August 14, 2018

Human-Style Text Parsing System

Bokai Chen

Xin Guan

HUMAN-STYLE TEXT PARSING SYSTEM

Introduction

The present disclosure is directed to a system that can be used to more effectively determine the content of text in messages (e.g., messages sent in an encoded format including the Unicode Transformation Format) by overcoming the shortcomings of existing text parsing systems. Existing computing systems read text messages from known encoded formats and compare, find, search, match, and distinguish text messages (e.g., character strings or a list of characters) based on the encoded values. Further, existing text parsing systems (e.g., spam filters used to detect unsolicited e-mail messages) use various methodologies, including black lists and white lists, to identify the content of text in messages and/or to classify the content according to a classification scheme. For example, existing systems may use the encoded text values to classify a message as unwanted (e.g., spam messages, phishing attacks, or other malicious or unwanted messages) as wanted (e.g., a normal message from a trusted source).

However, such methodologies often results in a failure to detect non-standard words particularly when messages are deliberately obfuscated as in the case of spam messages or some advertisements. This is in contrast with humans, who parse text messages in a manner different from that of computing systems and can determine the meaning of messages based on the appearance of the text as well as the context in which words are couched. Some ways in which humans parse text messages include visually parsing the text into a language that the reader is familiar with.

For example, if two characters are encoded differently but are visually similar (e.g., the letter “O” and the number “0”), a human reader can parse the character based on the context in which the character is presented (e.g., “hell0” with the number “0” replacing the letter “O” can

be parsed as the greeting "hello"). By way of further example, the set of characters "Βríαη" may appear like the name "Brian", the character "‥" may look like a period "." but is actually the Unicode character U+2024, "ONE DOT LEADER", and the character "ⅼ" may look like the letter "l" but is actually the Unicode character U+217C, "SMALL ROMAN NUMERAL FIFTY". Additionally, when presented with the set of characters "{{a}d}" a human reader may ignore the parentheses. Furthermore, a human reader can use their imagination and understanding of culture to parse text. For example, a reader in the United States can determine that the word "Teddy" is a synonym for a bear as well as being a personal name.

In general, existing computing systems often struggle to decipher ambiguous or deliberately obfuscated terms which enable spam e-mails or scam advertisements to escape identification and filtering by existing encoding-based parsing schemes. The present disclosure is directed to overcoming these shortcomings by more effectively parsing text in messages.

Summary

The present disclosure proposes to solve the challenges described above by providing a text parsing system that is able to more effectively determine the content of input text and serve as part of a system that can be used to detect and identify content including spam, phishing attacks, or other unwanted message content. Specifically, the text parsing system can receive input (e.g., data including encoded text and/or an image that depicts text that was received in an encoded format) via a communication channel (e.g., a communications network connected to another computing system or a communications interconnection within the text parsing system). The input can be parsed by a parsing component of the text parsing system by performing optical character recognition (OCR) on the input (e.g., by performing OCR on an image that depicts the decoded text). The parsing component can include a machine-learned model (e.g., a

convolutional neural network) and/or a rules based system that can generate an output including lists of words and information associated with the words (e.g., probability that a word is correctly identified, the language associated with the word, and related words or terms). Furthermore, the machine-learned model can be generated based at least in part on training that uses data in a supervised or unsupervised way to reduce loss calculated by a loss functional over multiple iterations. In some embodiments, the output of words can be ordered based on the respective confidence associated with each word.

Furthermore, the text parsing system can use a knowledge base to generate related findings (e.g., words, score, language and type of word) for the information produced by the parsing component. That is, the parsing component can produce a list of possible matchings (e.g., matched words, confidence, language and/or type of word) ordered by confidence or score and the text parsing system can use a knowledge base (e.g., knowledge graph) to generate related findings (words, score, language and type of word) for the list of possible matchings. The output of the text parsing system can be helpful in identifying spam message or scam advertisements that are not detectable using existing de-spam systems. In particular, the textual content included in the input can be identified and/or classified according to a classification scheme (e.g., unwanted versus wanted) at least in part on the basis of the related findings generated from the knowledge base.

Thus, by mimicking human visual perception of text (e.g., rather than processing within the encoding domain) the text parsing system can better understand and classify textual content that has been crafted to evade traditional filtering systems but remain comprehensible to a human viewer.

Detailed Description

FIG. 1 illustrates a schematic diagram of one embodiment of a computing system **100** in accordance with aspects of the present disclosure. In the embodiment shown in **FIG. 1**, the computing system **100** includes a network **102**, a network **104**, one or more remote computing devices **110**, a computing system **120**, and a computing system **130**.

The network **102** and the network **104** can include any type of communications network, including a local area network (e.g., intranet), wide area network (e.g., Internet), or some combination thereof and can further include any number of wired or wireless links. Communication over the network **102** or the network **104** can occur via any type of wired and/or wireless connection, using a variety of communication protocols (e.g., TCP/IP, HTTP, SMTP, or FTP), encodings or formats (e.g., HTML, XML), and/or protection schemes (e.g., VPN, secure HTTP, SSL). The network **102** can for example be used to exchange signals or data between the one or more remote computing devices **110** and the computing system **120**. Furthermore, the network **104** can for example be used to exchange signals or data between the computing system **120** and the computing system **130**.

As shown in **FIG. 1**, the one or more remote computing devices **110** can include a controller **112** that includes one or more processors **114** and one or more memory devices **116** associated with the one or more processors **114**. The one or more processors **114** can include any suitable processing device, including as a microprocessor, microcontroller, integrated circuit, logic device, or other suitable processing device. Similarly, the one or more memory devices **116** can include one or more computer-readable media, including, but not limited to, non-transitory computer-readable media, RAM, ROM, hard drives, flash drives, and/or other memory devices. Furthermore, the one or more remote computing devices **110** can generate one or more

outputs including signals or data (e.g., message data) including information associated with one or more messages including text (e.g., encoded text).

The communications interface **118** can be used to communicate (e.g., send) signals or data to one or more systems (e.g., the computing system **120**). By way of example, the one or more remote computing devices **110** can generate message data including text formatted according to the UTF-8 standard that is then sent to the computing system **120** via the network **102** using the communications interface **118**. In general, the communications interface **118** can include one or more transmitters, receivers, ports, circuits, and other interfaces for communicating digital information over a wired communication link, wireless communication link, or combination of wired and wireless communication links. As an example, the communications interface **118** can communicate data via a wired and/or wireless protocol including Bluetooth, IEEE 802.11, and/or WiMAX.

As shown in **FIG. 1**, the computing system **100** includes the computing system **120** that can include a controller **122** that includes one or more processors **124** and one or more memory devices **126** associated with the one or more processors **124**. The one or more processors **124** can include one or more features of the one or more processors **114**. Further, the one or more memory devices **126** can include one or more features of the one or more memory devices **116**.

In some embodiments, the one or more memory devices **126** can store information accessible by the one or more processors **124**, including computer-readable instructions that can be executed by the one or more processors **124**. The instructions can be any set of instructions that when executed by the one or more processors **124**, cause the one or more processors **124** to perform operations. For instance, the instructions can be executed by the one or more processors **124** to determine and/or parse the content of text in data (e.g., message data including one or

more messages or text) sent from the one or more remote computing devices **110**. Further, the one or more processors **124** can include one or more features of the one or more processors **114**. The one or more memory devices **126** can also store data for manipulation by the one or more processors **124**. Furthermore, the computing system **120** can generate one or more outputs based in part on one or more signals or data received by the communication interface **128**. For example, the computing system **120** can generate output based on data (e.g., message data including encoded text) received from the one or more remote computing devices **110**.

Further as shown in **FIG. 1** the computing system **120** can also include the communications interface **128** that can be used to communicate (e.g., send and/or receive) one or more signals or data with one or more computing devices including the one or more remote devices **110** and/or the computing system **130**. Further, the communications interface **128** can include one or more features of the communications interface **118** of the one or more remote computing devices **110**.

In accordance with aspects of the present disclosure, the controller **122** can, in one embodiment, be configured to determine the content of data (e.g., message data received from the one or more remote computing devices **110**). In some embodiments, the computing system **120** can use a parsing component **129** to determine the content of the message data. Further, the parsing component **129** can include one or more machine-learned models (e.g., a convolutional neural network) and/or rules based systems (e.g., systems using one or more rules to determine an output based on an input) to determine the content of the message data. The parsing component **129** can also include or perform various other types of optical character recognition techniques.

As further shown in **FIG. 1**, the computing system **100** can include the computing system **130**. The computing system **130** can include a controller **132** that includes one or more processors **134** and one or more memory devices **136** associated with the one or more processors **134**. The one or more processors **134** can include one or more features of the one or more processors **114**. Further, the one or more memory devices **136** can include one or more features of the one or more memory devices **116**. In some embodiments, the one or more memory devices **136** can store information accessible by the one or more processors **134**, including computer-readable instructions that can be executed by the one or more processors **134**. The instructions can be any set of instructions that when executed by the one or more processors **134**, cause the one or more processors **134** to perform operations. For instance, the instructions can be executed by the one or more processors **134** to determine the content of message data received from the computing system **120**. The instructions can be implemented in any combination of hardware and/or software. When software is used, any suitable programming, scripting, or other type of language or combinations of languages can be used to implement the teachings contained herein. The one or more memory devices **136** can also store data for manipulation by the one or more processors **134**. Furthermore, the computing system **130** can generate one or more outputs based in part on one or more signals or data (e.g.,. message data) received by the communication interface **138**.

Further as shown in **FIG. 1** the computing system **130** can also include the communications interface **138** for communicating (e.g., sending and/or receiving) one or more signals or data with one or more systems (e.g., the computing system **120**). The communications interface **138** can include one or more features of the communications interface **118**.

In accordance with aspects of the present disclosure, the controller **132** can be configured to use a knowledge base **139** to determine information associated with data (e.g., message data including encoded text) based on associated information in the knowledge base. Further, the knowledge base can include a structured data repository represented as a graph in which one or more portions of information are associated with other portions of information in the graph.

Referring now to **FIG. 2**, a flow chart illustrating one embodiment of a process **200** for parsing text is illustrated in accordance with aspects of the present subject matter. The operations of the process **200** can be performed by a computing system including one or more features of the computing system **120** and/or the computing system **130** that are depicted in **FIG. 1**. Although the operations of the process **200** are shown and described in a particular order, certain operations can be performed in a different order or at the same time.

As indicated in **FIG. 2**, at **202**, message data (e.g., data including encoded text) is received (e.g., received by the computing system **120** from the one or more remote computing devices **110**). The data can be received via one or more communications networks (e.g., the communications network **102**) and can include data that is encoded in accordance with various encoding formats including UTF-8, UTF-16, UTF-32, and/or US-ASCII.

At **204**, the content of the message data (e.g., the message data including encoded text) can be determined. For example, the computing system **120** can determine the content of the message data using the parsing component **129** which can include one or more machine-learned models trained to determine the content of message data. By way of further example, the parsing component **129** can include a rules based system that can determine the content of message of data based on a set of rules (e.g., if-then rules) associated with portions of text (e.g., individual characters or strings).

In some embodiments, the message data can be used by a computing system associated with a knowledge base (e.g., the knowledge base **139** of the computing system **130** depicted in **FIG. 1**) that includes a structured data repository that can be represented as a graph. For example, the knowledge base can include information associated with a time interval, person, event, organization, place, or object. Further, the information in the knowledge base can be represented as a graph in which individual portions of information can be associated with one or more other portions of information. In some embodiments, the information in the knowledge base can be structured according to various criteria and can include metadata that can be used to cluster and organize constituent portions of the knowledge base.

At **206**, output data can be generated based on the content of the message data. For example, the computing system **120** can generate output data including a list of potentially matched words (e.g., a list including the encoded text and one or more corresponding potentially matched words associated with each portion of encoded text). Further, each of the potentially matched words in the list can be associated with a confidence value (e.g., a probability that the potentially matched word accurately conveys the content of the encoded text), a language (e.g., a language associated with the encoded text), and/or a type of word (e.g., the word "dog" can be associated with an animal type or pet type).

In some embodiments, a computing system including a knowledge base (e.g., the knowledge base **139** of the computing system **130**) can be used to generate further output data including associated words (e.g., the word "dog" can be associated with the words "puppy" or "man's best friend"), a score (e.g., a score associated with the likelihood that an associated word corresponds to the encoded text), a language (e.g., determining whether the encoded text is

English or Chinese), and a type of word (e.g., determining whether a word is associated with a time interval, person, event, organization, place, or object).

In some embodiments, the output data can be communicated to one or more computing systems including spam detection systems or advertisement detection systems that can use the output data to respectively detect the occurrence of spam or advertisements.

Figures

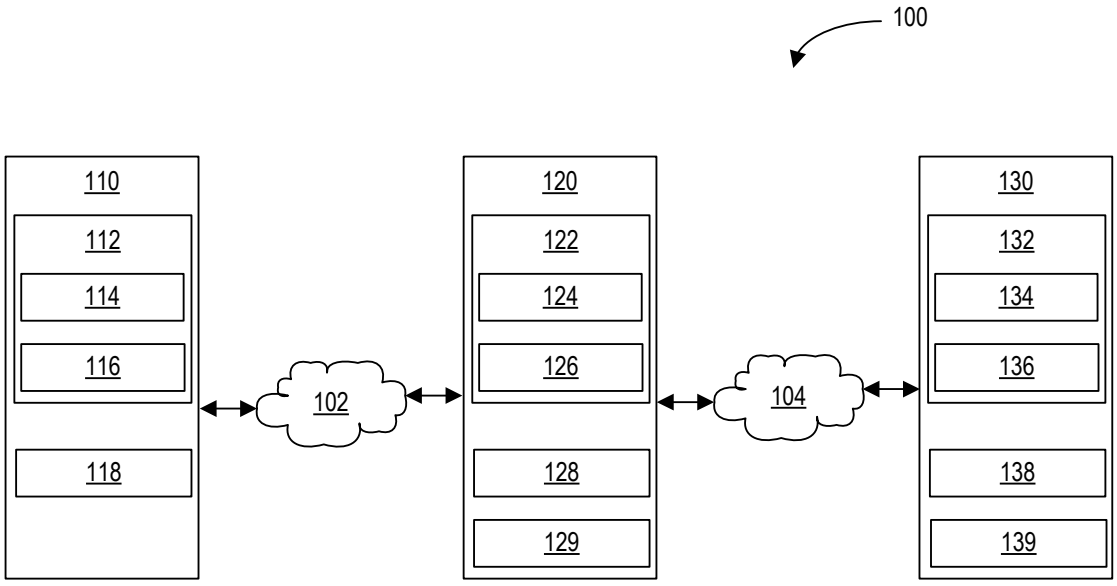

FIG. 1

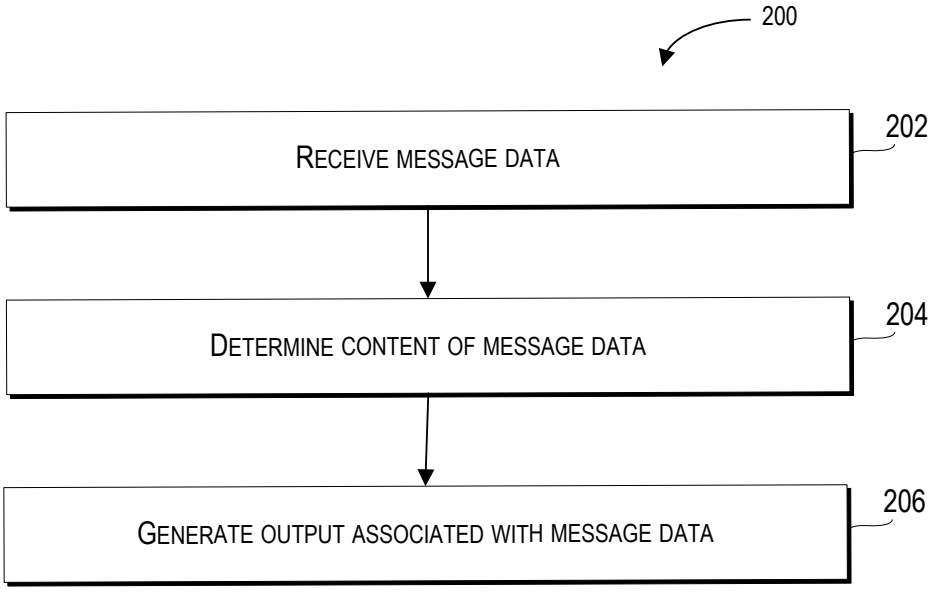

FIG. 2

Abstract

The present disclosure relates to a text parsing system and related method for accurately parsing the content of text in messages and providing an output that can be used by various systems including systems used to detect spam and advertising content. The text parsing system can include a computing system that can parse the content of text (e.g., using a computing system including a machine-learned model or a rules based text parsing system) and provide an output including a list of potential parsed words along with associated word types, language, and confidence of word matching. Furthermore, the text parsing system can further determine the content of text through use of a knowledge base that includes a structured data repository represented as a graph. The knowledge base can be used to generate further output associated with the content of the text including related information drawn from the knowledge base.