Technical Disclosure Commons

Defensive Publications Series

April 2020

Joint Graph-Based Reasoning For Interacting With A User

Anonymous

Joint Graph-Based Reasoning For Interacting With A User

ABSTRACT

After coreference resolution is completed by an automated assistant in an augmented reality (AR) device or smartphone, the automated assistant performs a joint graph-based reasoning method to conduct an intelligent dialog with a user. The joint graph-based reasoning method uses information from various data sources (such as a scene graph, memory graph, knowledge graph, etc.) that enables the automated assistant to provide responses to comments that are provided by the user during the dialog. The automated assistant performs the dialog with the user for shopping, visual question answering (VQA), or other interactive user activity.

KEYWORDS

- Knowledge graph
- Scene graph
- Coreference resolution
- Augmented reality (AR)
- Visual question answering (VQA)
- Visual dialog
- Automated assistant
- Artificial intelligence (AI)

BACKGROUND

In computational linguistics, coreference resolution refers to the process of resolving the references (e.g., nouns and pronouns) that are mentioned within a textual or audio dialog. A goal of coreference resolution is to properly interpret a phrase and context of the phrase, by

determining which pronouns or nouns in the phrase are referring to each other or to particular subjects.

An example phrase (which is used throughout this disclosure) might be the following, which is spoken by a user while viewing a product during online shopping: "Will this brown couch fit in our living room along the window?"

In the foregoing example, textual/audio coreference resolution techniques may generally be able to determine that "this" refers to "brown couch" (for "this brown couch") and that "our" refers to "living room" (for "our living room"). Moreover, even assuming that coreference resolution techniques are capable of additionally understanding the scene being viewed by the user, mapping the user's language (audio) to the "brown couch" being viewed, determining which living room or window that the user is thinking of, etc., there is still the challenge of being able to answer the user's ultimate question of whether or not the brown couch "will fit" in the living room along the window.

DESCRIPTION

This disclosure describes a reasoning module that is implemented in an augmented reality (AR) device (e.g., AR glasses) or in some other user device (e.g., a smartphone) of a user. The reasoning module works in cooperation with a multi-modal coreference resolution module that uses audio input, image input, and stored information (representing the user's memory) to perform coreference resolution. After the coreference resolution module resolves the references in the audio input, image input, and stored information, the coreference resolution module provides the resolved references to the reasoning module. The reasoning module uses the resolved references for engaging in a dialog with the user.

In operation, the coreference resolution module and the reasoning module are implemented by an automated assistant or other form of artificial intelligence (AI) module in the user device. Based on the resolved references, the reasoning module uses data from one or more knowledge sources to analyze the audio input, image input, and stored information (representing the user's memory) so as to generate an output to be presented to the user (e.g., an answer of whether the couch will fit) via visual question answering (VQA) or visual dialog with the user. The knowledge sources include a scene graph, memory graph, knowledge graph, and other sources of information.

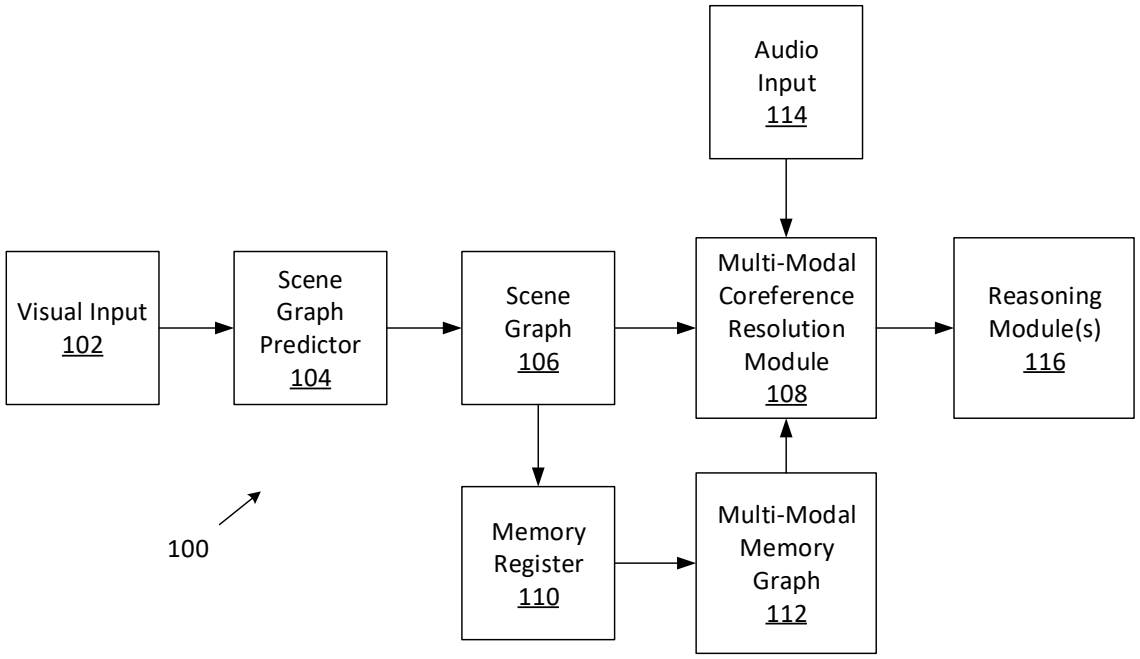

FIG. 1: System for performing reasoning based on multi-modal coreference resolution

FIG. 1 is a functional block diagram a system (100) that performs a method for performing reasoning based on multi-modal coreference resolution. The system receives visual input (102), in the form of a video or other type image. For instance, the image may be a scene that is being viewed by a user of AR glasses, an image posted on a web page being viewed by

the user, or other type of visual input. In the current example, the image is that of a couch on a web page of a furniture vendor, an actual couch that the user is viewing with AR glasses at a furniture store, or a virtual image of a couch in some room rendered by AR glasses, etc.

The visual input is provided to a scene graph predictor (104), which extracts the objects in the image and predicts their relationships. The scene graph predictor generates a scene graph (106), which is a data structure that arranges the logical and spatial relationships of the objects in the image. For example, information such as couch, brown, leather texture, floor, etc. form the branches, nodes, and leaves of a tree of the scene graph. The scene graph is provided as one input to a multi-modal coreference resolution module (108).

Furthermore, the scene graph is stored and indexed in a memory register (110) along with other datasets that represent the memory of the user, with user permission. The memory register also stores other datasets of experiences and actions of the user: calendar events, email content, photos, contact lists, recent phone calls/voicemails, web browsing history (e.g., recently viewed items), other scene graphs previously generated by the scene graph predictor, etc., all of which form the body of knowledge that make up the memory of the user. Thus, in the current example, the memory register stores and indexes a scene graph or a photo that shows the "living room" and the "window." The content(s) of the memory register can be used to generate a multi-modal memory graph (112), which serves as another input to the coreference resolution module.

Another input to the coreference resolution module is audio input (114). In the current example, the audio input is the question "Will this brown couch fit in our living room along the window?" that the user is asking the automated assistant while viewing the image of the brown couch.

In operation, the coreference resolution module receives the audio input and the scene graph for the visual input. The coreference resolution module parses the audio input, and generates and ranks all possible (m, r) pairs from the words in the audio input.

To map the visual input and the audio input with each other, the coreference resolution module traverses the branches/nodes/leaves of the scene graph to find the elements that pertain to the brown couch. When these elements are located in the scene graph, the coreference resolution module is able to conclude (e.g., resolve) that the "brown couch" uttered in the audio input co-refers to the brown couch that appears in the visual input.

Since the references of the living room and the window do not appear in the visual input, the coreference resolution module does not find any related branches/nodes/leaves of the scene graph that was generated from the visual input. Therefore, the coreference resolution module attempts to resolve these references from one or more of the memory graphs generated from the datasets stored in the memory register, so as to determine, for example, which "living room" and "window" that the user is referring to from the user's memory.

Having thus completed the coreference resolution to link the contents of the visual input, the audio input, and the memory register with each other, the coreference resolution module provides the linked contents to one or more reasoning module(s) (116). The reasoning module uses various tools, algorithms, and data to address the user's comments in the audio input, such as answering the question of "Will this brown couch fit in our living room along the window?"

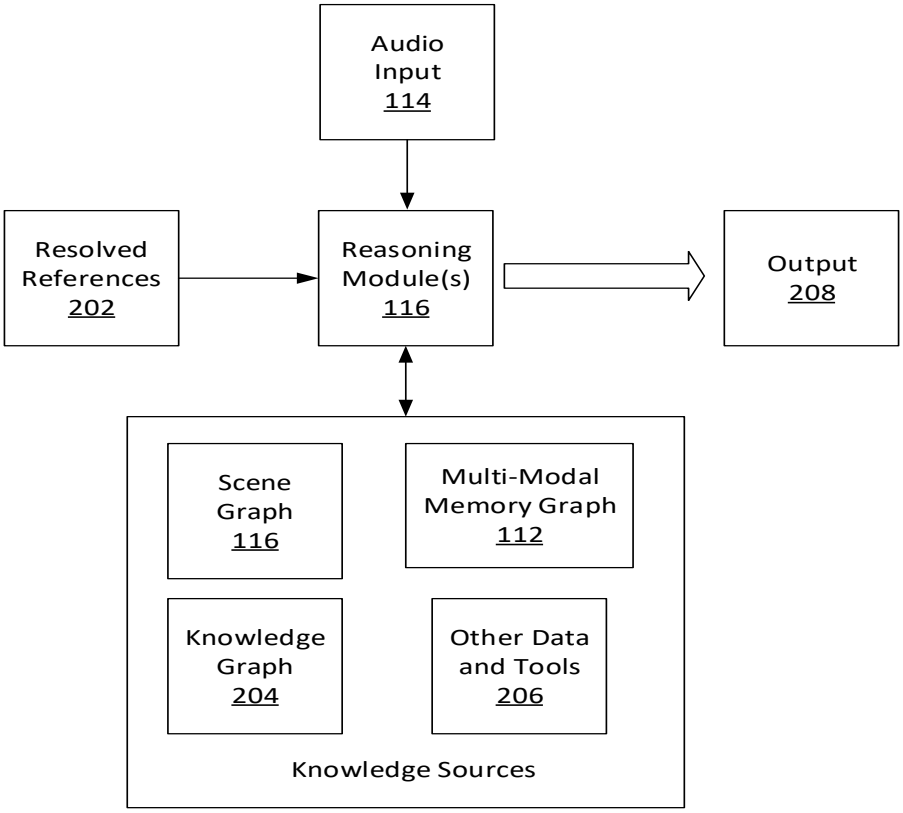

FIG. 2: Using multiple knowledge sources to perform reasoning

FIG. 2 depicts the reasoning module using various knowledge sources to perform reasoning that enables the automated assistant to engage in an intelligent dialog with the user. The reasoning module receives resolved references (202) from the coreference resolution module. The reasoning module also receives the user's audio input. The received references provide the reasoning module with query terms for searching the knowledge sources, while the audio input enables the reasoning module to identify a question/comment from the user than can be expanded in a dialog between the automated assistant and the user.

For example, the reasoning module parses the terminology in the audio input and determines that the user is asking a question that needs to be answered (e.g., "Will this brown couch fit…?") in a dialog with the user.

Also, and for example, the term “brown couch” is used by the reasoning module to traverse the scene graph to determine the properties of the couch in the scene (e.g., couch has four legs that sit on a floor, has pillows and a back rest, is brown in color, etc.). The reasoning module also uses “brown couch” to traverse a knowledge graph (204), so as to, for example, identify a vendor or website where the couch is available, determine dimensions and pricing for the couch, locate reviews for the couch, etc.

With respect to other references (terminology) in the audio input that have been resolved (e.g., living room and window), the reasoning module traverses the memory graph to obtain further information. For instance, the reasoning module determines from the memory graph (e.g. from information in the branches/nodes/leaves of the memory graph that have been distilled from the user’s stored photos of the living room) that the window runs along roughly half of a length of a wall of the living room, and extends from the floor to the ceiling.

The reasoning module also obtains further information from other data sources and tools (206). For instance, with user permission, the reasoning module can access the browsing history, calendar events, web pages, emails, etc. that are stored in the user’s device. As another example, the user’s device can include a measurement tool that calculates the dimension of an object in the camera’s viewfinder or in an image. As such, the reasoning module is able to determine the dimensions of the living room and window that the user is referring to.

The reasoning module can use several types of techniques to extract information from the scene graph, memory graph, knowledge graph, and other data/tools. Examples include graph walking techniques, graph neural network techniques, or other techniques to obtain information from multiple/joint knowledge sources.

Thus, knowing the question/comment being asked by the user (e.g., "Will this brown couch fit…?") and having extracted information from the various knowledge sources (e.g., length of the couch, dimension of the window and living room, etc.), the reasoning module is able to provide an output (208) to the automated assistant. Based on this output from the reasoning module, the automated assistant replies to the user with an answer such as "Yes, this couch will fit in the living room along the window, with 4 feet to spare on each side of the couch."

The couch example above represents one possible illustrative use scenario. In another use scenario, the automated assistant provides reminders that are triggered based on visual and/or audio input from the user. For example, the user is wearing AR glasses and looks inside the refrigerator at a carton of milk. The coreference resolution module performs multi-modal coreference resolution by resolving the image input with stored information of the user's memory. Then, based on the resolved references, the reasoning module accesses the scene graph, memory graph, knowledge graph, and other data/tools to generate an output, which the automated assistant presents to the user as a reminder of "It is Friday today. You should throw out the milk and other expired food in the refrigerator."

Another example use scenario involves the automated assistant attempting to trigger an action by the user, in response to multiple input sources. As an illustration, the user of the AR glasses is looking at a plant in a room. Another person in the room says, "That plant looks wilted." The automated assistant captures this video input and audio input from multiple persons, performs coreference resolution, and performs joint graph-based reasoning so as to output a message to trigger the user perform an action (e.g., "You should water that plant.").

Yet another example of a use scenario that involves a combination of multiple inputs, the user may be wearing AR glasses and looking at the engine in his car to diagnose a problem with the engine. The user asks, "What is wrong with this engine?" The coreference resolution module resolves the references to the "engine" in the image and audio inputs, with the user's memory of the engine represented in a memory graph. From these resolved references, the scene graph, memory graph, knowledge graph, and other data/tools, the reasoning module is able to determine the maintenance history of the car and a factory-specified maintenance schedule, and so the automated assistant presents a recommendation to the user, such as "Check the engine oil. The last oil change was 16 months ago; the manufacturer recommends changing the oil twice a year."

Other use scenarios are possible. Both the coreference resolution module and the reasoning module can be implemented using trained machine-learning models. If there is any data that is used by the coreference resolution module or the reasoning module that is of a personal or confidential nature to the user, then the use of the data can be allowed, restricted, or otherwise controlled in accordance with permissions provided by the user.

CONCLUSION

The techniques described herein enable an automated assistant to perform reasoning after completion of coreference resolution to link elements contained in visual input, in audio input, and in stored information that represents a memory of a user. The automated assistant includes a reasoning module that uses a scene graph, memory graph, knowledge graph, and other data/tools to generate an output that is used by the automated assistant to conduct an intelligent dialog with the user.