Natural Language Generation with Markov Chains and Grammar
TJHSST Senior Research Project
Computer Systems Lab 2009-2010

Sam Zhang

June 15, 2010

Abstract

Language can be generated stochastically using Markov Chain databases. In first order Markov Chain generation, a test corpus is analyzed for the probabilistic weights between each word and the word following it. Then a text is generated randomly by following these weights. This project explores the use of a second-order Markov Chain in conjunction with a semantic similarity model to create a de facto grammar such that the output has similar diction, syntax, and subjectively perceived tone as the source text, yet still different. **Keywords:** natural language generation, markov chain, grammar, computational linguistics, semantics

1 Introduction

By combining a corpus-based method of generating texts with semantic similarity, the accuracy of generated texts could perhaps be improved. To create a comprehensible text, this project uses a corpus-based method to gleam the basic grammatical rules and vocabulary, then using those, to piece together a new text.

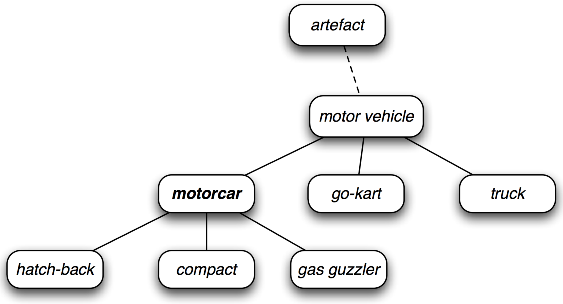

Figure 1: Fig 1: A hypernym/hyponym ontology. Source: http://nltkbook.googlecode.com

The ontological approach to semantics is relatively new compared to other linguistical studies, yet is already burgeoning due to its relevancy to contemporary semiotic environments, namely the internet, which is a complex ontology in itself. The use of the internet as a wealth of corpora is not a new idea, marketing research companies and sociologists have been using data mining techniques to determine online trends for years. Director of the World Wide Web Consortium Sir Tim Berners-Lee is strongly advocating the Semantic Web as the next step for the internet's evolution, in other words, Web 3.0. In the semantic web, the internet will become fundamentally interlinked with artificial intelligence, providing an automated source of information and help. The W3C endorses a new language for the internet to replace HTML, most notably the Web Ontology Language.

Existing word sense disambiguators have already reached a high degree of accuracy in statistical corpora analysis, for example in part-of-speech tagging. However, this project focuses on disambiguation on a lexical level, rather than sentential, to investiate the extremes to which a word's denotation can still be found. By forcing the computer to discern between subleties in word meaning using a minimum of context, this author will use a technque philosophically outlined by a book titled, "Ontological Semantics", which envisioned a network-based approach to semantics to augment or even replace the traditional statistical corpora method.

Part-of-speech tagging, if used with a machine learning method, can be potentially combined to give the Markov Chain generated text a grammatical structure. The difficulty lies in deciding which grammatical structure to use, and programming them into the system. Perhaps a similarly Markov-like grammatical reaping can be made on the harvest that is the corpus.

2 Development

2.1 Theory

Existing word sense disambiguators have already reached a high degree of accuracy in statistical corpora analysis, for example in part-of-speech tagging. However, this project focuses on disambiguation on a lexical level, rather than sentential, to investiate the extremes to which a word's denotation can still be found. By forcing the computer to discern between subleties in word meaning using a minimum of context, this author will use a techinque philosophically outlined by a book titled, "Ontological Semantics", which envisioned a network-based approach to semantics to augment or even replace the traditional statistical corpora method. Through ontological semantics, the least common hypernym can be determined between different senses of different words, and thus determining the most similar through an algorithm related to the hypernym distance and root distance. Natural Language Generation has progressed much since the millenium, but some argue that corpus-based methods remain as potent and relevant as ever (3). In a corpus-based method like the Markov Chain, each word is entered into a database with the subsequent word as an attached value. The more two words are paired, the more they are linked in the database. Thus to generate a text in the style of spam artists, one needs to jump from word to word randomly, with the database itself automatically organized by frequency. However, this creates a problem wherein the grammar is nonexistent and punctuation haphazard. To resolve this, this project will implement a grammatical structure to limit the possible continuations in the Markov Chain.

2.2 Languages and Modules

The author uses Python as the primary programming language, the natural language toolkit which facilitates corpora manipulation, and the Wordnet corpora, an ontology of common English words. Through a heuristical search across the lexical ontology of the words using such traversions as hyponymy, hypernymy, synonymy, antonymy, melonymy, and holonymy, the computer will be able to discover the average semantic distance from one word to all of the others. Perhaps this approach will be supplemented with a statistical analysis of corpora, as the Wordnet database of lexical relations could benefit with an update. The OpenCyc project is such a corpora that attempts to update with the current semiotic sphere, although some concerns have been raised by the computational linguistics community on its accuracy and scalability.

For the text generation, various corpora are used, and the author plans to set up an automatic text generator by attaching it with the internet and blog feeds. Currently, the most successful texts have been generated using various inaugural addresses throughout the history, for within the output one can see many of the semiotic constructs popular in the day.

2.3 Description of Project

This project employs semantic ontologies to perform dynamic word sense disambiguation on a lexically isolated set of words with only each other as reference. The program will heuristically search through the semantic net, an ontology of lexical relations, to determine the "distance" between the referents of the different words to determine a relevancy index for each word, dynamically based on its environment. The word with the smallest relevancy index is marked as the "odd one out", or the error. The ontological algorithm used traces the words upstream to their least common hypernym, and compares the distance of the least common hypernym to the root. This number is factored together with the distance of the words to each other (i.e. the average distance to the least common hypernym). With an index relating the similarity between two words, a metric now exists for large scale comparison between different sets of different words.

Then this project generates a text using similarity culled statistically from a text. By using a stochastic method, the original vocabularity of a corpus is maintained. The process is split into two parts: entry and generation. In entry, each word is entered into a database with the subsequent word entered as its value. In this case, the Markov Chain is second order, meaning that two words follow each entry, making for a more accurate rendition of the original text. In generation, the words are randomly picked from the previous words values. As the entry portion was not controlled for repetition, the frequency was self-controlled.

2.4 Tests or Analysis

The strength of the project's ontological traversal will be measured in two aspects: accuracy and time. The accuracy will be the aspect most focused on intially, while optimizations can occur after the accuracy is ensured. As a prototyping step, a graph could be created of the semantic network as the program conducts the bidirectional search. Currently, as per the Wu and Palmer algorithm that this project's algorithm is based on, the project has an 88/100 accuracy rating for determing the most similar words out of a pool, as related to human judgment.

The text generation lacks a grammar at this stage, but at a second order Markov Chain it still creates biting parodies of input corporas. Here are key inaugural addresses, recreated via this program, truncated after 100 words, throughout American history:

2005 – George W. Bush: In the truths of Sinai , the questions that come to those who feel its power , it burns those who feel its power , it burns those who feel its power , and the varied faiths of our soldiers . You have seen together . For as long as whole regions of the United States to seek and support the growth of democratic movements and institutions in every generation by reaffirming all that is the concentrated work of helping raise up free governments K the idealistic work of generations . The great objective of ending tyranny is the

1789 – George Washington: Than side no local prejudices or attachments , no separate views nor party animosities , will misdirect the comprehensive and equal eye which ought to be peculiarly conscious of his own deficiencies . In these honorable qualifications I behold the surest pledges that as on one side no local prejudices or attachments , no separate views nor party animosities , will misdirect the comprehensive and equal eye which ought to be suppressed . You will join with me , my error will be more consistent with those circumstances , and , in executing this task , I trust , in

1861 – Abraham Lincoln: Perpetual I do not choose now to specify particular acts of Congress as proper to be extended . This is the only substantial dispute . The people themselves can do this . Think , if not expressed , in the best way all our constitutional controversies , and not in any interior locality shall be delivered up " their oaths are unanimous . Now , if you can not remove our respective sections from each other nor build an impassable wall between them . But such is not our case . All members of Congress swear their support to the

1941 – FDR (Third Term): Those count of years , but by the count of years , but mostly plain people , who sought here , early and late , to find freedom more freely . The coordinate branches of the world . And yet we all understand what it is – the spirit of America , and in the Capital of the people of this democracy . For there is also the spirit of America , and to perpetuate the integrity of democracy . For action has been , and the mind , constricted in an alien world , lived on , the people

1937 – FDR (Second Term): Them these conditions of effective government shall be created and maintained . They will demand that these conditions of effective government shall be created and maintained . They hold out the clear hope that government within the separate States , and government of the chaos which followed the Revolutionary War ; they are making their country a good neighbor among the nations in its example of the United States can do the things the times require , without that aid , we have come far from the ideal ; and we will not listen to Comfort , Opportunism , and

1809 – James Madison: May for the past , as I trust , on any unwarrantable views , nor with large ones safe ; to liberate the public debts ; to foster a spirit of independence too just to invade the rights or the functions of religion , so wisely exempted from civil jurisdiction ; to maintain sincere neutrality toward belligerent nations ; to preserve in their full energy the other salutary provisions in behalf of private and personal rights , and to the advancement of its highest interest and happiness . But the source to which I am to tread lighted by examples

1801 – Thomas Jefferson: Trusted public order as his own personal concern . Sometimes it is proper you should understand what I deem the essential principles of our sages and blood of our felicities . About to enter , fellow citizens , unite with one heart and one mind . Let us restore to social intercourse that harmony and affection without which liberty and even life itself are but dreary things . And may that Infinite Power which rules the destinies of the right of election by the sword of revolution and reformation . The approbation implied by your suffrage is a great consolation

3 Discussion

Statistical corpora analysis is a computational linguistics tool that could augment the strictly ontological dynamic word-sense disambiguation. This author has experimented with streaming RSS feeds into corpora with moderate success, but has since discovered the OpenCyc ontology. The OpenCyc ontology contains a semiotic map of the current census reality of the English speaking world, but clearly such an ambitious project must have shortcomings. To develop RSS feed data into a lexical relations data, a simple parsing script can be written to find syntactical structures such as ”WORD1 is... of WORD2” to determine holonyms. Yet this author's research has uncovered the importance of the hypernym and hyponym structure to lexical similarity, to the point where the other lexical relations are insignificant. Thus the OpenCyc ontology could be used to determine such facts as whether or not George W. Bush and Barack Obama are more similar than Sarah Palin (which perhaps they are since they are both of the hypernym ”US Presidents”, although George W. Bush and Sarah Palin are both Republicans).

The statistical method of natural language generation has proven to be effective to a degree, reusing the original vocabulary in innovative way but not always in concordance with grammatical rules. Thus the next step is to construct a grammatical structure for the text generation to fit into. Perhaps a way to statistically determine this could be done in a similar fashion; by surveying grammatical structures in a corpus. To achieve this, a method of storing grammars would be necessary, as would part-of-speech tagging. As this is not the focus of this project, WordNet may provide some useful built in functions toward this goal.

4 Recommendations

Aside from its advantageous position at the intersection between consciousness studies and artificial intelligence, computational linguistics has a wide arrange of practical applications, from translation to intelligent computing. Specifically, ontological semantics are being researched as the backbone for Web 3.0, the movement to give the internet basic aspects of intelligence. That could produce a resounding impact on our life, like Asimov's MultiVac, so the ethical issues must be carefully analyzed. On a closer timeline, computational semantics is becoming closer related to neuroscience as scientists move to discover the neurological development of linguistic faculties. This project specifically would be crucial to search engine development, computer dictionaries, human-computer interaction, and the back-end of a speech-to-text filter. The next step for this project is the transition.

Other texts that could provide information once regenerated are political blogs, scientific texts, and internet dialog corpora.

5 Sources

C. Fellbaum, editor. WordNet: An Electronic Lexical Database (Language, Speech, and Communication). The MIT Press, 1998.

D. Fiaer. Using multilingual resources for building SloWNet faster. In Proc. 4th International WordNet Conference (GWC), Szeged, Hungary, 2008.

Feldman, J.A. 2004, "Computational cognitive linguistics"

A. Marchetti, M. Tesconi, F. Ronzano, M. Rosella, and F. Bertagna. Toward an architecture for the Global Wordnet initiative. In Proc. 3rd Italian Semantic Web Workshop, SWAP 2006. CEUR-WS.org, 2006.

M. MarszaBek and C. Schmid. Semantic hierarchies for visual object recognition. In IEEE Conference on Computer Vision Pattern Recognition, June 2007.

Reuters. Reuters Corpus, vol. 1: English Language, 1996-08-20 to 1997-08-19, 2000.

S. Niremburg, V. Raskin. 2004, "Ontological semantics"