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

March 15, 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 children's stories, and store these in a knowledge base that Picture Books and other NLP applications can utilize to do their tasks.

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

Contents

List of Figures

List of Tables

1 Research Description

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. In order for computers to achieve a level of expressiveness same as humans and be able to understand the world that we talk about, they 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. 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 moral characters 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') was 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.

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.

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 tool 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. 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.

At least 20 new semantic relations, resulting from those identified using RST, from analyzing the sample corpus, and from reviewing other works such as Mueller for modeling time and event occurrences, may be created in this research. Such additional relations include Happens(e, t) which represents that a fluent f holds at time t, and Terminates(e, f, t) which represents that if event e occurs at t then fluent f stops holding after t.

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.

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 through 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:

- There is an increase in the number of story variants that are generated by Picture Books.
- 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 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.

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 convey-

ing information and experience to listeners (Nakasone & Ishizuka, 2006), thus justifying the selection of this domain for the proposed research.

Various applications can utilize the knowledge extracted by the system from existing stories. The major beneficiary will be story generation systems that require a large body of common sense knowledge to do their tasks. Story generation systems are currently gaining grounds in both education and entertainment. Various research projects have shown that using stories can stimulate learning, and virtual environments where the user's choices have an effect on the flow of the story also promotes creativity and encourage user participation.

2 Review of Related Literature

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. One of the earliest works on this would be that 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 supervised, knowledge-intensive 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 pre-determined 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.

3 Research Methodology

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

3.1 Software Development Process

A variant of the Waterfall model, the Modified Waterfall model, seen in Figure 3.1, will be followed. The Modified Waterfall model includes a feedback mechanism for evaluating and validating the output of each phase in the software development life cycle. Its overlapping stages facilitate the modification and improvement of any aspect of the system immediately, from requirements specification, architectural design and detailed design to implementation and testing. It ensures that the milestones have been accomplished, and it provides flexibility in updating the output of previous stages when necessary.

3.2 Software Concept

During this stage, the research topic is defined and conceptualized. Related systems and papers on related concepts will also be gathered and reviewed. Consultations with the thesis adviser will be conducted regularly to monitor the progress of the project, to evaluate the current state of the technology in the country, and to decide on project objectives and scope.

During consultations, comments and suggestions will be considered. The omissions indicated on submitted documents will be immediately reassessed and changed, while the suggested readings will be read. Also, similar systems will be reviewed to fully understand and familiarize the depth of the research topic. Further research on conceptual relations and the different types will be done.

3.3 Requirements Analysis

During this stage, data gathering and research will be performed to identify the following: types of conceptual relations, architecture of relation extraction systems, algorithms for extracting conceptual relations. Additionally, the input corpus consisting of at least 30 children's stories will be gathered. The suitable programming language for the project will also be identified. Furthermore, interviews with

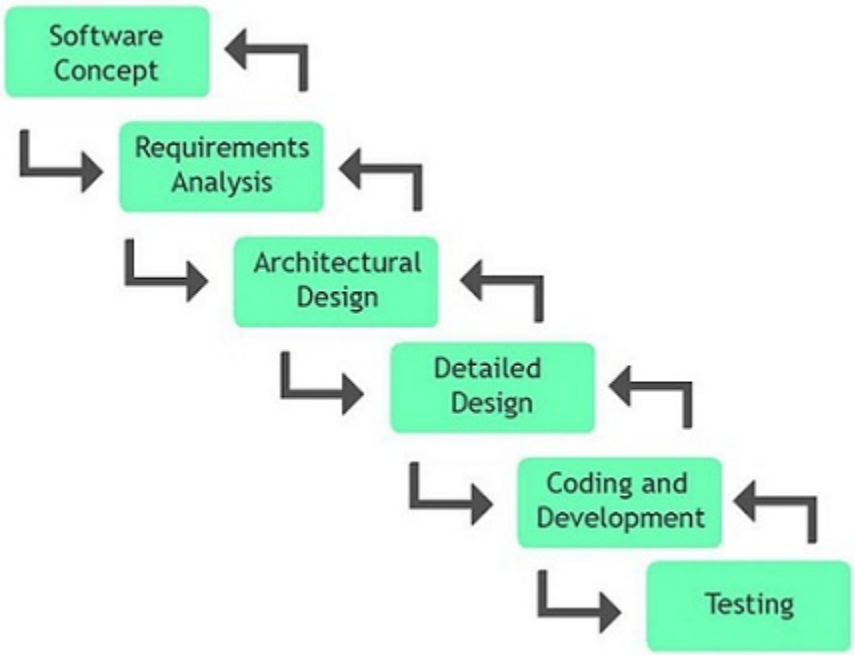

Figure 3.1: Modified Waterfall Model

Filipino language professors and linguists will be conducted.

After gathering the data, the requirements will be defined and analyzed to determine the objectives and scope of the system. The resulting requirements specification will be validated to ensure completeness of the software. Moreover, Use Case diagrams will be developed to help guide in creating the software by determining the appropriateness of the different functions in the system.

3.4 Architectural Design

In the Architectural Design stage, the overall function and subsystems will be identified. Also, relationships, dependencies, and interactions among the subsystems will be determined. The resulting architectural design will be reviewed to help understand the flow and design of the system more.

3.5 Detailed Design

At this stage, the classes and functions will be defined. The data structures and libraries to be used in the system will also be identified. The extraction templates

for the different conceptual relations will be done as well. Furthermore, algorithms for extracting conceptual relations will finally be formulated.

3.6 Coding and Development

The actual implementation of the software design will be done in the fifth stage, Coding and Development. Debugging and unit testing will be done regularly to ensure the efficiency and correctness of the software.

3.7 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 subsystem receives the correct input from the previous subsystem and generates the appropriate result for use by subsequent subsystems.

Test cases will be employed to check that all 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 PictureBooks. The generated story of PictureBooks after using the output semantic network will be evaluated. The output semantic network may also be evaluated by Filipino linguists to ensure their validity.

3.8 Documentation

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

3.9 Calendar of Activities

Table 3.1 shows a Gantt chart of the activities. Each bullet represents approximately one week worth of activity. As illustrated, there will be an overlapping of

activities to ensure that the Modified Waterfall model is observed. This is also to ensure that any omissions and modifications will be changed immediately. Moreover, the feedback mechanism of the said model allows easier evaluation of the system.

Table 3.1: Timetable of Activities

Activities (2010)	Jan	Feb	Mar	Apr	May	Jun	Jul
Software Concept	●●	●●●●	●●●●	●●●●			
Requirements Analysis			●●●	●●●●			
Architectural Design			●●	●●●●	●●		
Detailed Design				●●●	●●●●	●●	
Coding and Development					●●	●●●●	●●●●
Testing							●●●●
Documentation	●●	●●●●	●●●●	●●●●	●●●●	●●●●	●●●●

A Diagrams and Other Documentation Tools

An ontology is an artifact with a set of representational primitives to model knowledge for a particular domain. 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). Table A.1 lists some of these relations defined in Picture Books following the form $\langle relationship \rangle (\langle concept1 \rangle, \langle concept2 \rangle)$.

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 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.2.

Table A.1: ConceptNet semantic relationships (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)

Table A.2: 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.

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 "$\langle X \rangle isakindof \langle Y \rangle$", $thepossiblevaluesfor \langle X \rangle and \langle Y \rangle thatuserscanprovide$

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.*

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

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).

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.*

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).

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.*

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.