same shared collection of common sense knowledge about the basic relationships between things and events that nearly every person knows in order for them to achieve a level of expressiveness same as humans and be able to understand the world that we talk about. Such knowledge are represented as conceptual relations defining the relationship between two or more concepts in real life.

Recent creative text generation systems such as (Hong & Ong, 2009) have utilized a semantic network representation of concepts on common sense knowledge to identify relationships of words in human puns in order to generate computer puns. Another system, Picture Books (Solis, Siy, Tabirao, & Ong, 2009), generates stories with morals for children ages 4 to 6, by using a semantic ontology, patterned after ConceptNet (Liu & Singh, 2004a), containing conceptual knowledge about objects, activities, and their relationships in a child's daily life. The process of building and populating the Picture Books ontology required a lot of manual effort on the part of the proponents. Currently, the ontology contains 240 concepts and 369 relations, which were populated based on the themes that have been identified as relevant for the target age group.

Early Information Extraction (IE) systems have addressed the extraction of information from relatively small collections of well-structured documents such as newswire or scientific publications (Muslea, 1999). More recently, IE systems are focused on extracting facts from structured and unstructured documents for a particular domain, such as legal documents (Cheng, Cua, Tan, & Yao, 2008).

Although IE systems are capable of recognizing entities within documents (e.g. 'Renoir' is a 'Person', '25 Feb 1841' is a 'Date'), the relation between the entities (e.g., 'Renoir' was born on '25 Feb 1841') is not extracted, thus generating incomplete information that may be needed by certain applications (Banko & Etzioni, 2008). A variant of IE, Relation Extraction (RE), is the task of recognizing the assertion of a particular relationship between two or more entities in text.

The task of relation extraction is difficult, but relations such as hypernymy (IsA) and meronymy (PartOf) are often expressed using a small number of lexico-syntactic patterns (Hearst, 1992). Using a sample set of 500 sentences selected at random from an IE training corpus, Banko and Etzioni (2008) showed that many binary relationships are also consistently expressed using a compact set of relation-independent lexico-syntactic patterns.

The Artequakt project (Alani et al., 2003) also showed that it is possible to automatically acquire such relations from documents to populate an ontology. Working in the domain of artists, the Artequakt project identifies relations between entities of interest within sentences, following ontology relation declarations and lexical information. These relations are then used to populate an ontology with knowledge triples for use in the generation of biographies of artists.

To convey the ideas of a story, Nakasone and Ishizuka (2006) developed a storytelling ontology model by identifying relations between sentences in the story using the Rhetorical Structure Theory (RST) of Mann and Thompson (1987). As Knott and Dale (1994) pointed out, explicit and implicit relations hold between the sentences of a text, so that the content of one sentence might provide justification, elaboration or explanation for the content of another. These relations bind a text together to contribute to the overall comprehension of a story by the readers; for instance, whether understanding one text span (scene of a story) increases the reader's readiness to understand another scene, or whether understanding both spans allows the reader to recognize a particular semantic relation as holding between them. Certain discourse relations or cue phrases, such as "but", "so", "although", "more precisely" or "for example", are used to signal explicit relations between text spans.

Although both IE and RE have achieved significant progress in extracting facts and concepts in the domains of newspapers (Muslea, 1999), biographies (Alani et al., 2003), and legal documents (Cheng et al., 2008), limited work has been done on children's stories. Furthermore, since stories contain sequences of actions that characters perform or experience at various points in the story world, knowledge about how these events are ordered and the constraints under which they can occur must also be extracted.

1.2 Research Objectives

1.2.1 General Objective

To develop a methodology that automatically identifies and extracts the relations between everyday concepts and objects from children's stories and store them in a semantic network to provide ontological knowledge for Picture Books.

1.2.2 Specific Objectives

1. To collect a corpus of children's stories;
2. To analyze the English sentence structures in the corpus;
3. To derive a set of extraction patterns;
4. To develop a representation for modelling relations of every object common in children's stories;
5. To design and implement an algorithm for extracting conceptual relations automatically from the corpus, and;
6. To validate the resulting conceptual relations extraction tool through integration with Picture Books

1.3 Scope and Limitations of the Research

At least 30 children's stories will be collected to form the input corpus for the extraction tool to be developed. These will then be modified to remove the dialogues. Analysis of the English sentence structures in these stories will be performed to identify the types of relations that are present. This information will be used to derive a set of patterns or templates for extracting conceptual relations. A software extraction tool will be developed to use the extraction patterns to automatically locate instances of a known relation in the corpus.

One relation may be expressed in various ways in text. Consider the hypernymy (IsA) relation, wherein the following sentences are possible ways of expressing it:

The dog is a canine.
The dog is a kind of canine.

The dog, a canine, is
The dog is a type of canine.

Although lexico-syntactic extraction patterns mapped directly to relations, certain English sentence constructs require further analysis and decomposition in order to derive their corresponding relations. These include sentence structures containing conjunctions and embedded clauses, as shown in the examples below:

Cake is made of flour, sugar, and butter.
The boy is singing and the girl is dancing.
Anna, who is the queen, went to the market, while the king went to the mall.

Text structures in stories may contain rhetorical relations to reflect the semantic relations that may exist between concepts and events in a story. Using the Rhetorical Structure Theory (RST) of Mann and Thompson (1987), these relations may be identified and extracted to provide additional conceptual knowledge.

Extracted knowledge must be stored in a representation model that can be used by NLP systems, in this case, the Picture Books story generator. Part of the research will involve reviewing the design of the Picture Books ontology, which is patterned after the design of ConceptNet, to validate the presence of the appropriate relations against those identified from the collected corpus.

Since stories are sequences of events, their analysis may necessitate the creation of new relations to represent sequences of events, temporal relations between events, as well as the constraints under which certain events may take place. For example, during testing, evaluators noticed that one of the generated stories of Picture Books occurred at an inappropriate time; specifically, the first segment of the story that introduces the day, the place, and the main character, contained the following text:

The evening was warm. Ellen the elephant was at the school. She went with Mommy Edna to the school.

Since Picture Books' knowledge base currently does not provide relations about when certain events can occur, the main character went to school in the evening.

Nineteen (19) new semantic relations, resulting from those identified using RST, from analyzing the sample corpus, and from reviewing other works such as (Mueller, 2003) for modeling time and event occurrences, may be created in this research. The main relations, mostly from ConcepNet and Picture Books, include IsA, PropertyOf, PartOf, MadeOf, FirstSubeventOf, EventForGoalEvent, EventForGoalState, EventRequiresObject, EffectOf, EffectOfIsState, CapableOf, OftenNear, LocationOf and, UsedFor. Additional relations include Happens(e, t) which represents that a fluent *f* holds at time *t*, HasRole which represent character roles, RoleResponsibleFor which represent actions/events done by a specific role, TargetOf which represent the direct object of a verb and lastly, Owns which represent ownership.

Alani et al. (2003) noted that it is inevitable for duplicate and contradictory information to be extracted from the input corpus. But he further noted that handling such information is challenging for automatic extraction and ontology population approaches. Thus, this will not be considered in the current proposal.

Co-occurring events and objects are also imminent in extracting relations. Here is a sample sentence exhibiting co-occurring objects:

Cake is made of flour, sugar, and butter.

If MadeOf relations are to be extracted from this sentence, there will be three instances of it namely, MadeOf(cake, flour), MadeOf(cake, sugar) and MadeOf(cake, butter). In such cases, Picture Books cannot and will not be able to recreate the sentence above if the three relations will be used. Picture Bokks will make three separate sentences instead. In this research, Picture Books will not be modified anymore to recombine multiple relations to generate a single sentence.

Such is also the case for co-occurring events in Picture Books. Here is an example of a co-occurring event:

Anna went to the market, while waiting for her ride.

In the sentence above, the event *went to the market* happened simultaneously with the event *waiting for her ride.* This research can also recognize co-occurring events but Picture Books cannot replicate the same sentence if two relations containing the two events above will be used. Picture Books will still produce separate sentences for each event relation. Again, Picture Books will not be modified anymore to recombine multiple relations to generate a single sentence.

Mueller (1999) also noted that “story understanding goes beyond generating parse trees, disambiguating words, or filling templates, and includes the ability to answer arbitrary questions, generate paraphrases and summaries, fill arbitrary templates, make inferences, reason about the story, follow reasoning in the story, relate the story to general knowledge, and hypothesize alternative versions of the story.” Thus, aside from having a huge collection of common sense knowledge, a computer system must also be able to “make inferences about states and events not explicitly described in the text” (Mueller, 2003), by performing common sense reasoning using knowledge about the world. This requires a multi-representational model of this knowledge for the various realms of space, time, needs and feelings to be built, and will be beyond the scope of the current proposal.

Manual validation with the help of a linguist may be utilized to check that correct conceptual relations were extracted and stored in the ontology. Automated validation will also be performed by having Picture Books utilize the new knowledge in generating stories.

The following are indicators of a successful validation of the contents of the resulting ontology:

- Since there will be new relations introduced into Picture Books, its story planner may be modified to use the new relations such as HasRole and RoleResponsibleFor.
- There is a significant increase in the number of story variants that are generated by Picture Books. A story variant can be a story having a theme and ending similar to other stories but with a different plot. This can also be a story with differing number of lines.
- The length of the generated stories for older kids (i.e., 5 to 6-year old users), measured in terms of the number of sentences, also increases as additional information, such as new relation types and new relations for the existing types of relation, becomes available. Note that Picture Books currently placed a limit to the maximum number of sentences that will be generated for younger readers.
- The coherency of the generated stories will also increase, as new knowledge improves the narrative information presented to the reader. This will be determined with the help of linguists.

1.4 Significance of the Research

Researches in the field of natural language processing (NLP) seek to find ways to make human-computer interaction more fluent. But human-computer communication is hampered by the lack of a shared collection of common sense knowledge that people rely on when they communicate in order to understand each other. In order to make computers achieve the same level of expressiveness as humans, we must give them “a common language with richness that more closely approaches that of the human language” (Niles & Pease, 2001).

Although dedicated IE systems have been developed to extract information from various domains, this research is a first step towards extracting relations from children's stories. Storytelling is a natural and familiar means of conveying information and experience to listeners (Nakasone & Ishizuka, 2006), thus justifying the selection of this domain for the proposed research.

The knowledge base derived from extracting entities, concepts, and events can then be used for various applications. One such application is in story generation. Picture Books can use the knowledge base to generate more variants as well as longer stories. Currently, Picture Books generates stories containing 24 to 30 sentences, which not only vary according to the age of the reader, but is also dependent on the available relations between two concepts. Story generating applications can in turn be used for both educational and entertainment purposes.

In education, Riedl and Young (2004) applied narrative generation techniques to generate historical fictions for teaching history, which they defined as "the chronological record of significant events". Lester et al. (2007) explored integrating narratives into learning environments that teach microbiology to provide an "adaptive, effective pedagogy that is both motivating and meaningful".

In entertainment, story generation is applied to develop interactive fiction systems. Montfort (2009) defines interactive fiction as "a venerable thread of creative computing and a literary art". His Curveship project uses NLP techniques to create narratives in the virtual world, where the user directs the possible flow of the story. For his knowledge base, Montfort utilized a tree representation that describes the possible sequences of events and the relationship of events to one another, as well as models of objects in the virtual world. A similar system in the game area, Faade (Mateas & Stern, 2003) is a 3D interactive drama that makes use of artificial intelligence techniques to allow players to interact with the characters in the story by playing as one of the characters and typing textual commands that affect the flow and the outcome of the game (story). Young (2008) is also exploring the development of computational models to generate narratives for 3D virtual game environments, which are being considered as alternative approach to promote learning.

Story understanding system can also benefit from using the knowledge base. Story understanding requires an enormous amount of common sense knowledge, thus the question and answering system of Mueller (2007) has a limited scope focusing on modeling the spatial and temporal aspects of narratives involving one or two characters dining in a restaurant. He employed a combined technique using IE to extract key information about dining episodes, and common sense reasoning to build models of the dining episodes. The model is limited to only a single spatial layout consisting of the street, the dining room, and the kitchen, and further work can be done to extract information about the spatial layout from the text, and use this to construct models of room-scale space.

2 Review of Related Literature

This chapter elaborates on the related works and relation extraction systems. It also discusses on the different sets of semantic relations used by past systems. Lastly, it compares a variety well-known existing knowledge representations.

2.1 Information and Relation Extraction Systems

Over the years, there has been an increasing amount of interest in the automatic detection of semantic relations, with the goal of making computers understand text. The earliest works are those of Hearst (1992) and Berland and Charniak (1999).

Marking the start of the automatic acquisition of relations, Hearst (1992) developed a method that automatically extracts hyponyms (IsA) from a wide variety of texts. One example of this can be seen in the phrase, *Rizzy, a dog.* It shows a hyponymy relation between the words *Rizzy* and *dog.* In extracting hyponymy relations, she used a set of frequently occurring domain-independent lexico-syntactic patterns which undoubtedly define a hyponymy relationship. Though her method has shown encouraging results, it still had some drawbacks such as the ambiguity of some relations extracted. Because her patterns were based on sample sentences in the corpora and aimed to cover as much instances of the hyponymy relation as possible, some of the outputs were indicative of other types of relation. Lastly, she went on to suggest that her method can be used to automatically acquire other types of relation such as meronymy (PartOf).

Later that decade, Berland and Charniak (1999) used a statistical approach to find meronymy (PartOf) relations from a very large corpus. As an example, the phrase *the plot of the story* signifies a meronymy relation between the words *plot* and *story.* In determining such a relation, they used a method similar to Hearst (1992) by also using a pre-defined set of frequently occurring lexico-syntactic patterns. But instead of producing tuples which signify the relation, they focused on producing an ordered list of possible parts given a list of six seed words representing whole objects. The list includes book, building, car, hospital, plant and school. The plant seed word was added to the list to see if the algorithm can identify correct parts despite the ambiguity in the sense of the word. This experiment yielded accuracies lower than the five other seed words. They used statistical metrics to produce the ordered list of possible parts. Though they have stated that their comparable success against Hearst (1992) was due to the large corpora that they used, they were still not able to maximize their corpora to their advantage due to the limited number of wholes and patterns used. They produced a list with an accuracy of 55% for the top 50 parts and 70% for the top 20 parts overall.

Despite their efforts, Hearst (1992) and Berland and Charniak (1999) were not able to address the problem of ambiguity in their patterns and outputs. Cases of ambiguity may occur for patterns signifying a number of semantic relations. For example, *the room of the house* shows a meronymy (PartOf) relation while *the room of the boy* does not. Fortunately, Badulescu et al. (2006) also observed this from both works thus using it as his motivation in employing another approach which automatically extracts PartOf relations.

In tackling PartOf relations, Badulescu et al. (2006) used a knowledge-intensive and supervised method in contrast to what has been used by Berland and Charniak (1999). They trained the algorithm with manually annotated set of positive (indicative of meronymy) and negative (not indicative of meronymy) training samples to produce a decision tree and a set of rules. Particularly, they used C4.5 decision tree learning to produce the rules. After training, they were able to produce a comprehensive set of classification rules to cover almost all subtypes of PartOf relations. They then tested the said rules using two corpora and had an overall average precision of 80.95% and recall of 75.91%.

In comparison, Berland and Charniak (1999) used a few number of words to represent whole entities which have identifiable parts in their very large corpus. In addition, they limited themselves to single word entities and concepts. Badulescu et al. (2006), on the other hand, used an approach which utilizes WordNet and NERD to determine single and multiple word concepts in perspective thus making his approach more general. Lastly, instead of determining the parts of a predefined whole, their work can determine if two noun concepts are indeed part of a PartOf relation through the use of their decision tree and classification rules. Badulescu et al. (2006) also tried to replicate the testing done by Berland and Charniak (1999) in their work but because the corpora used were different, the same conditions cannot be applied.

The aforementioned systems aimed to extract specific relations present in an English text. But such relations, IsA and PartOf, though can be easily extracted, are not the only conceptual relations there is. In lieu of this, several systems have already extracted facts and relations openly from plain-texts (Agichtein & Gravano, 2000) (Banko & Etzioni, 2008), web documents (Alani et al., 2003) (Yates et al., 2007), legal documents (Cheng et al., 2008) and newspapers (Muslea, 1999).

Snowball (Agichtein & Gravano, 2000), an open relation extraction system, employed a novel strategy in generating patterns and extracting relational tables from plain-text documents, specifically newspaper articles. A training phase is done with minimal training samples from human users. The seed patterns are then used to extract new patterns and relation tuples. As part of its extraction process, the system statistically evaluates the newly generated patterns and tuples and retains only the reliable ones in the new iteration. The large-scale evaluation provides Snowball with a methodology to produce high-quality patterns. However, the system can only produce relational tables involving named-entities accurately labeled by Alembic, a third-party named-entity tagger employed by Snowball. An example of a relational table would be for ORGANIZATION and LOCATION pairs. Such a table can contain the pairs *Microsoft-Edmond* and *Boeing-Seattle* which shows that the organizations *Microsoft* and *Boeing* can be found in *Edmond* and *Seattle*, respectively. Though it is only correct to extract such relations, there are still those which do not only involve a couple of named-entities. Relations involving world states like that between morning and go to school, clearly shows that a relation can also be between named-entities and phrases. This scenario poses another limitation of Snowball which is similar to (Berland & Charniak, 1999). Another shortcoming of Snowball would be that it can only extract relations between two named-entities which is not always the case for conceptual relations.

Taking a different path in relation extraction systems, the Artequakt project (Alani et al., 2003) focused on the domain of artists' biographies and extracted conceptual relations in order to automatically generate biographical accounts of artists. In comparison to previous systems, this one did not use any predetermined extraction patterns per se and neither did it learn extraction patterns as a pre-process. Instead, the system just had a list of pre-determined ontology relations that it wants to extract along with its pair of concepts. In the whole process, the Artequakt project made use of third-party tools such as the Apple Pie Parser for syntactic analysis or part-of-speech tagging, GATE for entity recognition and WordNet to supplement GATE and to aid in actual relation extraction.

In extracting the relations, the unstructured web documents first goes through an entity recognition tool (GATE). WordNet is also used to supplement in case GATE fails to recognize any named-entity. The document then goes through the actual extraction phase wherein it gets decomposed into paragraphs and sentences. The part-of-speech of each word in a sentence is then labeled. After this, the main components of a sentence such as the subject, verb and object are identified. The system then uses the verb and entity pairs in each sentence and matches them with a corresponding ontology relation and concept pairs. In case of any linguistic variation, WordNet is used to increase the chance of matching with ontology relations and concepts. In its initial experiment, 50 web documents describing 5 artists were used. Promising results were shown as the system was able to extract at most 3 thousand unique conceptual relations with 85% precision and 42% recall on the average. Its low average recall was due to the varying cardinality of some relations. A high recall is preferred for relations with multiple cardinalities like *places_visited* while high precision is more preferred for relations with a single cardinality like that of *birth_place*.

Though this work has driven away from the usual use of templates in order to extract their target relations, it still boasts of its portability. The use of ontology relations instead of painstakingly specifying every single template for each target relation takes away the need to force-fit a relation extraction system to a specific domain.

In 2007, Yates et al. was able to develop an open information extraction system named TextRunner. It processes a corpus of heterogeneous web documents in a single pass without any human intervention. Though this system does not focus heavily on solving the problems faced by previous systems like portability but rather focus on the scalability of RE systems to the web, its novel contributions can still be considered a solution to such problems.

In developing the system, Yates et al. (2007) used the problems of automation, corpus homogeneity and scalability as motivations. This led to the development of some novel components such as the single pass extractor, self-supervised classifier, synonym resolution and query interface. The single pass extractor tags the sentences with their part-of-speech tags and noun-phrase chunks. Through the self-supervised classifier, it then checks for every pair of noun phrases that are not too far apart and determines whether or not there is a relationship between them. But before this can be done, the classifier has to be trained with positive and negative samples before it can accurately decide which among the noun phrase pairs has a relationship.

Since TextRunner (Yates et al., 2007) does not have a pre-determined set of relations unlike previous works, there is a high chance that the system extracts different tuples representing only one relation. To solve this problem, the system used Resolver to cluster the extracted tuples into sets of synonymous relations and entities.

In evaluating the system, a corpus of 9 million web documents was used. And with that, TextRunner was able to extract approximately 7.8 million well-formed tuples. Human reviewers evaluated some 400 randomly selected extracted tuples and determined that they were 80.4% correct. The system was then further compared to the performance of another traditional IE system, KnowItAll. After using a set of ten high-frequency relations, there were more correct relations extracted by TextRunner than KnowItAll.

In trying to improve TextRunner (Yates et al., 2007), Banko and Etzioni (2008) developed new systems in order to conduct a survey on the differences of open and traditional relation extraction. In these systems, the Conditional Random Fields model was used to label instances of a relation between all possible entity pairs. This is already an improvement from the Nave-Bayes classifier used by TextRunner which chooses tokens between entities heuristically and only predicts whether these indicated a relationship or not. Conditional Random Fields, on the other hand, is an undirected graphical model used to model multiple interdependent variables.

O-CRF, the new open relation extraction system, performs a self-supervised training as with TextRunner. It uses independent heuristics and applies them to the PennTreebank in order to obtain labeled relational tuples which are then described with features. Such features include part-of-speech tags, regular expressions, context words and the combination of features six words to the left and six words to the right of the labeled word. The context words used here include only closed classes like prepositions and determiners. Function words like verbs and nouns are not utilized as context words. The labeled relational tuples are then used to train the CRF. In extracting relations, O-CRF first does a single pass over the corpus and uses phrase chunking to identify entities. The CRF is then used to identify and label the relations occurring between entity pairs. As with TextRunner, O-CRF is also beset with duplicate relations. This was solved by applying the Resolver algorithm to predict if two relation strings refers to the same thing.

In order to make comparisons, R1-CRF, a system applying the same CRF model was developed. But this time, the traditional relation extraction paradigm is utilized. Though the same graphical model is used, there were some tweaks in order to comply with the traditional paradigm. A relation is given in advance and instead of training the CRF unsupervised, hand-labeled positive and negative samples are used. And unlike O-CRF, R1-CRF can use context words besides closed classes.

After evaluation, O-CRF showed 88.3% precision and 45.2% recall. These show promising results in using open relation extraction. However, the usage of such a paradigm will only be essential if the number of relations is big or unknown. This is also essential for extraction jobs concerning massive corpora. On the other hand, traditional relation extraction is more suitable for extraction jobs with a small number of target relations.

2.2 Semantic Relations

The interest in the automatic extraction of semantic relations in text has become one of the growing interests among researchers in the NLP community. And in recent years, a number of them applied different classification techniques on various domains. This, however, led to a variety of disjoint classification schemes which later on became a nuisance to the advancement of the field.

Way back in 1987, Mann and Thompson (1987) presented Rhetoric Structure Theory which describes major features of the organization of natural text. This descriptive theory is used linguistically to characterize the structure of natural text in terms of relations between parts of the text. It is a hierarchical structure which identifies both the transition point of a relation and the items related. Though it can be used for large corpora, its scope is limited to monologues only. Dialogues and spoken text, which are present in stories, are not handled by RST.

The relations in RST are mainly classified into two: nuclear-satellite and multinuclear. The nuclear-satellite relations can still be further classified as presentational or subject matter relations. Presentational relations are those which aim to increase inclination in the reader. An example of this would the Evidence relation which aims to increase the belief of the reader on the nucleus of the relation. Other than that, Motivation, Justify, and Background, among others, are also considered as presentational relations. Subject matter relations, on the other hand, aims to make the reader recognize the relation. Such relations include Condition, Circumstance, Elaboration, Purpose and Volitional cause, among others.

Years after RST, Knott and Dale (1994) conceptualized a set of coherence relations. But instead of treating relations as constructs used to describe a text, relations were thought of as constructs with psychological reality. Using this as motivation, Knott and Dale (1994) developed a bottom-up methodology to define a set of relations using cue phrases which is a concrete linguistic indicator of a relation in a text. Unlike most theorists who define relations between entities in a sentence, the relations described in this work are mostly those between the sentences of a text, thus implicit in nature. Such coherence relations are sometimes made explicit through the use of cue phrases like *for example* and *before*. The relations based on the cue phrases are divided into seven classes, namely: sequence, situation, causal/purpose, similarity, contrast/violated expectation/choice, clarifying and interruption.

In the domain of medicine, Rosario and Hearst (2001) defined a classification scheme for two-word noun compounds. Though their data was from MedLine, a collection of biomedical journals, the classes and relations defined in the study was made as general as possible. To be more specific, there was more granularity than those in case frames but the relations were also more general than the ones classified in traditional information extraction systems. In their classification scheme, there were actually 38 relations divided into 12 classes. General relations are also mixed with domain-specific ones. Examples of general relations include time of, frequency, instrument, object, topic and location, among others while those domain-specific ones include defect in location, person/center that treats, defect, research on and bind, among others.

Rosario et al. (2002) continued the study on semantic relations for noun compounds. But this time, a different classification scheme was used. Instead of their previous two-level hierarchy, they used the MeSH hierarchy which is a multi-level lexical hierarchy of classifying relations for noun compounds with 15 classes at the topmost level. Each of the 15 topmost classes corresponds mainly to a specific medical terminology or field like Anatomy, Biology, etc. This scheme presents classes which are more granular and more specific to the medicine field.

In 2003, Nastase and Szpakowicz presented a classification scheme for noun-modifier pairs in base noun phrases. This scheme is a two-level hierarchy classification of semantic relations for noun-modifier pairs. The hierarchy has 5 top-level classes and 30 bottom-level classes. Its 5 superclasses include causality, temporality, spatial, participant, and quality. Causality relations are mainly those which show cause-effect relations. For example, the base noun phrase cold virus will have a cause relationship between them since the head word virus caused the modifier cold. But other than the usual cause and effect relations, there is also the purpose relation which exists whenever the head word is meant for the modifier. Such is the case for the base noun phrase concert ground where the head word ground has the purpose of having a concert. Temporality relations, on the other hand, express time. One example is the frequency relation which holds whenever the head word occurs every time the modifier occurs. This is evident in the base noun phrase weekly mass. Spatial relations pertain to having the nature of space. Such is the case for outgoing call which shows a direction relation. Participant relations, unlike previous superclasses, include relations similar to semantic roles. One example of this would be the agent role which exists when the modifier performs the head word. The base noun phrase fan boycott signifies such a relation since fan performs the boycott. Lastly, the quality relations are those specifying content, manner and type, among others.

The same year, Alani et al. (2003) used a classification scheme very specific to the domain of artists' biographies. The ontology was derived from the CIDOC Conceptual Reference Model ontology and further modified by adding classes and relations needed to represent pieces of information appropriate for artists. Examples of such relations include date of birth, place of birth and inspired by, among others. These ontology relations are then utilized in generating artist biographies.

Instead of concentrating on classifying semantic relations for noun compounds or base noun phrases, Moldovan et al. (2004) specified a scheme in classifying relations for a range of phrases. This includes 35 classes of relations spanning at various syntactic levels. They were mostly derived from the list of relations specified in previous researches. However, it only contains the most frequently used relations in a large corpus. Some of the relations include possession, temporal, part-whole, is-a, cause, purpose, frequency, stimulus, manner and location, among others.

Concentrating more on the field of story generation, Nakasone and Ishizuka (2006) developed a storytelling ontology model using RST (Mann & Thompson, 1987). The ontology was made as generic as possible since most storytelling ontology models were defined and constrained by the way the events were linked and the nature of the narratives. Instead of constraining the model with such notions, the solution was more focused on how the narratives were organized and communicated to readers. Since the domain of the model is story generation, the ideas and events are to be focused on the concept of a conflict. Hence, the RST relations utilized were categorized into two: Conflict or Resolution relations. Conflict relations describe how the current state of the story is changed. Such relations include Contrast, Solutionhood, Elaboration, Consequence and Sequence. Resolution relations, on the other hand, describe how to understand the current state of the story. Examples of this type of relation include Background, Cause, Purpose and Result, among others.

And just recently, Hendrickx et al. (2009) developed a system which does a multi-way classification of semantic relations between a pair of nominals. But this time, instead of classifying all possible semantic relations, the focus was just on nine mutually exclusive domain-independent semantic relations with enough exhaustive coverage. The list includes Entity-Destination, Instrument-Agency, Product-Producer, Content-Container, Component-Whole, Entity-Origin, Cause-Effect, Member-Collection and Communication-Topic.

2.3 Knowledge Representations

Common sense knowledge acquisition is not new in the Natural Language Processing field. Over the years, several knowledge repositories or databases have been developed like WordNet, VerbNet, Cyc, FrameNet and ConceptNet. These repositories contain entries ranging from syntactic to semantic in nature. Though most, if not all, contain semantic relations, there are certainly differences on the relations they contain and how they are represented.

Begun in 1984, CYC aims to formalize common sense knowledge into a logical framework. It stores knowledge of every day concepts, objects and events in axioms. The assertions are both manually and automatically done by knowledge engineers at Cycorp assuming that they are already known in the world. In representing the assertions, a first-order predicate calculus, named CycL, with an extension of some second-order features is used. The knowledge base is partitioned into "microtheories" which are a bundle of assertions. Some microtheories are partitioned based on their common assumptions while some are partitioned based on a specific domain and level of detail. This mechanism allows Cyc to infer faster by focusing on a specific microtheory. Each time an inference is made, new assertions may be added into the knowledge base (source: cyc.com).

One of the forerunners and arguably the most popular among knowledge bases is WordNet. It is a general purpose semantic knowledge base started in 1985 at Princeton University. Its database consists of words, mostly nouns, verbs and adjectives. Each entry is structured into senses and associated using a small number of semantic relations such as the synonym, is-a and part-of relations. These relations are represented in WordNet as a semantic network with each word as a node and the relations as edges.

In 1998, Fillmore et al. (1998) developed FrameNet, a lexical resource containing frame-semantic descriptions of each English lexical item (noun, adjective and verb). The semantic domains that FrameNet covers are the following: health care, chance, perception, communication, transaction, time, space, body, motion, life stages, social context, emotion and cognition. The whole lexical database is composed of a lexicon, the frame database and the annotated example sentences. Each lexical entry contains some usual information like part-of-speech as well as formulas which describe how elements of a semantic frame can be recognized. FrameNet, as what was previously stated, also defines the argument structure of each entry in the lexicon through roles but instead of using case-roles or thematic roles, each argument is given a role name relative to a certain concept. The data structures used to represent the lexical entries along with their semantic frames were implemented using SGML.

VerbNet (Kipper, Dang, & Palmer, 2000) is another repository of semantic information but unlike WordNet, Cyc and ConceptNet, this repository is more focused on verbs and their semantics. It is primarily a verb lexicon using Levin verb classes to represent the lexical entries. As its semantic information, the lexical resource relates each verb's thematic roles and semantic predicates with syntactic frames and restrictions.

Though VerbNet has semantic information included in its verb lexical entries, it still differs from what WordNet, Cyc and ConceptNet has. The verb lexicon stores semantic roles and not semantic relations. Note that they are two different things though they are both semantic in nature. Semantic roles exist between a verb and its arguments while semantic relations may exist between any parts of speech.

Combining the structure of WordNet and the semantic richness of Cyc, ConceptNet (Liu & Singh, 2004b) is a large-scale common sense knowledge database aimed to optimize practical inferences over real-world texts. It adopted the semantic network knowledge representation of WordNet and included 17 additional relations such as EffectOf, SubEventOf and CapableOf. This will provide a richer semantic network compared to what WordNet already has. However, there are still differences on the relations they contain. In WordNet, relations are more formal and is assumed to always happen while in the case of ConceptNet, it relations are more informal and defeasible. This means that since ConceptNet is geared towards a more practical inference, its relations may not always happen. One example would be the part-of relation between dog and pet. A dog will always be a canine but not a pet.

Having a set of only 20 relations is not much of an advantage over Cyc since it provides more than 20 and with more detail. However, compared to the use of CycL as a knowledge base representation, ConceptNet's semantic network representation makes it easier to make practical inferences.

3 Research Methodology

This chapter discusses the systematic approach to be performed in order to accomplish the objectives of this research.

3.1 Data Gathering

During this stage, data gathering will be performed to identify the following: types of conceptual/semantic relations, appropriate semantic relations for the children's story domain, architecture of relation extraction systems and, algorithms for extracting conceptual relations. Additionally, the input corpus consisting of at least 30 children's stories will be gathered. Furthermore, interviews with English language professors and linguists will be conducted to verify the correctness of the input corpus. This means that if a corpus contains dialogues, among other considerations, it is deemed incorrect. Thus, a modification of the corpus will be done to address the issue. If possible, manipulation of the input corpus will be done in order to fit the requirements of the tools to be utilized. The extraction templates will also be defined by analyzing the sentence structures in a children's story.

There will be three types of modifications done on the corpus to remove dialogues. The first type of modification is done by transforming the dialogues into declarative sentences. The second type is the usual transformation of direct to indirect speech. And lastly, the third type of modification is the explicit addition of discourse markers whenever a dialogue is modified. After removing the dialogues, pronouns will also be removed and complex sentences will be simplified. Please see Appendix C for samples of each type of modification.

3.2 Requirements Specification

After gathering the data, the requirements will be defined and analyzed to determine the objectives and scope of the research. The resulting requirements specification will be validated to ensure completeness of the study and the tool to be developed. The final algorithm to be implemented should also be defined.

3.3 Architectural Design

In the Architectural Design stage, the different modules will be identified as well as the different external tools to be used. This includes a part-of-speech tagger, named-entity identifier, and text simplifier, among others. Other resources to be utilized will also be identified. Afterwards, the architectural design of the tool to be developed will be defined according to the final algorithm. Furthermore, the data structures which will represent the semantic relations in Picture Books will also be analyzed and designed.

3.4 Implementation

The actual implementation of the architectural design as well as the final algorithm will be done in this stage. Debugging and unit testing will be done regularly to ensure the efficiency and correctness of the tool and algorithm.

3.5 Testing

Testing will be done to ensure the quality and efficiency of the software. Unit testing for each subsystem will be performed. After doing so, integration testing will be performed to verify that each tool/subsystem receives the correct input from the previous tool/subsystem and generates the appropriate result for use by subsequent tools/subsystems.

Test cases will be employed to check that all tools/subsystems interact correctly. System and functional testing will also be performed to check the functionality and performance of system functions. Lastly, the outputs of the system will mainly be evaluated through the use of Picture Books. The generated story of Picture Books after using the output semantic network will be evaluated by employing the same evaluation technique done in Picture Books. The output semantic network may also be evaluated by English language linguists to ensure the validity of the relations.

3.6 Documentation

Throughout the entire process of implementing the algorithm, documentation will be done to track its progress. This is also to ensure that any changes and implementations in the requirements of the study will be reflected in the documents.

3.7 Calendar of Activities

Tables 3.1 and 3.2 shows a Gantt chart of the activities. Each bullet represents approximately one week worth of activity. The overlapping activities ensure that any omissions and modifications will be changed immediately.

Table 3.1: Timetable of Activities (Part 1)

Activities (2010)	Jan	Feb	Mar	Apr	May	Jun	Jul
Data Gathering	●●	●●●●	●●●●	●●●●			
Requirements Specification			●●	●●●			
Architectural Design				●●●	●●●●	●●	
Implementation					●●	●●●●	●●●●
Testing							●●●●
Documentation	●●	●●●●	●●●●	●●●●	●●●●	●●●●	●●●●

Table 3.2: Timetable of Activities (Part 2)

Activities (2010)	Aug	Sep	Oct	Nov	Dec
Data Gathering					
Requirements Specification					
Architectural Design					
Implementation	●●●●	●●●●			
Testing	●●●●	●●●●	●●●●	●●●●	
Documentation	●●●●	●●●●	●●●●	●●●●	●●●●

A Theoretical Framework

A.1 Semantic ontology and Semantic Relations

An ontology is an artifact with a set of representational primitives to model knowledge for a particular domain (Gruber, 2008). The representational primitives are classes or objects, attributes of the objects and relationship of each object. The design of the semantic ontology of Picture Books is patterned after ConceptNet (Liu & Singh, 2004a), a large-scale common sense knowledge base.

The nodes used by ConceptNet are of three general classes representing noun phrases, attributes, and activity phrases. A semantic relation connects two concepts while a semantic category classifies them. The semantic relations are binary relation types defined by Open Mind Commonsense project (Singh et al., 2002). Some of the ConceptNet relations are shown in Table A.1.

Table A.1: ConceptNet relations (Liu & Singh, 2004b) to be extracted with sample concepts

ConceptNet relations
IsA(dog, animal)
PropertyOf(apple, red)
PartOf(window, house)
MadeOf(sculpture, clay)
FirstSubeventOf(yawn, sleep)
EffectOf(become tired, sleepy)
CapableOf(ball, bounce)
LocationOf(seesaw, playground)
UsedFor(spoon, eat)

Aside from showing the list of ConceptNet relations to be extracted, Table A.1 also shows some sample concepts for each relation. As can be seen from the table, not all relations have the same concepts as part of their tuples. Table A.2 shows the different questions that each ConceptNet relation tries to answer and the different extraction templates that can be used to extract each one of them.

A.2 Picture Books

Picture Books generates a story for a given input picture that contain a background selected by the user from the background library, as well as the character

Table A.2: ConceptNet relations (Liu & Singh, 2004b) to be extracted with extraction templates

ConceptNet relations	Question answered	Extraction Template/s
IsA	What kind of thing is it?	<NP> is a kind of <NP> : A dog is a kind of canine. <NP> is a <NP> : A dog is a canine. <NP>, <NP>, is : The dog, a canine, is... <NP> is a type of <NP> : A dog is a type of canine.
PropertyOf	What properties does it have?	<Adjective> <Noun> : The red ball... <NP> is <AP> : The ball is red.
PartOf	What is it part of?	<NP> is a part of <NP> : A window is a part of a house. <NP> is part of <NP> : A window is part of the house. <NP> has <NP> : The house has a window. <Noun:Possessive> <Noun> : The house's window...
MadeOf	What is it made of?	<NP> is made of <NP> : A cake is made of flour. <NP> is comprised of <NP> : A cake is comprised of flour, eggs and yeast. <NP> form <NP> : Students form a class. <NP> become <NP> : Clay becomes sculptures.
CapableOf	What can it do?	<NP> <Verb> : The boy jumps. <NP> can <VP> : The boy can jump.
LocationOf	Where would you find it?	<NP> is at <NP> : The slide is at the playground. <NP> is located at <NP> : The swing is located at the playground. <NP> can be found at <NP> : The see-saw can be found at the playground. <NP> in <NP> : The sandbox in the playground....
UsedFor	What do you use it for?	<NP> is used for <VP> : A rolling pin is used for baking. <VP:Gerund> requires <NP> : Baking requires a measuring cup. <VP> with <NP>: Kisha hit with a bat.

and object stickers placed onto the background. The ontology is used to derive relations between concepts, which refer to objects in the picture as well as the theme associated by the system through the background. An excerpt of a generated story and the corresponding conceptual knowledge used is shown in Table A.3.

Table A.3: Excerpt from a story generated by Picture Books with corresponding conceptual knowledge

Line	Story Text	Conceptual Knowledge
1	Rizzy the rabbit was in the living room.	
2	She played near a lamp.	
		CapableOf(lamp, break) ***ConceptuallyRelatedTo***(break, break object)
3	Rizzy broke the lamp.	
		EffectOf(break object, be scared)
4	She was scared.	
	:	
5	Rizzy told Mommy Francine that Daniel the dog broke the lamp.	
	:	
		LastSubeventOf(break object, get punished)
6	He got punished.	
		LastSubeventOf(get punished, grounded) ***IsA***(grounded, punishment)
7	Mommy Francine told Daniel that he was grounded.	
		LastSubeventOf(grounded, cry)
8	He cried.	

In line 1, the main character (*Rizzy the Rabbit*) and the setting (*living room*) were determined from the character sticker placed onto the selected background by the user. In line 2, the object (*lamp*) may or may not be in the picture, but included in the generated story based on the theme that is associated to the background. In this example, the theme is *being honest* through admitting your mistake (that is, the main character must not lie about breaking the lamp).

Access to the ontology is needed to derive events that can happen next in the story, as shown in line 3, and the effects of the resulting event, shown in line 4. Line 5 is the starting point of the rising action, where the main character misbehaves (*told a lie*) and the subsequent events and effects of the misbehavior. All the knowledge needed by Picture Books to do its task were manually encoded by the proponents into the system, based on the identified background and themes, which are appropriate to the target age group. The knowledge in ConceptNet cannot be used directly as these are not suitable for the users of Picture Books. Thus, only some of the ConceptNet knowledge as well as relations were used to build the ontology of Picture Books. Table A.4 lists some of these relations defined in Picture Books following the form <relationship>(<concept1>, <concept2>).

Table A.4: Semantic relationships adopted from ConceptNet (Liu & Singh, 2004b) with sample concepts of Picture Books

Semantic Category	Semantic Relationships
Things	***IsA***(headache, pain) ***PropertyOf***(apple, healthy) ***PartOf***(window, pane) ***MadeOf***(toy car, clay)
Events	***FirstSubeventOf***(tell bedtime story, sleep) ***EventForGoalEvent***(go to grocery store, buy food) ***EventForGoalState***(clean up, be neat) ***EventRequiresObject***(play, toy)
Actions	***EffectOf***(become dirty, itchy) ***EffectOfIsState***(make friends, friendship) ***CapableOf***(toy car, play)
Spatial	***OftenNear***(sailboat, water) ***LocationOf***(teacher, school)
Functions	***UsedFor***(thermometer, check temperature)

A.3 ConceptNet

ConceptNet (Liu & Singh, 2004b) is a large-scale common sense knowledge database aimed to optimize practical inferences over real-world texts. It adopted the semantic network knowledge representation of WordNet and included 17 additional relations such as EffectOf, SubEventOf and CapableOf. This will provide a richer semantic network compared to what WordNet already has. However, there are still differences on the relations they contain. In WordNet, relations are more formal and is assumed to always happen while in the case of ConceptNet, it relations are more informal and defeasible. This means that since ConceptNet is geared towards a more practical inference, its relations may not always happen. One example would be the part-of relation between dog and pet. A dog will always be a canine but not a pet.

The ConceptNet semantic network was populated with concepts and relations through a distributed solution of acquiring common sense knowledge from the public using a web-based data entry mechanism of the Open Mind Common Sense (OMCS) project (Singh et al., 2002). OMCS employs both semi-structured and free-form data entry approaches. The semi-structured approach utilizes extraction patterns commonly used by IE systems. Each extraction pattern or template has slots that users can fill-up, and is mapped directly to a relation.

Given the template “*<X> is a kind of <Y>*”, the possible values for <**X**> and <**Y**> that users can provide and the corresponding hypernymy (IsA) relations that are acquired are shown in Table A.5.

Table A.5: Sample values to derive the hypernymy (***IsA***) relations

<X>	<Y>	Relations
Apple	Fruit	IsA(apple, fruit)
Ball	Toy	IsA(ball, toy)
Rose	Flower	IsA(rose, flower)

A.4 GATE

GATE (General Architecture for Textual Engineering) is a general-purpose infrastructure aimed for natural language software development. It also aims to reduce integration overheads. This is done through the provision of standard mechanisms of data communication for the the different software components. GATE also uses Java and XML as its platforms.

As a language engineering architecture, GATE provides processing resources with ANNIE as its main resource. ANNIE provides a set of reusable processing resources to facilitate language engineering tasks. It consists of the following resources: tokeniser, sentence splitter, POS tagger, gazetteer, finite state transducer or semantic tagger, orthomatcher and coreference resolver. The tokeniser splits a given text into simple tokens. The sentence splitter splits the text into sentences. The POS tagger tags each word or symbol with their specific part-of-speech tags. The gazetteer consists of lists like that of cities and organizations. It can also consist of lists of indicators, like titles and other designators. The orthomatcher performs coreference or entity tracking through the recognition of relations between entities. The coreference resolver detects identity relations between entities. Lastly, the semantic tagger consists of tailor-made rules written in JAPE language. These rules describe the patterns and annotations to be created. A JAPE grammar has a set of phases which consist of pattern rules. These phases run sequentially.

A.5 Relation Extraction Techniques

In extracting semantic relations, one technique is through the generation and use of extraction patterns. For each target relation, a set of extraction patterns are needed to handle all possible instances of a relation in a sentence.

Table A.6 shows other extraction patterns and the corresponding relations of ConceptNet.

Table A.6: Sample extraction patterns and corresponding ConceptNet relations

Extraction Pattern or Template	Relations
CAKE is a kind of FOOD.	IsA(cake, food)
CAKE is made of FLOUR.	MadeOf(cake, flour)
FLOUR is WHITE.	PropertyOf(flour, white)
The effect of DRINKING MILK is GOOD HEALTH.	EffectOf(drinking milk, good health)

From the examples above, an instance of an extraction pattern generates one relation. But sentences may contain conjunctive phrases, which in turn may result to multiple relations being learned, as shown in Table A.7 for the pattern “*<X> is made of <Y>*”. This will be explored further in this research.

Table A.7: Generating multiple relations from a single extraction pattern

Extraction Pattern or Template	Relations

CAKE is made of FLOUR, SUGAR, and MILK.	MadeOf(cake, flour)
	MadeOf(cake, sugar)
	MadeOf(cake, milk)

Part-of-speech tags may also be utilized to identify phrases and its constituents. For example, in Table A.8, the noun phrase used to fill the <*X*> variable in the *IsA* template has three components, namely an article (“*the*”), an adjective (“*sweet*”), and a noun (“*cake*”). Extracting this knowledge can lead to the relation *PropertyOf(cake, sweet).* This will be explored further in this research.

Table A.8: Utilizing POS tags for implicit relations

Input Sentence following a Template	Relations
The sweet cake is a dessert.	Explicit extraction pattern: IsA(dessert, cake) Implicit from POS tag: PropertyOf(cake, sweet)

The input stories may contain complex sentence structures, such as conjunctions and embedded clauses. Text simplification algorithms, employed in SimText (Damay, Lojico, Lu, Tarantan, & Ong, 2007) may be utilized to convert these sentence structures into simpler ones. Consider the sentence "*Anna, who is the queen, went to the market; meanwhile, the king went to the mall.*" By identifying and transforming this to three simpler sentences: "*Anna is the queen. She went to the market. Meanwhile, the king went to the mall.*", the following relations can be extracted. This will be explored further in this research.

IsPerson(Anna)
HasRole(person, queen)
HasRole(person, king)
CapableOf(person, go)
TargetOf(go, market)
TargetOf(go, mall)

A.6 Additional Relations

Aside from the 14 ConceptNet relations adopted by Picture Books, other semantic relations are also considered. The previous example shows some of the possible new relations that may be included in the output of the proposed system, namely:

- HasRole to designate that characters may play certain roles
- RoleResponsibleFor to model a specified role is responsible for a given task, e.g., the king rules a country
- TargetOf to model target objects of certain actions

One of the identified limitations in the current knowledge base of Picture Books is the lack of relations to denote event occurrences. Consider again the text:

The evening was warm. Ellen the elephant was at the school. She went with Mommy Edna to the school.

If appropriate relations are available, e.g., Happens to designate that an activity, such as going to school, can only happen at a certain time of day, such as morning, then the resulting text can be:

The morning was sunny. Ellen the elephant was at the school. She went with Mommy Edna to the school.

Certain granularities may be provided to the relations representing various aspects of time, namely season (planting can only occur during spring, snow can only fall during winter), month (Christmas in December, Valentine's in February), or even weeks, days, hours, and minutes.

Mueller (2003) made use of event calculus consisting of the following predicates to model event occurrences:

- Happens(e, t) represents that an event *e* happens at time *t*.
- HoldsAt(f, t) represents that a fluent *f* holds at time *t*.
- Initiates(e, f, t) represents that if event *e* occurs at *t* then fluent *f* starts holding after *t*.
- Terminates(e, f, t) represents that if event *e* occurs at *t* then fluent *f* stops holding after *t*.

However, in this research, only the Happens relation will be used and extracted. The other three relations defined by Mueller (2003) are not yet necessary.

Table A.9: Mapping of RST relations to ConceptNet relations

RST Relation	ConceptNet Relation
Cause (one event is the cause of another event)	EffectOf(event1, event2)
Background (one event serves as background information for the other)	EventForGoalEvent (clean up, be neat)
Example	InstanceOf

Nakasone and Ishizuka (2006) developed a concept representation model to convey ideas of a story, by identifying organizations of text structure using the Rhetorical Structure Theory of Mann and Thompson (1987). RST relations can then be mapped to existing ConceptNet relations, as shown in Table A.9. This will be explored in this research.

B Architectural Design

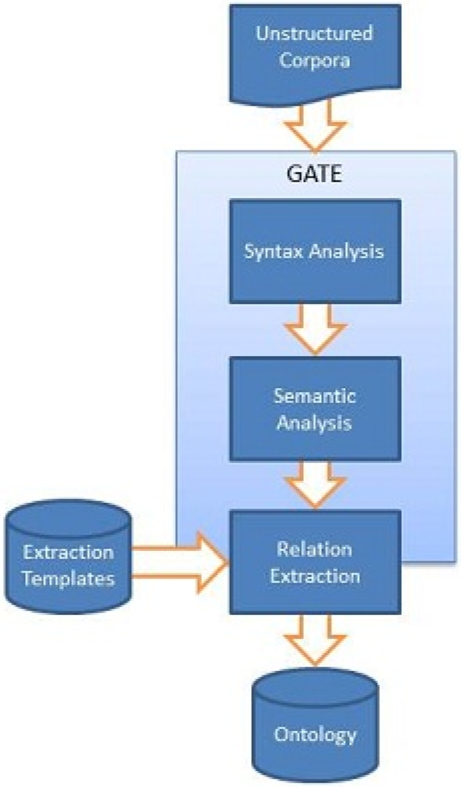

Figure B.1: Architectural Design

Shown in B.1 is the architectural design for the relation extraction system to be developed. The process requires an unstructured corpora as an input, which in this case would be a children's story. It has been manually validated and modified to fit the tool requirements. Some modifications include the removal of dialogues. Once the input has already been form-fitted, it goes through GATE for syntax analysis. It first splits the corpora into tokens and sentences. The corpora is then passed to the tagger which annotates each word or symbol with their part-of-speech tags. GATE is then used to recognize the named-entities in the corpora. The named-entities are used later in relation extraction to determine whether a pair of entities fit the extraction templates. After syntax analysis, the corpora then goes into semantic analysis wherein the words' morphology and tense are annotated. The sentences are also chunked into phrases/clauses to accommodate concepts involving multiple words. Once all preprocessing to the corpora has been done, it goes through relation extraction wherein relations found in a sentence and across sentences are extracted. This uses the extraction templates. The extraction is done in two phases. The first phase is done through the JAPE Transducer where simple relations are tagged and extracted. The second phase involves the extraction of event-related relations. After extraction, simple inferencing will be done to ensure that final relations will not have proper nouns and become as general as possible. Lastly, the extracted relations are stored in an ontology similar to that of Picture Books.

C Sample Children's Story

As mentioned in Chapter 3, each children's story in the corpus undergo three types of modifications to simulate different scenarios or story formats. And in order to illustrate the different modifications, here is an unmodified excerpt from one of the stories already in the corpus entitled "A Wild Weather Day".

It was a wild and windy day. The JumpStart ship was headed for Tree Fort Island.

Frankie was at the wheel. The sails flapped in the wind. The ship raced through the water.

"Why are we going so fast?" Pierre asked. "The wind is blowing us along on our adventure," CJ said. "Did you know that wind is just air that is strong and fast?"

"Like Frankie!" Pierre said. "He's strong and fast, too."

"Why is the sky getting so dark?" Pierre asked.

"I know why it's dark!" Eleanor said. "Clouds get dark when they fill up with tiny drops of water."

"Look! They're almost the same color as your bow," Pierre said.

A big drop of rain fell on Pierre's nose. "Oh, no!" he said. "It's starting to rain!"

"The rain is coming from the clouds," CJ said.

"The water in tne clouds got too heavy, and now it's falling down on us!"

"Just like when Hopsalot waters his garden," said Pierre.

The first type of modification is done by transforming the dialogues into declarative sentences. Here is the modified version of the excerpt above:

It was a wild and windy day. The JumpStart ship was headed for Tree Fort Island.

Frankie was at the wheel. The sails flapped in the wind. The ship raced through the water.

They are going so fast. The wind is blowing them along on their adventure. The wind is just air that is strong and fast.

Frankie is like the wind. He's strong and fast, too.

The sky is getting so dark.

Eleanor knows why it is dark. Clouds get dark when they fill up with tiny drops of water.

They are almost the same color as Eleanor's bow.

A big drop of rain fell on Pierre's nose. It is starting to rain!

The rain is coming from the clouds.

The water in the clouds got too heavy, and now it is falling down on them!

Just like when Hopsalot waters his garden.

The second type is the usual transformation of direct to indirect speech.

It was a wild and windy day. The JumpStart ship was headed for Tree Fort Island.

Frankie was at the wheel. The sails flapped in the wind. The ship raced through the water.

Pierre asked why they were going so fast. CJ said that the wind was blowing them along on their adventure. He also asked if they know that wind is just air that is strong and fast.

Pierre said like Frankie. Pierre said that Frankie was strong and fast, too.

Pierre asked why the sky was getting so dark.

Eleanor said that she knew why it is dark. She explained that clouds get dark when they fill up with tiny drops of water.

Pierre said look! They are almost the same color as Eleanor's bow.

A big drop of rain fell on Pierre's nose. He was shocked. It is starting to rain.

CJ said that the rain is coming from the clouds.

The water in the clouds got too heavy, and now it is falling down on them.

Pierre said that it is just like when Hopsalot waters his garden.

And lastly, the third type of modification is the explicit addition of discourse markers whenever a dialogue is modified.

The JumpStart ship was headed for Tree Fort Island because it was a wild and windy day.

Frankie was at the wheel. The sails flapped in the wind while the ship raced through the water.

They are going so fast because the wind is blowing them along on their adventure. The wind is just air that is strong and fast.

Frankie is like the wind because he's strong and fast, too.

The sky is getting so dark.

Eleanor knows why it is dark. Clouds get dark when they fill up with tiny drops of water.

They are almost the same color as Eleanor's bow.

A big drop of rain fell on Pierre's nose because it is starting to rain!
The rain is coming from the clouds.
The water in the clouds got too heavy, and now it is falling down on them!
Just like when Hopsalot waters his garden.

For each type of modification, there may be cases wherein a specific sentence or a line uttered by a character will be omitted. Such sentences can be interjections like "Oh my!" and "Alas!".

D Resource Persons

Ms. Ethel Ong
Adviser
College of Computer Studies
De La Salle University-Manila
`ethel.ong@delasalle.ph`

References

Agichtein, E., & Gravano, L. (2000). Snowball: Extracting relations from large plain-text collections. In *Proceedings of the Fifth ACM International Conference on Digital Libraries.*

Alani, H., Kim, S., Millard, D., Weal, M., Lewis, P., Hall, W., et al. (2003). Automatic Extraction of Knowledge from Web Documents. In *Proceedings of ISWC 2003 Workshop on Human Language Technology for the Semantic Web and Web Services.*

Badulescu, A., Girju, R., & Moldovan, D. (2006). Automatic Discovery of Part-Whole Relations. *Comput. Linguist.*, *32*(1), 83–135.

Banko, M., & Etzioni, O. (2008). The Tradeoffs Between Open and Traditional Relation Extraction. In *Proceedings of ACL Human Language Technology (HLT 2008)* (pp. 23–36).

Berland, M., & Charniak, E. (1999). Finding parts in very large corpora. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics* (pp. 57–64). Morristown, NJ, USA: Association for Computational Linguistics.

Cheng, T., Cua, J., Tan, M., & Yao, K. (2008). *Information Extraction for Legal Documents.*

Damay, J., Lojico, G., Lu, K., Tarantan, R., & Ong, E. (2007). Simplifying Text in Medical Literature. *Journal of Research in Science, Computing, and Engineering*, 37–48.

Fillmore, C., Baker, C., & Lowe, J. (1998). The Berkeley Framenet Project. In *Proceedings of the 17th international conference on Computational linguistics* (pp. 86–90). Morristown, NJ, USA: Association for Computational Linguistics.

Gruber, T. (2008). Ontology. In *Encyclopedia of Database Systems.*

Hearst, M. (1992). Automatic Acquisition of Hyponyms from Large Text Corpora. In *Proceedings of the 14th International Conference on Computational Linguistics* (pp. 539–545).

Hendrickx, I., Kim, S. N., Kozareva, Z., Nakov, P., Seaghdha, D., Pado, S., et al. (2009). SemEval-2010 Task 8: Multi-Way Classification of Semantic Relations Between Pairs of Nominals. In *Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions (SEW-2009), in conjunction with NAACL-HLT 2009* (pp. 94–99).

Hong, B., & Ong, E. (2009). Automatically Extracting Word Relationships as Templates for Pun Generation. In *Proceedings of the NAACL Human Language Technology 2009 Workshop on Computational Approaches to Linguistic Creativity* (pp. 24–31).

Kipper, K., Dang, H. T., & Palmer, M. (2000). Class-based construction of a verb lexicon. In *Proceedings of Seventeenth National Conference on Artificial Intelligence AAAI 2000.*

Knott, A., & Dale, R. (1994). Using Linguistic Phenomena to Motivate a Set of Rhetorical Relations. *Discourse Processes*, 35–62.

Lester, J., Rowe, J., & McQuiggan, S. (2007). Narrative Presence in Intelligent Learning Environments. In *Association for the Advancement of Artificial Intelligence Symposium on Intelligent Narrative Technologies 2007* (pp. 126–133).

Liu, H., & Singh, P. (2004a). Commonsense Reasoning in and over Natural Language. In *Proceedings of the 8th International Conference on Knowledge-Based Intelligent Information and Engineering Systems* (pp. 293–306).

Liu, H., & Singh, P. (2004b). Conceptnet – A Practical Commonsense Reasoning Tool-Kit. *BT Technology Journal*, 211–226.

Mann, W., & Thompson, S. (1987). *Rhetorical Structure Theory: A Theory of Text Organization* (Tech. Rep.). University of Southern California, Marina Del Rey, Information Sciences Institute.

Mateas, M., & Stern, A. (2003). Faade: An Experiment in Building a Fully-Realized Interactive Drama. In *Game Developers Conference, Game Design track.*

Moldovan, D., Badulescu, A., Tatu, M., Antohe, D., & Girju, R. (2004). Models for the semantic classification of noun phrases. In *Proceedings of the Computational Lexical Semantics Workshop at HLT-NAACL 2004* (pp. 60–67).

Montfort, N. (2009). Curveship: An Interactive Fiction System for Interactive Narrating. In *Proceedings of the NAACL Human Language Technology 2009 Workshop on Computational Approaches to Linguistic Creativity* (pp. 55–62).

Mueller, E. (1999). Prospects for In-Depth Story Understanding by Computer. *Journal of Cognitive Systems Research*, 307–340.

Mueller, E. (2003). Story Understanding through Multi-Representation Model. In *Proceedings of the HLT-NAACL 2003 Workshop on Text Meaning* (pp. 46–53).

Mueller, E. (2007). Modelling Space and Time in Narratives about Restaurants. *Literary and Linguistic Computing*, 67–84.

Muslea, I. (1999). Extraction Patterns for Information Extraction Tasks: A Survey. In *Proceedings AAAI-99 Workshop on Machine Learning for Information Extraction* (pp. 46–53).

Nakasone, A., & Ishizuka, M. (2006). Storytelling Ontology Model Using RST. In *Proceedings of IEEE/WIC/ACM International Conference on Intelligent Agent Technology (IAT 2006)* (pp. 163–169).

Nastase, V., & Szpakowicz, S. (2003). Exploring noun-modifier semantic relations. In *Fifth International Workshop on Computational Semantics (IWCS-5)* (pp. 285–301).

Niles, I., & Pease, A. (2001). Towards a Standard Upper Ontology. In *Proceedings of the International Conference on Formal Ontology in Information Systems* (pp. 2–9).

Riedl, M., & Young, R. M. (2004). A Planning Approach to Story Generation for History Education. In *Proceedings of the 3rd International Conference on Narrative and Interactive Learning Environments.*

Rosario, B., & Hearst, M. (2001). Classifying the semantic relations in noun-compounds via a domain-specific lexical hierarchy. In *Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing (EMNLP-01)* (pp. 82–90).

Rosario, B., Hearst, M., & Fillmore, C. (2002). The descent of hierarchy, and selection in relational semantics. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL-02)* (pp. 417–424).

Singh, P., Lin, T., Mueller, E., Lim, G., Perkins, T., & Zhu, W. L. (2002). Open Mind Common Sense: Knowledge Acquisition from the General Public. In *Odbase02.*

Solis, C., Siy, J. T., Tabirao, E., & Ong, E. (2009). Planning Author and Character Goals for Story Generation. In *Proceedings of the NAACL Human Language Technology 2009 Workshop on Computational Approaches to Linguistic Creativity.*

Yates, A., Cafarella, M., Banko, M., Etzioni, O., Broadhead, M., & Soderland, S. (2007). Textrunner: Open information extraction on the web. In *NAACL '07: Proceedings of Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations on XX* (pp. 25–26). Morristown, NJ, USA: Association for Computational Linguistics.

Young, R. M. (2008). Computational Creativity in Narrative Generation: Utility and Novelty Based on Models on Story Comprehension. In *Proceedings of the Association for the Advancement of Artificial Intelligence 2008 Sprint Symposium.*