QuickSum

Robert Wahlstedt

Quicksum

By Robert Wahlstedt

Abstract

Quick Summary is an innovate implementation of an automatic document summarizer that inputs a document in the English language and evaluates each sentence. The scanner or evaluator determines criteria based on its grammatical structure and place in the paragraph. The program then asks the user to specify the number of sentences the person wishes to highlight. For example should the user ask to have three of the most important sentences, it would highlight the first and most important sentence in green. Commonly this is usually the sentence containing the conclusion. Then Quick Summary finds the second most important sentence usually called a satellite and highlights it in yellow. This is usually the topic sentence. Then the program finds the third most important sentence and highlights it in red. The implementations of this technology are useful in a society of information overload when a person typically receives 42 emails a day (Microsoft). Another implication is meeting summary information from video for peace officer records and sports broadcasters. The paper also is a candid look at difficulty that machine learning has in literal textural translating. However, it speaks on how to overcome the obstacles that historically prevented progress. This research is different from other research attempts because it takes into account heuristics in the paragraph and treats the document as not just a list of disjointed sentences but also each sentence contributing meaning to other sentences until it achieves a pivotal point in document. Prior methods of a document summary generator included reducing redundant words or junction words until it develops a nuclear core. This paper proposes mathematical metadata criteria that justify the place of importance of a sentence. Just as tools for the study of relational symmetry in bioinformatics, this tool seeks to classify words with greater clarity. "Survey Finds Workers Average Only Three Productive Days per Week." Microsoft News Center. Microsoft. Web. 31 Mar. 2012.

Introduction

A piece of literature is similar to a piece of music. While the BBC has a recording of Tchaikovsky’s 1812 overture that lasts 17 minutes, most people remember only the theme. In this blog, we discuss what makes literature themes stand out amongst the rest of the piece through repetition, supporting variations of the theme. Many textbooks of text mining explain the author’s beliefs that word categories or parts of speech are important. These papers find summaries with the hope that by deleting irrelevant words until only the nuclear core is remaining. However, these authors overlook the importance of context. It is important to use Latent Semantic Analysis. This seeks to use structural knowledge of pragmatics. For example in order to get to the best diagnosis of a patient, it is important to take into account their medical histories and not just a few quotes. This method analyses the entire document before concluding so metaphors and words with multiple meanings are not confused. This blog proposes that by using metadata, or data about the data, achieves the goal of distinguishing repetition

of the theme and supporting variations also known as satellites. This is achievable by finding the roots of words through their origins, translating the word prefix, root, and suffix into metadata.

Designing with the Brain in Mind

In designing a program, it is necessary to understand what one has to work with. Imagine a machine with two cameras perceiving words. When a person reads a word like “unbuttoning” the human brain strips off the “un” and the “ing” and it is left with two morphemes, “but” and “ton”. There is a priming effect where a person is more inclined to think about a certain flow of words. For example, after casa, there sometimes is the word blanca like the movie. This is why we understand differing homograph meanings. A grapheme is a letter or a series of letters that map a phoneme in the target language. Imagine a web of words that links all words in existence. In the book, The Neuroscience of Language introduces a word-related functional web. It is a synchronous firing chain or synfire chain. At first glance, automata might appear to be similar to this web. However, a computer differs from a brain because even the multi-core processors are unicore in the sense that only a few threads at a time can be processed unified by a central process. The human brain is a democracy having processes happening simultaneously and the strongest impulse, usually by a collective number of neurons fires that is what the human brain thinks. Miguel Nicolelis argues in his book, Beyond Boundaries, there is not a separate part of the brain language support. However, a certain sequence of neurons that light up when a person reads about an object because it invokes an emotion. The book, The First Word, describes how non-humans animals could use simple language. Instances of non-humans are territorial warning signs and primitive inner-species communication. The article “How Dolphins Say Hello” says that dolphins only whistle when they are in a group of dolphins and identify themselves through echolocation.

This raises the question of how grammar comes about. A grammar being a set of rules in a language that allows us to communicate and understand. There is a debate between B.F. Skinner and Noam Chomsky. Skinner believes that a person learns language through association, the sight of things along with the sound of the word. This is reinforced to match conform to a dialect. Chomsky believes that a Martian scientist observing children in a single-language community would observe that language is very similar regardless of the culture and therefore innate. Although reading is unnatural, our brain wants to see patterns and group objects together. We recognize lines and shapes that we can recognize characters. We then group these characters into morphemes. We then put these morphemes into phrases, idioms, words, and sentences. These sentences form a paragraph. Similar to computers recognition is easy, recalling meaning is difficult.

Similar Research

Text mining projects at Universitat Politecnica de Catalunya depend on lexical categories for their lexical analysis tagging. In the book When you Catch and Adjective, Kill It suggests “According to grammarians, adjectives, nouns, verbs, and injunctions are considered “open” parts of speech because they shift functions … and because new words are continually added to their ranks. This proves that classification of words is not always necessary.” (Yagoda, 43). To make matters worse, a gerund is using a verb in its –ing form as a noun such as living. Although the English language has synoptically, there are two types of

words, content words and structure words. Structure words can be eliminated such as “hope that” “clearly” “strangely” ”indeed” ”conceivably” ”seriously” ”ultimately” ”theoretically” ”naturally” ”ironically” ”fortunately” ”incidentally” do not containing meaning in themselves however they send signals to listeners about words that are coming later. The word “like” is overused. The National Science Foundation did a study that demonstrated when a storyteller used the word “this” instead of “a” or “an” the person had better retention. Another concern is that these systems do not account for pragmatics that is the study of context is it situational.

Another criticism is that English is a moving target

English as a moving target: English is a West Germanic language that borrows its alphabet from Latin, adding three characters J, U, and W. During the Norman Conquest of England the French brought over a variation of Latin that morphed into English. Languages morph over time and words take on other meanings. For example, consider the common phrase Santa Claus. It began as “Sant Herr Hiclaes” in Dutch, transforming to “Santerclaes”, and eventually became the English “Santa Clause” today (McWhorter The Power of Babel, page 29). John Dryden admitted to translating his works to Latin to get the syntax to flow smoother (Lynch, 36). The two types of people who study are classified as prescriptive grammar and descriptive grammar. Scholars describe Samuel Johnson, the writer of A Dictionary of the English Language, as a descriptive grammar lexicographer. He realized there could be no establishment that could enforce grammar. Printers at the time noted that they could sell to more people their books if there was a standard way of speaking. Later, the success of The Oxford English dictionary was based because there were correspondence between the editors and the public. Today is no different. French President Sarkozy received a concerned note saying that the recipient was worried that French was on a downward spiral linguistically and Sarkozy should do something (Greene). China has the “Law of the People's Republic of China on the Standard Spoken and Written Chinese” that requires media personal and broadcasters a certain level of speaking proficiency. This ban has been lifted as of December 24th 2011. Linguists believe that all languages are related. Instead of being different “languages”, they are dialects of one common language that originated from the Persian Gulf.

Ambiguity of what is a Summary

A recent survey showed that if surveyors present a passage of text to a human subject they are likely to disagree on what passage should be highlighted as most important. This is because the subjects are biased by culture and project their preconceived notions on the passage. This shows the benefits of an artificial intelligence system that notices pronouns that a person who is caught up in emotions fails to recognize. In a study by James Pennebaker at the University of Texas Austin, he describes how natural language processors can detect who might win an election. In his book, the Hidden Life of Pronouns he accounts of how aids told John Kerry to use the pronoun “I” less. Should a person be of the same political party as John Kerry, they might overlook the fact that he was using words in this format. Linguists call this connotation lexical semantics, the study of what words denote.

How to Address Critics who question the standard of English

The Language Wars by Henry Hitchings discusses how it is amazing that we decided to uniformly spell words the way we do. Before computers he points out that even dictionaries were inconsistent. There was a survey with a computer where words were spelled phonetically and 50 percent of the words did not agree with the phonetic spelling. Steven Pickner tells us that 84 percent of the words are spelled in a way that conforms to patterns we can notice. As shown by these statistics, conformability is possible.

My Proposal for Metadata

It is necessary to study of metadata to correlate etymology with words. Etymology is the study of finding word roots and morphology is the study of word making. Today advances in bioinformatics tools have yielded a complex study of the human genome and have made great strides in finding out mutations that can create a higher probability of cancer. We should do so because we can learn about the human race as well as expand our knowledge of words. Here are some examples: nickname originally came from nekename containing the compound eke and name. Eke came from ēac that means also. It means an additional name. Hobby comes from the word hob and yn. Hob or hŏb means a threaded and fluted hardened steel cutter, resembling a tap, used in a lathe for forming the teeth of screw chasers, worm wheels (Webster's Revised Unabridged Dictionary) and yn means "to be". It means to spin time. Omelet comes from the French word omelette. Omelette comes from alemette that comes from alemelle that comes from la lemelle that means a thin plate like structure. We should know how to find the prefix, root, and suffix because we are able to understand why someone in history responded to situations. The written word often transcends a person's lifespan and the culture changes. A written manuscript can be likened to a fossil. For example, rabbit comes from the words robète that means to steal. This explains why Mr. McGregor thinks Peter Rabbit as such a nuisance. If you are a person that is a Christian, it might please you to know that Nicodemus comes from the word nincompoop. This comes from the Latin phase non-compos mentis meaning not of sound mind. People regard him as a hero of the Christian faith because he walked by faith and not by knowledge.

What does a computer have to do with this? Computers are able are machines that are set to carry out arithmetic or logical operations such as regular expressions or pattern matching. Given a rules file and a lexicon file, they are able to study morphology.

Examples of Metadata

A as in rag or Prague is big and i as in pin is small, e is somewhere in between. There has been some big vowel shifting in England between 1350 and 1500 known as the big vowel change so words like big and small have the wrong letters. Little and large are right. Pimpf means "a little boy", pimple means a little swelling, and pampers means a lot. Servant is in between.

Questions for Further Research

Although this project covers many bases, it is important to remind ourselves of what a sentence is. A sentence is as a flatbed truck here is nothing of value until different pieces by the author to the pile. These pieces consist of verbs, nouns, and conjunctions that are all linked together. There are different types of verbs including sensory verbs that express information an author receives from their senses.

There are action verbs that describe an action, for example, can you … for me. There are infinitive that such as “to be”. These words just hold the sentence together. The subject of the sentence is the one driving the truck. A sentence is made out of declarative that is a declaration that make statements. Interrogative is a question and are sometimes rhetorical. Imperative is a request or demand some action such as a how-to. Exclamations express strong emotions. Sentences can be either negative or positive based on logic in the sentence. Sentences are either active or passive. An active one is happening right now, and the passive occurred some time ago. Bearing this in mind we can question what to do about coordinate conjunctions. This is separate from having a comma in there because the phrases on the two sides of a semicolon are able to stand independently and complement each other often to contrast things. However, like a company, one is like a boss and the other one is a subordinate conjunction or subordinators. A noun class consists of a subject or predicate normative. There might also be a relative adjective clause. Sentences with semicolons are sometimes referred to as swinging gates. Also a consideration to error correction of the paragraph should be taken. Can the program shave off propositional phrases? How can the program spot and shave off introduction words that do not add meaning of a sentence and are before a comma. What about paragraphs that are not well formed and have preposition stranding or dangling infinitives or a dangling modifier? What if there is intentional fragments? What if sentences are missing elements such as “a” or “the”. What about paragraphs that are not unified? These do not revolve around one topic. What about meandering sentences? These could be long, rambling sentences for a powerful sensory experience. This is an intentional run-on-sentence.

Works Cited

Ciccarelli, Saundra K., and Glenn E. Meyer. *Psychology*. Upper Saddle River, NJ: Pearson Prentice Hall, 2006. Print.

Dehaene, Stanislas. *Reading in the Brain: The Science and Evolution of a Human Invention*. New York: Viking, 2009. Print.

Feldman, Ronen, and James Sanger. *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*. Cambridge: Cambridge UP, 2007. Print.

Greene, Robert L. *Ou Are What You Speak: Grammar Grouches, Language Laws, and the Politics of Identity*. Delacorte, 2011. Print.

Hitchings, Henry. *The Language Wars: A History of Proper English*. New York: Farrar, Straus and Giroux, 2011. Print.

Johnson, Jeff. *Designing with the Mind in Mind: Simple Guide to Understanding User Interface Design Rules*. Amsterdam: Morgan Kaufmann/Elsevier, 2010. Print.

Kenneally, Christine. *The First Word: The Search for the Origins of Language*. New York: Viking, 2007. Print.

Liberman, Anatoly. *Word Origins-- and How We Know Them: Etymology for Everyone*. Oxford: Oxford UP, 2005. Print.

Lynch, Jack. *The Lexicographer's Dilemma: The Evolution of "proper" English, from Shakespeare to South Park*. New York: Walker &, 2009. Print.

McWhorter, John H. *The Power of Babel: A Natural History of Language*. New York: Times, 2001. Print.

Nicolelis, Miguel. *Beyond Boundaries: The New Neuroscience of Connecting Brains with Machines--and How It Will Change Our Lives*. New York: Times /Henry Holt and, 2011. Print.

Pennebaker, James W. *The Secret Life of Pronouns: What Our Words Say about Us*. New York: Bloomsbury, 2011. Print.

Prado, Hercules Antonio Do., and Edilson Ferneda. *Emerging Technologies of Text Mining: Techniques and Applications*. Hershey, PA: Information Science Reference, 2008. Print.

Pulvermüller, Friedemann. *The Neuroscience of Language: On Brain Circuits of Words and Serial Order*. Cambridge [etc.: Cambridge UP, 2007. Print.

Sebranek, Patrick, Verne Meyer, and Dave Kemper. *Write for College: A Student Handbook*. Wilmington, MA: Write Source, Great Source Education Group, 2007. Print.

Woods, Geraldine. *English Grammar for Dummies*. New York: Hungry Minds, 2001. Print.

Yagoda, Ben. *When You Catch an Adjective, Kill It: The Parts of Speech for Better And/or Worse*. New York: Broadway, 2006. Print.