Cognitive Computing and Big Data Analytics
By Judith Hurwitz, Marcia Kaufman and Adrian Bowles

CHAPTER
3

Natural Language Processing in Support of a Cognitive System

One of the aspects that distinguish a cognitive system from other data-driven techniques is the capability to manage, understand, and analyze unstructured data in context with the questions being asked. In many organizations as much as 80 percent of the data that is collected and stored is unstructured. To make good decisions, these documents, reports, e-mail messages, speech recordings or images, and videos must be understood and analyzed to make good decisions. For example, in medical journals there are millions of articles published in a single year that can offer new treatment options. In the retail market, there are billions of social media conversations that are leading indicators of future trends. There is important information that is buried inside voice and video recordings that can have an impact on a variety of fields. Unlike structured database data, which relies on schemas to add context and meaning to data, unstructured information must be parsed and tagged to find the elements of meaning. Tools for this process of identifying the meaning of the individual words include categorization, thesauri, ontologies, tagging, catalogs, dictionaries, and language models.

In a cognitive system, the developer needs to generate and test hypotheses and provide alternative answers or insights with associated confidence levels. Often the body of knowledge used within the cognitive system is text-based. In this situation, Natural Language Processing (NLP) techniques interpret the relationships between massive amounts of natural language elements.

The availability of large amounts of unstructured content is critical to creating a meaningful model of information to support continuous learning. Keep in

mind, as discussed in Chapter 2, "Design Principles for Cognitive Systems," not all unstructured data is text. There is a requirement in some cognitive computing systems to support images, video, speech, and sensor data, depending on how the data will be used. Although the focus of this chapter is on the ability to use NLP techniques to support the continuous learning life cycle, other approaches are emerging to manage and process information that is not text-based.

The Role of NLP in a Cognitive System

NLP is a set of techniques that extract meaning from text. These techniques determine the meaning of a word, phrase, sentence, or document by recognizing the grammatical rules—the predictable patterns within a language. They rely, as people do, on dictionaries, repeated patterns of co-occurring words, and other contextual clues to determine what the meaning might be. NLP applies the same known rules and patterns to make inferences about meaning in a text document. Further, these techniques can identify and extract elements of meaning, such as proper names, locations, actions, or events to find the relationships among them, even across documents. These techniques can also be applied to the text within a database and have been used for more than a decade to find duplicate names and addresses or analyze a comment or reason field, for instance, in large customer databases.

The Importance of Context

Translating unstructured content from a corpus of information into a meaningful knowledge base is the task of NLP. *Linguistic analysis* breaks down the text to provide meaning. The text has to be transformed so that the user can ask questions and get meaningful answers from the knowledge base. Any system, whether it is a structured database, a query engine, or a knowledge base, requires techniques and tools that enable the user to interpret the data. The key to getting from data to understanding is the quality of the information. With NLP it is possible to interpret data and the relationships between words. It is important to determine what information to keep and how to look for patterns in the structure of that information to distill meaning and context.

NLP enables cognitive systems to extract meaning from text. Phrases, sentences, or complex full documents provide context so that you can understand the meaning of a word or term. This context is critical to assessing the true meaning of text-based data. Patterns and relationships between words and phrases in the text need to be identified to begin to understand the meaning and actual intent of communications. When humans read or listen to natural language text, they automatically find these patterns and make associations between words to determine meaning and understand sentiment. There is a great deal of ambiguity

in language, and many words can have multiple meanings depending on the subject matter being discussed or how one word is combined with other words in a phrase, sentence, or paragraph. When humans communicate information there is an assumption of context.

For example, imagine that a truck driver wants to use a cognitive system to plan a trip. He obviously needs to know the best route to travel. However, it would be even better if he could know what weather patterns are anticipated in the week of his trip. He also would like to anticipate any major construction projects that he should avoid. It would also be helpful to understand which routes prohibit trucks that weigh more than 10 tons. The truck driver may collect the answers to these questions. However, it would require him to access multiple systems, search different databases, and ask targeted questions. Even when the truck driver finds all the answers, they are not correlated together to provide the optimal travel route based on his requirements at a specific point in time. The same truck driver will have entirely different questions two weeks later. This time the truck driver may be planning his return after delivering merchandise, and he wants to build a vacation into his plans. The recipient (the truck driver) is required to bring context and understanding to the fragmented information he is gathering.

Now look at the example of a lung cancer specialist who is reviewing an MRI. Although some MRIs provide precise information to diagnose a problem, there are many shades of gray. The specialist may want to compare the MRI results to other patients that appear to have similar conditions. The specialist has treated lung cancer patients for many years and has certain hypotheses about the most appropriate treatments. However, one specialist cannot possibly keep up with all the new research and new treatments that are discussed in technical journals. That specialist needs to ask the cognitive system to look for the anomalies that appear in several MRIs. She may want to ask deeper questions to see what other specialists have experienced in treating the same type of lung cancer. She may want to ask for evidence and conduct a dialog with the cognitive system to understand context and relationships.

Both these examples point out the complexities of gaining insight from text and language. Written text often excludes history, definitions, and other background information that would help the reader understand more of the context for the text. The reader of the text brings his own level of experience to help understand the meaning. Therefore, humans use their understanding of the world to make the connections to fill in the context. Of course, depending on the level of knowledge and the expertise required for the text, some text may not be understood without additional information or training. NLP tools rely on rules of language to determine meaning and extract elements. When combining NLP tools in the context of cognitive systems, these tools have to work with a system where the data is dynamic by definition. This means that the system is designed to learn from examples or patterns; therefore language has to be interpreted based on context.

An NLP system begins with letters, words, and some predefined knowledge store or dictionary that helps to define what words mean. A word by itself lacks context. An NLP builds layers of contextual understanding by first looking to the left and right of that word to identify verb phrases, nouns, and other parts of speech. To build up the layers of understanding, the NLP can extract elements of meaning that can answer questions such as:

- Is there a date? When was the text generated?
- Who is speaking?
- Are there pronouns in the text? To whom or what do they refer?
- Are there references to other documents in the text?
- Is there important information in a previous paragraph?
- Are there references to time and place?
- Who or what is acting, and who/what is being acted upon?

What are the relationships of the entities (people, places, and things) to each other (usually indicated by verbs)? It is important to distinguish the actor and recipient of transitive verbs. (For example, who is doing the hitting and who is getting hit.)

There are many layers to the process of understanding meaning in context. Various techniques are used such as building a feature vector from any information that can be extracted from the document. Statistical tools help with information retrieval and extraction. These tools can help to annotate and label the text with appropriate references (that is, assigning a name to an important person in the text). When you have a sufficient amount of annotated text, machine learning algorithms can ensure that new documents are automatically assigned the right annotations.

Connecting Words for Meaning

The nature of human communications is complicated. Humans are always transforming the way language is used to convey information. Two individuals can use the same words and even the same sentences and mean different things. We stretch the truth and manipulate words to interpret meaning. Therefore, it is almost impossible to have absolute rules for what words mean on their own and what they mean within sentences. To understand language we have to understand the context of how words are used in individual sentences and what sentences and meanings come before and after those sentences. We are required to parse meaning so that understanding is clear. It is not an easy task to establish context so that those individuals asking questions and looking for answers gain insights that are meaningful.

THE HISTORY OF NLP

The desire to achieve techniques for transforming language has been around for decades. In fact, some historians believe that the first attempt to automate the translation from one language to the next occurred as early as the 17th century. From the 1940s to the late 1960s, much of the work in NLP was targeted to machine translation—translating between human languages. However, these efforts discovered a number of complexities that couldn't yet be addressed, including syntactic and semantic processing. The primary technique for translating in those years came through using dictionaries to look up words that would then be translated to another language—a slow and tedious process. This problem led computer scientists to devise new tools and techniques focused on developing grammars and parsers with a focus on syntax. The 1980s saw the evolution of more practical tools such as parsers and grammars to allow systems to better understand not just words but the context and meaning of those words. Some of the most important topics that were developed during the 1980s were the notions of word sense disambiguation, probabilistic networks, and the use of statistical algorithms. In essence, this period saw the beginning of moving from a mechanical approach to natural language into a computational and semantic approach to the topic. The trends in NLP in the past two decades have been in language engineering. This movement has coincided with the growth of the web and the expansion of the amount of automation in text as well as spoken language tools.

Understanding Linguistics

NLP is an interdisciplinary field that applies statistical and rules-based modeling of natural languages to automate the capability to interpret the meaning of language. Therefore, the focus is on determining the underlying grammatical and semantic patterns that occur within a language or a sublanguage (related to a specific field or market). For instance, different expert domains such as medicine or laws use common words in specialized ways. Therefore, the context of a word is determined by knowing not just its meaning within a sentence, but sometimes by understanding whether it is being used within a particular domain. For example, in the travel industry the word "fall" refers to a season of the year. In a medical context it refers to a patient falling. NLP looks not just at the domain, but also at the levels of meaning that each of the following areas provide to our understanding.

Language Identification and Tokenization

In any analysis of incoming text, the first process is to identify which language the text is written in and then to separate the string of characters into words (*tokenization*). Many languages do not separate words with spaces, so this initial step is necessary.

Phonology

Phonology is the study of the physical sounds of a language and how those sounds are uttered in a particular language. This area is important for speech recognition and speech synthesis but is not important for interpreting written text. However, to understand, for instance, the soundtrack of a video, or the recording of a call center call, not only is the pronunciation of the words important (regional accents such as British English or Southern United States), but the intonation patterns. A person who is angry may use the same words as a person who is confused; however, differences in intonation will convey differences in emotion. When using speech recognition in a cognitive system, it is important to understand the nuances of how words are said and the meaning that articulation or emphasis conveys.

Morphology

Morphology refers to the structure of a word. Morphology gives us the stem of a word and its additional elements of meaning. Is it singular or plural? Are the verbs first person, future tense, or conditional? This requires that words be partitioned into segments known as *morphemes* that help bring understanding to the meaning of terms. This is especially important in cognitive computing, since human language rather than computing language is the technique for determining answers to questions. Elements in this context are identified and arranged into classes. There are elements including prefixes, suffixes, infixes, and circumfixes. For example, if a word begins with "non-" it has a specific reference to a negative. There is a huge difference in meaning if someone uses the verb "come" versus the verb "came." Combinations of prefixes and suffixes can be combined to form brand new words with very different meanings. Morphology is also used widely in speech and language translation as well as the interpretation of images. Although many dictionaries have been created to provide explanations of different constructions of words in various languages, it is impossible for these explanations to ever be complete (each human language has its own context and nuances that are unique). In languages such as English, rules are often violated. There are new words and expressions created every day.

This process of interpreting meaning is aided by the inclusion of a lexicon or repository of words and rules based on the grammar of a specific language. For example, through a technique called parts of speech tagging or tokenization, it is possible to encapsulate certain words that have definitive meaning. This may be especially important in specific industries or disciplines. For example, in medicine the term "blood pressure" has a specific meaning, However, the words blood and pressure when used independently can have a variety of meanings. Likewise, if you look at the elements of a human face, each component may independently not provide the required information.

Lexical Analysis

Lexical analysis within the context of language processing is a technique that connects each word with its corresponding dictionary meaning. However, this is complicated by the fact that many words have multiple meanings. The process of analyzing a stream of characters from a natural language requires a sequence of *tokens* (a string of text, categorized according to the rules as a symbol such as a number or comma). Specialized *taggers* are important in lexical analysis. For example, an *n-gram tagger* uses a simple statistical algorithm to determine the tag that most frequently occurs in a reference corpus. The analyzer (sometimes called a *lexer*) categorizes the characters according to the type of character string. When this categorization is done, the lexer is combined with a parser that analyzes the syntax of the language so that the overall meaning can be understood.

The lexical syntax is usually a regular language whose alphabet consists of the individual characters of the source code text. The phrase syntax is usually a context-free language whose alphabet consists of the tokens produced by the lexer. Lexical analysis is useful in predicting the function of grammatical words that initially could not be identified. For example, there might be a word like "run" that has multiple meanings and can be a verb or a noun.

Syntax and Syntactic Analysis

Syntax applies to the rules and techniques that govern the sentence structure in languages. The capability to process the syntax and semantics of natural language is critical to a cognitive system because it is important to deduct inferences about what language means based on the topic it is being applied to. Therefore, although words may have a general meaning when used in conversation or written documents, the meaning may be entirely different when used in context of a specific industry. For example, the word "tissue" has different definitions and understanding based on the context of its use. For example, in biology, tissue is a group of biological cells that perform a specific function. However, a tissue can also be used to wrap a present or wipe a runny nose. Even within a domain context, there can still be word-sense ambiguity. In a medical context, "tissue" can be used with skin or a runny nose.

Syntactical analysis helps the system understand the meaning in context with how the term is used in a sentence. This syntactic analysis or parsing is the overall process for analyzing the string of symbols in a natural language based on conforming to a set of grammar rules. Within computational linguistics, *parsing* refers to the technique used by the system to analyze strings of words based on the relationship of those words to each other, in context with their use. Syntactical analysis is important in the question-answering process. For example, suppose you want to ask, "Which books were written by British women authors before the year 1800?" The parsing can make a huge difference in the accuracy of the answer. In

this case, the subject of the question is books. Therefore, the answer would be a list of books. If, however, the parser assumed that "British woman authors" was the topic, the answer would instead be a list of authors and not the books they wrote.

Construction Grammars

Although there are many different approaches to grammar in linguistics, construction grammar has emerged as an important approach for cognitive systems. When approaching syntactical analysis, the results are often represented in a grammar that is often written in text. Therefore, interpretation requires a grammar model that understands text and its semantics. Construction grammar has its roots in cognitive-oriented linguistic analysis. It seeks to find the optimal way to represent relationships between structure and meaning. Therefore, construction grammar assumes that knowledge of a language is based on a collection of "form and function pairings." The "function" side covers what is commonly understood as meaning, content, or intent; it usually extends over both conventional fields of semantics and pragmatics. Construction grammar was one of the first approaches that set out to search for a semantically defined deep structure and how it is manifested in linguistic structures. Therefore, each construction is associated with the principle building blocks of linguistic analysis, including phonology, morphology, syntactic, semantics, pragmatics, discourse, and prosodic characteristics.

Discourse Analysis

One of the most difficult aspects of NLP is to have a model that brings together individual data in a corpus or other information source so that there is coherency. It is not enough to simply ingest vast amounts of data from important information sources if the meaning, structure, and intention cannot be understood. Certain assertions may be true or false depending on the context. For example, people eat animals, but people are animals, and in general don't eat each other. However, timing is important to understanding context. For example, during the 18th century, cigarette smoking was thought to be beneficial to the lungs. Therefore, if someone were ingesting an information source from that period of time, it would assume that smoking was a good thing. Without context there would be no way to know that premise of that data was incorrect. Discourse is quite important in cognitive computing because it helps deal with complex issues of context. When a verb is used, it is important to understand what that verb is associated with in terms of reference. Within domain-specific sources of data you need to understand the coherence of related information sources. For example, what is the relationship between diabetes and sugar intake? What about the relationship between diabetes and high blood pressure? The system needs to be modeled to look for these types of relationships and context.

Another application for discourse analysis is the capability to understand the "voice of the customer" using sentiment analysis to determine the real feelings and intents being expressed by customers online. The ability to understand the full spectrum of customer issues is well suited for an NLP-focused application. This type of application helps bring together a lot of highly structured and less structured customer information to gain a full understanding of how that customer feels about the company. Is the customer happy? Can this customer get the right level of support? Does the vendor come across as a provider who understands his customer?

Pragmatics

Pragmatics is the aspect of linguistics that tackles one of the fundamental requirements for cognitive computing: the ability to understand the context of how words are used. A document, an article, or a book is written with a bias or point of view. For example, the writer discussing the importance of horses in the 1800s will have a different point of view than the writer talking about the same topic in 2014. In politics, two documents might discuss the same topic and take opposite sides of the argument. Both writers could make compelling cases for their point of view based on a set of facts. Without understanding the background of the writer, it is impossible to gain insight into meaning. The field of pragmatics provides inference to distinguish the context and meaning about what is being said. Within pragmatics, the structure of the conversation within text is analyzed and interpreted.

Techniques for Resolving Structural Ambiguity

Disambiguation is a technique used within NLP for resolving ambiguity in language. Most of these techniques require the use of complex algorithms and machine learning techniques. Even with the use of these advanced techniques, there are no absolutes. Resolution of ambiguity must always deal with uncertainties. We can't have complete accuracy; instead, we rely on the probability of something being most likely to be true. This is true in human language and also in NLP. For example, the phrase, "The train ran late," does not mean that the train could "run"; rather the train was expected to arrive at the station later than it was scheduled. There is little ambiguity in this statement because it is a commonly known phrase. However, others phrases are easily misunderstood. For example, examine the phrase, "The police officer caught the thief with a gun." One might decide that it was the police officer that used a gun to arrest the thief. However, it may well have been the thief was using the gun to commit a crime. Sometimes, the truth of meaning can be hidden inside a complicated sentence.

Because cognitive computing is a probabilistic rather than a deterministic approach to computing, it is not surprising that probabilistic parsing is one

way of solving disambiguation. Probabilistic parsing approaches use dynamic programming algorithms to determine the most accurate explanation for a sentence or string of sentences.

Importance of Hidden Markov Models

One of the most important statistical models for processing both image and speech understanding are *Markov Models*. Increasingly, these models are fundamental to understanding the hidden information inside images, voice, and video. It is now clear that it is complicated to gain a clear understanding of the meaning that is often hidden within language. While the human brain automatically understands how to cope with the fact that the real meaning of a sentence may be indirect, "The cow jumped over the moon" may seem like an impossible task if the sentence were read literally. However, the sentence refers to a song for young children and is intended to be unrealistic and silly. The human mind calculates the probability that this sentence is intended to be a literal action between the cow and the moon. The human understands through the context of their environment, which may dictate a specific interpretation.

The way systems interpret language requires a set of statistical models that are an evolution of a model developed by A.A. Markov in the early 1900s. Markov asserted that it was possible to determine the meaning of a sentence or even a book by looking at the frequency that words would occur in text and the statistical probability that an answer was correct. The most important evolution of Markov's model for NLP and cognitive computing is Hidden Markov Models (HMMs).

The premise behind HMMs is that the most recent data will tell you more about your subject than the distant past because the models are based on the foundations of probability. HMMs therefore help with predictions and filtering as well as smoothing of data. Hidden Markov Models (HMMs) are intended to interpret "noisy" sequences of words or phrases based on probabilistic states. In other words, the model takes a group of sentences or sentence fragments and determines the meaning. Using HMMs requires thinking about the sequence of data. HMMs are used in many different applications including speech recognition, weather patterns, or how to track the position of a robot in relationship to its target. Therefore, Markov models are very important for when you need to determine the exact position of data points when there is a very noisy data environment. Applying HMMs allows the user to model the data sequence supported by a probabilistic model.

Within the model, an algorithm using supervised learning will look for repeating words or phrases that indicate the meaning and how various constructs affect the likelihood that a meaning is true. Markov models assume that the probability of a sequence of words will help us determine the meaning. There are a number of techniques used in HMMs to estimate the probability that a

word sequence has a specific meaning. For example, there is a technique called maximum likelihood estimation that is determined by normalizing the content of a corpus.

The value of HMMs is that they do the work of looking for the underlying state of sentences or sentence fragments. Therefore, as the models are trained on more and more data they abstract constructs and meaning. The capability to generate probabilities of meaning and the state transition are the foundation of HMM models and are important in cognitive understanding of unstructured data. The models become more efficient in their ability to learn and to analyze new data sources. Although HMMs are the most prevalent method in understanding the meaning of sentences, another technique called maximum entropy is designed to establish probability through the distribution of words. To create the model, labeled training data is used to constrain the model. This classifies the data.

There are a number of approaches that are important in understanding a corpus in context with its use in a cognitive system. The next section examines some of the most important techniques that are being used.

Word-Sense Disambiguation (WSD)

Not only do you have to understand a term within an ontology, it is critical to understand the meaning of that word. This is especially complex when a single word may have multiple meanings depending on how it is used. Given this complexity, researchers have been using supervised machine learning techniques. A classifier is a machine learning approach that organizes elements or data items and places them into a particular class. There are different types of classifiers used depending on the purpose. For example, document classification is used to help identify where particular segments of text might belong in your taxonomy.

Often classifiers are used for pattern recognition and therefore are instrumental in cognitive computing. When a set of training data is well understood, supervised learning algorithms are applied. However, in situations in which the data set is vast and it cannot easily be identified, unsupervised learning techniques are often used to determine where clusters are occurring. Scoring of results is important here because patterns have to be correlated with the problem being addressed. Other methods may rely on a variety of dictionaries or lexical knowledge bases. This is especially important when there is a clear body of knowledge—in health sciences, for example. There are many taxonomies and ontologies that define diseases, treatments, and the like. When these elements are predefined, it allows for interpretation of information into knowledge to support decision making. For example, there are well known characteristics of diabetes at the molecular level, the evolution of the disease, and well-tested, successful treatments.

Semantic Web

NLP by itself provides a sophisticated technique for discovering the meaning of unstructured words in context with their usage. However, to be truly cognitive requires context and semantics. Ontologies and taxonomies are approaches that are expressions of semantics. In fact, the capability to combine natural language processing and the semantic web enables companies to more easily combine structured and unstructured data in ways that are more complicated with traditional data analytic methods. The semantic web provides the Resource Description Framework (RDF), which is the foundational structure used in the World Wide Web for processing metadata (information about the origins, structure, and meaning of data). RDF is used as a way to more accurately find data than would be found in a typical search engine. It provides the ability to rate the available content. RDF also provides a syntax for encoding the particular metadata with standards such as XML (Extensible Markup Language) that supports interoperability between data sources. One of the benefits of schemas that are written in RDF is that it provides one set of properties and classes for describing the RDF generated schemas. The semantic web is instrumental in providing a cognitive approach to gaining insights from a combination of structured and unstructured sources of information in a consistent way.

Applying Natural Language Technologies to Business Problems

Earlier this chapter mentioned two examples where professionals needed to gain insights from text and other unstructured data. NLP provides an important tool that enables humans to interact with machines. While we are at the early stages of cognitive computing, there are a number of applications that are emerging that take advantage of some of the capabilities of NLP in context with a specific use or market. IBM demonstrated with its Jeopardy! game challenge that it is possible to answer questions as they are being asked. This next section provides some examples of how NLP technologies can transform some industries by understanding language in context.

Enhancing the Shopping Experience

The most successful web-based shopping sites are those that create a satisfying customer experience. Too often, a customer comes to a site looking for a product based on specific requirements. Typical customers use a search capability to find what they are looking for. Customers may have specific requirements. “I want to purchase a brown and black sweater in size 12 that is not made of wool

and is made in a country that does not use child labor. The sweater should be delivered in no more than 5 days, and there should not be a shipping charge." Although it is possible to get an answer to each individual question, it will probably require the user to ask at least six different questions. In addition, some questions, such as if the manufacturer of the sweater has a history of using child labor, may require a series of questions. Users are required to bring all the answers together to complete their transaction. Using NLP text analytics tools within a cognitive context, it is possible to understand what users are asking and create a dialog to provide a positive and interactive experience between humans and machine. By evaluating the use of words and the pattern of use, customers can be satisfied.

Leveraging the Connected World of Internet of Things

As more and more devices, from cars to highways and traffic lights, are equipped with sensors, there will be the ability to make decisions about what actions to take as conditions change. Traffic management is a complex problem in many large metropolitan areas. If city managers can interact with sensor-based systems combined with unstructured data about events taking place (rallies, concerts, and snow storms), alternative actions can be looked at. A traffic manager may want to ask questions about when to reroute traffic under certain circumstances. If that manager can use an NLP interface to a cognitive system, these questions can be answered in context with everything from weather data to density of traffic to the time when an event will start. Individual domains such as traffic routing and weather predictions will each have their own Hidden Markov Models. In a cognitive system it is possible to correlate this data across domains and models. Matching this data with an NLP engine that interprets textual data can result in significant results. The NLP question and answer interface can help the human interact with this complex data to recommend the next best action or actions.

Voice of the Customer

The capability for companies to understand not only what their customers are saying but also how it will impact their relationship with that customer is increasingly important. One technique companies use to understand customer attitudes is sentiment analysis. *Sentiment analysis* combines text analytics, NLP, and computational linguistics in order to make sense of text-based comments provided by customers. For example, a company can analyze customer sentiment to predict sales for new product offerings from one of its divisions. However, a customer is not simply a customer for a single business unit. Many customers actually will do business with several different business units within the same company. Creating a corpus of customer data across business units can enable the customer service representative to understand all interactions with

customers. Many of these interactions will be stored in notes in customer care systems. These same customers may add comments to social media sites and send e-mail messages directly to the company complaining about problems. There are subtleties in how customers use language that need to be understood to get a clear indication of customer intent.

If the customer is sarcastic or uses the word "not" at the end of a sentence, it is not easily translated. Techniques such as Word Sense Disambiguation are used to decompose the words in a sentence and then provide a sense of the words in context. Word Sense Disambiguation and other fundamental NLP techniques can make the difference between understanding customer satisfaction and missing important signals. To be effective, a business needs to understand the true voice of the customer.

Increasingly, customers are making their preferences understood in new ways. For example, there is a growing requirement to understand the content of data from platforms such as YouTube where individuals provide their personal evaluations of products and services based on their experience. Techniques that understand not just the language but also the intent of those words are a critical part of understanding the voice of the customer. Interpreting not only what words the individual uses but also the order of those words and the intonation of their comments is important.

Although traditional text analytics offerings enable managers to understand words in context, they do not provide the context across lines of business, such as data from manufacturing or delivery systems. A business has to understand customers' attitudes about current problems, future requirements, and what competitors are lurking. To try to get deeper insights, businesses use the net promoter scores to determine how positively or negatively customers are feeling. However, without a true cognitive approach, it is typical that a company will miss some key words and phrases that might include a completely different interpretation of the customers' perceptions of the company. This is why techniques such as Hidden Markov Models may be very important in understanding what a customer is really trying to say about a company or its products and services.

Sentiment Analysis is different across industries. For example, the type of clues that you look for in healthcare will be very different than the types of words that will be meaningful in retail. For example, in healthcare the word "hot" may be an indication of a fever. However, in retail, "hot" may refer to a popular product. Document categorization, ontologies, and taxonomies are important in understanding the difference and making sense of words in context.

In addition to looking for clues in documents, companies rely heavily on information from social media to assess what customers are saying to the company and to other customers. These messages may not always mean what they seem to be saying. For example, a Twitter message that says, "This company sure knows how to treat its customers . . . I wish." is a negative comment. This

is why it is important to use NLP tools for text analytics and sentiment analysis to truly understand your customers. These same tools can be used for competitive intelligence. These tools can determine if there is more discussion about an emerging company in your market that should not be ignored. Chapter 6, "Applying Advanced Analytics to Cognitive Computing," discusses advanced analytics in detail.

Fraud Detection

One of the most important applications for NLP and cognitive computing is the detection of fraud. Traditional fraud detection systems are intended to look for known patterns from internal and external threat databases. Determining risk before it causes major damage is the most important issue for companies dealing with everything from hackers to criminal gangs stealing intellectual property. Although companies leverage firewalls and all sorts of systems that put up a barrier to access, these are not always effective. Smart criminals often find subtle techniques that go under the radar of most fraud detection systems. Having the capability to look for hidden patterns and for anomalies is critical to preventing an event from happening.

In addition, leveraging thousands of fraudulent claims documents, an insurance company can be better prepared to detect subtle indications of fraud. NLP-based cognitive approaches can enable the user to ask questions related to the corpus of data that has been designed based on a model of both acceptable and unacceptable behavior. This corpus can be fed with new information about detected schemes happening somewhere in the world. Understanding not only the words but also their context across many data sources can be applied to fraud prevention. Understanding word sense in complex documents and communications can be significant in preventing fraud.

Summary

Natural Language Processing is one of the technologies that enables humans to understand the meaning of unstructured data. The ability to not just ask questions but to have an ongoing dialog is key to the value of NLP in context with cognitive computing. As you know, there isn't one single, right answer to just about any question in the world. We make conclusions and judgments based on the information we have available. We also make decisions based on the context of that information. This is not an easy challenge. Not all data is text and words. Increasingly, you access content with data embedded in images, videos, speech, gestures, and sensor data. In this case, deep learning techniques are needed to analyze this type of unstructured data.

We are faced with a world in which there is an unending source of data that only grows by the hour. There are new techniques to analyze that data and there are new methods for putting the pieces together. The human mind has the uncanny ability to make the connections between seemingly unrelated events. But humans are flawed in how much information they can find and then ingest at the same time. NLP, when used in combination with machine learning and advanced analytics, can help humans leverage the depth and breadth of human knowledge in new ways.