Natural Language Processing with Grammar

Application

Sanaa Mironov

Computer Science Graduate Student

CMSC 676 Information Retrieval

University of Maryland, Baltimore County

0. Abstract

This paper will inspect the current literature review of how natural language process architectures with machine learning and highlight the information retrieval application in-app such as Grammarly. The task is to classify existing papers according to their information retrieval, such as content categorization, contextual extraction, sentiment analysis, document summarization.

I. Introduction

Writing is key to human self-expression. Whether it is an email, journal, article, text message, or social media post, it is communicated through written text. Sentiment analysis through Natural Language Processing (NLP) has become essential to speed up the writing process through applications like Grammarly, Google Doc, and Writefull, to name a few.

Natural Language Processing (NLP) is a subset of artificial intelligence designed to understand, interpret, and manipulate human language. It has many purposes like predictive text, text categorization, translation, grammar checking, or topic classification. [1]

Furthermore, machine learning (ML) applies algorithms that teach machines how to learn and improve from experience automatically without being explicitly programmed. [11] NLP using ML models can learn representations of language and semantical similarities from raw data.

In comparison to traditional statistical techniques, Information retrieval (IR) is the field that focuses on the structure, analysis, organization, storage, searching, and retrieval of information. IR analysis collections of documents to satisfy a user query. [4]

Traditional writing programs that use only grammar rules or spell check are fixed to only the elements within a sentence; these rules are limiting since language is tricky and prevents mistakes a writer could make. However, applications like Grammarly use machine learning with

a variety of natural language processing approaches. [5] The models look at the context of words and sentences to assess what is needed and, based on this, can give meaningful feedback.

For example, Grammarly spots that something is mistaken in a text. It often gives the author the option to make their text follow a writing convention such as email, academic, business, or creative. Once the convention is set, alternatives to errors within a sentence are given, making Grammarly a perfect application for writing enhancement. Language correctness is a challenge because of the messiness and ambiguity inherent to language that NLP tries to solve.

This paper explains a brief introduction to the natural language process and some approaches utilized in these models in the context of how Grammarly uses this to enhance writing.

II. Statistical VS Symbolic Approach within NLP

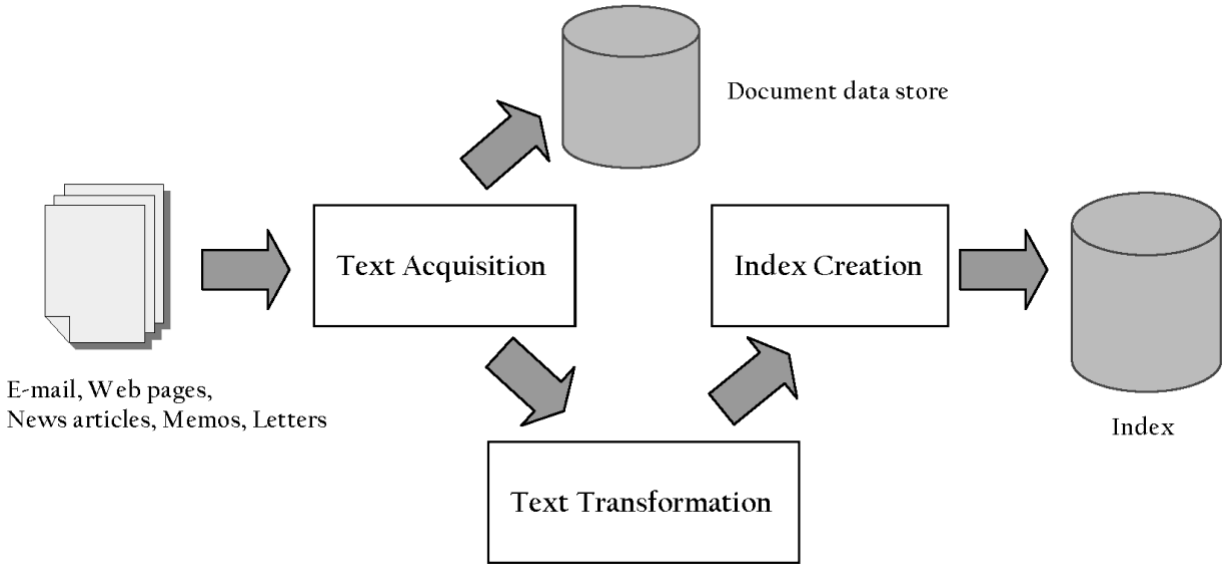
The natural language process approach falls into two categories based on the type of tasks it performs: symbolic and statistical.

- A statistical approach in NLP is similar to an information retrieval application and can use machine learning algorithms to learn language phenomena, but it has limitations. Some benefits of using machine learning algorithms for this task include control over the quality of training data, feature engineering, and creating new features.[6]
- On the other hand, a symbolic approach consists of rules in the language that model different language phenomena. [3] The Symbolic approach allows control over the change of rules by creating new ones and testing the results to figure out the errors with the picked rules.

III. Statistical Processing of Natural Language Process

Human language is complex; models like information retrieval algorithms define the vocabulary of terms before binding things together in the Natural Language process to make it more

meaningful. Information retrieval is a field concerned with the structure, analysis, organization, storage, searching, and retrieval of information." [4] The following section represents the classical model of information retrieval systems and is characterized from each document's set of keywords, known as the terms index.



[Image from slides of Dr.Pearce, and textbook][4]

The document processing model involves the following stages:

a) Document pre-processing: preparing the documents for parameterization by tokenizing each word within the document, removing stop words, stemming, or lemmatization.

b) Measures: Once relevant terms have been identified and weight has been added, it becomes time to index the tokens.

c) Applying learning algorithms: Once the processing of the document has been done, using a machine-learning algorithm to learn language phenomena to increase process time for the natural language processing model to learn.

Tokenization

Tokenization is the process of taking a whole document and separating the text into small chunks called tokens. The token can be words, character, or sub-word.

Stop Words Removal

Some words in the English language give little meaning or context to a sentence. [8] These words are called stop words and include language articles, pronouns, and prepositions such as "and," "the," "or." Removing relevant words allows analyzing words with more meaning. However, since there is no standard list of stop words, removing relevant information modifies the context in a given sentence.

Stemming

A suffix is a collection of letters placed at the end of a root word, and a prefix is a collection of letters placed before the root word resulting in a change to the meaning of the root word. The stemming process removes suffixes and prefixes from a root word and allows for similar words to match each other. [2] However, removing prefixes changes the meaning because suffixes can create or expand a word to form a new meaning. For example, happy and unhappy do not mean the same thing. A solution to this issue would be to create a list of common suffixes and prefixes that can be removed, resulting in keeping the original meaning of the root word.

Lemmatization

Lemmatization removes inflectional endings and returns the word to its dictionary form. [2] This process seems close to stemming. However, the approach to reverting to a root word is different and requires more computational power. The algorithm uses a detailed dictionary to figure out how to remove the inflectional endings. For example, words like "running," "runs," and "ran" are all forms of the word "run," so "run" is the root word to all the previous words. Another

difference between lemmatization and stemming is that it can discriminate between homonyms by taking part-of-speech parameters to a word. For example, the word "bat" can mean the animal or the wooden club used in baseball.

Bag of Words

A bag-of-words model tracks the occurrence of words within a document by counting all the words in a piece of text. Information regarding the word structure means the order of words has been removed because it is not relevant or needed for this type of analysis. [4] The model includes two things- creating a vocabulary of known words and measuring the presence of known words. The main drawback of a bag-of-words model is the absence of semantic meaning and context relating to the document. To solve this dilemma would be to rescale the frequency of words and how often they appear within the text. One approach would be to assign weights to each word within the text. This approach is called Term Frequency - Inverse Document Frequency (TFIDF). Term frequency is the method of awarding higher weight to rare words while giving a lower weight to familiar words. [4] However, bag-of-words with term frequency still lack semantics and context within the document because meaning is lost when the order of words is ignored.

Machine Learning Models

Machine learning (ML) applies algorithms that teach machines how to learn and improve from experience automatically without being explicitly programmed. [11] NLP using ML models can learn representations of language and semantical similarities from raw data. Some models that are being used to analyze the raw data imported from a corpus include:

1. Dictionary-based or Knowledge-based model

2. Supervised learning - these are learning models that make predictions based on labeled training data. Usually, the data is labeled by an expert in the domain.

3. Semi-supervised learning - Learning model that makes a prediction based on mixed data where some data is labeled while others are not.

4. Unsupervised learning - There is no output variable in unsupervised learning methods to guide the learning process, and algorithms explore data to find patterns.

IV. Linguistic Processing of Natural Language Processing

This approach is based on applying different techniques and rules that explicitly encode linguistic knowledge: symbolic, grammar-based, and rule-based rules to detect NPS.[10]

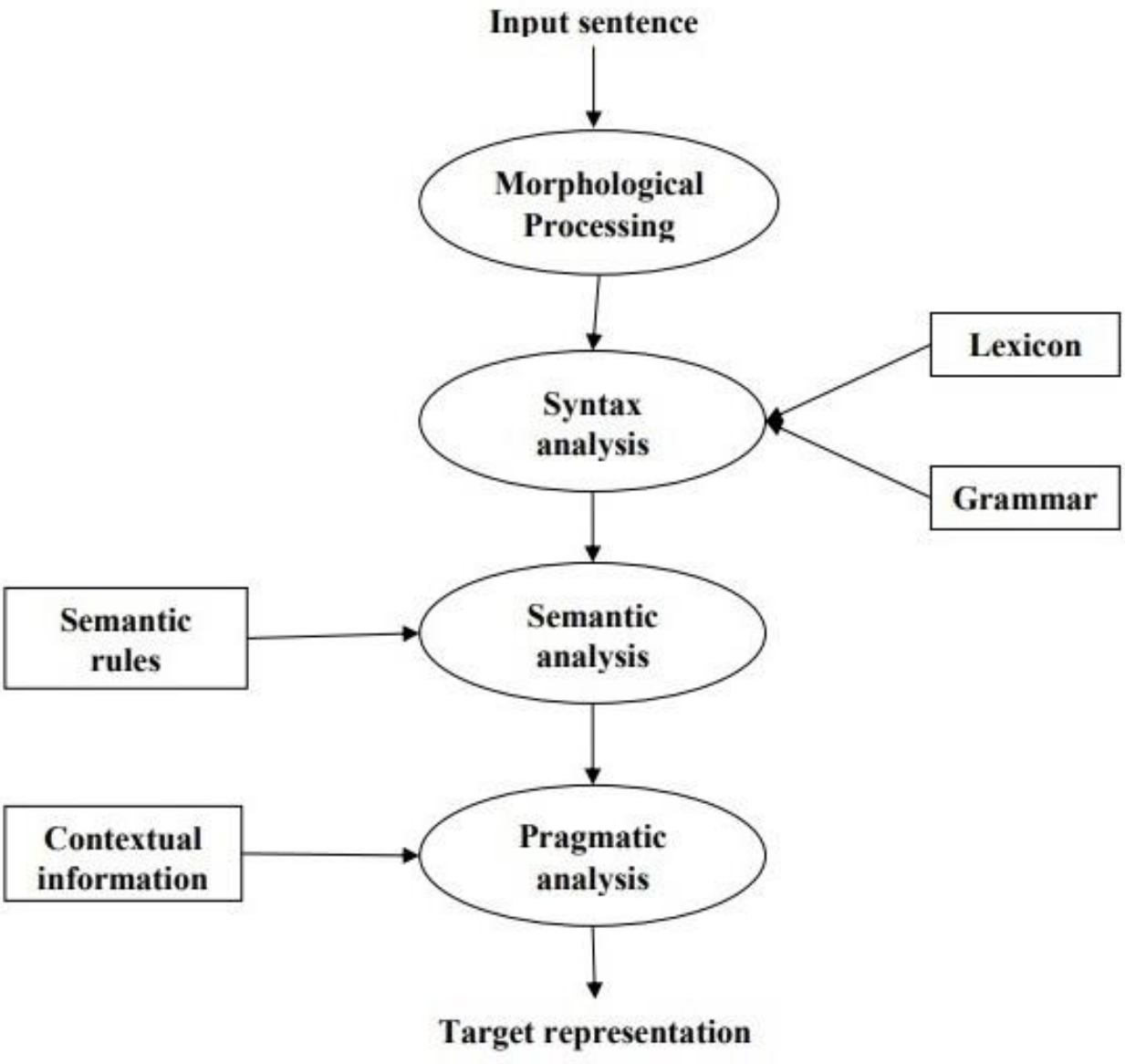
NLP models need to do three things in analyzing written text:

1) learn the recurrent patterns of texts;

2) recognize when the written language is not following the patterns;

3) suggest a change to the written language so that it follows the text patterns.

The text is analyzed through different linguistic levels. These levels include morphology, syntax, semantics, and pragmatics. [5]



Phase 1: Morphology

Morphology studies the form of words. The morphological analysis of a word assigns each word to a grammatical category.[9] Each word is characterized based on if it is a noun, verb, or adjective.

Phase 2: Syntax

The syntax of a language is the rules and the arrangement of words and phrases to create well-formed sentences in a language. [8] The syntax parser examines if a sentence is well-formed or not; sentences are broken into a structure to show the syntactic relationship between different words within the text.

Phase 3: Semantics

Semantics is the branch of linguistics dealing with the meaning of words, precisely the dictionary meaning of a word, phrase, or sentence. [8]

Semantic analyzers look for how to extract meaningfulness from the text. However, the complexity of semantic ambiguity within language causes many issues. Some of these issues with texts come from the polysemy, synonymy, antonymy, hyponymy, or hypernymy of words within the language.

This data can be learned from a corpus where the definition of each word is correctly identified. Lexical semantics deals with definitions of words as units, while compositional semantics examines how words merge to form more significant meanings. Word sense disambiguation (WSD) tries to distinguish the sense of a polysemic word in a given sentence.

Phase 4: Pragmatics

Pragmatics is the branch of linguistics dealing with language in how contexts of a text contribute to meaning - thinking of not what is said but how it is said, in context. [8] A pragmatic analyzer will choose between these possibilities when analyzing text to determine which interpretation is meant. One area that causes issues is the distinction between sarcasm and irony. To make the distinction between sarcasm and irony easier is to employ a classifier. This classifier would be trained on a data set that helps it form a distinction between the two based on word frequency or

other techniques. Not enough research has been done to analyze irony and sarcasm within a text to be feasible. Adding to that the fact that many human beings (including programmers) do not understand the concepts in the first place.

Ambiguity in NLP

The natural language process comes with many challenges when dealing with complex human language. The main issue is the ambiguity of language. Words can be spelled the same but with different meanings. These ambiguities include lexical, syntactic, semantic, and pragmatic. [8]

Deep Learning Model

Grammarly uses a mix of machine learning, natural language process, and deep learning models to make their product the best on the market. Deep learning algorithms or neural networks are built with multiple layers of interconnected neurons, emulating how the human brain works.[10] The input can be any of the classic machine learning algorithms like supervised, unsupervised, semi-supervised learning. However, Deep learning models usually perform better than other machine learning algorithms for complex problems like examining human language. It requires massive sets of data to make accurate predictions.

VI. Conclusion

This paper explains the natural language process and the different elements involved in various types of evaluation of human language in applications like Grammarly to enhance individual writing. I elaborated and compared the various processing techniques between statistical and linguistic processing in the natural language process. The statistical approach looked at probabilities of language phenomena. In comparison, a symbolic approach consists of rules in the language that model different language phenomena. Then looked at how these different

approaches use other subset methods of artificial intelligence like machine learning and deep learning to enhance their prediction.

Grammarly helps any person who is not naturally inclined to writing become a better writer. It does so by supplementing the person's abilities with natural language processing techniques.

References:

[1] Agarwal, S. (2017). Information Retrieval in Natural Language Processing. Retrieved February 2021, from https://medium.com/@shubhamagarwal328/information-retrieval-in-natural-language-processing-part-1-96c9a62fdb5f

[2] Balakrishnan, Vimala and Lloyd-Yemoh, Ethel (2014) *Stemming and lemmatization: A comparison of retrieval performances.* In: Proceedings of SCEI Seoul Conferences, 10-11 Apr 2014, Seoul, Korea. http://eprints.um.edu.my/13423/

[3] The Computer Journal, Volume 35, Issue 3, June 1992, Pages 268–278, https://doi.org/10.1093/comjnl/35.3.268

[4] Croft, W. Bruce, et al. Pearson Education, Inc., 2015.

[5] "How We Use AI to Enhance Your Writing: Grammarly Spotlight." Grammarly Spotlight: How We Use AI to Enhance Your Writing, 17 May 2019, www.grammarly.com/blog/how-grammarly-uses-ai/.

[6] Lewis, D., & Park, K. J. (1993, July). Natural language processing for Information Retrieval. Retrieved February, 2021, from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.14.1921&rep=rep1&type=pdf

~~[7] Mari Vallez; Rafael Pedraza-Jimenez. *Natural Language Processing in Textual Information Retrieval and Related Topics* [en linea]. "Hipertext.net", num. 5, 2007. http://www.hipertext.net~~

[8] Rindflesch, T. C., & Aronson, A. R. (1993). Semantic processing in information retrieval. *Proceedings. Symposium on Computer Applications in Medical Care*, 611–615.

[9] SANJUAN, E. (2018). Turing tests in Natural Language Processing and Information Retrieval. Retrieved February 13, 2021, from https://dev.termwatch.es/esj/turing_tests_esj.pdf

[10] Weiwei Guo, Huiji Gao, Jun Shi, and Bo Long. 2019. Deep Natural Language Processing for Search Systems. *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval* (*SIGIR'19*). Association for Computing Machinery, New York, NY, USA, 1405–1406. DOI:https://doi.org/10.1145/3331184.3331381

[11] Freitg Dayne. Information Extraction from HTML:: Application of a General Machine Learning Approach. AAAI-98 Proceedings