October 02, 2017

ASSISTANT TEXT NORMALIZATION

Google Inc.

ASSISTANT TEXT NORMALIZATION

ABSTRACT

A virtual, intelligent, or computational assistant (e.g., also referred to simply as an "assistant") is described that is configured to perform text normalization when converting text to speech (e.g., when synthesizing audio data for output to a user). The assistant may perform text normalization by determining how pronounce a particular set of characters (e.g., word, homonyms, number, date, acronym, abbreviation, etc.) based on the context in-which the particular set of characters is used. For instance, when performing text to speech on the text "1233 St. Andrew St." (e.g., when reading an address aloud), the assistant may determine that the first use of the set of characters "St." should be pronounced as "saint" as it is a prefix of a street address and that the second use of the set of characters "St." should be pronounced as "street" as it is a suffix of a street address.

DESCRIPTION

Assistants execute on counter-top devices, mobile phones, automobiles, and many other type of computing devices. Assistants output useful information, responds to users' needs, or otherwise performs certain operations to help users complete real-world and/or virtual tasks. Some assistants may perform text to speech (TTS) operations to read text aloud by playing audio data synthesized based on the text. In regular person-to-person speech, it may be considered proper for some text to be pronounced differently based on context.

The example system shown in FIG. 1 provides an assistant that pronounces text differently based on context when performing TTS. For example, when the assistant is reading

aloud text that includes a particular set of characters, the assistant may determine how to pronounce the particular set of characters based on the context in-which the particular set of characters is used in the text.

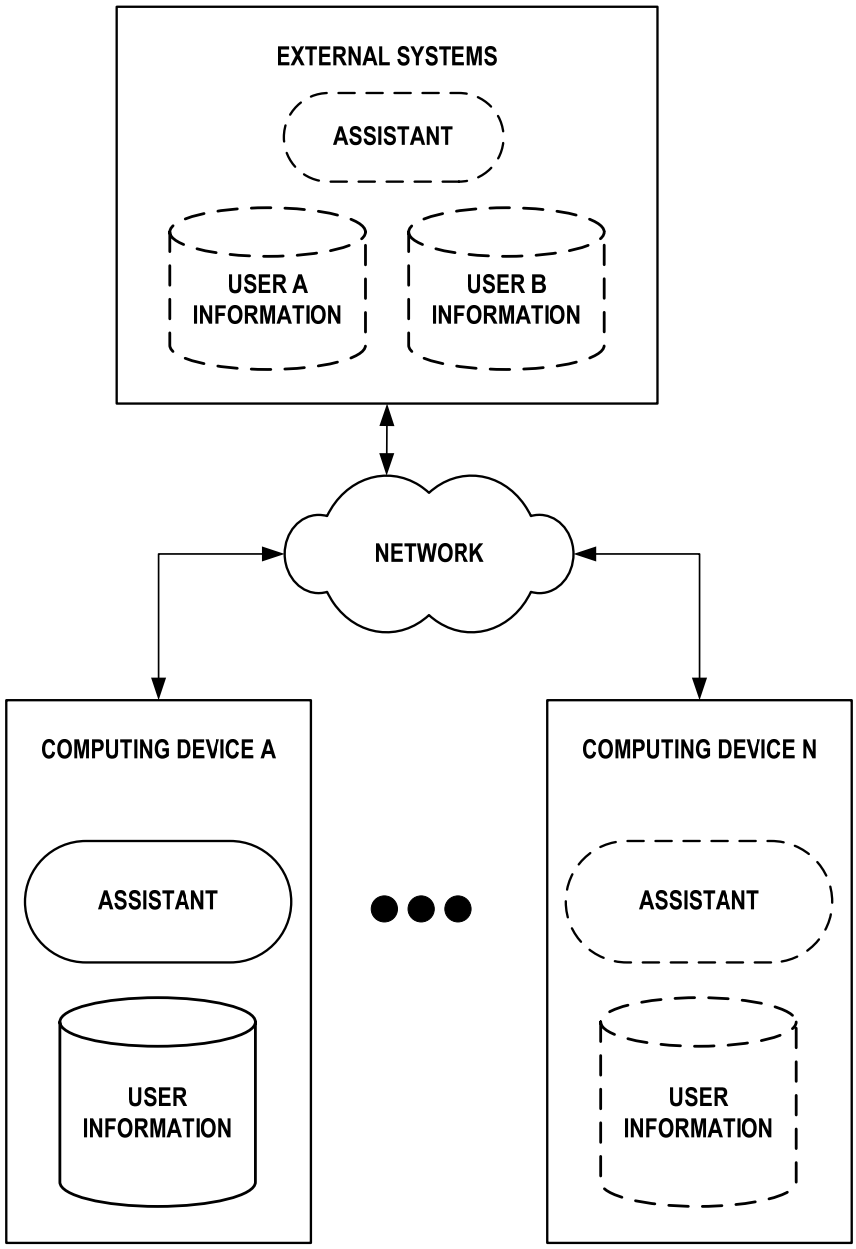

FIG. 1

The system of FIG. 1 includes one or more external systems and computing devices A–N communicating across a network with each of computing devices A–N executing an assistant

that performs operations involving groups of people. The network of FIG. 1 represents a combination of any one or more public or private communication networks, for instance, television broadcast networks, cable or satellite networks, cellular networks, Wi-Fi networks, broadband networks, and/or other type of network for transmitting data (e.g., telecommunications and/or media data) between various computing devices, systems, and other communications and media equipment. Computing devices A–N represent any type of computing device, or other system that is configured to execute an assistant and communicate on a network. The external systems represent any type of cloud computing environment, mainframe, server, or other computing system that is configured to support the assistants executing at computing devices A–N.

Computing devices A–N can be personal computing devices. In some examples, the external systems and/or computing devices A–N may be shared assets of multiple users. Examples of computing devices A–N mobile phones, tablet computers, wearable computing devices, countertop computing devices, home automation computing devices, laptop computers, desktop computers, televisions, stereos, automobiles, and any and all other type of mobile and non-mobile computing device that is configured to execute an assistant. For example, computing device A may be a countertop assistant device and computing device N may be a mobile phone or automobile infotainment system.

An assistant executes across any combination of external systems one or more of computing devices A–N to provide assistant services to users of computing devices A–N. Examples of assistant services include: setting up reminders, creating calendar entries, booking travel, online ordering, sending messages or other communications, reading text aloud, controlling televisions, lights, thermostats, appliances, or other computing devices, providing

navigational instructions, or any other conceivable task or operation that may be performed by an assistant.

As a user interacts with the assistant, the assistant may obtain personal information about the user. Examples of personal information include: habits, word or phrase selections, voice samples, routines, preferences, notes, lists, contacts, communications, interests, location histories, and other types of user information. After receiving explicit permission from the user, the assistant may store, the personal information at user information data stores and in the course of providing assistant services, make use of the personal information stored at the user information data stores.

The external systems and computing devices A–N and the assistant treat the information stored at the information stores so that the information is protected, encrypted, or otherwise not susceptible to unauthorized use. The information stored at the information data stores may be stored locally at each of computing devices A–N and/or remotely (e.g., in a cloud computing environment provided by the external systems and which is accessible via the network of FIG. 1).

Further to the descriptions below, a user may be provided with controls allowing the user to make an election as to both if and when the assistant, the computing device, or the computing systems described herein can collect or make use of supplemental data (e.g., user information or contextual information about a user's social network, social actions or activities, profession, a user's preferences, or a user's current location), and if and when the user is sent content or communications from a server. In addition, certain data may be treated in one or more ways before it is stored or used, so that personally identifiable information is removed. For example, a user's identity may be treated so that no personally identifiable information can be determined

for the user, or a user's geographic location may be generalized where location information is obtained (such as to a city, ZIP code, or state level), so that a particular location of a user cannot be determined. Thus, the user may have control over what supplemental data is collected about the user, how that supplemental data is used, and what supplemental data is provided to the user.

In operation, the assistant may perform TTS to read text aloud to a user using text normalization. For instance, when reading text aloud to user A, the assistant may use context to determine how to pronounce various sets of characters included in the text.

The assistant may use several techniques to determine pronunciation based on context. For instance, the assistant may apply a set of rules based on the type of text being read aloud. As one example, where the text being read aloud is an address, the assistant may use a set of address pronunciation rules to determine how to pronounce various sets of characters in the address. As another example, when the text being read aloud is a short message written by another user, the assistant may use a set of short message pronunciation rules to determine how to pronounce various sets of characters in the short message.

The assistant may apply the rules in a variety of manners. As one example, the assistant may apply the rules based on user instructions. As another example, the assistant may automatically apply the rules based on user interactions and/or training sets.

As discussed above, the assistant may use a set of address pronunciation rules to determine how to pronounce various sets of characters in an address. An example set of address pronunciation rules may specify that the set of characters "st." before a street name is to be pronounced "saint", that the set of characters "st." after a street name is to be pronounced "street", that the character "W" before or after a street name is to be pronounced "west" (similar for other cardinal directions such as "S" pronounced as "south", "SW pronounced as "south

west", etc.), the "#" character is to be pronounced as "number" (e.g., "suite #300" is to be pronounced as "suite number three-hundred"), numbers followed by an ordinal (e.g., 1st, 2nd, etc.) are to be pronounced as ordinal numbers (e.g., "1st" is pronounced as "first", "2nd" is pronounced as "second", etc.), numbers that are even multiples of one-hundred or one-thousand are to be pronounced as such (e.g., "300" is to be pronounced as "three-hundred", "7000" is to be pronounced as "seven-thousand", etc.), numbers that are not even multiples of one-hundred or one-thousand are to be pronounced digit-by-digit (e.g., "307" is to be pronounced as "three zero seven" or "three oh seven", "7526" is to be pronounced as "seven five two six", etc.).

As discussed above, the assistant may use a set of short message pronunciation rules to determine how to pronounce various sets of characters in a short message. An example set of short message pronunciation rules may specify that the "#" character is to be pronounced as "hashtag" (e.g., "#dinner" is to be pronounced as "hashtag dinner"), various acronyms are to be spelled out as their component letters (e.g., "lol" is to be pronounced as "lol"), various acronyms are to be expanded (e.g., "smh" is to be pronounced "shake my head"), etc.

The pronunciation rules may be stored in a lookup table, and identified for use based on word order, first word, or word combinations in the target phrase to be spoken. The appropriate pronunciation rules may also or alternatively be identified by evaluating the phonetics of a target phrase. Machine learning techniques may be applied to identify likely words or combinations to which pronunciation rules are to be applied, with frequency of use of phrases, transcription of user commands, and other inputs used to assist with training of models for the machine learning. Furthermore, the use of pronunciation rules may be based on the vocalization model being employed for TTS. That is, for an assistant voice output with one accent or dialect, a certain first

set of pronunciation rules may be utilized, while for a different accent or dialect of the voice assistant a second, different set of pronunciation rules may be utilized.

By using text normalization, the assistant may be able to read text aloud in a less mechanical manner and/or sound like an actual person. As such, the assistant may allow for less awkward and smoother user interactions. The above examples are just some use cases for the assistant architecture shown in FIG. 1, the assistant architecture has many other applications and use cases.