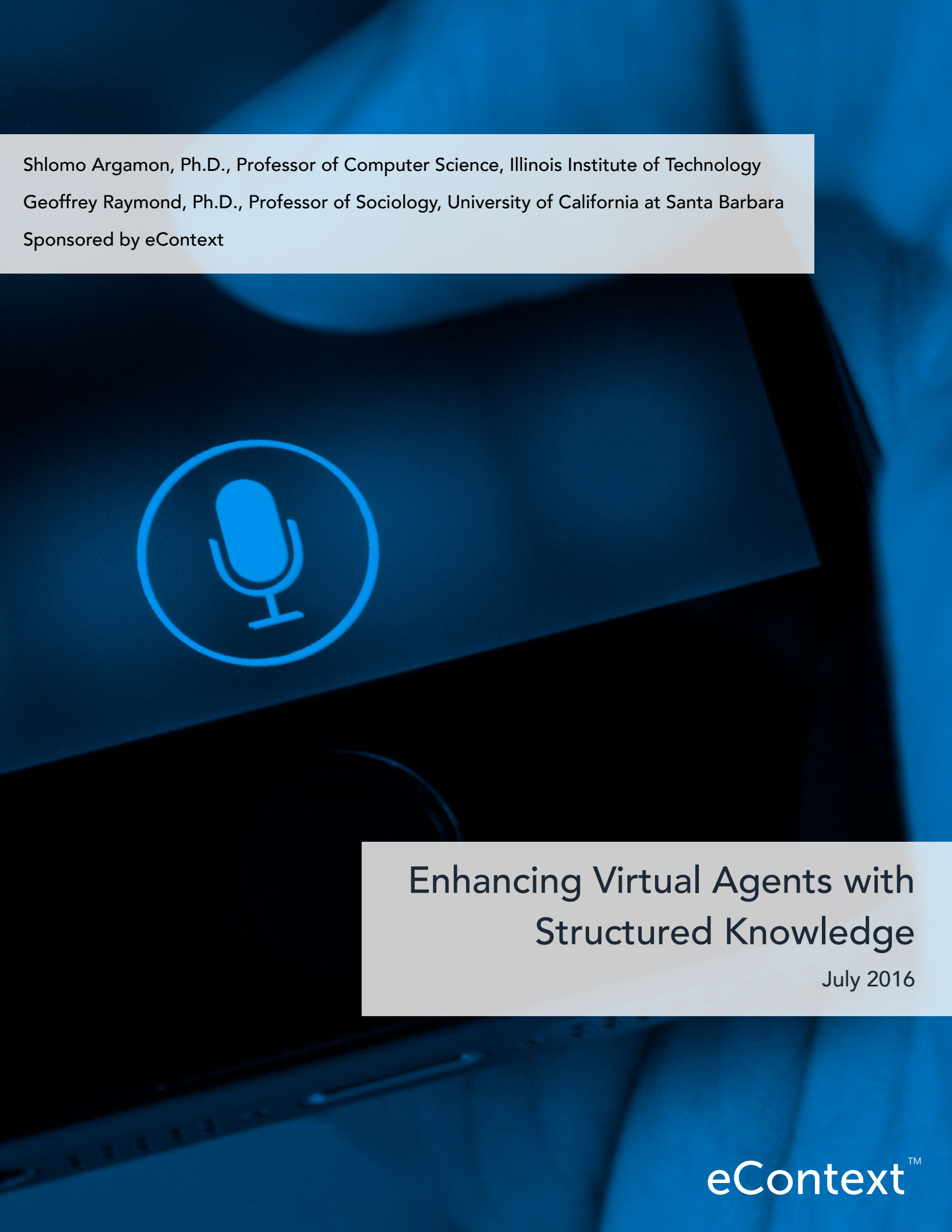
Shlomo Argamon, Ph.D., Professor of Computer Science, Illinois Institute of Technology

Geoffrey Raymond, Ph.D., Professor of Sociology, University of California at Santa Barbara

Sponsored by eContext



Enhancing Virtual Agents with Structured Knowledge

July 2016

eContext™

Contents

Executive Summary

Virtual Agent (VA) technology is poised to transform the role of computers and information technology. Its share of consumer attention and wallets is set to explode as it becomes ubiquitous in peoples' everyday lives. This transformation will not be limited to smartphones. VA technology is increasingly incorporated into embedded devices for ecommerce, including the Internet of Things.

The key advantage to voice-activated Virtual Agents over other human-computer interfaces is that interaction can be *natural, hands-free, and faster* – people can interact with a VA in the same way that they would interact with a person. Conversational interaction is the primary interface humans use with one another to manage their affairs. When VAs become capable of nearly human-level conversational interaction, the physical interface of most computer technology will essentially disappear.

The critical factors for wide-scale acceptance of VA technology are *reliability* and *user satisfaction*. Achieving improvements in this area will require developing improved methods for intelligent understanding of user queries and knowledge processing. This is not simply a technological problem with a purely technological solution, but one of technology connecting with fundamental human social structures.

Transformative VA technology will need to represent and effectively use knowledge about the world, about its users, about typical tasks, about conversational structure, about conversational errors and repairs, and so forth. Machine learning and data analytics can help, but structured knowledge bases curated by human experts will be critical to achieving these advances.

VAs will need to incorporate deep and broad taxonomies, including fully articulated *ontologies*, into existing methods for question understanding and answering. VAs will need to understand a larger variety of semantic relations represented between concepts as well as inference rules that allow complex reasoning to take place. VAs will model their users' interests, goals, plans, and worlds, and will use this information to anticipate their needs, understand their requests, and be natural in conversation with them.

Owners of VA technologies will be partnering or investing in taxonomies and ontologies, and other large enterprises should establish a solid VA strategy to take advantage of this disruptive technology.

Introduction

> ***"In the U.S., on our mobile app in Android, one in five queries – 20% of our queries – are voice queries, and that share is growing."***
>
> –Sundar Pichai, CEO, Google, Google I/O Keynote, 2016

"In the not-too-distant future, users will no longer have to contend with multiple apps; instead they will literally talk to digital personal assistants such as Apple's Siri, Amazon's Alexa or Google Assistant."

– Mark O'Neill, Research Director at Gartner, June 20, 2016 Gartner Press Release

Gartner predicts that by 2019 at least a quarter of households in developed countries will use Virtual Agents (VAs). A 2016 global survey commissioned by Nuance showed that 89 percent of consumers want to engage in conversation with virtual assistants to quickly find information instead of searching through Web pages or a mobile app on their own.

VAs are software systems with the capability of conversing with users via voice recognition and synthesis, or text chat. Typically, humans can talk four times faster than they can type. According to research from KPCB (2016 Internet Trends Report), already 25 percent of searches performed on Windows 10 task bar are voice searches, and on Baidu, 10 percent of queries come through speech. This growth has spurred the development of new products specifically hosting VAs, such as Amazon Echo, Amazon Dot, and Google Home.

Historically, the modes of technological interaction that have been most disruptive – PCs, search, messaging, social, and mobile – impact high-tech companies first, then are eventually adopted by other commercial organizations who intend to keep pace with consumer behavior and contemporary monetization channels. Today, Alibaba, Alphabet, Amazon, Apple, Baidu, Facebook,IBM, Microsoft, and Tencent are rapidly increasing their R&D budgets to VA solutions, while traditional companies in retail, healthcare, finance, and travel are still focusing on their social or mobile strategies.

These traditional companies should urgently establish a solid VA strategy, as its disruptive reach dramatically escalates. The number of users moving their commercial interactions to VA technology is poised to increase enormously once it can offer extensive personalization, seamless ease-of-use,better language understanding, and can establish trust in a wide variety of applications.

Virtual Agent technology will fundamentally transform the role of computers and information technology in the economy and society. This influence will not be limited to smartphones or other mobile devices. VA technology is increasingly incorporated into embedded devices for ecommerce, including the Internet of Things.

The notable offerings in this market are:


Siri


Cortana


Google Assitant


Alexa


Viv

Other companies, with more specialized offerings, include:


X.ai


Nuance


next IT


Wit.ai


eGain


Duer


Creative Virtual


Indisys


Speaktoit


Anboto


Artifical Solutions


CodeBaby

Technological factors, therefore, will play a major role in the development of this market.

The main challenges are:

1. Correctly understanding the context of the question or request, *as a means of identifying the user's intent*.
2. Retrieving the correct answer that fulfills that intent from the most appropriate knowledge base, source, or vendor.

Achieving improvements in this area will require developing new methods for semantic understanding of user queries and knowledge processing. This is not simply a technological problem with a purely technological solution, but one of integrating technology into fundamental aspects of human social structures.

Virtual Agent technology will be one of the main drivers of information technology over the next decade because conversation is the most basic, natural, and intuitive mode of social organization for people whether they are working, playing, shopping, or connecting with others; it is the next progressive step in reducing the artificiality of our human-computer interactions.

Technical Challenges

Social Assistance with Personal Projects

Although this paper focuses on technology for answering questions, it is important to keep in mind that, as with most conversational exchanges, questions and their answers are best understood as vehicles for social action. Humans use a variety of grammatical forms to accomplish a very broad range of actions: requesting, reporting, complaining, greeting, and others that are more subtle and complex (such as "confirming an allusion," see Schegloff, 1996). Even business, sales, and marketing application agents, both human and virtual, must be capable of dealing with a very wide variety of actions.

The most common activities users pursue with Virtual Agents are requests for action or information and current systems are closest to being able to recognize and respond to them. However, the approach we developed below anticipates the broader range of actions humans engage in, and can be generalized to these other types of actions as systems develop.

Questions and Requests

Virtual Agents are designed to both **provide information** and **fulfill requests**. These different types of actions can be expressed in multiple grammatical forms:

Interrogatives (grammatical questions)

"What is the capital of Nepal?"

"Can you schedule a meeting with John?"

Declarative statements

"I wonder what the capital of Nepal is."

"I need a meeting with John."

Imperatives

"Tell me the capital of Nepal."

"Schedule a meeting with John."

Thus, effective language processing for Virtual Agents must determine what the user is trying to *accomplish* overall, not just the form of what is said. The agent must consider how each interaction fits in with the user's overall **intent**. It is not just about language, per se; broader knowledge about the user and the world must be brought into play.

But most user needs cannot be achieved through simple "one-off" interactions, and demand extended conversation. Virtual Agents need the ability to engage effectively and naturally. This ability involves complex inferences about the real world provided by natural language understanding (NLU) and developments in artificial intelligence. This is the medium term key to the industry.

Questions: The Easy and the Hard

Easy Questions

First, consider what these agents already do well. A question which is well-formed grammatically, pronounced clearly, and stated in terms that exactly match information in the database available to the Virtual Agent can be answered immediately and well. Answering questions like these is a relatively straightforward process:

Siri, where's the nearest Apple store?

Alexa, what's the price of The Divergent Series, Insurgent on instant video?

The words in these questions are not utterly unambiguous, but the phrases clarify meaning. "Apple" may be an electronics firm, a fruit, or a music company, but "Apple store" has one, more commonly-used meaning. "The Divergent Series" refers to many books and films (as well as a mathematical concept), in several formats, but "The Divergent Series Insurgent instant video" narrows the possibilities to just one thing.

Hard Questions
Questions become more difficult to answer when stated in a less direct manner, put in words that are not in the answer database, or when the words used have more than one possible meaning:

> *Siri, where can I get my screen fixed?*
>
> *Alexa, how much for the new Divergent?*

The user's intent may be the same, but the agent's task is more complex. The word "screen" has many possible meanings: computer, phone, TV, sun, mosquito, phone, preliminary interview. Asking where to get it fixed narrows the possibilities, and some of the possible meanings are used much more frequently than others, but not enough to confidently narrow the question to one clear meaning. And some of the necessary elements to identify the right answer are missing. If the intent was to fix a phone screen, for example, it may be necessary to know the brand of phone and whether it is under warranty.

Understanding the phrase "the new Divergent" requires not just knowing that it refers to a movie (or a book) series rather than just a dangling adjective, but also knowledge about what is "new" – knowing that the last Divergent book was published long before the most recent movie is necessary to know that the user means the movie.

VA technologies are starting to go beyond statistical speech-to-text and information-retrieval methods. Contextually meaningful responses will require VAs with articulated knowledge about the possible meanings of words and phrases, connected with each other and with the real-world context. They must continue to improve their language models, but also start to gather information to create richer situational awareness, understand both individual and general context, and become attuned to the intent of the questioner.

Technological Market Drivers

> ***"Every single conversation is different. Every single context is different… we want to understand your context…I think computing is poised to evolve beyond just phones. It will be in the context of a user's daily life. It will be on their phones, devices they wear, in their cars, and even in their living rooms."***
>
> –Sundar Pichai, CEO, Google, Google I/O Keynote, 2016

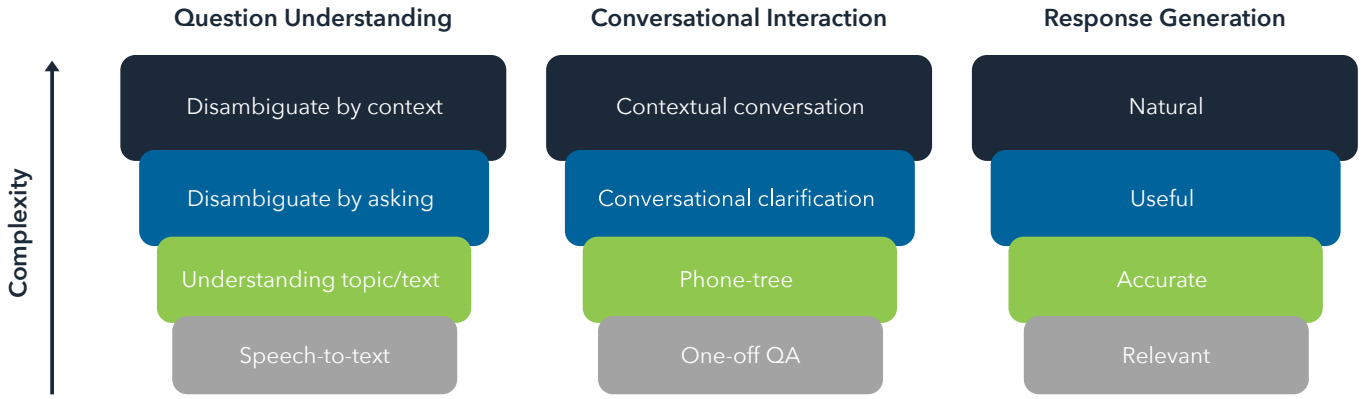
The three key technology dimensions for VA development are:

1. **Question understanding:** Generating an accurate internal representation of the user's question or request.

2. **Response generation:** Generating a relevant, accurate, and useful response for the user. This may involve finding and explaining information or invoking an action on a device.

3. **Conversational interaction:** Interacting with the user. At a minimal level, this involves simply responding to requests; more advanced capabilities involve maintaining multi-turn conversations with the user, as appropriate.

Each of these dimensions defines a spectrum of sophistication, which define a three-dimensional space for VA technology, depicted in the figure below. Question **understanding** defines a range from basic speech-to-text capability through contextual understanding of the meaning of the question. **Interaction** begins with simple one-off question answering, progressing towards more sophisticated modes include phone-tree-like verbal menus, conversational questioning of the user to clarify information, and naturalistic conversation that uses full contextual information. Finally, **responses** generated should be relevant to the question, with improvements leading to more fully accurate, useful, and natural responses.



VA technologies are starting to reach more sophisticated levels of processing in all three areas. There are three key technological barriers that need to be overcome to achieve qualitative improvements:

Understanding

Using the full context to understand the meaning of user queries. This includes the conversational context, as accounting for the month, the location, and even device and the time of day, and also knowledge about the specific user including age, gender, and their topical search history. VAs will need to resolve ambiguity, either by referencing its context or background knowledge, or by actively seeking user confirmation.

Response Relevance

Ensuring that responses are contextually relevant to the intent. This will require understanding the user's actual need, not just the language or grammar used, and evaluating different potential responses that may satisfy that need.

Conversational Structure

For VAs to be widely accepted as intelligent, they will need to understand how to structure a conversation that makes users feel comfortable. Several aspects of extended interaction are critical:

1. Follow up questions as part of disambiguation and clarification of the user's needs.

2. Driving complex multi-turn actions, such as booking travel across dates, times, destinations, flights, and hotels.

3. Detecting and repairing a temporarily derailed conversation. VAs must detect when the user and agent are not on the same page, or when the user wants to "edit" something they already said. The VA will need to know enough about the expected structure of the conversation, troubles it may encounter and how to recognize them, and how to initiate a conversational repair. Even if the repair is just to say "I'm sorry, we seem to have gotten off-track. Shall we try again from the beginning?" Detecting such errors can greatly improve the naturalness of conversation with a VA.

Let's look at each of these areas in more detail and consider how advances in structured knowledge can significantly improve user experiences.

Question Understanding

The first phase in any question-answering system is processing the input query into some representation of its "meaning," that is, a form that can be used to find an answer.

This meaning consists of at least two components: the question "type" and the question "topic." The *type* informs the VA what sort of an answer to seek; for example: "How many people live in the UK?" is a *quantitative* question, "What is a gene?" is a *definition* question, and "How do I knit a sweater?" is a *procedural* question. Knowing the type of a question helps to narrow the field of acceptable answers, and enables the system to structure the response appropriately.

The *topic* represents what the question is about. A topic representation may be as simple as just the set of topical words in the question, but more sophisticated representations can greatly improve results. For example, the representation of "*What pharmaceutical companies are using Oracle?*" should refer to knowledge that pharmaceuticals, medicines, and drugs (in one sense) are the same thing, and that Oracle is a technology company.

Even common, highly commercial questions can present topical ambiguity. Understanding the set of likely topics allows the system to prioritize certain interpretations over others, based on additional meta data or qualifying interactions.

Phrase	Topical Possibilites
Where can I buy oil?	Castor Oil; Coconut Oil; Crude Oil; Essential Oils; Heating Oil, Mineral Oil; Motor Oil; Olive Oil
Where can I buy a monitor?	Baby Monitors; Blood Pressure Monitors; Computer Monitors; Heart Rate Monitors; Monitor Lizards; Studio Headphones
Where can I buy windows?	Automotive Windows; Bay Windows; Double Hung Windows; Microsoft Windows; Stained Glass; Windows Phones
Where can I buy a Jaguar?	Arctic Cat Snowmobiles; Atari Jaguar Video Game Consoles; Aquarium Fish; Jaguar Automobiles; Jaguar Guitars; NFL Fan Apparel; Schwinn Cruiser Bicycles; Wild Felines
Where can I buy a polo?	Casual Shirts; Polo Mallets; Polo Mints; Ralph Lauren Apparel; Perfumes & Fragrances; Volkswagen Automobiles

Data courtesy eContext.

Question Types

Determining the type of a question requires knowing what the different types of questions are. This requires the classification of question types, each with associated constraints on the kinds of answers acceptable for that type of question.

There are not yet any widely accepted taxonomies of question types. Most systems use something ad hoc, often based on one of two long-standing taxonomies, neither developed for purposes of question answering: Graesser's taxonomy of questions in tutoring sessions[1] and Bloom's taxonomy of educational objectives[2].

A good question type taxonomy will be fine-grained enough to give strong constraints on the possible answers to a question, to improve relevance and accuracy, and will also have clear and accurate criteria for determining the type of a given question. In some cases, the type of question is pretty clear, as in "Who shot JFK?" where the word "who" tells us that we want to identify a person; other cases are more difficult. The word "what," for example, says little about the type of answer:

1. What is the population of Indonesia? (quantity, of people)
2. What is the capital of Congo? (city name)
3. What does the cheapest tablet go for? (price, in currency)
4. What is wrong with the cable company that they keep overcharging me? (explanation)

[1] *Question Asking During Tutoring,* Arthur C. Graesser and Natalie K. Person, American Educational Research Journal, Spring 1994, Vol. 31 No. 1, pp. 104 - 137

[2] Bloom's Taxonomy of Educational Objectives, Bloog, B.S., Engelhart, M.D; Furst, E.J., Hill, W.H., Handbook I: Cognitive domain. New York: David McKay Company

The taxonomy must not simply classify questions by their grammatical types, but rather by what the questioner's intent is — what they are trying to accomplish by asking the question. This is what constrains the form of relevant responses.

In a more sophisticated Virtual Agent system, questions and requests must be classified into "user actions." Developing a taxonomy of such actions together with an accurate classification method will be a vital ingredient to VA development. This will require analysis of actual questions and answers in context. Machine learning will be a key component of classifying questions to the right type in the taxonomy. However, fully automated techniques do not yet give high enough accuracy for deployment, so some level of expert human involvement will provide cognitive context.

In subsequent work, systems will need to expand the range of utterance formats that they can accept and generate. VAs must be able to handle directives, declaratives, and even multi-unit forms such as stories and telling, as well as the wide range of actions these accomplish, such as reporting, evaluating, revising, complaining, and so on.

Question Topics

In addition, it is essential to determine the topic of the question. This is now often done by the collection of keywords in the question (a *bag-of-words*), perhaps expanded through a lexical resource such as Wordnet, or mathematically modeled using statistical models of word meaning, such as vector-space models like latent semantic indexing or more sophisticated word-embedding models like word2vec.

Syntactic analysis can help to create a more fine-grained understanding of the topic. The bag-of-words approach applied to "Can dogs be allergic to dust?" would look like a question about allergies to dogs and dust as much as a question about dog allergies. Syntax is needed to disambiguate.

Another key notion in question topic is that of *focus*. Consider the question "Who sells coffee that pays good wages?" The question could be either "What coffee shops pay good wages?" or "Where can I buy/get fair trade coffee?" The system must determine if the focus of the question is the seller or the coffee. However, even given that information, background knowledge is necessary to frame the question properly. If the focus is the seller, what are the other alternatives? The system must recognize someone interested in "wages" is looking for a job, and that certain types of sellers, like bricks-and-mortar retailers, will likely have open positions more than large eCommerce retailers. This requires a taxonomy of relevant world knowledge about objects and attributes.

Parenthetically, it should be noted that improvements in speech recognition, intonation, and prosody may help resolve some of the ambiguities associated determining focus.

Question Ambiguity

> ***"You know, I think that most people underestimate the difference between 95% accurate speech recognition, which is maybe where we are, and 99%. 99% is not an incremental, 4% is improvement — it's a game changer. It's the difference between you barely using it — maybe what you do now — versus you using it all the time."***
>
> —Andrew Ng, Chief Scientist, Baidu, Bloomberg West, May 23, 2016

One of the difficulties in question understanding, as in natural language processing in general, is *ambiguity*, which arises at all levels of processing. Speech-to-text can be foiled by homophones; often knowing the general topic of the question is the only way to disambiguate; consider:

> *I need new sneakers — where can I get a pair/pear?*
> *I'd like some fruit — where can I get a pair/pear?*

Knowledge of typical information needs or specific search queries can even help with this:

> *How do I clean a flu/flue/flew?*

In this case, a request for cleaning a *flu* or cleaning a *flew* would be quite rare compared to cleaning a *flue*. Exploring multiple homophonic words from text-to- speech and then selecting among them based on contextual awareness can increase accuracy.

And even when the specific words are recognized correctly by speech-to-text, knowledge of topic categories is critical to help with disambiguation; consider:

> *I'd like to gamble — where is the Taj Mahal? (Atlantic City)*
> *I'd like to travel — where is the Taj Mahal? (Agra, India)*

What is needed is a taxonomy or ontology of possible topic categories, together with statistics of how often users request information on particular topics and combinations of topics, with certain words and phrases, in given contexts.

To illustrate this, suppose a user asks for a "nearby jaguar park." Does the user mean a place to park a car or to view wild cats? The VA can disambiguate the sense of this request by matching the taxonomic categories for the different words in the query – "jaguar" and "park" – with each other and with commonly used concepts. If we consider taxonomic concept similarity[3] (on a scale of 0.0 to 1.0) for this example, we get:

Concept Similarity					
User Input	Arctic Cat Snowmobiles	Wild Felines	Wildlife Sanctuaries	Jaguar Automobiles	Parking Garages
Nearby Jaguar Park	0.0543	0.5921	0.6306	0.1301	0.0956

This shows how a conceptual taxonomy can bring clarity to a query which is ambiguous on a purely lexical level. The taxonomic structure also permits consideration of other concepts along hierarchical pathways, enabling the prediction that "nearby jaguar parks" are also similar to "Wolf Sanctuaries," a sub-type of Wildlife Sanctuaries, as well as super-types like "Wildlife Facilities," or "Animal Disease Research Institutes."

Such information is quite valuable to question answering systems and Virtual Agents. Even if questions are to be primarily processed by humans, as currently in Facebook's *M* service (working in tandem with its text understanding engine, DeepText), comprehending the topic categories of a question allows it to be routed to specialists in the relevant topics, and to help people deal with the question appropriately. This will become more of a common practice as more service and retail organizations adopt chatbots as a useful type of interface. Using this information, the bot can better establish deep context, and successfully interface with the most relevant third party service or vendor on the web, instead of relying on one or two of the most generic sources. This means that although Wikipedia may return **relevant** answers on the 'Cubs 2016 starting rotation', they may not be as **useful** as the answers that could be supplied by ESPN or the MLB.

Question Context

> ***"If we can understand text, we can help people connect and share in a lot of different ways."***
>
> – Hussein Mehanna, Director of Engineering, Facebook,
> Interview with Mike Murphy, Quartz News, June 1, 2016

[3]Using eContext's taxonomy.

As noted, knowing the context of a question is essential to properly understanding it, most easily seen in the cases of ambiguity ("Where is the Taj Mahal?") or under-specification ("How do I get there?") Disambiguation is currently based on general statistics of the frequency of different kinds of questions. Systems may assume that "Taj Mahal" means the mausoleum rather than the casino, regardless of what the user actually wants. To do better, systems will need to understand information about the situation, including the individual user's location and recent activity, while tracking the topics used in conversation and the interests of similar users.

An explicated taxonomy of possible topics can enable such tracking, as illustrated in the following figure. With a classification of the conversational topic, the same question can be interpreted correctly in very different ways, depending on the context.

Different interpretations of "Where Can I Get a Triumph?" based on knowing the contextual topic category.

Conversation 1	Conversation 2	Conversation 3
What stores stock the Adios Boost? What about the Newton Gravity?	Are there any Yamaha dealers nearby? What about Harley-Davidson?	Where can I find a good pet grooming place? Do dogs get asthma?
Where Can I Get a Triumph?		
Running Shoe: Saucony Triumph	*Motorcycle:* Triumph	*Dogs:* Triumph Dog Food

Similarly, topic tracking can help with determining question focus. If someone just asked for job applications or résumé templates, then the focus of "Who sells coffee that pays good wages?" is probably the seller-as-employer, based on the contextual category of "Jobs."

The same idea applies to modeling user preferences. The topical categories mentioned by a user can form a profile of the user's interests. Systems would need to combine information about a user's general interests with the specific context of a particular conversation, but having a taxonomic representation is necessary.

Semantic Response Relevance

Users demand VAs return accurate answers to their questions, relevant to their needs, and useful for them in context. Current systems do this reasonably well for simple common questions such as "Where is the nearest gas station?" or "When did Abraham Lincoln die?" but fail for more complex questions such as "Does New York or Chicago have more pizza shops?", "How does someone contract Leukemia?", "What does the United Nations do?", or "What caused the housing bubble?"

Long-term, improvements will require sophisticated language understanding methods that extract detailed representations of the meaning of documents as well as of questions. To achieve this, large-scale structured (labeled) knowledge is a vital ingredient, while at the same time will immensely improve response accuracy, relevance, and usefulness at all stages of the process. This is both by improving overall question understanding, as well as by injecting useful information into the response construction process in various ways.

At the input level, a topic taxonomy can be used to index known questions and their answers, which will enable VAs to reformulate an original question in different ways and make it easier to find a correct answer. For example, if the question "Where can I find black high-tops?" is already in the VA's knowledge database with a high-quality answer, connected to a given topic category (for example, HIGH-TOP SNEAKERS) other similar questions that can be classified to the same category could be matched with that known question, enabling them to be answered directly as well.

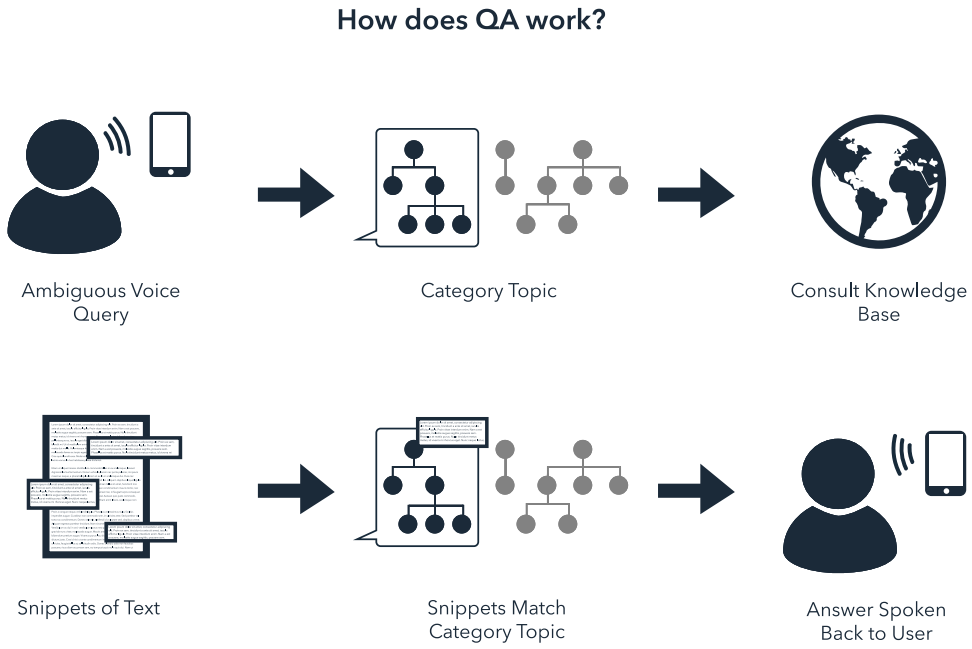
So questions such as:

1. *Where can I find black hi-tops?*
2. *Where can I find black above-the-ankle sneakers?*
3. *Where can I find black high Chuck All Stars?*
4. *Where can I find black Freestyles?*

would all be mapped to the same high-quality answer. The virtual assistant could answer questions by searching a database to lookup a question and retrieve the corresponding answer. This would significantly expand the list of possible questions that can be answered correctly by a Virtual Agent, improving its retrieval rate.

Adding the topic category or categories to an input query can improve retrieval of likely relevant documents, by limiting attention to those collections most likely to be relevant and by increasing the scores of likely relevant documents. Thus, for the user asking "Where can I get a Triumph?" from the example above with their preceding interactions classified to the PET FOOD category, a VA would retrieve information from web services (including APIs) it knows to offer pet supplies, or look specifically for stores selling "Triumph" whose product descriptions are classified to PET FOOD.

Similarly, answer extraction can be improved by picking out those phrases that are most relevant to the topic categories as well as to the question itself, as below:



For example, if the user asks a VA "What canyons are ridden in the Tour de France?", the system can return a more useful answer by extracting snippets from its knowledge base about the brand of Canyon bicycles, one of the highly confident topic classifications of the input query, rather than geographical descriptions of the route the Tour de France follows.

Finally, knowing the topic, as well as the type, of the question can be used to filter and rank potential answers by how well they match the kind of answer and the topics that the user desires.

For example, if the user asks a VA "Where can I see Quantico?", it is crucial to understand that Quantico can refer to the topic of geographic places or to the topic of television shows. The question type, "where" often refers to a physical location — however, the addition of the action "to see" (without any conflicting signals like "where can **I go to see**") weights the question type away from geography, and should return the user information about the ABC television network as more confident than the place in Virginia.

Relevance and Usefulness

Besides helping to provide more accurate answers, using taxonomies to model context can improve relevance and usefulness beyond what is normally seen today. Currently, relevance and usefulness can be improved by using simple external context, such as location; "Find me a good pizza shop" can use GPS to find a nearby shop as opposed to one in another neighborhood or city.

However, by modeling the user's interests expressed in the current conversation, remembering the user's general interests, and structuring these within a consistent topic taxonomy, VA systems will be able to better predict what answers to a question will be most relevant and useful to a user in a specific situation.

Consider "Where can I get a reasonably priced attractive suit?" To give a truly relevant answer, a VA must know the user's gender and age, whether a business suit or swimsuit is meant, whether the user is looking for an online or brick-and-mortar purchase, and what "reasonably priced" means to the user. None of these can be answered by analysis of just the input question, no matter how sophisticated. However, if the system tracks and classifies that the user was previously talking about swimming, beaches, online shopping, or specific upscale or downscale brands, a more relevant and useful answer can be constructed.

Follow-Up and Intent Disambiguation

Even the best modeling of conversational context will not resolve every ambiguity and lack of specificity, however. At times, VA systems will need to reference multiple contextual information sources, as well as general background knowledge. If the user asks for "Jets scores" and the VA knows the user is located in Winnipeg, Manitoba, their intent is probably hockey information. If the user asks for "Jets scores" and is in New York, their intent is probably football information—unless it's between February and June, when the NHL season is active and the NFL season is completed.

At other times, VAs need to interact with their users and ask clarifying questions. Consider the previous example with ambiguous focus, "Who sells coffee that pays good wages?" A simple follow-up question might be "Would you like to see nearby barista job listings?" but this will be hard to understand for the user who is not thinking about the need of the system to determine the question's focus. A better follow-up would be "Are you interested in buying some fair trade coffee, or are you interested in jobs in coffee shops with high employee satisfaction?"

To get to that naturalistic response, the VA would need to know that "good wages" is a conceptual attribute valued by both job-seekers and shoppers looking for certain products; the VA would need to know that "good wages" on a career level is linked to employee satisfaction, but on a product level its linked with fair international trade. The system would thus need to have a linked taxonomy of concepts and be able to recognize when terms in a question refer to those concepts.

Relevance and Personalization through Structured Knowledge

> ***"We want, over the next five or ten years, to take on a road map to try to understand everything in the world semantically and map everything out. These are the big themes for us and what we are going to try and do over the next five or ten years"***
>
> —Mark Zuckerberg, CEO, Facebook, TechCrunch Keynote, September 11, 2013

The function of structured knowledge, disambiguation, and follow-up goes beyond simple lexical clarifications, and can help at all levels of the process, including speech-to-text. Because successful taxonomies operate with controlled vocabularies or positive/negative business rules, deploying these rules would improve speech-to-text accuracy. Consider the question:

Where can I get pens?

In many accents (particularly in the southern United States), “pen” and “pin” sound very similar. Is the system sure which one was said? If not, it would use its taxonomy with knowledge of grammar to validate the concept.

A hypothetical flow of the VA may be:

1. The grammatical structure indicates that the word is a noun or part of a noun phrase.
2. The sound of the word is not clear, but the highest confidence options include “pen” and “pin.”
3. The VA consults its topical taxonomy to define the scope of possible meanings:
    a. The topical understanding of “Pen” is associated with noun-based concepts of: writing implements, enclosed spaces, style of handwriting, style of authorship…
    b. The topical understanding of “Pin” is associated with noun-based concepts of: small devices for fastening pieces of cloth, metal rods for holding machine parts together, jewelry, badges…
4. The VA applies knowledge of local context of the interaction to narrow the likely meanings:
    a. If the full phrase spoken to the VA was: “I need to order some glass head p_ns,” applying a topical knowledge base would identify that “glass head pens” do not exist as part of the topic hierarchy, but “glass head pins” do.
5. The VA applies knowledge of the user’s profile context to narrow the likely meanings:
    a. If the full phrase spoken to the VA was: “I need to order some ball-point p_ns,” the local context is no longer helpful because there are accurate topical understandings for both “ball-point pens” and “ball-point pins.”
    b. A pure statistically driven VA might fall back on frequency of usages from all digital interactions to form the list of likely meaning, which would probably weight toward “pens.” However, a more sophisticated VA would recognize if the user had, in a recent prior conversation, asked about other sewing topics, or has a high interaction history in topics of fiber crafts. This user can be delivered their desired answer, in a way that creates bond and trust with the VA for understanding their overall goals.

If these considerations all together do not suffice to disambiguate, the VA would use its topical understanding to pose a clarification question:

> *Did you want something for writing or sewing? Or for something else?*

The process need not end there. If the request is for ball-point pens, the system has identified a broad level of meaning in the taxonomy. It can go on to ask questions that identify more specific details of the request, such as:

> *What brand would you like?*
> *What color ink would you like?*

In order to ask these questions, the VA must have prior knowledge of options, and use it to organize its thinking. Or, even more desirably, some clarifying questions can be skipped if there is sufficient information in the previous conversation to draw the answers from, and reconsider previously ambiguous statements through the contextual lens of the now-established goal.

Finally, the VA should remember key aspects of the disambiguating conversation, when they imply important background knowledge needed to understand future interactions. For example, if the user responded to "Did you want something for writing or sewing?" with "I don't sew!" the system should remember that sewing is much less likely to be relevant for that particular user.

Conversational Structure

> ***"I'm super excited about artificial intelligence, but we like to say that there are probably a dozen or half a dozen miracles needed to really build these out to be truly intelligent things (bots)...getting to that future where we have that truly intelligent thing we can have a conversation with I think is years away.***
>
> —Facebook CTO, Mike Schroepfer, Bloomberg West TV, April 14, 2016

Virtual Agents need to be able to interact with their users, and not just provide one-off answers. To converse effectively, VAs must produce both questions and responses that are **relevant** and **natural**; to do this, they must understand the broad sweep of the full conversation, not just the current user request.

The VA must recognize how earlier parts of a conversation inform interpretation of later parts of the same conversation. The VA must distinguish which parts of a conversation cohere as part of a larger unit, and which parts are discrete, or one-off, question-answer pairs.

The VA should possess generic knowledge about how conversations are constructed. Many kinds of conversations are built on the skeleton of a *script* (Schank & Abelson 1977). Scripts establish a rough sequence of actions and information exchange in a particular context. By identifying what scripts are relevant to a conversation, a VA can better constrain the possible interpretations of user utterances and generate more relevant and helpful utterances of its own.

A system capable of conversational interaction will also require a system for tracking where the VA and the user are in an unfolding project, and the ability to chunk the sub-units or elements out of which it is built.

To accomplish this, VAs can draw on the intersection of three basic forms of social organization used to manage complex courses of action:

Sequence Organization:

Speakers use basic grammatical forms to compose questions that initiate the project and distinct units of it ("How can I…", "When is…") By contrast they use reduced, or parasitic, grammatical forms with "and prefacing" to pose questions that continue in progress units ("And how much is that?"). They also distinguish between initiating and responsive actions, thereby enabling the parties to move between leading and following.

Practices for Referring to Places, Persons, Time, and Objects

Speakers draw on alternative practices for managing initial and subsequent references to places, persons, dates, times, and things (Schegloff, 1996). The VA may present the user with a variety of specific flight departure times as the initial reference, and the user may respond that they are interested in "the early morning one," indicating a subsequent reference, rather than a request for an alternative time. VAs can use these references as a method for indicating whether an utterance continues an in-progress sequence or initiates a new one.

Variations in the Pitch, Pace, and Volume of the Talk

Studies suggest that utterances initiating new sequences tend to be louder than the preceding talk and have distinct prosodic patterns (Goldberg, 2004). By contrast, talk within a sequence tends to match the volume of preceding turns.

Another key is understanding users' goals and the typical plans they use to achieve them. This will enable systems to better predict what users are likely to say, improving comprehension and appropriate responses. Such goals are often not expressed explicitly, so VAs will need to identify implied goals – for example, if a user remarks that a flight has "a lot of transfers," they are implying the goal of taking a non-stop flight.

This sort of understanding of the user's goals and plans in a conversation are essential.

Chatbots

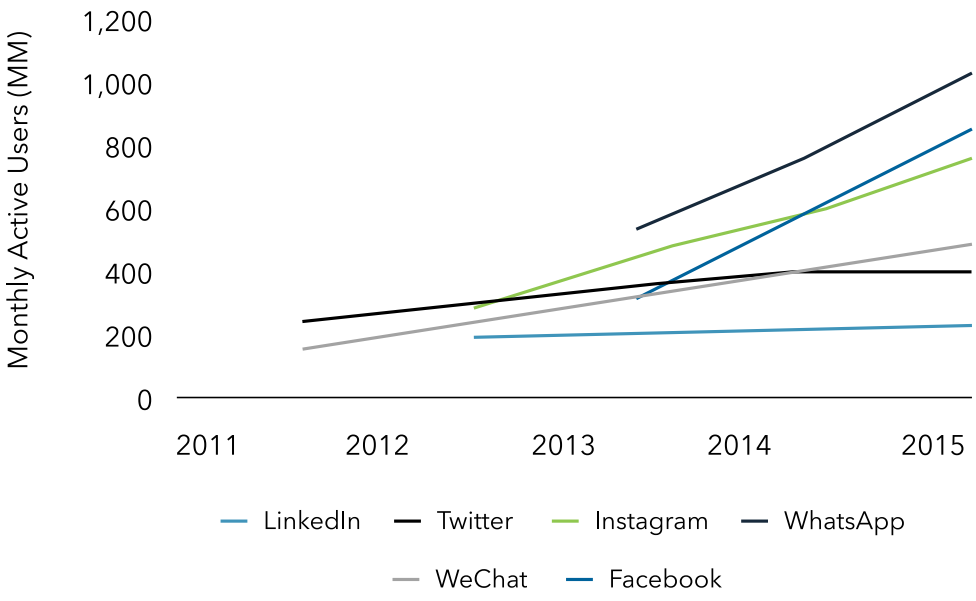
> ***"In China there are more bots put on WeChat every day then there are websites put on the internet. Said another way, WeChat is the internet in China."***
>
> –Ted Livingstone, CEO KIK Messenger , TechCrunch Disrupt, May 11, 2016

> ***"Some 31% of Chinese WeChat users buy from retailers."***
>
> –Mary Meeker, Partner, KPCB, Code Conference, June 1, 2016

Messenger apps like Line, Viber, QQ, and WeChat offer chatbots. Kik recently debuted bots from the Microsoft Bot Framework. It would seem inevitable that Apple's iMessage and Google's just-announced messenger app, Allo, (May 18, 2016) would eventually be open to bots. Apple announced during WWDC (June 13, 2016) that it will give developers access to its Messages app (as well as Siri). According to Ted Livingston, CEO of Kik Messenger, 40% of all U.S. teens use Kik Messenger each month. Internet web chat is ranked alongside social media as the most popular contact channel by Generation Y (born 1981 - 1999). In other words, more popular than email or smartphones. According to David Pierce, senior staff writer at *Wired* magazine, (June 14, 2016), there are more people using message services than there are on social networks and messaging is the "interface of the future." The table below shows the total monthly active users on selected social networks and messengers, 2011 - 2015 (2016 Internet Trends report).

Messaging Continues to Grow Rapidly
Leaders: WhatsApp/ Facebook Messenger/ WeChat

Monthly Active Users on Select Social Networks and Messengers, Global, 2011-2105



> ***"Chatbots will fundamentally revolutionize how computing is experienced by everybody...so pretty much everyone today who is building applications whether they be mobile apps or desktop apps or websites will build bots as the new interface."***
>
> — Satya Nadella, Microsoft CEO, Worldwide Partner Conference, July 11, 2016

Consider a company developing a chatbot for ecommerce sales and customer support. Whether deployed on an existing framework like Skype, Kik, Telegram, Facebook Messenger, or Slack, or in their own app or site, the bot should be able to help users find the products or services they want, but also know when it is appropriate to recommend or cross-sell, by understanding the user's goals and how they are or are not satisfied at each stage of the conversation. As contextual technologies evolve, understanding user preferences, chatbots in essence become intelligent, ie., smartbots. This is already being seen in niche verticals such as health and finance.

Developing all of the knowledge bases and algorithms needed to attain naturalistic conversation is a significant technical challenge, though much research progress has been made. The industry is already gathering and analyzing large amounts of conversational data to extract patterns that can be used to construct script libraries and goal/plan representations. Longer range progress will depend on fundamental advances in systems that can represent and reason about discourse, goals, and plans.

The Role of Taxonomies

The kinds of taxonomies that will be useful will include background knowledge and topics, as well as taxonomies of question/action types, conversation turns, errors, and disambiguation/repair strategies. Improvements in the input data available, both from better analysis of prosodic patterns and from better fundamental natural language processing, will also improve results markedly.

As we have seen, a vital factor for Virtual Agent technology is creating a variety of structured knowledge, including question types, question topics, local user interests, real world context, types of speech acts, conversational structures, user goals, and so on.

The backbone of any structured knowledge representation is a *taxonomy* — specifying a set of concepts in a hierarchical organization as the fundamental terms of discourse (Davis, Shrobe, and Szolovits 1993). Further structure and relationships can be represented in a more complete *ontology*, which can enable more sophisticated inference. But the core ingredient is taxonomy; without a solid taxonomy, no other aspects of the structured knowledge will be as useful, or easily scalable.

What should we seek in a taxonomy? Seth Grimes of the Alta Plana Corporation, has set forth criteria for good taxonomies for text analytics; roughly the same criteria apply for taxonomies underlying Virtual Agents. Adapted from Grimes (2014), we have:

1. **Domain Suitability**
   A system designed for use in medicine, covering pharmaceuticals, diseases, clinical symptoms, treatments, and anatomy would be an odd choice for use in agents geared to helping customers of a retail chain.

2. **Scope**
   Even a suitable choice may not be a best choice. For instance, a taxonomy that captures hospitality terminology will fall short in analysis of travel reviews if it lacks food service and restaurant coverage. General taxonomies are useful, but must be specific enough in the target domain.

3. **Precision**
   Detail enables exact classification, the ability to differentiate based on fine-grained characteristics. Look for a level of precision that provides complete domain coverage (breadth) and enumerates all variations and attributes (depth).

4. **Accuracy**
   Simply put, is the software or taxonomy publisher's work correct? An important question is whether the taxonomy is built fully automatically, or whether it involves human curation for higher accuracy.

5. **Currency**
   Is the taxonomy or model frequently refreshed with new categories, nodes (whether companies, brands, products, or people), and attributes?

6. **Ease of Implementation**
   Does the method work 'out of the box,' across all subjects — a distinctive advantage — or is model training (e.g., for machine learning) or rule-writing (language-engineering approaches) required?

It is common to find VAs relying on taxonomies or ontologies from Wikipedia and dbPedia, Freebase (now Wikidata), Wordnet, Schema.org, and potentially other specialized sources.

However, the most effective taxonomies are those which provide a consistent and common operation across data source, input types, styles of speech, and linguistic practices.

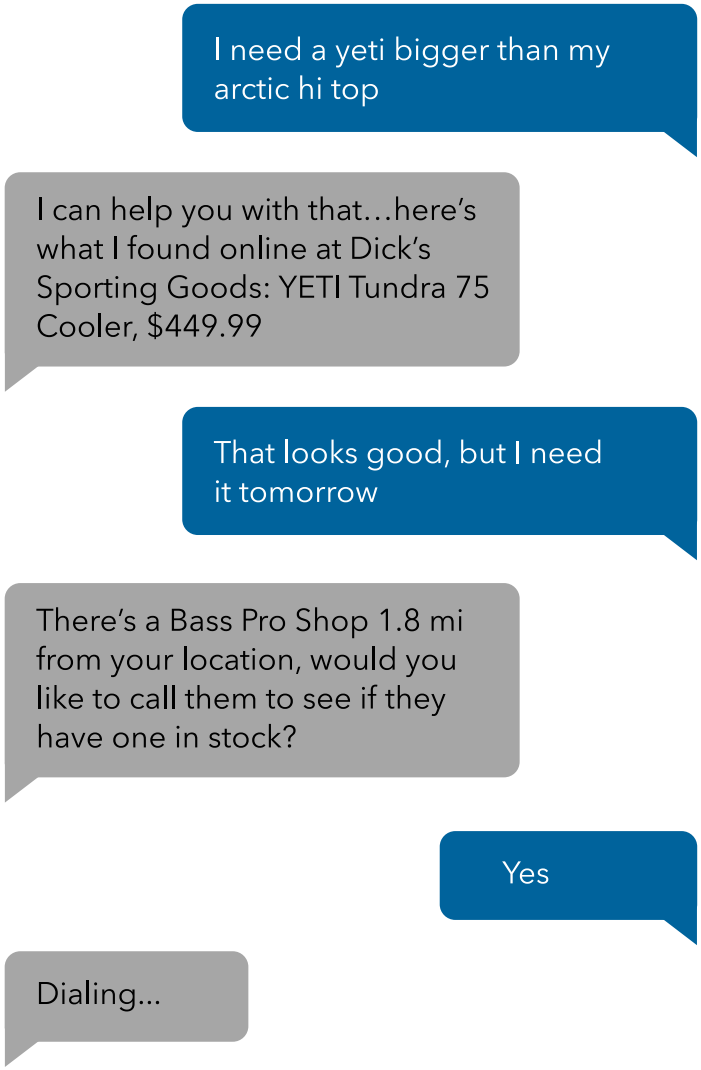
A taxonomy that can universally structure and enrich the direct interaction from a user, as well as the other contextual cues that might be available to the VA — app usage, social network engagement, or content consumption, for example — will provide a richer and more consistent experience.

These abilities are also tested through canonical versus colloquial language use. For example, understanding that "RG3," "RGIII," or "rg three" all refer to NFL Quarterback Robert Griffin III, and do so with a higher confidence than they refer to a model of Lamborghini or a move in chess.

A taxonomy that offers strong links between taxonomic nodes also provides a distinct advantage. These links may be hierarchical, where nodes are nested in clear supertype/subtype relationships, or ontological, where nodes by links that represent a variety of semantic relations that support various kinds of inferences. With linked data, the system can combine a series of weak signals to provide evidence for another, stronger, signal, giving the system better understanding and the user an improved experience.

Especially for personal experiences, a taxonomy that can be easily or automatically expanded is extremely useful. A system that can be individually tailored, adding to or reorganizing its nodes and connections based on the habits of user, will create more trust with the VA. The user will feel encouraged to interact with the VA more often and in more scenarios, rather than contorting or compromising their natural habits to fit within the rigid boundaries of what the VA originally offered.

A taxonomy-powered agent will understand products, services, and consumer needs at multi-levels, ranging from abstract to specific. A taxonomy-powered agent will understand brands, product classes, products, components, and attributes. A taxonomy-powered agent can bridge the divide between the user's intent and the vendor which will satisfy the user. Through contextual, semantically structured knowledge, a VA can support progressively difficult queries:



Without a clear understanding of context, a "yeti" could have been understood to be a mythical creature, a crustacean, an airline, a car brand, a bicycle brand, a microphone, or a cooler. But taxonomy-powered agents understand that "arctic hi top" refers to the Arctic Zone Hi-Top lunch box; and by classifying vast numbers of search and social queries, they understand lunch boxes are more frequently mentioned in relation to Yeti Coolers than any of the other possible interpretations.

Together with an understanding of conversational structure and typical users' goals, the system can successfully complete the exchange to the user's satisfaction.

Establishing accurate context, and aligning the intent of the user with the proper vendor of the desired product or service, allows the taxonomy-powered agent to leverage specialized services and other structured knowledge across the web. A taxonomy-powered agent can more easily translate and standardize the highly varied language of user inputs with the rigid expectations of existing product and service vendors.

The Route Forward

> ***"The opportunity here and the excitement should be around these digital assistants...this is a new platform play, it is a race to a single interface."***
>
> – Gary Morgenthaler, Partner, Morgenthaler Ventures
> (seed investor in Siri and Viv Labs), Bloomberg West TV, June 13, 2016

The analysis in this paper has one fundamental conclusion with respect to the main technological challenge that the Virtual Agent market confronts over the next decade or so: The key is structured knowledge and inference, producing rich context.

A Virtual Agent needs to represent and effectively use knowledge about the world, about its users, about typical tasks, about conversational structure, about conversational errors and repairs, and so forth. While machine learning and data analysis can help and will be essential, high-quality knowledge bases curated by human experts are well suited to improving VA technology to create more human-like exchanges.

Philosophers of language, cognitive scientists, and social psychologists agree that establishing the relevant "context" for understanding human utterances and actions poses significant analytical problems. And yet human agents manage this task routinely.

Virtual Agents need to perform this task at a high, if not human, level to be accepted as natural conversational partners and thus to realize the promise of conversational interfaces. The key will be to attain such levels of performance without having to achieve full human-level intelligence, by finding shortcuts for modeling the most relevant aspects of conversational context.

The present growing market of chat and messenger bots will encourage more companies and developers to participate in Virtual Agent-like exchanges. These Agents will start out with very narrow knowledge bases and conversational limits, constrained to only the products or services the organization offers: ordering a bouquet of flowers, or instructing which toppings to put on a pizza. Users deviating from these tightly scripted interactions will meet ungraceful errors, and will be funneled back into the script. But the data collected through these interactions will be valuable, providing any developer a massive testing ground of individuals and how they ask questions, make requests, and forge conversation patterns.

Structured knowledge, in the form of taxonomies and ontologies, will power more accurate, more responsive voice assistants, virtual agents, and conversational user interfaces. Players that are able to incorporate this key technological ingredient effectively into their systems will likely dominate, particularly due to speed to contextual understanding.

VAs built using these principles will support a wide variety of highly desirable applications, including scalable online shopping support agents; a factual research assistant; a diagnostic agent for health and medical concerns; banking services; a financial planner that can make recommendations based on market analysis as well as individual preferences; or a truly intelligent personal assistant who will remember who you are, what you want, what you do, etc.

In the longer term, Virtual Agents will model their users' interests, goals, plans, and worlds with greater accuracy and precision, and will use this information to anticipate their needs, understand their requests, and be natural in conversation with them. As they interact with their users more like humans, users will adjust as well, and treat Virtual Agents more as conversational partners.

A fully articulated VA as described above has the possibility to become the first "infomediary" as predicted by senior McKinsey consultants John Hagel III and Mark Stinger in their book *Net Worth* published 17 years ago (1999, HBS Press). The company that first manages to develop this technology to this level will therefore be well-poised to achieve a large share of this enormous growing market, as they enable users to protect themselves from intrusive advertising messages and to get intuitive and helpful solutions in their daily lives, including task management, research, ecommerce, and more.

Related Research

R. Davis, H. Shrobe, and P. Szolovits. What is a Knowledge Representation? AI Magazine, 14(1):17-33, 1993.

Schank, R. C., & Abelson, R. P. (1977). *Scripts, plans, goals, and understanding: An inquiry into human knowledge structures. Psychology Press.*

Emanuel A. Schegloff: "Reflections on Studying Prosody in Talk-in- Interaction," *Language and Speech*, 41:3/4, 1998, 235-63.

Grice, Paul (1975). "Logic and Conversation." In Cole, P.; Morgan, J. *Syntax and semantics*. 3: Speech acts. New York: Academic Press. pp. 41–58.

Emanuel A. Schegloff: "Some Practices for Referring to Persons in Talk-in-Interaction: A Partial Sketch of a Systematics" in B. Fox (ed.), Studies in Anaphora (Amsterdam: John Benjamins, 1996), 437-85.

Goldberg, Jo Ann (2004). Amplitude shift: A mechanism for the affiliation of utterances in conversational interaction. In Gene Lerner (Ed), \; John Benjamins. pp. 257-298.

Allied Market Research (2016). World Intelligent Virtual Assistant Market Opportunities and Forecasts, 2014 - 2020. http://www.researchandmarkets.com/research/wfgf62/world_intelligent

Authors

Shlomo Argamon
Dr. Shlomo Argamon is Professor of Computer Science and Director of the Master of Data Science at Illinois Institute of Technology, in Chicago, and is a Fellow of the British Computer Society. His research focuses on developing computational methods for style-based analysis of natural language using machine learning and shallow lexical semantic representations, exploring applications in intelligence analysis, forensic linguistics, biomedical informatics, and humanities scholarship. He has published over 100 scientific articles on machine learning and computational linguistics, and is the co-editor of *Computational Methods in Counterterrorism* (Springer, 2009) and *The Structure of Style* (Springer, 2010).

Geoffrey Raymond
Dr. Geoffrey Raymond is Professor of Sociology at the University of California, Santa Barbara. His research interests include the study of talk and other conduct in interaction and research methods. His publications have appeared in American Sociological Review, Social Psychology Quarterly, Research on Language and Social Interaction, and Language in Society, and his books include Talk and Interaction in Social Research Methods, (co-edited with P. Drew and D. Weinberg); Conversational Repair and Human Understanding, (co-edited with M. Hayashi and J. Sidnell), Units of Talk – Units of Action, (co-edited B. Szczepek Reed).

Sponsor

eContext™
Always On Topic

This paper was sponsored by eContext. The world's largest semantic text classification engine, eContext classifies text in real time to any of its 450,000 topic categories within a hierarchical structure. Powered by a dynamically growing proprietary knowledge base of over 55 million curated rules and billions of documents and online consumer interactions, eContext has invested more than 1 million hours of R&D to date. eContext's 25 verticals cover all sectors of the consumer digital experience: commerce, chat, social, news, entertainment, and more. eContext's service accepts inputs from 35 languages and can be used to inject natural-language intelligence into voice-activated assistants, topically map path-to-purchase and consumer journey, predict user behavior, classify videos, validate image recognition, and structure training corpuses. Clients include Ask.com, Datasift, Kantar Media, and Publicis Groupe. eContext is owned by metasearch company Info.com. For more information, visit www.econtext.com.

180 N. Michigan Ave.
Suite 1400
Chicago, IL 60601

9 Belgrave Road
Westminster, London SW1V 1QB
United Kingdom

+312-477-7310
hello@econtext.com
www.econtext.com

+44 (0)20 7834 5000