2014

Alternative Approaches to Correction of Malapropisms in AIML Based Conversational Agents

Walter A. Brock
Nova Southeastern University, wbrock@nova.edu

NSUWorks Citation

Walter A. Brock. 2014. *Alternative Approaches to Correction of Malapropisms in AIML Based Conversational Agents.* Doctoral dissertation. Nova Southeastern University. Retrieved from NSUWorks, Graduate School of Computer and Information Sciences. (20)
http://nsuworks.nova.edu/gscis_etd/20.

Alternative Approaches to Correction of Malapropisms in AIML Based Conversational Agents

By
Walter A Brock

A dissertation submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy
In
Information Systems

Graduate School of Computer and Information Sciences
Nova Southeastern University
2014

An Abstract of a Dissertation Submitted to Nova Southeastern University
in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy

Alternative Approaches to Correction of Malapropisms in AIML Based Conversational Agents

By
Walter A Brock
October 2014

The use of Conversational Agents (CAs) utilizing Artificial Intelligence Markup Language (AIML) has been studied in a number of disciplines. Previous research has shown a great deal of promise. It has also documented significant limitations in the abilities of these CAs. Many of these limitations are related specifically to the method employed by AIML to resolve ambiguities in the meaning and context of words. While methods exist to detect and correct common errors in spelling and grammar of sentences and queries submitted by a user, one class of input error that is particularly difficult to detect and correct is the malapropism. In this research a malapropism is defined a "verbal blunder in which one word is replaced by another similar in sound but different in meaning" (“malapropism,” 2013).

This research explored the use of alternative methods of correcting malapropisms in sentences input to AIML CAs using measures of Semantic Distance and tri-gram probabilities. Results of these alternate methods were compared against AIML CAs using only the Symbolic Reductions built into AIML.

This research found that the use of the two methodologies studied here did indeed lead to a small, but measurable improvement in the performance of the CA in terms of the appropriateness of its responses as classified by human judges. However, it was also noted that in a large number of cases, the CA simply ignored the existence of a malapropism altogether in formulating its responses. In most of these cases, the interpretation and response to the user’s input was of such a general nature that one might question the overall efficacy of the AIML engine. The answer to this question is a matter for further study.

Acknowledgements

As anyone who has ever attempted a doctoral dissertation will attest, this has been one of the most difficult challenges of my life. Frankly it could not have happened without the help and support of a great many people.

First among these must be my advisor Dr. Sumitra Mukherjee who provided guidance and excellent advice throughout this process, whose patience I am sure was tested to the limit, but most importantly, who believed I could do this!

I must thank my wonderful and amazing wife, Brenda, who was there every step of the way and who kept me on track even as I battled with cancer. No words here can express just how much I owe to her, but it is no exaggeration to say that this dissertation would not and could not have happened without her love and support.

My three children have been marvels through this entire process. They have been patient when Dad had to work on his paper. I must single out my son Aaron, whom God has gifted with an amazing mind. It was he who stepped into my study one day, after I had worked on a problem with an algorithm for two full days, looked over my shoulder and said “Hey Dad, don’t you see the pattern there?” He then proceeded to point out the set of relationships in the data, which had eluded me for what seemed like an eternity.

My thanks also to Dr. Amon Seagull and Dr. Michael Laszlo whose concise and insightful feedback was instrumental in making this research a success.

Special thanks must go to Ms. Amy Hickman and Mr. Benjamin Shorb who so graciously volunteered to act as judges and classified many hundreds of sentences as part of this research.

Finally to my brothers and “big sister” whose encouragement and prayers when I thought I might never complete this process was a huge boost.

Thank You All!

Table of Contents

Chapters

List of Tables

Tables

List of Figures

Figures

Chapter 1

Introduction

Background/Introduction

The automated processing of natural (i.e., human) languages by computers has been the subject of extensive and ongoing research for decades (Bitter, Elizondo, & Yang, 2009; Sebastiani, 2002). Within the field of Natural Language Processing (NLP), Conversational Agents (CAs), often referred to as chat bots or 'chatterbots', have seen considerable study. The application of CAs has been studied in a wide-ranging array of disciplines. Examples include psychological research, language and mathematics tutoring, public planning, economic policy, e-commerce, student advising, cultural heritage and even law enforcement (Augello, Pilato, & Gaglio, 2010; Boden, Fischer, Herbig, & Spierling, 2006; Ghose & Barua, 2013; Hossain, Rahman, Tran, & Saddik, 2010; Hubal et al., 2008; Lundqvist, Pursey, & Williams, 2013; Mascari et al., 2010; McMahan, 2010; Mikic Fonte, Rial, Juan, & Nistal, 2009; Morales-Rodríguez, González, Juárez, Huacuja, & Flores, 2010; Giovanni Pilato et al., 2004; Shawar & Atwell, 2007; Soliman & Guetl, 2013).

Problem Statement

One method for the creation of CAs, which has seen extensive use over the past decade, is Artificial Intelligence Markup Language (AIML) (Wallace, 2009). In fact, AIML has achieved several successes for use in the creation of domain specific CAs.

However, this potential is limited by several factors inherent in the structure of AIML. Specifically:

1. AIML based CAs are easily led off topic.
2. They have little or no memory of a conversation.
3. They are very poor at discerning the emotional intent of the user.
4. They often repeat themselves when given undifferentiated (highly similar) input or ambiguous input queries.
5. They often have difficulty deriving the context of a conversation.

(Augello, Pilato, Vassallo, & Gaglio, 2009; Neves, Barros, & Hodges, 2006; Schumaker & Chen, 2010).

Some of these limitations relate specifically to the method employed by AIML to resolve ambiguities in the meaning and context of words. Several methods exist to detect and correct common errors in spelling and grammar of sentences and queries submitted by a user (Sebastiani, 2002). One class of input error that is particularly difficult to detect and correct is the malapropism (Bolshakov & Gelbukh, 2003). Early research into this area, such as that by Hirst & St-Onge (1998) and Budanitsky & Hirst (2001) used a relatively loose definition of malapropism that included common errors in spelling and even some grammatical errors. Bolshakov & Gelbukh (2003) use a much more precise definition of the word malapropism as a "verbal blunder in which one word is replaced by another similar in sound but different in meaning" (“malapropism,” 2013). They then point out that this form of error is particularly difficult to detect and correct. It is this more narrowly focused definition that will be used here.

The disambiguation method employed by AIML is known as the Symbolic Reduction, denoted in the markup language by the SRAI tag. A search of the literature has produced no other studies similar to this one. This research sought to determine the effectiveness of AIMLs Symbolic Reduction at detecting and correcting malapropisms by comparing the performance on an AIML CA using Symbolic Reductions with and without the aid of two other approaches whose performance has been well studied and documented. The first of these approaches involved using a measure of semantic distance to detect the malapropism (Chiru, Cojocaru, Rebedea, & Trausan-Matu, 2010) and the second utilized an n-gram approach (Wilcox-O'Hearn, Hirst, & Budanitsky, 2008).

To understand this, a more thorough explanation of the structure and operation of AIML is needed. AIML is essentially an implementation of Case Based Reasoning (CBR) (Breese & Heckerman, 1996; Kolodner, 1992). AIML stores its knowledge (the cases) as a series of 'categories'. Each category in turn contains a 'pattern' and a 'template'. The AIML engine will read this knowledge base and construct "a memory-resident directed graph of all the AIML patterns" using a component called the Node Mapper (Schumaker & Chen, 2010). When an AIML CA recognizes an input pattern, it will select a response from the corresponding template using the Node Mapper, a component of the Graph Master. Figure 1 illustrates the basic architecture of an AIML engine (Freese, 2007; Schumaker & Chen, 2010; Wallace, 2003).

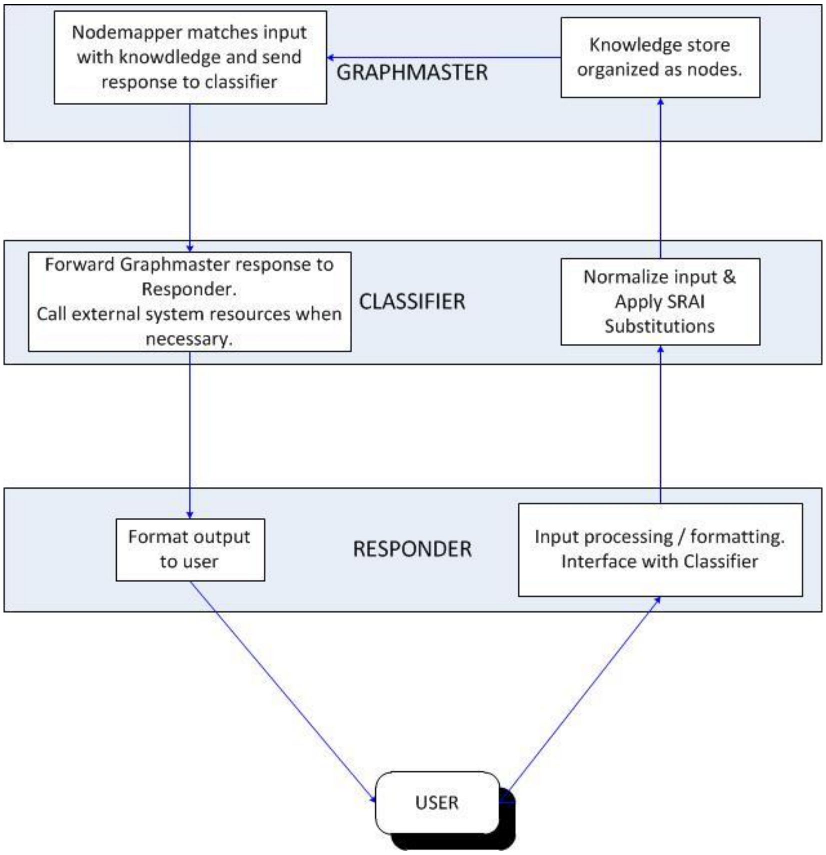
Nodemapper matches input with knowdledge and send response to classifier
GRAPHMASTER
Knowledge store organized as nodes.
Forward Graphmaster response to Responder.
Call external system resources when necessary.
CLASSIFIER
Normalize input & Apply SRAI Substitutions
Format output to user
RESPONDER
Input processing / formatting. Interface with Classifier
USER

Figure 1: Basic AIML Architecture

The responder receives input from the user and formats output to display to the user. The classifier first performs a series of normalization steps. For instance, punctuation is removed and all characters are converted to upper case. Disambiguation is then performed by processing the input through a series of specialized AIML categories known as symbolic (or safe) reductions (denoted using the <SRAI> tag in AIML) (Freese, 2007; Pothuru, 2003; Schumaker & Chen, 2010; Wallace, 2003).

To illustrate the operation of a symbolic reduction consider the following AIML category excerpted from the current default AIML knowledge set:

```
<category>
   <pattern>ON WHAT OCCASION *</pattern>
   <template><srai> when <star /></srai></template>
</category>
```

Figure 2: AIML Code Snippet Retrieved from http://code.google.com/p/aiml-en-us-foundation-alice/downloads/detail?name=aiml-en-us-foundation-alice.snapshot.zip

Let us assume the user has entered the sentence “On what occasions do you celebrate?” After normalization, the sentence would be rendered as “ON WHAT OCCASION DO YOU CELEBRATE”. Note that the punctuation has been removed and the plural form of occasions has been changed to the singular. The change from plural to singular would have been handled by a separate category. After application of the above reduction, the sentence will now be rendered as “WHEN DO YOU CELEBRATE”. It is this simplified phrase that will now be passed to the Graph Master for matching (Freese, 2007; Schumaker & Chen, 2010; Wallace, 2003).

Symbolic Reductions can be called recursively in order to reduce complex input queries to a series of phrases that can then be recognized by the Graph Master. Symbolic Reductions can also perform limited word substitutions for synonyms and common spelling errors. However, these reductions are limited only to those that have been specifically programmed into the AIML knowledge set. The programming of AIML categories is a very labor intensive activity. Improvements and corrections require manual intervention. Symbolic Reductions can only match those patterns that have been anticipated by the writers. Any input not correctly interpreted by the reductions will result in questionable and even nonsensical replies. These problems suggest the use of alternative methods for disambiguating the input to the CA, specifically as it applies to the correction of malapropisms (Freese, 2007; Schumaker & Chen, 2010; Wallace, 2003).

After classification, the input is then passed to the Graph Master, which will use a somewhat modified depth-first search algorithm to find a match between the input and the category patterns. This search is simplified by the use of simple wild cards in the pattern syntax. It is further simplified by the observation that despite the fact that there may be several tens of thousands of categories to search, there are only around two thousand words that might be found in the first position of a phrase. For this reason the patterns are indexed according to their first word and the search only descends into the patterns themselves once the first word has been identified (Freese, 2007; Schumaker & Chen, 2010; Wallace, 2003).

For every possible malapropism, syntax error, or error in grammar, the programmer (or “bot master”) would have to create a specific AIML category to correct it. For malapropisms, the task becomes even more difficult as the bot master must anticipate the exact sentence and use in order to catch the error. Figure 3 shows an AIML category for the sentence “The flood damage was so bad they had to evaporate the city” (“What Are Malapropisms,” n.d.). The use of the word “evaporate” where “evacuate” should have been used is a classic example of a malapropism.

```
<category>
  <pattern>* they had to evaporate the *<pattern>
  <template>Perhaps you meant “evacuate”?</template>
  <srai>THEY HAD TO EVACUATE<star/></srai>
</category>
```

Figure 3: AIML Code Snippet for a malapropism

The result of course, is an AIML CA that would require an enormously large, manually created set of reductions in order to be able to carry on even a simple conversation. Wallace contends that a sufficiently large AIML set (created by the community at large) would contain enough such categories to handle nearly all of the input possibilities it might encounter (Wallace, 2003). The fact that such an AIML set does not (yet) exist after more than a decade of effort gives a hint at the scope of the effort involved in this undertaking (Graesser, 2011; Lundqvist et al., 2013; Schumaker & Chen, 2010). It is reasonable then to suspect that an alternate approach to the disambiguation function may be more satisfactory.

Dissertation Goal

While the AIML safe reductions are elegant in their simplicity, in reality it is difficult to implement a set of safe reductions large enough to be completely effective. However, in recent years there has been considerable research into alternative approaches to disambiguation using a machine learning approach.

In order to address some of the limitations concerning disambiguation in regards to the recognition of malapropisms within AIML enumerated above, this research evaluated the use of two different alternative approaches. As illustrated in Figure 4 below, this involved the insertion of an additional processing step between the AIML responder and classifier. The purpose of this new step (here called the *pre-classifier*) was to detect and correct malapropisms before the input was sent to the classifier.

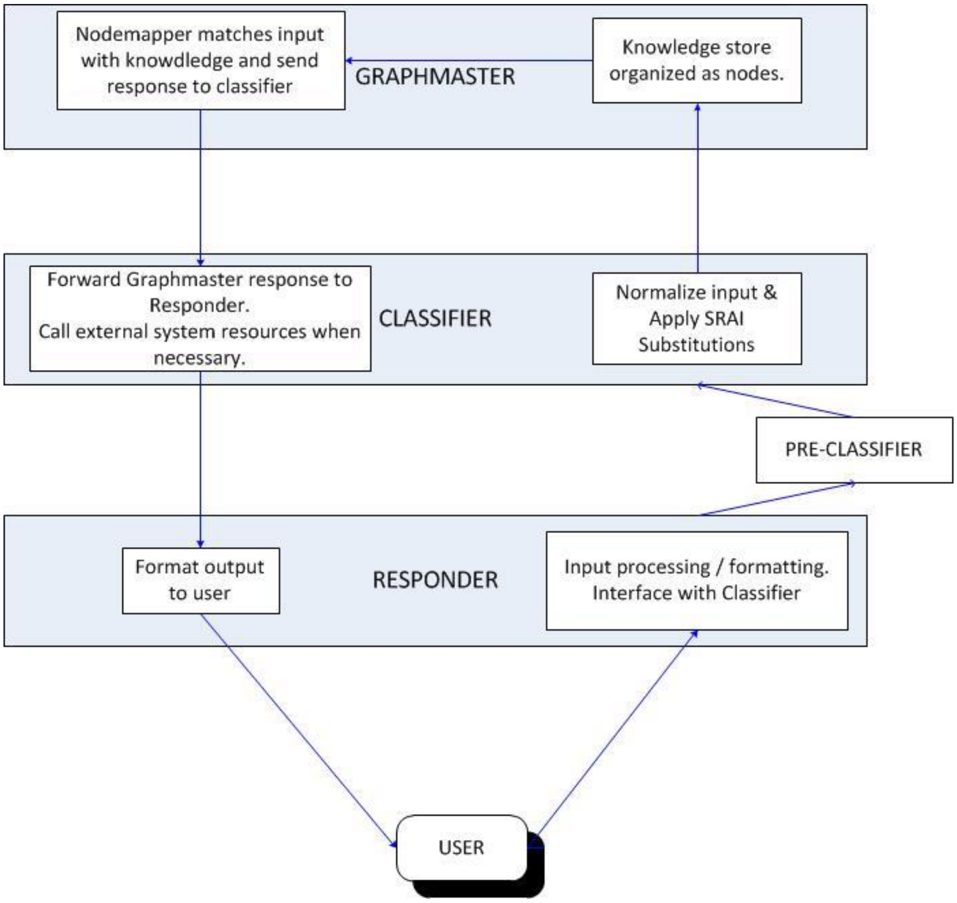
Nodemapper matches input with knowdledge and send response to classifier
GRAPHMASTER
Knowledge store organized as nodes.
Forward Graphmaster response to Responder.
Call external system resources when necessary.
CLASSIFIER
Normalize input & Apply SRAI Substitutions
PRE-CLASSIFIER
Format output to user
RESPONDER
Input processing / formatting.
Interface with Classifier
USER

Figure 4: AIML architecture modified with a Pre-Classifier

The first of these two approaches implements a measure of semantic distance utilizing WordNet in a manner similar to the lexical cohesion algorithm (Budanitsky & Hirst, 2006). The fundamental premise here is that words in a coherent sentence will all be closely related in terms of their semantic distance. However, the existence of malapropisms within a sentence would interrupt the cohesion of the sentence and this interruption should be detectable as an increase in semantic distance.

The second approach is based on a tri-gram model (Wilcox-O'Hearn et al., 2008). In this approach, the input sentence is broken down into a series of tri-grams, which are then compared to a weighted tri-gram corpus (in this case the Microsoft N-gram web service). Tri-grams with probabilities lying outside a predetermined cutoff value are considered candidates for correction. The correction candidate tri-gram is assumed to contain the suspected malapropism. Several researchers, most notably Islam & Inkpen, have reported very good results with this approach (Bolshakov, 2005; Chiru et al., 2010; Islam & Inkpen, 2009).

Candidate malapropisms are then corrected by selecting an entry from a paronyms dictionary (Chiru et al., 2010). For this research, paronym is defined as "words similar to each other in letters, sounds or morphs" (Bolshakov & Gelbukh, 2003). This definition differs from common use in that paronyms, as defined here, might or might not have a similar meaning to the original word. The search of the paronyms dictionary is then augmented by a second search looking for words similar in spelling and number of syllables. The candidate word, which gives either the lowest semantic distance, in the semantic distance model, or the best probability, in the case of the tri-gram model, is then selected as the replacement word (Hirst & Budanitsky, 2005; Mihalcea, Corley, & Strapparava, 2005; Mihalcea & Csomai, 2005; Pedler, 2007).

Research Questions

Given the problem as stated and that the effectiveness of the AIML Symbolic Reductions have never been studied, this research has attempted to answer the following:

1. Will the use of a semantic distance based disambiguation algorithm, using WordNet, in addition to the symbolic reduction approach currently used in AIML lead to better performance, as measured by accuracy and precision, on

a specific set of input queries containing malapropisms as compared to AIML Symbolic Reductions alone?

2. Will the use of a tri-gram based disambiguation algorithm, in addition to the symbolic reduction approach currently used in AIML lead to better performance, as measured by accuracy and precision, on a specific set of input queries containing malapropisms as compared to AIML Symbolic Reductions alone?

Relevance and Significance

Conversational agents have been extensively researched for decades. One of the earliest conversational agents to receive wide attention was ELIZA. Though deceptively simple in its programming, ELIZA was able to effectively mimic a therapist talking to users about their problems. In reality, ELIZA used two simple ideas to accomplish this. It looked for keywords and also used portions of input sentences and its responses to users. Despite this simplicity, the convincing nature of ELIZA's conversations convinced many that truly intelligent computer systems would be available in less than a generation (Norvig & Russel, 2010; Weizenbaum, 1966, 1967).

In recent years, researchers in many disciplines have studied the use of CAs in a multitude of applications. Perhaps the most thoroughly explored application of CAs is as a virtual tutor. The concept of an artificially intelligent computer system able to help teach and guide students is compelling. Arthur Graesser and his colleagues have reported on successes with AutoTutor, a CA targeted specifically at tutoring college students in computer hardware and operating systems. Yet as recently as 2011, Graesser has reported that AutoTutor's results have been limited by its inability to effectively deal with the ambiguities of language (Graesser & Jackson, 2006; Graesser, Wiemer-Hastings, Wiemer-Hastings, & Kreuz, 1999; Graesser, 2011). Kerly and her colleagues looked at several applications of CAs in e-Learning, concluding that their use would continue to grow as the technology improved (A Kerly, Hall, & Bull, 2007; Alice Kerly, Ellis, & Bull, 2009). Hubal, et al. looked at the use of CA tutors both for prisoners and teenagers, reporting similar results (Hubal et al., 2008).

AIML based CAs have also been studied in the tutoring role. For instance, T-BOT, Q-BOT and TQ-BOT were designed to be integrated into learning management systems, such as Moodle to assist and track student progress (Mikic, Burguillo, Rodriguez, Rodriguez, & Llamas, 2008; Mikic Fonte et al., 2009). The use of CAs as teaching aids in clinical psychology and medicine has also seen a great deal of attention (Dickerson, Johnsen, Raij, & Lok, 2005; Heller, Procter, & Mah, 2005; Morales-Rodríguez et al., 2010; Veletsianos, Heller, Overmyer, & Procter, 2010). An interesting application of Knowledge Space Theory to a CA tutor was done by Pilato, Pirrone & Risso (2008). Their results with a domain specific tutoring CAs are very promising.

Another area of investigation has been the use of CAs as an instrument to disseminate information and enhance communication in a crisis situation, especially in regards to terrorism (Reid, Qin, Chung, & Xu, 2004; R. Schumaker, Ginsburg, Chen, & Liu, 2007; Schumaker & Chen, 2007, 2010). While all of the researchers cited above reported varying degrees of success, they all shared one thing in common in their conclusions. The CAs they employed required further research to address limitations in language, especially in regards to ambiguities.

Barriers and Issues

To the author's knowledge, no previous attempt to study the effectiveness of the AIML symbolic reductions or the efficacy of enhancing them with alternate algorithms can be found in the literature. The previous implementations of the algorithms under study here do exist. This is essentially a new work that required re-implementing small portions of the AIML engine.

An extensive search of extant corpora determined that no corpus specifically created to study the problem of malapropisms exists today. In order to facilitate this research, a corpus consisting of 317 sentences containing malapropisms was created.

Assumptions

As no studies of the Symbolic Reduction functions in AIML have been found in the literature, the assumption is made that the SRAI function is at least somewhat effective. Well over a decade of use by the AIML community as a whole provides the basis for this assumption.

Definition of Terms

Several of the acronyms used within this document are defined here. They are further defined, along with other important terms, with the narrative of this document.

AIML:	Artificial Intelligence Markup Language
CA:	Conversational Agent
ES:	Error Score
SRAI:	Symbolic Reductions in Artificial Intelligence [markup language]

Chapter 2

Review of the Literature

Foundational Material

Even before Tim Berners-Lee, Hendler & Lassilo's (2001) seminal article, The Semantic Web, research into Natural Language Processing (NLP) was extensive. Numerous approaches to the problem have been explored with varying degrees of success (Sebastiani, 2002). Research exploring the creation of domain specific ontologies, the use of Web Ontology Language (OWL) and Resource Description Framework (RDF) is extensive. However, such structured ontologies can be extremely difficult to build and, like AIML, have great difficulties dealing with the ambiguities inherent in natural language (De Maio, Fenza, Loia, & Senatore, 2010).

Lee, Jian & Huang (2005) attempted to address the problem of understanding ambiguous language through the application of fuzzy logic. Their approach started from a set of traditional or 'crisp' domain ontologies created by domain experts. These ontologies covered Chinese parts of speech as well as terms and concepts related to Chinese weather reporting. Using a software stack consisting of five distinct layers they first parsed the inputs taken from Chinese weather reports. They then extracted meaningful terms using Chinese Parts of Speech (POS) and domain specific dictionaries. From there, they derived the fuzzy relationships between the extracted terms by calculating membership probabilities (the probability that the terms were related) based on word similarity, parts of speech similarity and semantic distance. The calculated probabilities were then used in a sentence generator to build summaries of the weather reports.

Conversational Agents

Early work in Conversational Agents

The study of Conversational Agents, also known as ‘chatbots’ or ‘chatterbots’, capable of engaging in real-time, natural language discussion with human beings has been a goal for as long as electronic computers have existed. Alan Turing first described the now famous ‘Turing test’ in 1950. The purpose of the test was to determine if a computer had developed the ability to ‘think’ as humans do. Debate has raged to this day on exactly how to define what it means for a computer to ‘think’, but the debate tends to follow two broad philosophical lines. The first line are the behaviorists, who believe that if a computer can convincingly behave as a human would (at least in terms of language and conversation) than the computer system will have met the goal. The second line of thought is the idea that to be ‘intelligent’ a computer must be able to learn to reason and demonstrate an understanding of its environment, much the way human beings do (Deryugina, 2010; Norvig & Russel, 2010).

The first CA to become widely popular was ELIZA, developed by Joseph Weizenbaum in 1966. Eliza was deceptively simple in its construction. The program would parse input sentences looking for keywords. It would then insert these keywords into a series of programmed sentences, which it would use to elicit further responses from its human conversant. The dialogue was programmed to mimic the types of questions a psychotherapist might be expected to ask. ELIZA was deceptive in its simplicity and its ability to fool uninitiated users into thinking they were conversing with a real therapist. This led many to believe that the behaviorist's approach might turn out to be the most productive avenue of research (Deryugina, 2010; Norvig & Russel, 2010; Weizenbaum, 1966, 1967).

Since ELIZA, the development of CAs has occurred apace with Natural Language Processing in general and CAs have become progressively more and more sophisticated. Even so, more than 60 years after Alan Turing first proposed his Turing Test, only one CA has ever passed a Turing test. A CA known as "Eugene Goostman" passed the Turing Test at the *Turing Test 2014* competition held at the University of Reading in England (McCormick, 2014; "Turing Test Success Marks Milestone in Computing History," 2014). Almost immediately following the publication of this claim there was a firestorm of criticism. An article posted on IEEE Spectrum summed it up this way (Ackerman, 2014):

> "Almost immediately, it became obvious that rather than proving that a piece of software had achieved human-level intelligence, all that this particular competition had shown was that a piece of software had gotten fairly adept at fooling humans into thinking that they were talking to another human, which is very different from a measure of the ability to 'think.' "

Though the Turing test is not universally accepted as the best method to judge whether or not a computer system can behave like a human, it remains the most commonly used test. The most public form of the Turing test is the annual Loebner prize competition, held every year since 1990. The Loebner Gold Medal and a $100,000 cash prize, which is to be awarded to the first 'chat bot' to pass the Turing test, remains unclaimed (Deryugina, 2010; Wallace, 2009; "What is the Loebner Prize?," 2013).

AIML Based Conversational Agents

As noted in Chapter 1, AIML is a popular approach to the challenge of creating a useful CA, which attempts to extract semantic and contextual meaning from natural language through the use of structured knowledge. This approach is very similar in concept to ontology and uses an XML format somewhat similar to RDF. However, AIML is less rigorous in its structure than RDF. It is intended to be more easily accessible to non-expert users and is easier to process by low cost computer systems. However, as previously noted, creating large knowledge bases in AIML is still a daunting task. AIML knowledge bases are generally built by hand and AIML suffers from the same difficulties interpreting ambiguities that RDF and OWL based knowledge bases do. The original AIML CA, called ALICE for Artificial Linguistic Internet Computer Entity won the Loebner prize competition three times in 2000, 2001 and 2004. However, despite its success, AIML (and ALICE) is not without its limitations (Neves et al., 2006; Schumaker & Chen, 2010; Wallace, 2003, 2009; "What is the Loebner Prize?," 2013).

There have been numerous attempts to address the limitations inherent in AIML. Research in this area is ongoing. Neves, Barros & Hodges (2006) proposed iAIML, which extends AIML in two ways. First, they added rule sets based on Conversational Analysis Theory (CAT). The intent of these rule sets was to deduce the intention of the user interacting with the system. Second, they attempted to address the difficulty in creating and maintaining large AIML sets (knowledge bases) by incorporating a more rigorously defined structure in AIML itself. The use of CAT as a basis for the creation of AIML rules is interesting, but the general AIML community has not adopted their work. Freese (2007) demonstrated a successful method of translating ontologies expressed in

RDF directly into AIML in an effort to automate the creation of domain specific AIML sets.

Schumaker, et al., (2007) explored the use of a dialogue system using a CA for knowledge acquisition. In their system, the user interacts with the chat bot in order to teach it knowledge about a specific domain. The user can then review and edit the transcripts of this session in order to correct errors made by the chat bot in parsing this knowledge. Their results indicated that their system was better than general conversation in capturing domain knowledge. Cho & Chun (2007) examined a method of detecting the emotional intent of the user by analyzing certain keywords and how they were used in the conversation.

Perhaps the greatest body of work on AIML CAs belongs to Augello, Pilato and their colleagues who have sought to address the limitations in AIML (and extend its functionality) by adding what may be termed a *higher reasoning* function or what they have termed “conceptual similarity relationship layers” to AIML. They attempt to discern a semantic relationship between user input and semi-structured data (such as Wikipedia), which would then drive responses from the chat bot. Over the years they have experimented with a host of other knowledge sources such as CyC or lexicons, such as WordNet. However, these efforts have failed to make effective use of these resources other than to quote facts related to user input statements or to substitute words in the response which could be done by referencing a thesaurus (Agostaro, Augello, Pilato, Vassallo, & Gaglio, 2005; Augello et al., 2009; Augello, Scriminaci, Gaglio, & Pilato, 2011; Augello, Vassallo, Gaglio, & Pilato, 2008; G. Pilato, Augello, Vassallo, & Gaglio, 2007; Giovanni Pilato, Augello, & Gaglio, 2011; Giovanni Pilato, Augello, Vassallo, & Gaglio, 2008; Pirrone, Pilato, Rizzo, & Russo, 2007).

At their core, all of the CAs developed along this line of research share a common architecture (See Figure 5 below). The knowledge source under study is used to extend the knowledge base of the standard AIML Graph Master in a way that will allow the CA to retrieve more data relevant to the input query. Unfortunately, this architecture does not give the CA any greater ability to deal with ambiguities in the input query. It can be seen from Figure 5, that while intriguing; their approach differs significantly from the approach under study here (Augello et al., 2009; Coursey, 2004; G. Pilato et al., 2007; R. Schumaker et al., 2007).

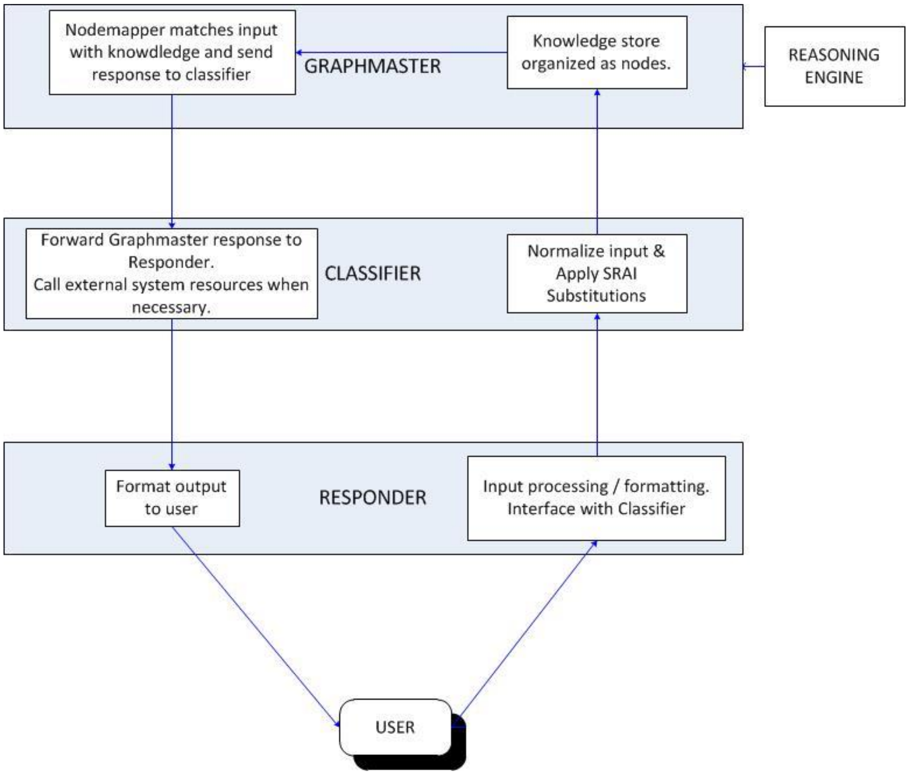
Nodemapper matches input with knowdledge and send response to classifier
GRAPHMASTER
Knowledge store organized as nodes.
REASONING ENGINE
Forward Graphmaster response to Responder. Call external system resources when necessary.
CLASSIFIER
Normalize input & Apply SRAI Substitutions
Format output to user
RESPONDER
Input processing / formatting. Interface with Classifier
USER

Figure 5: AIML architecture extended with a reasoning engine

In their study, Schumaker & Chen (2010) noted several specific strengths and weaknesses of conversational agents. Specifically, they noted that combining both conversational and domain specific knowledge bases led to an increase in user satisfaction. They also corroborated earlier studies showing that users tend to use chat bots as knowledge retrieval tools in similar fashion to search engines (Moore & Gibbs, 2002). Their findings also indicated that shorter input queries led to more satisfactory results and that longer inputs often led to off-topic or poorly related responses. This indicates that as the query length increases, the agents are less able to deal with ambiguity and subsequently lose the context of the conversation. Closely related to this conclusion are their findings concerning question types. Questions of type 'What are', 'What do' and 'Are' rated very high in user satisfaction while questions of the type 'Why' and 'What' scored the lowest (Schumaker & Chen, 2010).

Word Sense Disambiguation & Malapropisms

Semantic Distance

Numerous algorithms and methodologies have been proposed to perform word sense disambiguation and to identify malapropisms. The majority of these algorithms depend on some form of measure of the relationship between words or *semantic distance*. There is little agreement on a precise definition of *semantic distance*. This is due in part to the fact that the definition itself is partially derived from the approach taken in making the measurement. Technically it is the inverse of semantic relatedness, a measure of the similarity of the meaning between two words or phrases (Budanitsky & Hirst, 2006; Goranson, 2005; Szarvas, Vincze, Farkas, Móra, & Gurevych, 2012). It has become common place in the literature to use the term *semantic distance* as a generic moniker referring to both *semantic distance* and *semantic relatedness* (or similarity). Semantic distance and semantic relatedness are not actually the same. Algorithms that represent the relationship between two words as a progression through a graph, or tree, are actually measures of *semantic distance*. Algorithms representing the relationship between two words through comparison of actual information content such as dictionary definitions are more precisely referred to as measures of *semantic relatedness*. The phrase *semantic distance* will be used throughout the remainder of this document.

The idea that the 'conceptual distance' (in terms of their meaning) between two terms could be measured as the distance between these terms on a hierarchical graph was apparently first proposed by Rada & Bicknell in their work using MeSH (Medical Subject Headings) (Rada, Mili, Bicknell, & Blettner, 1989).

Figure 6 on page 21 is a conceptual view of how semantic distance might be measured. The figure contains a small subset of ‘synset’ graph taken from WordNet. The graph attempts to visualize the relationship between the words ‘red’ and ‘mountain’ and ‘red’ and ‘naturally’ as used in the following 2 sentences:

> The light from the storm painted the mountain red.
>
> Because of its iron content, the earth in the mountain was naturally red.

The first path (the dashed lines) shows 10 steps from ‘red’ to ‘mountain’, while the second path (the solid lines) shows 9 steps from ‘red’ to ‘nature’. This would indicate that the word ‘red’ is somewhat more closely related (semantically) to ‘nature’ than to ‘mountain’. In reality, it is much more difficult to measure semantic distance. The simplified view in Figure 6 does not include the myriad of interconnections between the thousands of words in WordNet or the many variations of meaning common in English words. This has led to significant work attempting to improve measures of semantic distance (Budanitsky & Hirst, 2001; Fellbaum, 1998; Jurafsky & Martin, 2000; Pedersen & Michelizzi, 1998).

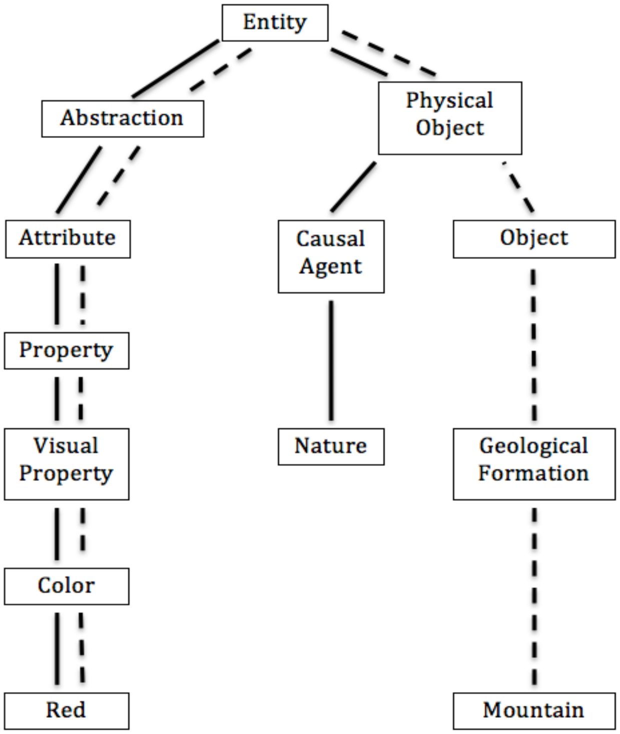

Figure 6: Conceptual View of Semantic Distance

Recent research in this area has centered on a group of six algorithms that have demonstrated effectiveness in prior research (Budanitsky & Hirst, 2006; Hessami, Mahmoudi, & Jadidinejad, 2011; Navigli & Lapata, 2010; Sinha & Mihalcea, 2007).

Three of these algorithms, Resnik (Resnik, 1995), Lin (Lin, 1989) and Jiang & Conrath (Jiang & Conrath, 1997) require a large corpus, specific to the type of text under study, from which statistical data can be extracted to derive the Information Content (IC) values. Of these three, Budanitsky & Hirst (2006) demonstrated superior results with

Jiang & Conrath, indicating that there are considerable further opportunities for study. However, as no such corpus specific to the CA under study here exists, these three algorithms will not be considered for this research.

Another two of these algorithms utilize a sense tagged tree graph (most often derived from WordNet) to calculate the semantic distance between pairs of words. Those words located more closely together are then considered semantically related. Leacock & Chodorow (1998) (LeC) proposed a similarity measure between two words, *a* and *b*, calculated as:

$$Similarity = \frac{-log(length(a,b))}{2} * D \tag{1}$$

Here, *D* is the maximum depth of the tree and *length* is the node count along the graph from *a* to *b*.

Wu & Palmer (1994) (WuP) proposed that semantic relation (as viewed on a WordNet tree graph) for two words, *a* and *b*, could be calculated as:

$$\frac{2 * depth(LCS(a,b))}{depth(a) + depth(b)} \tag{2}$$

Here LCS is the Least Common Subsumer, that is, the lowest common node in the tree graph.

Lesk (1986) proposed a measure of semantic relatedness expressed as a measure of the overlap of the dictionary definitions of two given words.

A similar approach to Lesk is the Gloss Vector. However, “Gloss Vector measure is based on second order co–occurrences” (Patwardhan & Pedersen, 2006). That is to say, the Gloss Vector measure will look beyond just the overlap in the dictionary definition by looking at overlap with closely related words as well (specifically using WordNet).

Though there is some variation between implementations, the basic steps for performing disambiguation using any of the algorithms described above can be summarized as:

1. *Tokenization* – Essentially this involves separating the input text into individual words, punctuation, numbers, etc.
2. *Word stemming* – Reducing words to their root form (and possibly separating compound words).
3. *Parts of speech tagging* – Identifying the parts of speech; noun, verb, adjective, adverb.
4. *Word sense disambiguation.* - Utilize sentence context, parts of speech, etc. to determine word sense.
5. *Calculate Semantic similarity (or distance)* – Utilizing one of the algorithms defined above.
6. *Calculate the similarity score* – A normalized total of the similarity scores of all of the word pairs in the input (Augello et al., 2009; Budanitsky & Hirst, 2006; Mihalcea et al., 2005; Mihalcea & Csomai, 2005; Pedler, 2007).

Hirst and St-Onge, (1998), identified the detection and correction of malapropisms as a separate and more difficult task from the more general research being done on the detection of spelling and grammar. They proposed that a phrase or sentence could be looked at as part of a ‘lexical chain’. This lexical chain has a property called cohesion, which is defined by the semantic distance between the words in the chain. WordNet was used in their measure of semantic distance. They proposed that malapropisms would be detectable as interruptions to the cohesion of the lexical chain or, in other words, an increase in the measure of semantic distance in comparison to other words in the chain. Their experiments met with modest success in detecting and correcting malapropisms. Hirst and Budanitsky (2005) reported incremental improvements using refinements on this approach.

N-Gram Probabilities

A completely different approach to the detection of malapropisms involves the use of n-gram distributions and their probabilities. Of the various n-gram candidates, the tri-gram seems to have garnered the most success. Mays, Damerau, and Mercer (1991) did some of the first experiments using tri-grams. Their results met with only limited success based in large part to the limited size of the corpora available to them at the time. Wilcox-O'Hearn, Hirst, and Budanitsky (2008) revisited this line of research comparing it to their own previous work on lexical cohesion and found that the tri-gram model using a much larger corpus was able to outperform their own previous research.

Islam and Inkpen (2009) achieved even better results (though their work was not strictly limited to malapropisms) with the tri-gram model by using Google's Web-1T corpus. This corpus is based on a set of over one trillion words, representing Google searches collected over a one-month period in January of 2006. Chiru, et al. (2010) took a different approach. In their approach lexical cohesion was used to identify malapropisms. Correction candidates were identified using a paronyms dictionary and a live Google search. Candidates that achieved a high probability for lexical cohesion were considered correct.

Figure 7 illustrates how n-gram probabilities (in this case, tri-grams) can be used to detect a malapropism. The sentence containing a malapropism is broken down into groups of three words. Sentence markers ("<s> and </s>") are used to complete the groupings for words at the beginning and end of each sentence. Note that the use of sentence markers actually results in bi-grams at the beginning and end of each sentence. Then the probabilities for each group are calculated. In the example in Figure 7, the conditional probabilities were obtained using the Microsoft N-Gram corpus available from Microsoft Research. The relative probabilities are shown for simplicity. In this case, the higher the score (as depicted in the histogram), the more likely it is that the tri-gram contains an error. This "Error Score" is actually the probability as expressed as a power of ten. Thus, the higher the Error Score, the lower the probability that the tri-gram will appear in the corpus. In this case, the tri-gram phrase 'had to evaporate' exceeds the predetermined error threshold and is considered to be a candidate for a malapropism (Manning & Schutze, 1999; Wang, Thrasher, & Viegas, 2010).

```
TRI-GRAMS                                            ERROR SCORE

<s> They                                                 ■■■■■■■
<s> They had                                                ■■■■
    They had to                                               ■■
         had to evaporate                           ■■■■■■■■■■■■
             to evaporate the                                  ■
                evaporate the city                      ■■■■■■■■
                          the city </s>                  ■■■■■■■
                              city </s>                  ■■■■■■■
```

Figure 7: Tri-Grams with their probabilities

Bolshakov and Gelbukh (2003) have also achieved significant results using a somewhat similar approach to the tri-gram. However, rather than calculating the probabilities that words would appear together as a three word group they instead used co-locations. Co-locations work under the assumption that semantically related words will appear physically close to each other and that given a sufficiently large corpus the probabilities of these co-locations can be calculated with sufficient accuracy to be useful in identifying correction candidates.

Chapter 3

Methodology

Introduction

This chapter will discuss the specific methods and algorithms that were used to create the input data, test the conversational agent and create the two versions of the pre-classifier, which were used to enhance the conversational agent. The first of the pre-classifiers utilizes a semantic distance algorithm for malapropism detection. The second pre-classifier uses a tri-gram search algorithm for malapropism detection. The methods used by both pre-classifier implementations to correct the malapropism will then be discussed. The procedures for processing the data and analyzing the results will then be given. This chapter will end with a summary.

Preparing the Input Data

Malapropism Corpus

In their search for a suitable corpus for their work on malapropisms, Budanitsky & Hirst (2005) made the observation "no such corpus of naturally occurring malapropisms exists". A review of the literature since 2005 has shown that no such corpus has yet been created. Thus, the creation of a corpus of naturally occurring malapropisms was necessary. Several sources for sentences containing malapropisms have been identified. These consist mostly of books and articles that cite malapropisms found in English literature, pop culture, and broadcast media. While most of these tend to be humorous in nature, the majority of them do represent genuine malapropisms as previously defined. A corpus of 317 individual sentences containing malapropisms appropriate for this study was created from these sources. In most cases, the sources also provided the correct words. In the few cases where correct words were not provided by the sources, the author selected a correct word. The judges then verified the selections. A small computer program was used to select 25 sentences at random from the overall corpus to serve as training data. The other 292 sentences comprised the actual test corpus.

In order to facilitate its use with the CA to be studied, the following format for the malapropism corpus was used:

```
[Sentence with malapropism] [Corrected sentence] [Bad word, corrected word]
```

This format was chosen to more readily facilitate processing of the corpus programmatically for both input to the CA and for evaluation of the results. Since the pre-classifiers used in this study were designed to deal with sentences containing only a single malapropism. Sentences containing two or more malapropisms were split into separate sentences, each containing a single malapropism. An attempt was made to exclude any overtly offensive malapropisms from the corpus (Baisely, 2000; MacHale, 2006; Norman, 1985; Toseland, 2007; “What Are Malaopropisms,” n.d.)

Tri-Gram Corpus

In recent years both Google and Microsoft have made available very large corpora extracted from their search engines. The corpora are derived from a large body of English in its normal usage. These corpora are essentially a collection of the searches and their results from a one-month period on either Google or Bing. Each of the corpora measure in the trillions of words and are available in multiple languages. They have made it possible to extract meaningful probabilities of n-grams of almost any size. Recent work by Islam & Inkpen (2009) has shown that the use of tri-grams from the Google Web-1T corpus was effective in identifying malapropisms. However, the Google Web-1T corpus is available only on DVD, thus the entire corpus must be loaded onto a local server and indexed before it can be used.

The corpus available from Microsoft research is of comparable size and can be accessed using a simple web based API, making it a nearly trivial task to include searches for tri-gram probabilities in research projects. There are two versions of the Microsoft corpus. The first is based on search results of June 2009 and the second from April 2010. The June 2009 data set is the larger and more complete of the two. This is because the April 2010 data set was more narrowly focused on metadata. The June 2009 data set was chosen because it contains a much larger volume of general English language usage (Wang et al., 2010).

The Conversational Agent

The Conversational Agent was created using a readily available open source AIML engine known as “program-O” (https://github.com/Program-O/Program-O). Program-O is a web-based CA written in PHP and stores the AIML knowledge set in a MySQL database. This CA used an unmodified standard AIML knowledge set from https://code.google.com/p/aiml-en-us-foundation-alice/. As mentioned, this knowledge set was used in its original form without any edits or modifications and was loaded into the database using the standard tools distributed with Program-O. Program-O was chosen due to the ease with which the pre-classifiers can be implemented.

First Pre-Classifier Using Semantic Distance

The first implementation relied on a measure of semantic distance using WordNet (Fellbaum, 1998). Sinha & Mihalcea (2007) reported the best results with the Wu & Palmer (WuP) and Lesk algorithms. However, the Wu & Palmer algorithm works only with words of the same part of speech and therefore would be of limited use in this context. The Lesk and Gloss Vector algorithms, as described in Patwardhan & Pedersen (2006) and Sinha & Mihalcea (2007), have therefore been chosen since both will work across parts of speech. As noted previously, these two algorithms are more precisely called measures of *semantic relatedness* because they both work by comparing the overlap in information content between two words. These algorithms were implemented using the WordNet::SenseRelate::AllWords Perl package. (Pedersen & Kolhatkar, 2009). This implementation involved processing the input sentences with the pre-classifier where malapropisms were detected using semantic distance measured using WordNet::SenseRelate::AllWords. The detection process consisted of calculating the semantic distance (or relatedness) scores using both the Lesk and Glass Vector algorithms. From these two scores, a weighted composite score favoring the Lesk score was used.

This algorithm assumes each sentence contains at least one malapropism. This assumption works well for the malapropism corpus, but would require enormous computational resources if used on a source of common text, such as a newspaper article.

The word (or words) with the lowest semantic relatedness scores are assumed to be malapropisms. For each of these words, one or more correction candidates are chosen, as described on page 32. A single candidate (probable correct word) was then chosen (using the method described on page 33). If none of the candidate words produces a superior semantic relatedness score, the sentence is then considered to be correct in its original form. Since all of the sentences in the corpus do contain malapropisms, any sentence considered already correct by this algorithm actually represents a failure to detect the malapropism. The corrected sentences were forwarded onto the AIML CA for processing. A further modification to the CA involved a simple alteration to the output so that the chosen AIML category was included with the output sentence. This facilitated the detection of AIML pickup lines.

Second Pre-Classifier Using Tri-Gram Probabilities

The second implementation of the pre-classifier utilized the tri-gram model described earlier (Islam & Inkpen, 2009; Wilcox-O'Hearn et al., 2008).

The process of detecting and correcting malapropisms with the tri-gram algorithm is as follows:

1. The input is first broken down into a series of progressive tri-grams as described in Chapter 2.
2. For each tri-gram, a search is made through the corpus to determine the Error Score (ES). The ES represents the probability that the three words will appear together in common use. This is actually the conditional probability that given the first two words in the tri-gram, the next word

will appear after it. The values returned by the Microsoft N-Gram Service are actually base 10 logarithms of the raw probabilities.

3. In terms of absolute value, any tri-gram with an ES under a threshold value, determined during the training phase, of 5.21, was considered correct. This equates to a probability of $10^{-5.21}$ (or 6.166 x 10^{-6}) that given the first two words in the tri-gram, the third word will be the one following.
4. Tri-grams with an ES above the threshold were considered to contain errors. Specifically, since the ES is based on a conditional probability, it is assumed that the last word of the tri-gram is the malapropism and it is that word which was processed for correction (using the process described below).
5. The corrected sentence was then passed to the AIML CA for processing.

Finding the Correct Word Once the Malapropism Has Been Identified

Once the malapropism was identified, a search was made for candidates for the correct word. Several methods have been proposed for this process. These include Google searches (Bolshakov, 2005), paronyms dictionaries, essentially lists of commonly misused words and their corrections (Chiru et al., 2010), various searches based on various other attributes of the word, such as morphology and edit distance (Jurafsky & Martin, 2000), as well as various ways of searching through the WordNet hierarchy itself (Chiru et al., 2010; Hirst & Budanitsky, 2005). For this research, the paronyms dictionary approach was easily implemented using the freely available machine-readable version of

Common Errors in English Usage (Brian, 2008). Thus, correction candidates can be found through a simple search of the paronyms dictionary.

The paronyms search was augmented by a search based on various other word attributes. This involved searching through a dictionary (or other corpus) for a list of words that fall with an edit distance (Levenshtein distance) of one to three edits (depending on the number of syllables) of the word to be replaced. The result was a list of words, which were evaluated as possible replacements (Jurafsky & Martin, 2000; Norvig & Russel, 2010).

For the semantic distance pre-classifier, each candidate word was inserted into the original input sentence in place of the suspected malapropism and the semantic relatedness score or Error Probability was recalculated. After all candidate words were tested, the word having the best semantic distance score was selected. For the tri-gram pre-classifier, each candidate word was inserted into the original tri-gram and a new query against the Microsoft N-gram service was made. The word which produced the lowest error score was then selected.

Training the Pre-Classifier

A group of 25 sentences selected at random from the malapropism corpus along with their corrections was used in training the pre-classifiers. Specifically, the method for calculating the threshold values for the semantic distance measure was tested. The most effective cut-off value (in terms of determining whether a word is a malapropism) for the Error Score in the tri-gram measure was determined. Finally, methods for testing correction candidates were evaluated.

Processing Sentences from the Corpus

Since AIML has no specific malapropism detection mechanism, but instead relies on Symbolic Reduction to correct for any errors in the input, this research looked only at malapropism *correction*. Corrections made by Symbolic Reductions in AIML had to be inferred based on the classification of the responses. Thus this problem was fundamentally a classification problem. Sentences from the malapropism corpus were sent to the CA one at a time. For each sentence, the CA's response and the AIML category chosen by the CA were recorded.

Retrieval of results

The CA will always respond to a query (an input sentence). The responses will fall into one of three classifications:

- A default response (or 'pickup line' in AIML nomenclature). This is the response given by the CA when there are no matches to any of the AIML categories. The 'pickup lines' are designed to move the user on to a different topic. So called 'pick-up-lines' are randomly selected from a single default category within the AIML knowledge set.
- A category match that is classified appropriate by the judges in the context of the input sentence.
- A category match that is classified as nonsense by the judges in the context of the input sentence.

For each sentence, the judges will view the sentence sent to the CA and the response sent back by the CA. The judges will then assess whether the input sentence was Appropriate (that is to say, "Made Sense") or Nonsense. They will also assess whether the response from the CA was Appropriate (in response to the input sentence) or Nonsense. It is these assessments made by the judges that are then used to determine whether an input sentence has been "corrected" and if the CA responded in a way that indicated it had "understood" the input sentence.

Figure 8 will help to put this in a visual context:

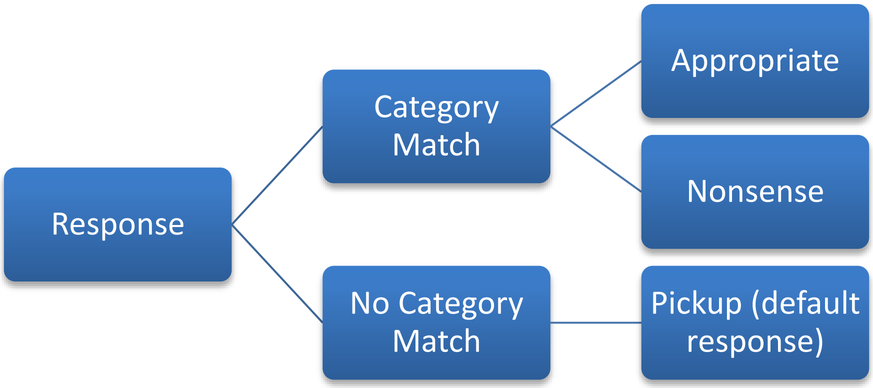

Figure 8: Classification of Responses

The case where no category match is made can be classified programmatically by simply checking for the default response. Any time the CA responds with a pick-up line, the selection of the default category due to the lack of a category match is recorded in the CAs logs, making it a straightforward task to classify these responses.

Two human judges made the decision whether the category matches were classified as Appropriate or Nonsense. In order to reduce the possibility of bias, the pickup responses were included in the data sent to the judges for classification. Thus, for all pickup line responses, there was a secondary classification of Appropriate/Nonsense assigned by the judges.

The CA processed the sentences in four distinct data sets:

1. The set of sentences containing uncorrected malapropisms processed with no pre-classifier.
2. The set of sentences with all malapropisms corrected and processed with no pre-classifier.
3. The set of sentences containing malapropisms, processed through the semantic distance based pre-classifier
4. The set of sentences containing malapropisms, processed through the tri-gram based pre-classifier.

The success of the CA in correcting for the malapropism was determined in two ways. The first was to simply analyze the sentences leaving the pre-classifiers to see if each contained the correct word as noted in the corpus. Second, it was possible to infer the success of the CA by analyzing the response classifications of the judges. The analysis of the judges' classifications made allowances for sentences that had been corrected using meaningful and semantically correct words that did not necessarily match the correct word noted in the corpus.

Results for each set of tests were represented using confusion matrices. Predictions were extrapolated from the results on the training data and these predictions were then compared to the actual results of the tests.

This resulted in a set of four matrices. There were two matrices for the unmodified CA (one for sentences containing malapropism and one for the corrected sentences).

There was one matrix each for the CA with the semantic distance pre-classifier and for the CA with the tri-gram pre-classifier. The confusion matrix format used for this purpose is shown in Table 1.

Table 1: Confusion Matrix

Response Classification		Predicted		
		Appropriate	Nonsense	Pickup
Actual	Appropriate	n_{aa}	n_{an}	n_{ap}
	Nonsense	n_{na}	n_{nn}	n_{np}
	Pickup	n_{pa}	n_{pn}	n_{pp}

In order to look more closely at the individual classifications a series of 2x2 matrices was derived. For example, in terms of appropriate responses, a 2x2 confusion matrix derived from the one above would look like this:

Table 2: Confusion Matrix for Appropriate Responses

		Predicted	
		Appropriate	NOT Appropriate
Actual	Appropriate	$n_1 = n_{aa}$	$n_2 = n_{ap} + n_{an}$
	NOT Appropriate	$n_3 = n_{pa} + n_{na}$	$n_4 = n_{pp} + n_{pn} + n_{np} + n_{nn}$

From this matrix, the following measures were defined: *True Positive Rate*, *False Positive Rate*, *True Negative Rate*, *False Negative Rate* and *Precision.* This gives us the following:

$$TPR = \frac{n_1}{n_1 + n_2} \tag{4}$$

$$FPR = \frac{n_3}{n_3 + n_4} \tag{5}$$

$$TNR = \frac{n_4}{n_3 + n_4} \tag{6}$$

$$FNR = \frac{n_2}{n_1 + n_2} \tag{7}$$

$$Precision = \frac{n_1}{n_1 + n_3} \tag{8}$$

The final measure, *Accuracy*, gives an overall measure of how closely the predictions based on the training set matched the actual results:

$$Accuracy = \frac{n_1 + n_4}{n_1 + n_2 + n_3 + n_4} \tag{9}$$

Once all relevant statistics were calculated, results were analyzed and reported with special attention to comparisons between the unenhanced and enhanced CAs. The results were also analyzed for any unexpected responses such as detection of errors and malapropisms not intentionally placed in the malapropism corpus.

From this data it was also possible to make an inference concerning the overall effectiveness of the SRAI model without enhancement as assumed in Chapter 1.

Lastly, a function was added to the CA to record the processor load and memory usage and time taken for each query. This allowed for a comparison of the computational resources used by the original CA using SRAI with the CAs enhanced with the two pre-classifiers.

Resources

Overall, the resources required for this research were not extensive. They included:

1. A server capable of running all of the software described previously. For this purpose the author's employer allowed the use of CentOS Linux version 6.4 server running in a virtual machine hosted on a Dell R410 server. The Virtual Computer was assigned

8 processor cores (at 2.3 Ghz each) and 8 Gigabytes of memory. Permission to use this server was obtained after it became apparent that the author's own Mac OS X based iMac computer would not be sufficient for this work.

2. Server software. The software for this research was implemented using the Apache web server (version 2.2.22), MySQL database (version 5.6.12) database server and PHP (version 5.3.15) for running the Program-O CA.
3. This server also ran the WordNet::SenseRelate::AllWords server under Perl version 5.12.4 installed via CPAN.
4. Access to the Microsoft N-Gram server over the Internet was provided using a web based API from Microsoft. http://blogs.msdn.com/b/webngram/
5. All program development for the pre-classifiers and modifications to the Program-0 code was done using the NetBeans IDE (version 7.3).
6. Two of the author's colleagues volunteered to act as judges for the classification of CA responses.

Chapter 4

Results

Introduction

In this chapter, the data gathered during the classification of the CA responses to the sentences in the test corpus are presented. The sentences in the test corpus were processed a total of four times. The uncorrected forms of the sentences were run three times: against the unmodified CA, the CA modified with the semantic distance based pre-classifier and the CA modified with the tri-gram based pre-classifier. The corrected forms of the sentences were processed once against the unmodified CA. Computer performance data gathered during the processing of the CA are also presented.

Raw Totals

There were 292 sentences in the test corpus. The corpus was processed four times using the following methods:

1. The unmodified CA responding to uncorrected input sentences (that is, all input sentences contained malapropisms). This represents the worst-case scenario.
2. The CA using the semantic distance based pre-classifier responding to uncorrected input sentences.
3. The CA using the tri-gram pre-classifier responding to uncorrected input sentences.

4. The unmodified CA responding to corrected sentences (that is, no input sentences contained malapropisms). This represents the best-case scenario.

Table 3 summarizes the totals for the classifications across the four data sets. Each CA response was classified as “Appropriate” or “Nonsense” by the judges. Moving from the worst-case scenario on the left toward the best-case scenario on the right the data sets show that the number of CA responses classified as Appropriate steadily increases with the tri-gram pre-classifier demonstrating a slight improvement over the semantic distance pre-classifier.

Table 3: Classification of CA Responses

	First set Uncorrected	Third set Semantic PC	Fourth set Tri-Gram PC	Second set Corrected
Appropriate	151	177	179	196
Nonsense	141	115	113	96
Correction contains “correct” word as noted in corpus		20	51	

Specific Results for the Four Test Sets

When looking at the third row of Table 3 it would appear as though the tri-gram pre-classifier significantly outperformed the semantic distance Pre-classifier. However, the judges’ classifications in row one would seem to indicate otherwise. In analyzing the corrections made by the pre-classifiers it was found that often a semantically correct word that did not necessarily match the “correct” word noted in the corpus was chosen. In the limited (single sentence) context of the research, the judges often classified the sentences containing these words as “Appropriate”.

Table 4 shows the overall results for detection and correction of the two pre-classifiers based solely on comparison to the corpus. From this data it can be seen that the tri-gram pre-classifier clearly outperformed the semantic distance pre-classifier. It also appears that both did a very poor job of correcting the malapropisms. However, as noted above, these results do not take into account corrections that do not match the corpus but that are semantically correct and make sense in the context of the sentence in which they are placed.

Table 4: Detection and Correction Results Based On Comparison to Corpus

	Detected	Corrected
Semantic Distance	186	20
Tri-Gram	212	51

Before examining the results of the actual classifications it is important to examine the level of agreement between the judges. Tables 5 through 8, below, show the classification (of the CA response) totals for each judge across the four data sets with the Kappa values for each. In all four instances the Kappa values indicate a “moderate” level of agreement. While the level of agreement is not high, it is consistent across all 4 data sets. (Landis & Koch, 1977)

Table 5: Agreement Between Judges, Data Set 1.

	Judge 2 Appropriate	Judge 2 Nonsense
Judge 1 Appropriate	74	19
Judge 1 Nonsense	58	141
Kappa ≈ 0.45		

Table 6: Agreement Between Judges, Data Set 2.

	Judge 2 Appropriate	Judge 2 Nonsense
Judge 1 Appropriate	104	2
Judge 1 Nonsense	90	96
Kappa ≈ 0.42		

Table 7: Agreement Between Judges, Data Set 3.

	Judge 2 Appropriate	Judge 2 Nonsense
Judge 1 Appropriate	109	22
Judge 1 Nonsense	46	115
Kappa ≈ 0.54		

Table 8: Agreement Between Judges, Data Set 4.

	Judge 2 Appropriate	Judge 2 Nonsense
Judge 1 Appropriate	99	15
Judge 1 Nonsense	65	113
Kappa ≈ 0.47		

In order to generate predictions for the confusion matrices, three basic assumptions were used:

1. For the first data set (unmodified CA responding to uncorrected sentences) the assumption was that since all input sentences contained malapropisms, all CA responses would be classified as “Nonsense”. In fact, only approximately 48% where so classified. This would seem to indicate that the CA has an inherent ability to deal with malapropisms in its original form.
2. For the second data set (unmodified CA responding to sentences with no malapropisms) the assumption was made that all CA responses would be classified as “Appropriate”. The judges classified approximately 67% of the responses as “Appropriate”. Clearly, the CA (at least using the default AIML set) is not able to correctly understand all of the input.
3. For the third and fourth data sets (modified CA responding to uncorrected sentences) the assumption was made that for each sentence in which the pre-classifier replaced a suspected malapropism (“corrected” the sentence)

the CA response would be classified as appropriate. Conversely, for each sentence in which the pre-classifier did not find a suspected malapropism, the CA response would be classified as “Nonsense”. Both of the pre-classifiers were remarkably similar in their results. The judges classified approximately 39% of the responses as “Nonsense”. This is an improvement of almost 10% over the unmodified CA and only slightly worse than the best-case scenario of the 4th data set.

From these three assumptions and the data from the judges’ classifications, four confusion matrices were derived.

Table 9: Confusion Matrix - No Pre-Classifier - With Malapropisms

		Predicted		
		Appropriate	Nonsense	Pickup
Actual	Appropriate	0	140	11
	Nonsense	0	119	22
	Pickup	0	0	0

Table 10: Confusion Matrix - No Pre-Classifier - Without Malapropisms

		Predicted		
		Appropriate	Nonsense	Pickup
Actual	Appropriate	181	0	5
	Nonsense	75	0	8
	Pickup	0	3	20

Table 11: Confusion Matrix - Semantic Distance Pre-Classifier

		Predicted		
		Appropriate	Nonsense	Pickup
Actual	Appropriate	66	103	1
	Nonsense	30	54	4
	Pickup	2	4	28

Table 12: Confusion Matrix - Tri-Gram Pre-Classifier

		Predicted		
		Appropriate	Nonsense	Pickup
Actual	Appropriate	119	51	0
	Nonsense	71	13	9
	Pickup	0	5	24

From these four matrices, more specific matrices for each of the classifications can be derived.

Unmodified CA with Malapropisms

Table 13, Table14 and Table 15 show the matrices for the first data set. In Table 13, note that since the assumption was made that all responses would be classified as “Nonsense”, the “Appropriate” column under “Predicted” is 0. Thus, there were no true or false positives. Slightly more than one half of the input sentences produced a response classified as “Appropriate”. This gives an accuracy rate for this set of predictions of approximately 48%

Table 13: Responses Classified Appropriate - First Data Set

		Predicted	
		Appropriate	NOT Appropriate
Actual	Appropriate	0	151
	NOT Appropriate	0	141

TPR =	0
FPR =	0
TNR =	1
FNR =	1
Accuracy =	0.482876712

In Table 14, the prediction that all responses would be classified as "Nonsense" had an accuracy rate of only about 44.5%.

Table 14: Responses Classified Nonsense - First Data Set

		Predicted	
		Nonsense	NOT Nonsense
Actual	Nonsense	119	22
	NOT Nonsense	140	11

TPR =	0.843971631
FPR =	0.927152318
TNR =	0.072847682
FNR =	0.156028369
Accuracy =	0.445205479

In Table 15, the assumption was made that none of the input sentences would result in "Pickup" responses. However, there were in fact 33 such responses out of the total of 292. This prediction turned out to be approximately 88.7% accurate.

Table 15: Responses Classified Pickup - First Data Set

		Predicted	
		Pickup	NOT Pickup
Actual	Pickup	0	33
	NOT Pickup	0	259

TPR =	0
FPR =	0
TNR =	1
FNR =	1
Accuracy =	0.886986301

Tables 13 through 15 have established the worst-case scenario.

Unmodified CA with No Malapropisms

Tables 16 through 18 show classification results for the second data set. Here the assumption that all responses would be classified as “Appropriate” worked well, producing an accuracy of slightly more than 70 percent for both the “Appropriate” (Table 16) and “Nonsense” (Table 17) classifications. As was the case with the first data set, the predictions for the “Pickup” classifications (Table 18) were very accurate at almost 95%.

Table 16: Responses Classified Appropriate - Second Data Set

		Predicted	
		Appropriate	NOT Appropriate
Actual	Appropriate	181	5
	NOT Appropriate	75	31

TPR =	0.97311828
FPR =	0.70754717
TNR =	0.29245283
FNR =	0.02688172
Accuracy =	0.726027397

Table 17: Responses Classified Nonsense - Second Data Set

		Predicted	
		Nonsense	NOT Nonsense
Actual	Nonsense	0	83
	NOT Nonsense	3	206

TPR =	0
FPR =	0.014354067
TNR =	0.985645933
FNR =	1
Accuracy =	0.705479452

Table 18: Responses Classified Pickup - Second Data Set

		Predicted	
		Pickup	NOT Pickup
Actual	Pickup	20	13
	NOT Pickup	3	256

TPR =	0.606060606
FPR =	0.011583012
TNR =	0.988416988
FNR =	0.393939394
Accuracy =	0.945205479

Modified CA Using Semantic Distance Pre-Classifier

Tables 19 through 21 summarize the classification results for the first of the two pre-classifiers. This third data set, for the semantic distance based pre-classifier, shows some interesting results. The accuracy rate for the “Appropriate” (Table 19) predictions is almost identical to the first data set. The rate for the “Nonsense” (Table 20) predictions is only slightly better.

Table 19: Responses Classified Appropriate - Third Data Set

		Predicted	
		Appropriate	NOT Appropriate
Actual	Appropriate	66	104
	NOT Appropriate	30	58

TPR =	0.388235294
FPR =	0.340909091
TNR =	0.659090909
FNR =	0.611764706
Accuracy =	0.480620155

Table 20: Responses Classified Nonsense - Third Data Set

		Predicted	
		Nonsense	NOT Nonsense
Actual	Nonsense	54	34
	NOT Nonsense	107	97

TPR =	0.613636364
FPR =	0.524509804
TNR =	0.475490196
FNR =	0.386363636
Accuracy =	0.517123288

This pre-classifier produced several interesting corrections to sentences in the corpus. For example, when presented with the sentence “It was a crushing crow” (from the training corpus), the pre-classifier produced “It was a cussing crow”. While this elicited chuckles from the judges, it also resulted in a response classified as “Appropriate” from the judges. In a later run of the training corpus, the pre-classifier produced “It was a trashing crow” which resulted in a response classified as “Nonsense” by the judges. What is interesting to note here is that in both cases, the CA responded with exactly the same reply of “Oh, I get it”.

Table 21 shows that the accuracy rate of the “Pickup” prediction was very good.

Table 21: Responses Classified Pickup - Third Data Set

		Predicted	
		Pickup	NOT Pickup
Actual	Pickup	28	6
	NOT Pickup	2	253

TPR =	0.823529412
FPR =	0.007843137
TNR =	0.992156863
FNR =	0.176470588
Accuracy =	0.972318339

Modified CA Using Tri-Gram Pre-Classifier

Tables 22 through 24 (Page 51) summarize the classification results for the second of the two pre-classifiers. The results here show a measurable improvement over the performance of the semantic distance based pre-classifier. The accuracy rate for the "Appropriate" predictions is about 58%. The rate of the "Nonsense" predictions is slightly over 53%. Although these rates are not terribly good, they are reliably over 50%. What is more interesting is looking at these rates in the context of the raw totals presented earlier in Table 3. In the raw totals, semantic distance based pre-classifier resulted in a total of 177 "Appropriate" responses, while the tri-gram based pre-classifier resulted in 179 "Appropriate" responses. This difference of only two "Appropriate" responses would seem to indicate that the tri-gram based pre-classifier and the semantic distance based pre-classifier were nearly identical in their performance. Two factors would appear to account for this difference. First, the tri-gram based pre-classifier produced fewer "Pickup" classifications. Second, the tri-gram based pre-classifier resulted in few "corrected" sentences that produced responses marked as "Nonsense".

Table 22: Responses Classified Appropriate - Fourth Data Set

		Predicted	
		Appropriate	NOT Appropriate
Actual	Appropriate	119	51
	NOT Appropriate	71	51

TPR =	0.7
FPR =	0.581967213
TNR =	0.418032787
FNR =	0.3
Accuracy =	0.582191781

Table 23: Responses Classified Nonsense - Fourth Data Set

		Predicted	
		Nonsense	NOT Nonsense
Actual	Nonsense	13	80
	NOT Nonsense	56	143

TPR =	0.139784946
FPR =	0.281407035
TNR =	0.718592965
FNR =	0.860215054
Accuracy =	0.534246575

Table 24: Responses Classified Pickup - Fourth Data Set

		Predicted	
		Pickup	NOT Pickup
Actual	Pickup	24	9
	NOT Pickup	5	254

TPR =	0.727272727
FPR =	0.019305019
TNR =	0.980694981
FNR =	0.272727273
Accuracy =	0.952054795

Comparison of Results

Comparing the results of the worst-case scenario (first data set) to the other three data sets produced the results shown in Table 25 which shows the number of responses from the first data set that changed in any of the other three. The important point here is that the CA did NOT change its response substantially, that is to say, it chose its response based on the *same* AIML category chosen for the response in the first data set.

The best-case results for the unmodified CA show that for the overwhelming number of corrected input sentences used in the fourth data set, the judges classified the CA response as “Appropriate” despite the fact that the AIML response category had not changed. The results for the two pre-classifiers are much as expected given the prediction results presented earlier.

Table 25: Classifications that changed while AIML Categories Did Not

	Nonsense -> Appropriate	Appropriate -> Nonsense
Unmodified CA	59	13
Semantic Distance Pre-Classifier	47	20
Tri-Gram Pre-Classifier	54	19

Figure 9 below shows another interesting result from an analysis of the CA responses. The response category numbers from all four data sets were sorted according to frequency. Number 98856 is the “Pickup” category and was the second most common response category chosen. However, category number 91855, the most commonly chosen response category, is so general in its nature as to be nearly a “Pickup” category itself.

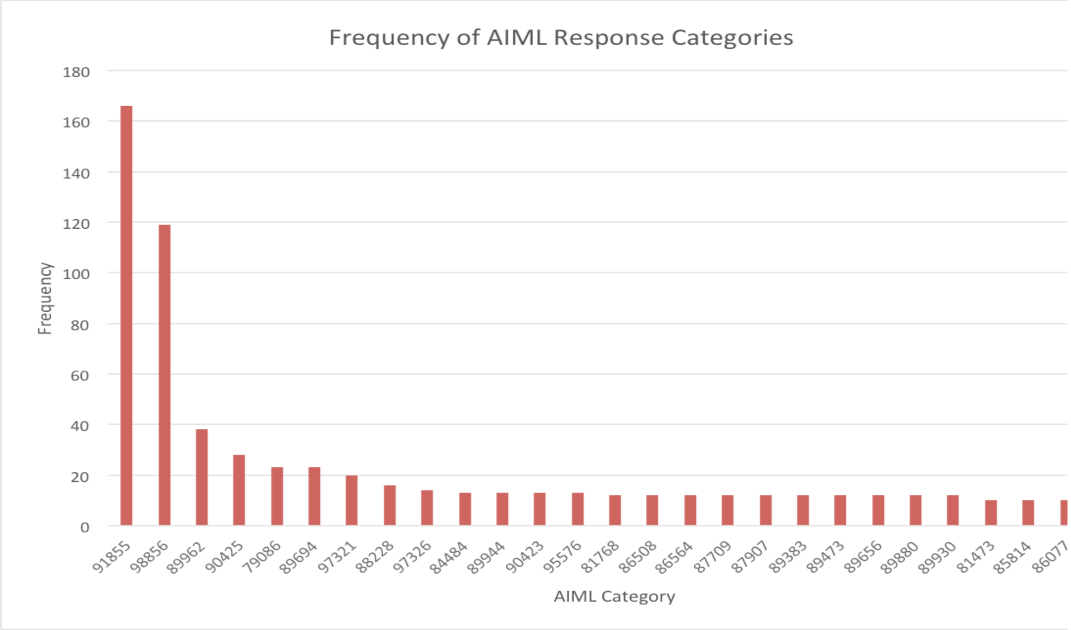

Figure 9: Frequency of AIML Categories[1]

Figure 10 (Page 54) shows the AIML code for category 91855. This code shows the use of the AIML symbolic reduction to reduce personal statements of fact or preference. As an example, the input sentence "I'd rather soak in a sub than take a shower." Will be matched by this category and may elicit a response such as "That's good information". Technically, the "Pickup" category is only chosen when there is NO category match at all. So by this definition, the category shown in Figure 10 is not a "Pickup" category. Yet, the generality of this category will cause it to completely ignore nearly all errors in grammar and syntax, including malapropisms.

[1] The reader will note that these category numbers are merely the category identifiers from the database used in this research and have no relationship to the original AIML data sets (in which the categories are not numbered).

```
<category>
       <pattern>I *</pattern>
       <template>
              <random>
                <li>Why?</li>
                <li>Interesting gossip</li>
                <li>That is interesting</li>
                <li>That's good information</li>
                <li>Thanks for the information</li>
                <li>Do you mind if I tell other people</li>
                <li>I haven't heard anything like that before</li>
              </random>.
              <think>
                  <srai>PUSH <person>YOU <star/></person></srai>
              </think>
       </template>
</category>
```

Figure 10: Most Common AIML Category

This phenomenon illustrated in Figure 9 and Figure 10 helps to explain some disparity in the raw results. Table 26 shows that the number of responses classified as “Pickup” improves by only 10 sentences from the worst-case to the best-case data set. Yet, there is an increase of 45 responses marked “Appropriate”.

Table 26: Analysis of CA random responses

	No. Pickup Lines	Change from 1st set	No. Appropriate	Change from 1st set
1st set	33	-	151	-
2nd set	34	+1	177	+26
3rd set	29	-4	179	+28
4th set	23	-10	196	+45

The judges were also asked to tag the input sentences with Y or N answer to: "MAKES SENSE (Y/N)". Note that in Table 27, as the number of input sentences tagged as “Makes Sense” increases, the number of responses classified as “Appropriate” increases as well. Table 27 is perhaps the best overall evidence of the relative performance of the semantic distance based pre-classifier (third data set) and the tri-gram based pre-classifier (fourth data set) showing how the increasing number of "correct" input sentences results in a greater number of responses classified as "Appropriate".

Table 27: Input Sentences vs. Reponses Classification

	Number Makes Sense	Number Appropriate
First Data Set	29	151
Third Data Set	65	177
Fourth Data Set	106	179
Second Data Set	284	196

Computational Performance

Tables 28 through 30 summarize the computational performance of the CA with and without the pre-classifiers. In all cases, peak memory usage was nearly identical. This was driven by the server configuration in which the PHP interpreter was limited in its maximum memory consumption. In terms of CPU resources, there was only a small increase from the unmodified CA in Table 28 to the CA using the tri-gram pre-classifier in Table 30. Note also that the total time required to process the entire corpus was less than one hour for both the unmodified CA and the CA using the tri-gram pre-classifier.

Table 28: Computation Performance - Unmodified CA

	Total Seconds	**Peak Memory Usage (bytes)**	**Total CPU Tics**
	2471	46,390,528	625,466
Total Hours:	0.687		

The CA using the semantic distance pre-classifier (Table 29) required dramatically greater resources. Note here that the total time to process the corpus was nearly 60 hours! In many cases, individual sentences required processing times in the range of 15 to 20 minutes per sentence.

Table 29: Computation Performance - CA with Semantic Distance Pre-Classifier

	Total Seconds	**Peak Memory Usage (bytes)**	**Total CPU Tics**
	214285	46,390,528	412,799,153
Total Hours:	59.1		

Table 30: Computation Performance - CA with Tri-Gram Pre-Classifier

	Total Seconds	**Peak Memory Usage (bytes)**	**Total CPU Tics**
	3468	46,390,528	100,333,748
Total Hours:	0.96		

The tri-gram pre-classifier performed slightly better (in terms of correcting malapropisms) than the semantic distance pre-classifier while consuming a small fraction of the computational resources and time.

Chapter 5

Conclusions

Conclusions

This research examined the ability of an AIML based CA to correct malapropisms present in input sentences. This included determining whether the addition of one of two pre-classifiers to the CA would improve this ability. The answer to whether the CA modified with a semantic distance based pre-classifier would indeed perform better than the unmodified CA is a qualified yes. While the CA did indeed perform better at correcting malapropisms, the increase in computational resources, especially the dramatic increase in processing time combined with a marginal gain in performance renders this solution of very little use.

The answer to whether the CA modified with a tri-gram based pre-classifier would indeed perform better than the unmodified CA is yes. Not only did the tri-gram based pre-classifier perform better than the semantic distance based pre-classifier, but it did so with a minimal increase in computational resources.

Implications

Several more questions were raised and several problems were identified in the process of answering these questions.

While the tri-gram based pre-classifier did provide a measurable improvement in the correction of malapropisms, the improvement was only moderate. Moreover, an analysis of the CA responses showed that in a large number of cases, the CA did not actually correct for the malapropism, but instead ignored it, choosing response categories that were very broad and generalized. While this approach may be satisfactory to hobbyist built “chat-bots” used for entertainment, such an approach cannot be said to be truly “understanding” the input sentences. In fact, the approach is little different from that used by ELIZA for fooling users into believing they were having a “real” conversation when in fact they were not. This argument is essentially the same fundamental criticism leveled against the ‘Goostman’ CA referred to earlier. While it is beyond the scope of this research to make conclusions on the overall usefulness of CAs using techniques similar to that studied here, the approach is called into question.

Another question raised is the possibility of bias on the part of the judges. The fact that the second through fourth data sets showed a marked increase in responses classified as “Appropriate” when the number of AIML category responses changed very little would tend to lend credence to this conclusion. However, the bias may be more “natural” than one at first supposes. Recent research in the field of neuroscience has challenged earlier assumptions on the way in which human beings process language. The long held belief that our brains work from the bottom up, that is to say, that we first process the meanings of the individual words and then put these meanings together to understand the entire sentence, then paragraph, etc. may not be true. It seems that as people hear or read, their minds are trying to predict what comes next, based on the words that are seen (or heard). The mind then works "down" to retrieve the details. Thus, it may be that the “arrival” of expected words as one mentally processes a stream of words leads to a natural inclination to regard the responses as “Appropriate”. It is interesting to note that the tri-gram based pre-classifier more closely mimics this process than does the semantic distance pre-classifier (Dikker, Silbert, Hasson, & Zevin, 2014).

It should also be noted that this research was applied against the ALICE CA in its “out of the box” configuration, which contains 98,854 categories. By contrast the commercial “Silver Edition” of the ALICE CA contains over 120,000 AIML categories (Wallace, 2010). It is quite possible that such a CA, having many more specific categories, would respond differently to malapropisms. Context is also an issue. In this research the CA looked only at individual sentences without surrounding context. It is possible to configure a CA to track a conversation and "remember" the previous few input sentences. Putting the input sentences in the context of a longer conversation could also affect the responses.

Recommendations

Neither of the pre-classifiers studied here was capable of dealing with compound malapropisms (two or more malapropisms in the same sentence). For a CA to be able to deal with “real world” conversations, development of this ability would be necessary.

Another factor impacting the performance of the pre-classifiers are the probabilities used to determine when a given word is considered a malapropism or not. These probabilities are critical to the performance of the pre-classifiers and the methods used to estimate them represent an area for future research.

One notable contribution of this research beyond the results of the study itself was the creation of a corpus of malapropisms. Earlier work on malapropisms has involved the artificial creation of malapropisms by programmatic means (Hirst & Budanitsky, 2005).

Summary

Finally, as stated earlier, the author has been unable to locate any previous research documenting the effectiveness of the “engine” that underlies AIML. This research has shown that, at least in its “out of the box” form, the AIML CA is only marginally capable of correcting malapropisms. Further, the addition of the tri-gram based pre-classifier did indeed improve the CAs ability to correct malapropisms. Ample opportunities exist for further research and improvement, including improved methods for estimating the probabilities used in the tri-gram based pre-classifier and variations on the composite methods used to detect and correct malapropism. The ability to correct multiple malapropisms within a single sentence would be a great benefit as well.

Appendix A

Training Corpus

A bunch of fair-feather fans
A bunch of fair-weather fans
[feather, weather]

Bad news could make your socks sink.
Bad news could make your stocks sink.
[socks, stocks]

Here comes the street vendor, hawking his hares.
Here comes the street vendor, hawking his wares.
[hares, wares]

Here comes the street vendor, walking his wares.
Here comes the street vendor, hawking his wares.
[walking, hawking]

He's doggy and sedated, but he'll recover.
He's groggy and sedated, but he'll recover.
[doggy, groggy]

May I pour a pup for you?
May I pour a cup for you?
[pup, cup]

The couch? The dog has hovered it with hair.
The couch? The dog has covered it with hair.
[hovered, covered]

A bit out of the bay, though.
A bit out of the way, though.
[bay, way]

What a ducky dog!
What a lucky dog!
[ducky, lucky]

And, of course, A Sale Of Two Cities, by "Chuck" Dickens.
And, of course, A Tale of Two Cities, by "Chuck" Dickens.
[Sale, Tale]

It was a crushing crow
It was a crushing blow.
[crow, blow]

You were lighting a liar in the quadrangle.
You were lighting a fire in the quadrangle.
[liar, fire]

He had to use a fire distinguisher.
He had to use a fire extinguisher.
[distinguisher, extinguisher]

This is unparalyzed in the state's history.
This is unparalleled in the state's history.
[unparalyzed, unparalleled]

O, he will dissolve my mystery!
O, he will resolve my mystery.
[dissolve, resolve]

I was swimming in the sea when an octopus came out of the water and wrapped his testicles around me.
I was swimming in the sea when an octopus came out of the water and wrapped his tentacles around me.
[testicles, tentacles]

From an early age my son wanted a career in law enforcement so he became a detective in the police farce.
From an early age my son wanted a career in law enforcement so he became a detective in the police force.
[farce, force]

When I was a little girl I was vaccinated against polo.
When I was a little girl, I was vaccinated against polio.
[polo, polio]

The optometrist told me I have cadillacs on my eyes.
The optometrist told me I have cataracts on my eyes.
[cadillacs, cataracts]

George Bush is undergoing his second term of pregnancy.
George Bush is undergoing his second term of presidency.
[pregnancy, presidency]

Scientists should be given lots of money so to build lots of lavatories in which to do their work.
Scientists should be given lots of money so to build lots of laboratories in which to do their work.
[lavatories, laboratories]

My husband was in expensive care at the hospital when he was ill.
My husband was in intensive care at the hospital when he was ill.
[expensive, extensive]

My car gets 30 miles to the galleon.
My car gets 30 miles to the gallon.
[galleon, gallon]

In the stomach your food gets mixed up with the ghastly juices.
In the stomach your food gets mixed up with the gastric juices.
[ghastly, gastric]

Much of the world’s rubber is produced in malaria.
Much of the world’s rubber is produced in Malaysia.
[malaria, Malaysia]

Appendix B

Malapropism Corpus

A bunch of wear-weather fans
A bunch of fair-weather fans
[wear, fair]

Bad news could make your stocks stink.
Bad news could make your stocks sink.
[stink, sink]

Bill Gates is the leaky lord of computer desktops.
Bill Gates is the geeky lord of computer desktops.
[leaky, geeky]

Bill Gates is the geeky gourd of computer desktops.
Bill Gates is the geeky lord of computer desktops.
[gourd, lord]

Billy will now sing Dutiful Dreamer.
Billy will now sing Beautiful Dreamer.
[Dutiful, Beautiful]

Billy will now sing Beautiful Breamer.
Billy will now sing Beautiful Dreamer.
[Breamer, Dreamer]

Did you ever watch the Dinah Door Show?
Did you ever watch the Dinah Shore show?
[Door, Shore]

Did you ever watch the Shine a Shore Show?
Did you ever watch the Dinah Shore show?
[Shine a, Dinah]

Did you meet the father of the fried?
Did you meet the father of the bride?
[fried, bride]

Did you meet the bother of the bride?
Did you meet the father of the bride?
[bother, father]

Enough with the monk mail already!
Enough with the junk mail already!
[monk, junk]

Enough with the junk jail already!
Enough with the junk mail already!
[jail, mail]

Have you read Marcel Proust's In Search Of Lost Lime?
Have you read Marcel Proust's In Search of Lost Time?
[Lime, Time]

Have you read Marcel Proust's In Search Of Tossed Time?
Have you read Marcel Proust's In Search of Lost Time?
[Tossed, Lost]

He's groggy and cemented, but he'll recover.
He's groggy and sedated, but he'll recover.
[cemented, sedated]

I actually look forward to the dyer drills.
I actually look forward to the fire drills.
[dyer, fire]

I actually look forward to the fire frills.
I actually look forward to the fire drills.
[frills, drills]

I'd rather soak in a sub than take a shower.
I'd rather soak in a tub than take a shower.
[sub, tub]

I'd rather toke in a tub than take a shower.
I'd rather soak in a tub than take a shower.
[toke, soak]

I'd rather soak in a tub than shake a shower.
I'd rather soak in a tub than take a shower.
[shake, take]

I'd rather soak in a tub than take a tower.
I'd rather soak in a tub than take a shower.
[tower, shower]

I had a nasty encounter with the deadroom door.
I had a nasty encounter with the bedroom door.
[deadroom, bedroom]

I had a nasty encounter with the bedroom bore.
I had a nasty encounter with the bedroom door.
[bore, door]

I need to empty out this dirty daughter.
I need to empty out this dirty water.
[daughter, water]

I need to empty out this wordy water.
I need to empty out this dirty water.
[wordy, dirty]

I stayed home, written with a nasty rash.
I stayed home, smitten with a nasty rash.
[written, smitten]

I stayed home, smitten with a nasty smash.
I stayed home, smitten with a nasty rash.
[smash, rash]

I throw myself on your majesty's face and favor.
I throw myself on your majesty's grace and favor.
[face, grace]

I throw myself on your majesty's grace engraver.
I throw myself on your majesty's grace and favor.
[engraver, and favor]

I banged my bed.
I banged my head.
[bed, head]

I hanged my head.
I banged my head.
[hanged, banged]

Let's whip up a pasty paddock for the horses.
Let's whip up a hasty paddock for the horses.
[pasty, hasty]

Let's whip up a hasty haddock for the horses.
Let's whip up a hasty paddock for the horses.
[haddock, paddock]

May I poor a cup for you?
May I pour a cup for you?
[poor, pour]

One lump, or loo?
One lump or two?
[loo, two]

One tump, or two?
One lump or two?
[tump, lump]

Now you'll taste the hair of the hog.
Now you'll taste the hair of the dog.
[hog, dog]

Now you'll taste the dare of the dog.
Now you'll taste the hair of the dog.
[dare, hair]

Opposition to the waging of wombat has grown.
Opposition to the waging of combat has grown.
[wombat, combat]

Opposition to the caging of combat has grown.
Opposition to the waging of combat has grown.
[caging, waging]

Ow, my bunny bone!
Ow, my funny bone!
[bunny, funny]

Ow, my funny phone!
Ow, my funny bone!
[phone, bone]

Schumacher couldn't maintain the fast face he'd established.
Schumacher couldn't maintain the fast pace he'd established.
[face, pace]

Schumacher couldn't maintain the past pace he'd established.
Schumacher couldn't maintain the fast pace he'd established.
[past, fast]

She gave us a slight slave, then quickly disappeared.
She gave us a slight wave, then quickly disappeared.
[slave, wave]

She gave us a white wave, then quickly disappeared.
She gave us a slight wave, then quickly disappeared.
[white, slight]

She is not merely missed by us all.
She is not dearly missed by us all.
[merely, dearly]

She is not dearly dissed by us all.
She is not dearly missed by us all.
[dissed, missed]

Thanks for pointing out the broken link, Keith. I'll excise the decay, host-paste.
Thanks for pointing out the broken link, Keith. I'll excise the decay, post-haste
[host-paste, post-haste]

That's such a pretty pimple you've got.
That's such a pretty dimple you've got.
[pimple, dimple]

That's such a dirty dimple you've got.
That's such a pretty dimple you've got.
[dirty, pretty]

The couch? The dog has covered it with care.
The couch? The dog has covered it with hair.
[care, hair]

The Pitching Post is an excellent steak place in Casmalia.
The Hitching Post is an excellent steak place in Casmalia.
[Pitching, Hitching]

The Hitching Host is an excellent steak place in Casmalia.
The Hitching Post is an excellent steak place in Casmalia.
[Host, Post]

A whit out of the way, though.
A bit out of the way, though.
[whit, bit]

There must be a million stars in the sty.
There must be a million stars in the sky.
[sty, sky]

There must be a million scars in the sky.
There must be a million stars in the sky.
[scars, stars]

The success of e-commerce depends on having thrusted third parties.
The success of e-commerce depends on having trusted third parties.
[thrusted, trusted]

Things look bleak for the Gang of Gore.
Things look bleak for the Gang of Four.
[Gore, Four]

Things look bleak for the Fang of Four.
Things look bleak for the Gang of Four.
[Fang, Gang]

This ranch could use a strong-headed strangler like you.
This ranch could use a strong-headed wrangler like you.
[strangler, wrangler]

This ranch could use a wrong-headed wrangler like you.
This ranch could use a strong-headed wrangler like you.
[wrong-headed, strong-headed]

What a lucky log!
What a lucky dog!
[log, dog]

What a wearable waste of food.
What a terrible waste of food.
[wearable, terrible]

What a terrible taste of food.
What a terrible waste of food.
[taste, waste]

What's a smart smeller like you doing in a place like this?
What's a smart feller like you doing in a place like this?
[smeller, feller]

What's a fart feller like you doing in a place like this?
What's a smart feller like you doing in a place like this?
[fart, smart]

What shall we make of this puff tapestry?
What shall we make of this puff pastry?
[tapestry, pastry]

What's this jumpy junk doing in my food?
What's this lumpy junk doing in my food?
[jumpy, lumpy]

What's this lumpy lunk doing in my food?
What's this lumpy junk doing in my food?
[lunk, junk]

Well I certainly won't wake your word for it.
Well I certainly won't take your word for it.
[wake, take]

Nights In White Satin was a big hit for the Moody Blues.
Knights in White Satin was a big hit for the Moody Blues.
[Nights, Knights]

Will plights of passion put you on the right track?
Will fits of passion put you on the right track?
[plights, fits]

Will fits of fashion put you on the right track?
Will fits of passion put you on the right track?
[fashion, passion]

You are Mary, Queen of Squats?
You are Mary, Queen of Scotts?
[squats, Scotts]

You are Mary, Keen of Scotts?
You are Mary, Queen of Scotts?
[Keen, Queen]

You look delicious in that silly suit.
You look delicious in that frilly suit.
[silly, frilly]

You look delicious in that frilly fruit.
You look delicious in that frilly suit.
[fruit, suit]

You've endured such a strong life. I wish you pate peace.
You've endured such a strong life. I wish you great peace.
[pate, great]

You've endured such a strong life. I wish you great grease.
You've endured such a strong life. I wish you great peace.
[grease, peace]

Don't sweat the sweaty things.
Don't sweat the petty things.
[sweaty, petty]

Don't pet the petty things.
Don't sweat the petty things.
[pet, sweat]

Silly Rabbit, Trix are for trids.
Silly rabbit, Trix are for kids.
[trids, kids]

Silly Rabbit, kicks are for kids.
Silly rabbit, Trix are for kids.
[kicks, Trix]

Silly Rabbi, Trix are for kids.
Silly rabbit, Trix are for kids.
[Rabbi, rabbit]

And, of course, A Tale of Two Cities, by "Duck" Dickens.
And of course, A Tale of Two Cities, by "Chuck" Dickens.
[Duck, Chuck]

And, of course, A Tale of Two Cities, by "Chuck" Chickens.
And of course, A Tale of Two Cities, by "Chuck" Dickens.
[Chickens, Dickens]

His speech was a half-warmed fish
His speech was a half-warmed dish.
[fish, dish]

A blushing blow
A crushing blow.
[blushing, crushing]

The Lord is a loving leopard.
The Lord is a loving shepherd.
[leopard, shepherd]

The Lord is a shoving shepherd.
The Lord is a loving shepherd.
[shoving, loving]

May I sew you to another seat?
May I show you to another seat?
[sew, show]

May I show you to another sheet?
May I show you to another seat?
[sheet, seat]

When addressing a gathering of English farmers, the Reverend Spooner said that he was pleased "to address so many sons of soil."
When addressing a gathering of English farmers, the Reverend Spooner said that he was pleased "to address so many sons of toil."
[soil, toil]

When addressing a gathering of English farmers, the Reverend Spooner said that he was pleased "to address so many tons of toil."
When addressing a gathering of English farmers, the Reverend Spooner said that he was pleased "to address so many sons of toil."
[tons, sons]

You have missed my history lectures; you have tasted a whole term.
You have missed my history lectures; you have wasted a whole term.
[tasted, wasted]

You have missed my history lectures; you have wasted a whole worm.
You have missed my history lectures; you have wasted a whole term.
[worm, term]

You have hissed my history lectures; you have wasted a whole term.
You have missed my history lectures; you have wasted a whole term.
[hissed, missed]

You have missed my mystery lectures; you have wasted a whole term.
You have missed my history lectures; you have wasted a whole term.
[mystery, history]

You will leave Oxford on the next down drain.
You will leave Oxford on the next down train.
[drain, train]

You will leave Oxford on the next town train.
You will leave Oxford on the next down train.
[town, down]

You were fighting a fire in the quadrangle.
You were lighting a fire in the quadrangle.
[fighting, lighting]

Is the dean dizzy?
Is the dean busy?
[dizzy, busy]

Is the bean busy?
Is the dean busy?
[bean, dean]

Dad says the monster is just a pigment of my imagination.
Dad says the monster is just a figment of my imagination.
[pigment, figment]

He's a wolf in cheap clothing.
He's a wolf in sheep's clothing.
[cheap, sheep's]

Michelangelo painted the Sixteenth Chapel.
Michelangelo painted the Sistine chapel.
[Sixteenth, Sistine]

My sister has extra-century perception.
My sister has extra-sensory perception.
[century, sensory]

Don't is a contraption.
Don't is a contraction.
[contraption, contraction]

Flying saucers are just an optical conclusion.
Flying saucers are just an optical illusion.
[conclusion, illusion]

A rolling stone gathers no moths.
A rolling stone gathers no moss.
[moths, moss]

Let's get down to brass roots.
Let's get down to brass tacks.
[roots, tacks]

Their father was some kind of civil serpent.
Their father was some kind of civil servant.
[serpent, servant]

You can lead a horse to manure but you can't make him drink.
You can lead a horse to water but you can't make him drink.
[manure, water]

The flood damage was so bad they had to evaporate the city.
The flood damage was so bad they had to evacuate the city.
[evaporate, evacuate]

Ease my ears
Ease my fears
[ears, fears]

A lack of lies
A pack of lies
[lack, pack]

A pack of pies
A pack of lies
[pies, lies]

It's pouring with pain
It's pouring with rain
[pain, rain]

It's roaring with rain
It's pouring with rain
[roaring, pouring]

Save the sails
Save the whales
[sails, whales]

Wave the whales
Save the whales
[wave, save]

It is beyond my apprehension.
It is beyond my comprehension.
[apprehension, comprehension]

Listen to the blabbing brook
Listen to the babbling brook.
[blabbing, babbling]

Cardial - as in cardial arrest.
Cardial - as in cardiac arrest.
[cardial, cardiac]

Marie Scott...has really plummeted to the top.
Marie Scott...has really risen to the top.
[plummeted, risen]

He's on 90...10 away from that mythical figure.
He's on 90...10 away from that magical figure.
[mythical, magical]

Unless somebody can pull a miracle out of the fire, Somerset are cruising into the semi-final.
Unless somebody can pull a miracle out of the air, Somerset are cruising into the semi-final.
[fire, air]

We cannot let terrorists and rogue nations hold this nation hostile or hold our allies hostile.
We cannot let terrorists and rogue nations hold this nation hostage or hold our allies hostage.
[hostile, hostages]

The police are not here to create disorder, they're here to preserve disorder.
The police are not here to create disorder, they're here to preserve order.
[disorder, order]

He was a man of great statue.
He was a man of great stature.
[statue, stature]

Republicans understand the importance of bondage between a mother and child.
Republicans understand the importance of bonding between a mother and child.
[bondage, bonding]

Well, that was a cliff-dweller.
Well, that was a cliff-hanger.
[cliff-dweller, cliff-hanger]

If Gower had stopped that cricket ball he would have decapitated his hand.
If Gower had stopped that cricket ball he would have incapacitated his hand.
[decapitated, incapacitated]

We seem to have unleased a hornet's nest.
We seem to have unleashed a hornet's nest.
[unleased, unleashed]

This series has been swings and pendulums all the way through.
This series has been swings and misses all the way through.
[pendulums, misses]

Be sure and put some of those neutrons on it.
Be sure and put some of these croutons on it.
[neutrons, croutons]

It's got lots of installation.
It's got lots of insulation.
[installation, insulation]

Oftentimes, we live in a processed world
Oftentimes, we live in a complex world
[processed, complex]

They have miscalculated me as a leader.
They have misunderstood me as leader.
[miscalculated, misunderstood]

I am mindful not only of preserving executive powers for myself, but for predecessors as well.
I am mindful not only of preserving executive power for myself, but for successors as well.
[predecessors, successors]

We need an energy bill that encourages consumption.
We need an energy bill that encourages conservation.
[consumption, conservation]

We are making steadfast progress.
We are making steady progress.
[steadfast, steady]

Promise to forget this fellow - to illiterate him from your memory.
Promise to forget this fellow - to obliterate him from your memory.
[illiterate, obliterate]

He is the very pineapple of politeness!
He is the very pinnacle of politeness!
[pineapple, pinnacle]

I have since laid Sir Anthony's preposition before her.
I have since laid Sir Anthony's proposition before her.
[preposition, proposition]

Oh! it gives me the hydrostatics to such a degree.
Oh! it gives me the hysterics to such a degree.
[hydrostatics, hysterics]

I hope you will represent her to the captain as an object not altogether illegible.
I hope you will represent her to the captain as an object not altogether eligible.
[illegible, eligible]

She might reprehend the true meaning of what she is saying.
She might comprehend the true meaning of what she is saying.
[reprehend, comprehend]

She's as headstrong as an allegory on the banks of Nile.
She's as headstrong as an alligator on the banks of the Nile.
[allegory, alligator]

I am sorry to say that my affluence over my niece is very small.
I am sorry to say that my influence over my niece is very small.
[affluence, influence]

He can tell you the perpendiculars.
He can tell you the particulars.
[perpendiculars, particulars]

Nay, no delusions to the past - Lydia is convinced;
Nay, no allusions to the past - Lydia is convinced.
[delusions, allusions]

This very day, I have interceded another letter from the fellow.
This very day, I have intercepted another letter from the fellow.
[interceded, intercepted]

I thought she had persisted from corresponding with him.
I thought she had desisted from corresponding with him.
[persisted, desisted]

His physiognomy so grammatical!
His phraseology so grammatical!
[physiognomy, phraseology]

I am sure I have done everything in my power since I exploded the affair.
I am sure I have done everything in my power since I exposed the affair.
[exploded, exposed]

I am sorry to say, she seems resolved to decline every particle that I enjoin her.
I am sorry to say, she seems resolved to decline every article that I enjoin her.
[particle, article]

If ever you betray what you are entrusted with...you forfeit my malevolence forever.
if ever you betray what you are entrusted with...you forfeit my benevolence forever.
[malevolence, benevolence]

Your being Sir Anthony's son would be sufficient accommodation.
Your being Sir Anthony's son would be sufficient recommendation.
[accommodation, recommendation]

I've been reading a very interesting book about General Rommel who commanded Hitler's pansy division in North Africa.
I've been reading a very interesting book about General Rommel who commanded Hitler's panzer division in North Africa.
[pansy, panzer]

Mussolini's followers were facetious.
Mussolini's followers were fascists.
[facetious, fascists]

I'm not superstitious, but I like to read my horrorscope in the newspaper every day.
I'm not superstitious, but I like to read my horoscope in the newspaper every day.
[horrorscope, horoscope]

My father was a wonderful musician, he played the baboon in the symphony orchestra.
My father was a wonderful musician, he played the bassoon in the symphony orchestra.
[baboon, bassoon]

I was swimming in the sea when a big octobus came out of the water and wrapped its tentacles around me.
I was swimming in the sea when a big octopus came out of the water and wrapped its tentacles around me.
[octobus, octopus]

The bible tells us not to lay up trousers on earth.
The bible tells us not to lay up treasures on earth.
[trousers, treasures]

The wise man brought gifts of gold, frankenstein and myrrh.
The wise men brought gifts of gold, frankincense, and myrrh.
[frankenstein, frankincense]

Judas asparagus was one of the twelve apostles.
Judas Iscariot was one of the twelve apostles.
[asparagus, Iscariot]

Jews worship in a cinemagogue.
Jews worship in a synagogue.
[cinemagogue, synagogue]

Solomon had 500 wives and 700 cucumbers.
Solomon had 500 wives and 700 concubines.
[cucumbers, concubines]

Jiminy Crocket fought at the Alamo.
Davey Crocket fought at the Alamo.
[Jiminy, Davey]

I was very worried when I was told that my husband the musician had a tumor
on his brain, but thank heavens it turned out to be non-militant.
I was very worried when I was told that my husband the musician had a tumor on his brain, but thank heavens it turned out to be malignant.
[non-militant, malignant]

Once when I was in court the judge asked who was making the allegations and I told him that I was the alligator.
Once when I was in court the judge asked who was making the allegations and I told him that I was the declarant.
[alligator, declarant]

From an early age my son wanted a career in law enforcement so he became a defective in the police force.
From an early age my son wanted a career in law enforcement so he became a detective in the police force.
[defective, detective]

I'm very afraid of being attacked by a stranger some dark night, so I'm taking a course in the marital arts.
I'm very afraid of being attacked by a stranger some dark night, so I'm taking a course in the martial arts.
[marital, martial]

I don't really like going up stairs in big department stores because I'm afraid of travelling on the alligators.
I don't really like going up stairs in big department stores because I'm afraid of travelling on the escalators.
[alligators, escalators]

I prefer to do my Christmas shopping in November because then I don't have to mangle with the terrible crowds.
I prefer to do my Christmas shopping in November because then I don't have to mingle with the terrible crowds.
[mangle, mingle]

My musician husband used to play the hobo in an orchestra.
My musician husband used to play the oboe in the orchestra.
[hobo, oboe]

In my English class in school I learned about the bowels which are a, e, i, o and u.
In my English class in school I learned about the vowels which are a, e, i, o and u.
[bowels, vowels]

I don't take risks for other people, I'm not putting my head in a moose for anybody.
I don't take risks for other people, I'm not putting my head in a noose for anybody.
[moose, noose]

My sister never married, she remained a sphincter all her life.
My sister never married, she remained a spinster all her life.
[sphincter, spinster]

I just love chicken, turkey and other foul dinners.
I just love chicken, turkey and other fowl dinners.
[foul, fowl]

I went to the doctor and he injected me with an epidemic needle.
I went to the doctor and he injected me with an epidermic needle.
[epidemic, epidermic]

I went to the supermarket and bought a cartoon of orange juice.
I went to the supermarket and bought a carton of orange juice.
[cartoon, carton]

I'm hoping to go to Africa for my annual vaccination.
I'm hoping to go to Africa for my annual vacation.
[vaccination, vacation]

My mother-in-law is seriously ill. She collapsed and is still in a comma.
My mother-in-law is seriously ill. She collapsed and is still in a coma.
[comma, coma]

My next door neighbor goes in and out to the city to work everyday. He’s a computer.
My next door neighbor goes in and out to the city to work everyday. He’s a commuter.
[computer, commuter]

When invaders came, ancient tribes used to light a deacon on top of a hill to warn others.
When invaders came, ancient tribes used to light a beacon on top of a hill to warn others.
[deacon, beacon]

Atoms join together to form monocles.
Atoms join together to form molecules.
[monocles, molecules]

A triangle with an angle bigger than ninety degrees is called obscene.
A triangle with an angle bigger than ninety degrees is called obtuse.
[obscene, obtuse]

Juniper is the largest of the planets.
Jupiter is the largest of the planets.
[Juniper, Jupiter]

A centimeter is an insect with a hundred legs.
A centipede is an insect with a hundred legs.
[centimeter, centipede]

The earth makes a complete resolution every twenty-four hours.
The earth makes a complete revolution every twenty-four hours.
[resolution, revolution]

A good public speaker should always breath with his diagram.
A good public speaker should always breath with his diaphragm.
[diagram, diaphragm]

At my father's funeral, the coffin was carried by six polar bearers.
At my father's funeral, the coffin was carried by six pal bearers.
[polar, pal]

At my father's funeral, the coffin was carried by six pal bears.
At my father's funeral, the coffin was carried by six pal bearers.
[bears, bearers]

One of my greatest pleasures in life is to have a giraffe of wine with a good meal.
One of my greatest pleasures in life is to have a carafe of win with a good meal.
[giraffe, carafe]

If someone makes veiled suggestions about me I always ask them what they are incinerating.
If someone makes veiled suggestions about me, I always ask them what they are insinuating.
[incinerating, insinuating]

I think the law is too laxative on criminals.
I think the law is too lenient on criminals.
[laxative, lenient]

I do a bit of sailing on the river and I always keep a boat tied up at the dwarf.
I do a bit of sailing on the river and I always keep a boat tied up at the wharf.
[dwarf, wharf]

I may not be very good at grammar, but at least I can puncture a sentence.
I may not be very good at grammar, but at least I can punctuate a sentence.
[puncture, punctuate]

My son went to visit his grammar during vacation.
My son went to visit his grandma during vacation.
[grammar, grandma]

My son is training to be a doctor with the hopes of becoming a sturgeon.
My son is training to be a doctor with the hopes of becoming a surgeon.
[sturgeon, surgeon]

I'd love to bag a few peasants on my shooting holiday in Scotland.
I'd love to bag a few pheasants on my shooting holiday in Scotland.
[peasants, pheasants]

I find it very convenient to buy eggs in a cartoon.
I find it very convenient to buy eggs in a carton.
[cartoon, carton]

I always hope for the best - I'm an eternal octopus.
I always hope for the best - I'm an eternal optimist.
[octopus, optimist]

I'm planning to take up art classes - so I have bought myself a weasel.
I'm planning to take up art classes - so I have bought myself an easel.
[weasel, easel]

I don't want my cat to have anymore kittens so I'm having it sprayed.
I don't want my cat to have anymore kittens so I'm having it spayed.
[sprayed, spayed]

I'm thinking of paying a visit to the chiropractor because I have a bazooka on my foot.
I'm thinking of paying a visit to the chiropractor because I have a bunion on my foot.
[bazooka, bunion]

The circulatory system contains the veins, the archeries, and the capillaries.
The circulatory system contains the veins, the arteries, and the capillaries.
[archeries, arteries]

Bodies are sent to a mortuary to be mortgaged.
Bodies are sent to a mortuary to be embalmed.
[mortgaged, embalmed]

Many organs in the body have ducks leading from them.
Many organs in the body have ducts leading from them.
[ducks, ducts]

Some people break out in spots if they eat shellfish. They have a terrible
allegory.
Some people break out in spots if they eat shellfish. They have a terrible allergy.
[allegory, allergy]

My doctor couldn't make up his mind about my condition so he insulted a specialist.
My doctor couldn't make up his mind about my condition so he consulted a specialist.
[insulted, consulted]

When I was a little girl I was intoxicated against polio.
When I was a little girl, I was inoculated against polio.
[intoxicated, inoculated]

When I was a little girl I was inoculated against polo.
When I was a little girl, I was inoculated against polio.
[polo, polio]

If you get too wet you can die of ammonia.
If you get too wet you can die of pneumonia.
[ammonia, pneumonia]

It was so quiet in the house you could hear a mouse dropping.
It was so quiet in the house, you could hear a pin dropping.
[mouse, pin]

I check prices at my local supermarket I'm amazed at the way baked beans flatulate.
I check prices at my local supermarket I'm amazed at the way baked beans fluctuate.
[flatulate, fluctuate]

World War II was an event unparalysed in human history.
World War II was an event unparalleled in human history.
[unparalysed, unparalleled]

My little nephew is always getting throat infections, so he's going to the hospital to have his asteroids removed.
My little nephew is always getting throat infections, so he's going to the hospital to have his adenoids removed.
[asteroids, adenoids]

I'm very glad to hear the miner's strike has been settled by holding a ballet at the pits.
I'm very glad to hear the miner's strike has been settled by holding a ballot at the pits.
[ballet, ballot]

I couldn't guess what I was getting for my birthday, but I worked it out by a
process of illumination.
I couldn't guess what I was getting for my birthday, but I worked it out by a process of elimination.
[illumination, elimination]

I always boil water before drinking it in order to putrefy it.
I always boil water before drinking it in order to purify it.
[putrefy, purify]

General Washington sent the calvary up the hill into battle.
General Washington sent the cavalry up the hill into battle.
[calvary, cavalry]

My uncle was an old soldier who rose to the rank of corpuscle.
My uncle was an old soldier who rose to the rank of corporal.
[corpuscle, corporal]

When I'm writing a letter I like to leave a one inch virgin all around the page.
When I'm writing a letter I like to leave a one inch margin all around the page.
[virgin, margin]

I'm very concerned about the damage done by topical typhoid storms.
I'm very concerned about the damage done by tropical storms.
[topical typhoid, tropical]

People often try to make an escape coat out of me for things I haven't done.
People often try to make an escape goat out of me for things I haven't done.
[coat, goat]

I've been reading about a dreadful disease called sleeping sickness brought on by the bite of the sexy fly.
I've been reading about a dreadful disease called sleeping sickness brought on by the bite of a tsetse fly.
[sexy, tsetse]

I'm going to get some insect spray to get rid of the aunts in my house.
I'm going to get some insect spray to get rid of the ants in my house.
[aunts, ants]

My grandfather was in the infamy during the war.
My grandfather was in the infantry during the war.
[infamy, infantry]

I won't eat any food that has conservatives in it.
I won't eat any food that has preservatives in it.
[conservatives, preservatives]

The philosopher I admire most is Pluto, the ancient Greek.
The philosopher I admire most is Plato, the ancient Greek.
[Pluto, Plato]

I watch every space shuttle launch from Cape Carnival.
I watch every space shuttle launch from Cape Canaveral.
[Carnival, Canaveral]

Catholicism and Prostitution are the two main religions in America.
Catholicism and Protestantism are the two main religions in America.
[Prostitution, Protestantism]

My daughter has just walked down the isle to get married.
My daughter has just walked down the aisle to get married.
[isle, aisle]

I have just cut my hand on a silver of glass.
I have just cut my hand on a sliver of glass.
[silver, sliver]

All of the guests threw graffiti at the bride and groom.
All of the guests threw confetti at the bride and groom.
[graffiti, confetti]

My husband has bought me a waist disposal system.
My husband has bought me a waste disposal system.
[waist, waste]

My little grandson has to go for therapy because he has a speech predicament.
My little grandson has to go for therapy because he has a speech impediment.
[predicament, impediment]

We should always be on our guard against illnesses because some viruses can lie doormat for years.
We should always be on our guard against illnesses because some viruses can lie dormant for years.
[doormat, dormant]

The guests were invited to the wedding conception after the ceremony.
The guests were invited to the wedding reception after the ceremony.
[conception, reception]

Shakespeare wrote tragedies, comedies, and hysterectomies.
Shakespeare wrote tragedies, comedies, and histories.
[hysterectomies, histories]

As one of the three witches said, "Bubble, Bubble, toilet trouble".
As one of the three witches said, "Bubble, Bubble, Toil and trouble."
[toilet, toil and]

I'd like to live to be a hundred-years old and become a centurion.
I'd like to live to be a hundred-years old and become a centenarian.
[centurion, centenarian]

I excuse my verbal lapses by saying I am metamorphically speaking.
I excuse my verbal lapses by saying I am metaphorically speaking.
[metamorphically, metaphorically]

When I was younger I was hoping to marry a rich typhoon.
When I was younger I was hoping to marry a rich tycoon.
[typhoon, tycoon]

The chateau my husband stayed in during the war had a French widow in every bedroom.
The chateau my husband stayed in during the war had a French window in every bedroom.
[widow, window]

My son has gone to Europe but I hope to be incommunicado with him soon.
My son has gone to Europe but I hope to be in communication with him soon.
[incommunicado, in communication]

My husband is studying for a doctorate and soon he will submit his doctoral faeces.
My husband is studying for a doctorate and soon he will submit his doctoral thesis.
[faeces, thesis]

I went to the doctor and he told me I had sick as hell anemia.
I went to the doctor and he told me I had sickle cell anemia.
[sick as hell, sickle cell]

One of the most perplexing religious doctrines is that of the emasculated conception.
One of the most perplexing religious doctrines is that of the Immaculate Conception.
[emasculated, immaculate]

One of the most perplexing religious doctrines is that of the immaculate deception.
One of the most perplexing religious doctrines is that of the Immaculate Conception.
[deception, conception]

My favorite fast food is hamburglars.
My favorite fast food is hamburgers.
[hamburglars, hamburgers]

Some of my favorite flowers are coronations.
Some of my favorite flowers are carnations.
[coronations, carnations]

All three of my teenage boys have acme on their faces.
All three of my teenage boys have acne on their faces.
[acme, acne]

I'd love to have a quaint old house with ivory growing up the walls.
I'd love to have a quaint old house with ivy growing up the walls.
[ivory, ivy]

My husband gave me a 1-carrot diamond for our anniversary.
My husband game me a 1-carat diamond for our anniversary.
[carrot, carat]

When you eat, your food passes through your elementary canal.
When you eat, your food passes through your alimentary canal.
[elementary, alimentary]

Nothing should be taken for granite when writing a doctoral paper.
Nothing should be taken for granted when writing a doctoral paper
[granite, granted]

My brother has a job at the vegetable market at a celery of $30,000.
My brother has a job at the vegetable market at a salary of $30,000.
[celery, salary]

One of the highlights of my trip to Russian was a visit to the gremlin.
One of the highlights of my trip to Russia was a visit to the Kremlin.
[gremlin, Kremlin]

The first time I saw a snake I was absolutely putrefied.
The first time I saw a snake I was absolutely petrified.
[putrefied, petrified]

My father was illegitimate, he couldn't read or write.
My father was illiterate, he couldn't read or write.
[illegitimate, illiterate]

I admit I'm pretty fat, but I don't agree with my doctor when he says I'm obeast.
I admit I'm pretty fat, but I don't agree with my doctor when he says I'm obese.
[obeast, obese]

I don't know how to move my prawns when playing chess.
I don't know how to move my pawns when playing chess.
[prawns, pawns]

I was looking for legal representation so I picked a lawyer at ransom from the phone book.
I was looking for legal representation so I picked a lawyer at random from the phone book.
[ransom, random]

When I went to Egypt, I saw the stinks in the desert.
When I went to Egypt, I saw the sphinx in the desert.
[stinks, sphinx]

When I went to Egypt, I saw the sphinx in the dessert.
When I went to Egypt, I saw the sphinx in the desert.
[dessert, desert]

I had a severe inflection until the doctor gave me a shot of antibiotics.
I had a severe infection until the doctor gave me a shot of antibiotics.
[inflection, infection]

I had a severe infection until the doctor gave me a shot of peninsula.
I had a severe infection until the doctor gave me a shot of penicillin.
[peninsula, penicillin]

Criminals are usually castrated in prison.
Criminals are usually incarcerated in prison.
[castrated, incarcerated]

I really miss the company of my recently diseased husband.
I really miss the company of my recently deceased husband.
[diseased, deceased]

Grease is just a spot on the map.
Greece is just a spot on the map.
[Grease, Greece]

Canada is sparsely copulated.
Canada is sparsely populated.
[copulated, populated]

Some South American countries are bandana republics.
Some South American countries are banana republics.
[bandana, banana]

The four seasons are salt, pepper, vinegar, and mustard.
The four seasonings are salt, pepper, vinegar, and mustard.
[seasons, seasonings]

Japanese girls dress in commodes.
Japanese girls dress in kimonos.
[commodes, kimonos]

Columbus circumcised the world in a forty-foot clipper.
Columbus circled the world in a forty-foot clipper.
[circumcised, circled]

I'm very afraid of floods, earthquakes, and other catechisms of nature.
I'm very afraid of floods, earthquakes and other catastrophes of nature.
[catechisms, catastrophes]

I saw a very exciting film the other night where a caveman was attacked by a giant thesaurus.
I saw a very exciting film the other night where a caveman was attacked by a giant tyrannosaurus.
[thesaurus, tyrannosaurus]

I need to watch my carbohydrates since I am diabolic.
I need to watch my carbohydrates since I am diabetic.
[diabolic, diabetic]

At my wedding I carried a bouquet of my favorite flowers - enemas.
At my wedding I carried a bouquet of my favorite flowers - anemones.
[enemas, anemones]

My favorite books to read are science friction.
My favorite books to read are science fiction.
[friction, fiction]

My husband goes to the golf curse every Saturday morning to play.
My husband goes to the golf course every Saturday morning to play.
[curse, course]

If I were on television I bet I would win an Enema Award,
If I were on television I bet I would win an Emmy Award.
[Enema, Emmy]

The snack I like best is cream cheese with a beagle.
The snack I like best is cream cheese with a bagel.
[beagle, bagel]

I have to go to the hospital to get a sex ray.
I have to go to the hospital to get a x-ray.
[sex ray, x-ray]

I love putting explanation marks at the end of a sentence.
I love putting exclamation marks at the end of a sentence.
[explanation, exclamation]

My cousin is a sealiac and has to have a glutton free diet.
My cousin is a celiac and has to have a glutton free diet.
[sealiac, celiac]

My father was injured during the war when he was hit with sharpnel.
My father was injured during the war when he was hit with shrapnel.
[sharpnel, shrapnel]

Reference List

Ackerman, E. (2014). Can winograd schemas replace turing test for defining human-level AI? Retrieved September 28, 2014, from http://spectrum.ieee.org/automaton/robotics/artificial-intelligence/winograd-schemas-replace-turing-test-for-defining-humanlevel-artificial-intelligence/?utm_source=roboticsnews&utm_medium=email&utm_c

Agostaro, F., Augello, A., Pilato, G., Vassallo, G., & Gaglio, S. (2005). A conversational agent based on a conceptual interpretation of a data driven semantic space. In S. Bandini & S. Manzoni (Eds.), *Proceedings of the 9th Congress of the Italian Association for Artificial Intelligence, Milan, Italy, September 21-32, 2005.* (pp. 381–392). Milan, Italy: Springer. doi:10.1007/11558590_39

Augello, A., Pilato, G., & Gaglio, S. (2010). An intelligent advisor to suggest strategies in economic policy decisions. In *2010 International Conference on Complex, Intelligent and Software Intensive Systems* (pp. 734–739). IEEE. doi:10.1109/CISIS.2010.75

Augello, A., Pilato, G., Vassallo, G., & Gaglio, S. (2009). A semantic layer on semi-structured data sources for intuitive chatbots. In *International Conference on Complex, Intelligent and Software Intensive Systems* (pp. 760–765). IEEE. doi:10.1109/CISIS.2009.165

Augello, A., Scriminaci, M., Gaglio, S., & Pilato, G. (2011). A modular framework for versatile conversational agent building. In *2011 International Conference on Complex, Intelligent, and Software Intensive Systems* (pp. 577–582). IEEE. doi:10.1109/CISIS.2011.95

Augello, A., Vassallo, G., Gaglio, S., & Pilato, G. (2008). Sentence induced transformations in “Conceptual” spaces. In *2008 IEEE International Conference on Semantic Computing* (pp. 34–41). IEEE. doi:10.1109/ICSC.2008.74

Baisely, W. E. (2000). Wayne’s Spoonerism Page. Retrieved June 01, 2013, from http://web.archive.org/web/20000816044617/http://www-oss.fnal.gov/~baisley/spooners.html

Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The Semantic Web. *Scientific American*, *284*(5), 34–43. doi:10.1038/scientificamerican0501-34

Bitter, C., Elizondo, D. a., & Yang, Y. (2009). Natural language processing: a prolog perspective. *Artificial Intelligence Review*, *33*(1-2), 151–173. doi:10.1007/s10462-009-9151-4

Boden, C., Fischer, J., Herbig, K., & Spierling, U. (2006). CitizenTalk: application of chatbot infotainment to e-democracy. *Technologies for Interactive Digital Storytelling and Entertainment*, 370–381.

Bolshakov, I. A. (2005). An experiment in detection and correction of malapropisms through the web. In A. Gelbukh (Ed.), *Proceedings of the 6th International Conference on Intelligent Text Processing and Computational Linguistics, CICLing 2005* (pp. 803–815). Mexico City, Mexico: Springer Berlin / Heidelberg. doi:10.1007/b105772

Bolshakov, I. A., & Gelbukh, A. (2003a). On detection of malapropisms by multistage collocation testing. In *Proceedings of the 8th Int. Conference on Application of Natural Language to Data Bases, NLDB-2003* (pp. 28–41). Burg, Germany.

Bolshakov, I. A., & Gelbukh, A. (2003b). Paronyms for accelerated correction of semantic errors. *International Journal on Information Theories & Applications*, *10*, 198–204.

Breese, J. S., & Heckerman, D. (1996). Decision-theoretic case-based reasoning. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, *26*(6), 838–842. doi:10.1109/3468.541343

Brian, P. (2008). *Common Errors in English Usage: The Book* (2nd ed., p. 304). Sherwoord, OR.: William, James & Company.

Budanitsky, A., & Hirst, G. (2001). Semantic distance in WordNet : An experimental , application-oriented evaluation of five measures. In *Workshop on WordNet and Other Lexical Resources, Second Meeting of the North American Chapter of the Association for Computational Linguistics* (pp. 29–34). Pittsburgh, PA.

Budanitsky, A., & Hirst, G. (2006). Evaluating WordNet-based measures of lexical semantic relatedness. *Computational Linguistics*, *32*(1), 13–47. doi:10.1162/coli.2006.32.1.13

Chiru, C., Cojocaru, V., Rebedea, T., & Trausan-Matu, S. (2010). Malapropisms detection and correction using a paronyms dictionary, a search engine and Wordnet. In *Proceedings of the Fifth International Conference on Software and Data Technologies, ICSOFT 2010, Volume 2.* (pp. 364–373). Athens, Greece: INSTICC.

Cho, A., & Chun, R. (2007). Emotion & domain concept enhancements to Alicebot. In H. Arabnia, M. Yang, & J. Yang (Eds.), *Proceedings of the 2007 International Conference on Artificial Intelligence, ICAI 2007, Volume II, June 25-28, 2007, Las Vegas, Nevada, USA* (pp. 852–858). Las Vegas: CSREA Press.

Coursey, K. (2004). *Living in CyN: Mating AIML and CyC Together with Program N. English*.

De Maio, C., Fenza, G., Loia, V., & Senatore, S. (2010). Knowledge structuring to support facet-based ontology visualization. *International Journal of Intelligent Systems*, *25*(12), 1249–1264. doi:10.1002/int.20451

Deryugina, O. V. (2010). Chatterbots. *Scientific and Technical Information Processing*, *37*(2), 143–147. doi:10.3103/S0147688210020097

Dickerson, R., Johnsen, K., Raij, A., & Lok, B. (2005). Evaluating a script-based approach for simulating patient-doctor interaction. In *Proceedings of the International Conference of HumanComputer Interface Advances for Modeling and Simulation (2005)* (pp. 79–84).

Dikker, S., Silbert, L. J., Hasson, U., & Zevin, J. D. (2014). On the same wavelength: Predictable language enhances speaker-listener brain-to-brain synchrony in posterior superior temporal gyrus. *Journal of Neuroscience*, *34*(18), 6267–6272. doi:10.1523/JNEUROSCI.3796-13.2014

Fellbaum, C. (Ed.). (1998). *WordNet: An Electronic Lexical Database (Language, Speech, and Communication)* (p. 423). Cambridge, MA.: The MIT Press.

Freese, E. (2007). Enhancing AIML bots using semantic web technologies. In *Proc. of Extreme Markup Languages* (pp. 1 – 27). idealliance.

Ghose, S., & Barua, J. J. (2013). Toward the implementation of a topic specific dialogue based natural language chatbot as an undergraduate advisor. In *2013 International Conference on Informatics, Electronics and Vision (ICIEV)* (pp. 1–5). Dhaka: IEEE. doi:10.1109/ICIEV.2013.6572650

Goranson, H. (2005). Semantic distance and enterprise integration. In P. Bernus & M. Fox (Eds.), *Knowledge Sharing in the Integrated Enterprise* (Vol. 183, pp. 39–52). Boston: Springer Boston. doi:10.1007/0-387-29766-9_4

Graesser, A. C. (2011). Learning, thinking, and emoting with discourse technologies. *The American Psychologist*, *66*(8), 746–57. doi:10.1037/a0024974

Graesser, A. C., & Jackson, G. T. (2006). Applications of human dialogue tutoring to AutoTutor: An intelligent tutoring system. *Revista Signos: Linguistic Studies*, *39*(60), 31–48. doi:10.4067/S0718-09342006000100002

Graesser, A. C., Wiemer-Hastings, K., Wiemer-Hastings, P., & Kreuz, R. (1999). AutoTutor: A simulation of a human tutor. *Cognitive Systems Research*, *1*(1), 35–51. doi:10.1016/S1389-0417(99)00005-4

Heller, B., Procter, M., & Mah, D. (2005). Freudbot: An investigation of chatbot technology in distance education. In P. Kommers & G. Richards (Eds.), *Proceedings of World Conference on Educational Multimedia, Hypermedia and Telecommunications* (pp. 3913–3918). Montreal, Canada: ED-MEDIA.

Hessami, E., Mahmoudi, F., & Jadidinejad, A. H. (2011). Unsupervised graph-based Word Sense disambiguation using lexical relation of WordNet. *IJCSI International Journal of Computer Science*, *8*(6), 225–230.

Hirst, G., & Budanitsky, A. (2005). Correcting real-word spelling errors by restoring lexical cohesion. *Natural Language Engineering*, *11*(1), 87–111. doi:10.1017/S1351324904003560

Hirst, G., & St-onge, D. (1998). Lexical chains as representations of context for the detection and correction of malapropisms. In C. Fillbaum (Ed.), *WordNet: An Electronic Lexical Database* (pp. 305–332). Cambridge, MA.: Massachustts Institute of Technology.

Hossain, S. A., Rahman, A. S. M. M., Tran, T. T., & Saddik, A. El. (2010). Location aware question answering based product searching in mobile handheld devices. In *2010 IEEE/ACM 14th International Symposium on Distributed Simulation and Real Time Applications* (pp. 189–195). IEEE. doi:10.1109/DS-RT.2010.28

Hubal, R. C., Fishbein, D. H., Sheppard, M. S., Paschall, M. J., Eldreth, D. L., & Hyde, C. T. (2008). How do varied populations interact with embodied conversational agents? Findings from inner-city adolescents and prisoners. *Computers in Human Behavior*, *24*(3), 1104–1138. doi:10.1016/j.chb.2007.03.010

Islam, A., & Inkpen, D. (2009). Real-Word spelling correction using Google Web 1T 3-grams. In *EMNLP '09 Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3* (pp. 1241–1249). Singapore: Association for Computational Linguistics.

Jiang, J., & Conrath, D. (1997). Semantic similarity based on corpus statistics and lexical taxonomy. In *Proceedings of the International Conference Research on Computational Linguistics (ROCLING X).* (pp. 1–15). Taiwan.

Jurafsky, D., & Martin, J. (2000). *Speech and Language Processing: An Introcuction to Natural Language Processing, Computational Linguistics, and Speech Recognition* (1st ed.). Englewood Cliffs, NJ: Prentice Hall.

Kerly, A., Ellis, R., & Bull, S. (2009). Conversational agents in E-Learning. In T. (Nottingham T. U. Allen, R. (Stratum M. L. Ellis, & M. (University of G. Petridis (Eds.), *Applications and Innovations in Intelligent Systems XVI* (pp. 169–182). London: Springer-Verlag London. doi:10.1007/978-1-84882-215-3_13

Kerly, A., Hall, P., & Bull, S. (2007). Bringing chatbots into education: towards natural language negotiation of open learner models. *Knowledge-Based Systems*, *20*(2), 177–185. doi:10.1016/j.knosys.2006.11.014

Kolodner, J. L. (1992). An introduction to case-based reasoning. *Artificial Intelligence Review*, *6*(1), 3–34. doi:10.1007/BF00155578

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159–174. doi:10.2307/2529310

Leacock, C., & Chodorow, M. (1998). Combining local context and WordNet sense similarity for word sense disambiguation. In C. Fellbaum (Ed.), *WordNet, An Electronic Lexical Database* (pp. 265–283). Cambridge, MA.: The MIT Press.

Lee, C.-S., Jian, Z.-W., & Huang, L.-K. (2005). A fuzzy ontology and its application to news summarization. *IEEE Transactions on Systems, Man, and Cybernetics. Part B, Cybernetics : A Publication of the IEEE Systems, Man, and Cybernetics Society*, *35*(5), 859–80. doi:0.1109/TSMCB.2005.845032

Lesk, M. (1986). Automatic sense disambiguation using machine readable dictionaries. In *Proceedings of the 5th annual international conference on Systems documentation - SIGDOC '86* (pp. 24–26). New York, New York, USA: ACM Press. doi:10.1145/318723.318728

Lin, D. (1989). An information-theoretic definition of similarity. In *Proceedings of the 15th international conference on Machine Learning (Vol. 1)* (pp. 296–304).

Lundqvist, K. O., Pursey, G., & Williams, S. (2013). Design and implementation of conversational agents for harvesting feedback in eLearning systems. In *Scaling up Learning for Sustained Impact* (pp. 617–618). Springer Berlin / Heidelberg. doi:10.1007/978-3-642-40814-4_79

MacHale, D. (2006). *A Decapitated Coffee Please & Other Great Malapropisms*. Douglas Village, Cork: Merier Press.

malapropism. (2013). Retrieved January 03, 2013, from http://www.britannica.com

Manning, C., & Schutze, H. (1999). *Foundations of Statistical Natural Language Processing*. Cambridge, MA.: The MIT Press.

Mascari, G. F., Maniscalco, U., Moltedo, L., Moscati, P., Pilato, G., Pitolli, L., Toffoli, G. (2010). Adaptive semantics of complex information/services networks: A case study from cultural heritage. In *2010 International Conference on Complex, Intelligent and Software Intensive Systems* (pp. 1105–1110). Krakow, Poland: IEEE. doi:10.1109/CISIS.2010.125

Mays, E., Damerau, F. J., & Mercer, R. L. (1991). Context based spelling correction. *Information Processing & Management*, *27*(5), 517–522. doi:10.1016/0306-4573(91)90066-U

McCormick, D. (2014). Virtual tween passes turing test. Retrieved September 28, 2014, from http://spectrum.ieee.org/tech-talk/robotics/artificial-intelligence/virtual-tween-passes-turing-test

McMahan, B. (Minnesota S. U. (2010). An automatic dialog system for student advising. *Journal of Undergraduate Research*, *10*, 1–12.

Mihalcea, R., Corley, C., & Strapparava, C. (2005). Corpus-based and knowledge-based measures of text semantic similarity. *North*, *21*(1), 775–780.

Mihalcea, R., & Csomai, A. (2005). SenseLearner: word sense disambiguation for all words in unrestricted text. In *Proceedings of the ACL 2005 on Interactive poster and demonstration sessions - ACL '05* (pp. 53–56). Morristown, NJ, USA: Association for Computational Linguistics. doi:10.3115/1225753.1225767

Mikic, F. a., Burguillo, J. C., Rodriguez, D. a., Rodriguez, E., & Llamas, M. (2008). T-Bot and Q-Bot: A couple of AIML-based bots for tutoring courses and evaluating students. In *2008 38th Annual Frontiers in Education Conference* (pp. S3A–7–S3A–12). Ieee. doi:10.1109/FIE.2008.4720469

Mikic Fonte, F. A., Rial, Juan, C. B., & Nistal, M. L. (2009). TQ-Bot : An AIML-based tutor and evaluator bot. *Journal Of Universal Computer Science*, *15*(7), 1486–1495.

Moore, R., & Gibbs, G. (2002). Emile: Using a chatbot conversation to enhance the learning of social theory. *Univ. of Huddersfield, Huddersfield, England*.

Morales-Rodríguez, M., González, J. B., Juárez, R. F., Huacuja, H. F., & Flores, J. M. (2010). Emotional conversational agents in clinical psychology and psychiatry. In G. Sidorov, H. Aguirre, & C. Arturo and Reyes García (Eds.), *Advances in Artificial Intelligence* (pp. 458–466). Berlin / Heidelberg: Springer Berlin / Heidelberg. doi:10.1007/978-3-642-16761-4_40

Navigli, R., & Lapata, M. (2010). An experimental study of graph connectivity for unsupervised word sense disambiguation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *32*(4), 678–92. doi:10.1109/TPAMI.2009.36

Neves, A. M. M., Barros, F. A., & Hodges, C. (2006). iAIML: A mechanism to treat intentionality in aiml chatterbots. In *2006 18th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'06)* (pp. 225–231). IEEE. doi:10.1109/ICTAI.2006.64

Norman, P. (1985). *Your Walrus Hurt The One You Love*. Long Acre London: Garden House.

Norvig, P., & Russel, S. (2010). *Artificial Intelligence: A Modern Approach* (3rd ed.). Upper Saddle River, NJ: Prentice Hall.

Patwardhan, S., & Pedersen, T. (2006). Using WordNet-based context vectors to estimate the semantic relatedness of concepts. In *Proceedings of the EACL 2006 Workshop on Making Sense of Sense: Bringing Computational Linguistics and Psycholinguistics Together* (pp. 1–8). Trento, Italy.

Pedersen, T., & Kolhatkar, V. (2009). WordNet :: SenseRelate :: AllWords - A broad coverage word sense tagger that maximizes semantic relatedness. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Demonstration Session* (pp. 17–20). Boulder, Colorado: Association for Computational Linguistics.

Pedersen, T., & Michelizzi, J. (1998). WordNet :: Similarity - Measuring the relatedness of concepts. In *HLT-NAACL--Demonstrations '04 Demonstration Papers at HLT-NAACL 2004* (pp. 38–41). Association for Computational Linguistics.

Pedler, J. (2007). *Computer Correction of Real-word Spelling Errors in Dyslexic Text. ReCALL*. Citeseer.

Pilato, G., Augello, A., & Gaglio, S. (2011). A modular architecture for adaptive chatBots. In *2011 IEEE Fifth International Conference on Semantic Computing* (pp. 177–180). IEEE. doi:10.1109/ICSC.2011.68

Pilato, G., Augello, A., Vassallo, G., & Gaglio, S. (2007). Sub-symbolic semantic layer in cyc for intuitive chat-bots. In *International Conference on Semantic Computing (ICSC 2007)* (pp. 121–128). IEEE. doi:10.1109/ICSC.2007.37

Pilato, G., Augello, A., Vassallo, G., & Gaglio, S. (2008). Sub-Symbolic knowledge representation for evocative chat-bots. In *INTERDISCIPLINARY ASPECTS OF INFORMATION SYSTEMS STUDIES* (Vol. 7, pp. 343–349). Physica-Verlag HD. doi:10.1007/978-3-7908-2010-2_42

Pilato, G., Pirrone, R., & Rizzo, R. (2008). A kst-based system for student tutoring. *Applied Artificial Intelligence*, *22*(4), 283–308. doi:10.1080/08839510801972785

Pilato, G., Vassallo, G., Augello, A., Vasile, M., Gaglio, S., & Dipartimento, D. (2004). Expert chat-bots for cultural heritage. In *IX Convegno della Associazione Italiana Intelligenza Artificiale Proc. of. Workshop Interazione e Comunicazione Visuale nei Beni Culturali* (p. 15). Perugia, Italy.

Pirrone, R., Pilato, G., Rizzo, R., & Russo, G. (2007). Semantics driven interaction using natural language in students tutoring. In *Knowledge-Based Intelligent Information and Engineering Systems* (Vol. 4694, pp. 720–727). Springer Berlin / Heidelberg. doi:10.1007/978-3-540-74829-8_88

Pothuru, R. K. (2003). *Agent-based architecture for web deployment of multi-agents as conversational interfaces*. University of North Texas.

Rada, R., Mili, H., Bicknell, E., & Blettner, M. (1989). Development and application of a metric on semantic nets. *Ieee Transactions On Systems Man And Cybernetics*, *19*(1), 17–30. doi:10.1109/21.24528

Reid, E., Qin, J., Chung, W., & Xu, J. (2004). Terrorism knowledge discovery project: a knowledge discovery approach to addressing the threats of terrorism. In *Intelligence and Security Informatics Second Symposium on Intelligence and Security Informatics, ISI 2004, Tucson, AZ, USA, June 10-11, 2004. Proceedings* (pp. 125–145). Tucson: Springer Berlin / Heidelberg. doi:10.1007/978-3-540-25952-7_10

Resnik, P. (1995). Using information content to evaluate semantic similarity in a taxonomy. In C. Mellish (Ed.), *IJCAI-95 - Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence* (Vol. 1, pp. 448–453). Montreal, Canada.

Schumaker, R., Ginsburg, M., Chen, H., & Liu, Y. (2007). An evaluation of the chat and knowledge delivery components of a low-level dialog system: The AZ-ALICE experiment. *Decision Support Systems*, *42*(4), 2236–2246. doi:10.1016/j.dss.2006.07.001

Schumaker, R., & Chen, H. (2007). Leveraging question answer technology to address terrorism inquiry. *Decision Support Systems*, *43*(4), 1419–1430. doi:10.1016/j.dss.2006.04.007

Schumaker, & Chen, H. (2010). Interaction analysis of the ALICE chatterbot: A two-study investigation of dialog and domain questioning. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, *40*(1), 40–51. doi:10.1109/TSMCA.2009.2029603

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys*, *34*(1), 1–47. doi:10.1145/505282.505283

Shawar, B. A., & Atwell, E. (2007). Fostering language learner autonomy through adaptive conversation tutors. In *Proceedings of the The fourth Corpus Linguistics conference*. Birmingham, England: University of Birmingham.

Sinha, R., & Mihalcea, R. (2007). Unsupervised graph-basedword sense disambiguation using measures of word semantic similarity. In *International Conference on Semantic Computing (ICSC 2007)* (pp. 363–369). Ivine, CA: IEEE. doi:10.1109/ICSC.2007.87

Soliman, M., & Guetl, C. (2013). Implementing intelligent pedagogical agents in virtual worlds: Tutoring natural science experiments in OpenWonderland. In *2013 IEEE Global Engineering Education Conference (EDUCON)* (pp. 782–789). Berlin: IEEE. doi:10.1109/EduCon.2013.6530196

Szarvas, G., Vincze, V., Farkas, R., Móra, G., & Gurevych, I. (2012). Cross-genre and cross-domain detection of semantic uncertainty. *Computational Linguistics*, *38*(2), 335–367. doi:10.1162/COLI_a_00098

Toseland, M. (2007). *The Ants Are My Friends*. London: Portico Books.

Turing test success marks milestone in computing history. (2014). Retrieved September 28, 2014, from http://www.reading.ac.uk/news-and-events/releases/PR583836.aspx

Veletsianos, G., Heller, R., Overmyer, S., & Procter, M. (2010). Conversational agents in virtual worlds: Bridging disciplines. *British Journal of Educational Technology*, *41*(1), 123–140. doi:10.1111/j.1467-8535.2009.01027.x

Wallace, R. S. (2003). *The Elements of AIML Style*. ALICE A. I. Foundation, Inc.

Wallace, R. S. (2009). The anatomy of A.L.I.C.E. In R. Epstein, G. Roberts, & G. Beber (Eds.), *Parsing the Turing Test* (pp. 181–210). Springer Berlin / Heidelberg.

Wallace, R. S. (2010). Superbot - The easy way to create your own custom bot. Retrieved June 22, 2014, from http://www.alicebot.org/superbot.html

Wang, K., Thrasher, C., & Viegas, E. (2010). An overview of Microsoft Web N-gram corpus and applications. *Proceedings of the NAACL …*, (June), 45–48.

Weizenbaum, J. (1966). ELIZA - a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, *9*(1), 36–45. doi:10.1145/365153.365168

Weizenbaum, J. (1967). Contextual understanding by computers. *Communications of the ACM*, *10*(8), 474–480. doi:10.1145/363534.363545

What Are Malaopropisms. (n.d.). Retrieved from http://www.fun-with-words.com/mala_explain.html

What is the Loebner Prize? (2013). Retrieved September 08, 2013, from http://www.loebner.net/Prizef/loebner-prize.html

Wilcox-O'Hearn, A., Hirst, G., & Budanitsky, A. (2008). Real-word spelling correction with trigrams: A reconsideration of the Mays, Damerau, and Mercer model. In *9th International Conference*, *CICLing 2008, Haifa, Israel, February 17-23, 2008* (Vol. 4919, pp. 605–616). Springer Berlin / Heidelberg. doi:10.1007/978-3-540-78135-6_52

Wu, Z., & Palmer, M. (1994). Verbs semantics and lexical selection. In *Proceedings of the 32nd annual meeting on Association for Computational Linguistics* - (pp. 133–138). Morristown, NJ, USA: Association for Computational Linguistics. doi:10.3115/981732.981751