Derek Erwin
Jay Kaiser
Kenny Kha
Connor Russell
CSCI-B351
12/7/16

Natural language processing (NLP) is the interaction between humans and technology through language. NLP is related to computational linguistics, which deals with using technology to solve language issues found in linguistics. For an expert interview, Chung-chieh Shan, or Ken for short, was asked a variety of questions involving NLP, ranging from topics of current work being done in the field to the future of NLP as a whole. Ken wrote his PhD in Harvard on the integration and comparison of semantic elements of natural and programming languages.

As a graduate student at Harvard, Ken devised a program that was able to search through his college's course catalog using user-defined keywords in an attempt to expedite the class-searching process for his fellow students. Because this program was well-utilized and appreciated by many of his peers, this remains one of his favorite projects in NLP. However, as evidenced by the many NLP projects found online today, this type of application is not unique nor particularly difficult to complete using today's technology, showing the field's swift growth over the past decade.

In recent years there has been a rapid expansion of the number of openly-available implementations that heavily rely on natural language processing. These include commonly used programs like automatic translation tools such as Google Translate, spam filters present in all major email companies, grammar and spelling checkers found in word document processors, and information retrieval-based search engines such as Google or Bing. However, most notable of these NLP applications from a utilitarian perspective are virtual assistants that can be found in many prominent technological sources: Siri in Apple products, Google Now in Android phones, and Cortana in Windows phones and computers. Virtual assistants incorporate a wide variety of NLP processes that together allow for a user to ask questions verbally to the assistant, who then is able to grammatically parse and retrieve an answer for the query.

A similar technology to consumer virtual assistants is IBM's Watson. Watson is a cognitive technology that uses information retrieval and natural language processing to "think" in a similar way to humans. Watson has the ability to build its knowledge by looking at structured data like databases and unstructured data, human-targeted data sources, on the Internet. Although Watson participated in Jeopardy and was able to beat 74-time Jeopardy champion Ken Jennings, it is not able to do this all by itself. In order for Watson to be effective, humans had to provide Watson with certain information so it can begin to train itself on real-world data it gleaned through parsing and determining keywords in both questions and text. Watson then performs a search to determine what information will be useful in answering the question, using NLP techniques described in

more detail below. Possible answers, called candidate answers, are derived from the information, and then another similar search is performed to gather evidence in favor of the answers. Finally, Watson merges the candidate scores using machine learning algorithms to arrive at final confidence levels for each answer. Watson can then reply with the answer that holds the highest confidence value. Although Watson was able to win in Jeopardy, he was not perfect in his understanding of the questions presented to him, demonstrating the challenges of natural language processing. Watson is a state-of-the-art virtual assistant that has the potential to help millions in various fields like healthcare, law, business, and other industries; however, much further work is needed to design a machine that can replicate human-language understanding at similar levels to that of a human.

Natural language is an extremely complex field that consists of a wide variety of interacting factors, each of which must be represented and well accounted for during processing if speech is to be fully understood as well as it is by native human speakers. These range across all levels of the linguistic spectrum, from phonetics and phonology at the basic level all the way up to semantics and pragmatics at the upper echelon. Each of these linguistic sub-fields can be represented and derived using distinct computational linguistic methods, described below through an example. If any of these levels of linguistic phenomena are misrepresented or perform poorly at their given task, sentential information will either be misconstrued or blatantly crash.

When a user asks a question to a decent virtual assistant, the following processes at a *minimum* must all interact correctly and perform accurately to retrieve an appropriate response. So for example, if a user were to ask his device to "*Find me a nearby restaurant that has cheap Chinese food*," the following processes (though not necessarily in the order listed) must all be performed:

Speech-to-text processing is the process where the assistant reads in the acoustic waves taken from the user and converts them into a phonetic format that can be further parsed. Using statistical models, the correct words are retrieved and properly segmented and finally displayed to the user. There exist many factors, however, that work to make this process more difficult to accurately complete. Background noise is universal, be it from weather, other speakers, music, or mere background ambiance, and the assistant cannot misunderstand the user nor incorporate elements from these distractions into its parse either.

Token segmentation is the process where given a stream of speech or a body of text, words are properly isolated and distinguished. In languages like English, where spaces are inserted between words, this process is made simpler. However, in languages like Chinese or Japanese where spaces are non-existent, this process becomes more difficult. Even in languages like English, there exist exceptions; for example, names like "New York City" and idioms like "kick the bucket" must be isolated as unique entities, as if parsed word-for-word, information would be lost.

Part-of-speech (POS) tagging is the process where word tokens are each given their respective parts-of-speech (like noun, verb, adjective, determiner, etc.) that can be

used in further parsing. The type and number of POS tags differs across languages, where the morphological (word-structure) complexity of the language can rapidly expand the number of required parts-of-speech that can be applied to it. For example, languages of simpler word-complexity but complex syntactic structure (like English and Chinese) usually have between 50 and 200 POS tags; however, highly inflective languages (like Hungarian or Korean) can have more than 2000 tags that must all be accounted for and accurately described (Harris et al. 2000). Though there are universal parts-of-speech across languages, differing guidelines provide differing analyses of POS tags for a given language, further complicating this description (Santorini 1990).

Word-sense disambiguation (WSD) is the process where words whose meaning is ambiguous due to multiple possible senses that could be inferred without context are narrowed down to isolate the single correct one. Though it could be assumed that any given word doesn't have more than a handful of possible word-senses, in reality the number of senses can be far more than possibly expected. For example, according to WordNet, a free WSD source for English created by Princeton, the word "fall" has a *minimum* of 44 different senses, 32 of which are verbs and 12 nouns. Though this is a higher number of senses than most English words, this ambiguity must be accounted for and accurately described. POS tagging goes hand-in-hand with cases of word-sense disambiguation where lexical ambiguity of part-of-speech of importance.

Named-entity recognition (NER) is the process where real-world entities of a specific nature are isolated and given a label, be it name, country, culture, date, etc. In cases of languages like English, this process is alleviated by the fact that proper nouns are capitalized, allowing for easier recognition. However, in languages like German, *every* noun is capitalized, and in eastern Asian languages like Korean and Chinese, there is no distinction like capitalization in their given writing system; other methods are necessary to account for these differences.

Relationship extraction is the process where the relationships between entities are inferred through context. This could range from familial relations to ownership between objects and their owners to even pronouns to their referent. There exist many forms of ambiguity that arise from this type of relationship of words in languages like English. For example, in the sentence "George bought his boat yesterday," *his* can refer either to George or to another male either mentioned earlier or not yet mentioned. Because this ambiguity exists for native speakers as well, this process is exceptionally hard is some contexts to resolve computationally.

Syntactic parsing is the process where the relationship between individual words in a given sentence are found and graphically parsed in some grammatical formalism, like influential linguist Noam Chomsky's popular Generative Grammar. In this manner, subjects, direct and indirect objects, and various modifiers are determined and distinguished. In languages like English where syntactic structure is extremely inflexible, this process is relatively simple; however, in languages with looser word order like *most* other languages in the world, to languages with no set word order like Latin, this process

becomes vastly more difficult and must incorporate other processes like morphological segmentation to ensure passable accuracy.

After all these steps are performed, as well as a small variety of others not mentioned, the final step of information retrieval is performed by the virtual assistant. This process relies on pragmatic information that is inherent in the intention of the speaker that is unspoken and implied in every sentence. For example, *"find me a nearby restaurant that has cheap Chinese food"* really tells the virtual assistant to "search online for open restaurants (that will still be open for at least another hour) within a mile radius (or more if the user has shown evidence of owning a car) of a user's current location (taking into account their direction of movement) that serve Chinese food (or food of similar ethnic or fusion background) for under 15 dollars (per person not including drinks) and that does not require a reservation. All this information is well-understood for a human speaker of language, but it must be inferred by the computer to allow proper results to be found.

However, there exists a major problem inherent to every natural language across every corner of the planet: ambiguity. Because many aspects of language are ambiguous, there exist many potential issues that plague every step of information parsing and retrieval in NLP. For humans, ambiguity in natural language serves an important purpose that helps to improve efficiency in communication and decrease the overall effort necessary to linguistic systems (Piantadosi 2012); computers, on the other hand, must be fed information manually, and any potential confusion between various aspects of a sentence cannot be as easily accounted for. Below, a variety of ambiguous sentences and sentence-pairs have been listed, with a description of the ambiguity present. The most important message to take from this list is the fact that not only do computers have difficulty in parsing these sentences, but for many of the sentences below, humans show difficulty as well, proving that even the brains of humans--brains far more advanced than the processors of machines--are not powerful enough for these types of ambiguity.

The horse raced past the barn fell.	*Hidden passivized relative clause.*
Time flies like an arrow. **Fruit flies like a banana.**	*Conflation of parts-of-speech for "time flies like" and "fruit flies like".*
The cat the man the boy loved hated died.	*Recursion of depth two resulting in non-parsability for most native speakers.*
Buffalo buffalo buffalo Buffalo buffalo buffalo Buffalo buffalo.	*"Buffalo" is an animal, verb, and city; unpronounced relative clause.*
Kill somebody. **Kill a process.**	*Differing degrees of the effect of "kill".*
Grandpa *kicked the bucket* yesterday.	*Idiomatic expression cannot be directly parsed.*

If only to expand the difficulty computers face when tasked with parsing and understanding natural language, human languages also tend to be very messy, both in speech and in text. As humans form sentences and speak them aloud, ideas change suddenly and sentence structure is updated to fit this. This means that in the speech of any human from any part of the world, there exist a variety of ungrammatical sentential elements, including repetition, run-on sentences, sentence fragments, changes of topic and purely ungrammatical words and phrases, all of which must be understood by a listener. Humans are able to process these changes in real time, and in most cases, the presence of these factors does little to deter understanding. However, for a computer that must slowly parse sentences bottom-up word-by-word, sudden changes can easily confuse and end a sentence parse in failure.

On the other hand, though written text tends to be more grammatical and well thought-out for a variety of media sources, as errors can be caught and revised before the text is published and parsed by a machine, there exist glaring exceptions to this rule. For example, text on the internet, especially text between peers in informal circumstances like on forums and social media, can hold little to no correct sentence structure, spelling, capitalization, etc. This type of text is often filled with non-words as well, like emoticons, emojis, hashtags, memes, and web addresses, all of which must be handled by a competent parser no matter the severity. Moreover, there are a variety of languages around the world where no form of writing system exists; this lack of data and the aforementioned difficulties present in many forms of text show that written speech does not necessarily have to be easier to deal with for a computer than in spoken speech.

However, there is one factor of language that provides a silver lining to all these factors. Despite the fact that there are so many issues that complicate NLP, language itself has limitations that can be found and restricted. For example, across all languages there are a variety of universal faculties, one of which is the ability to recurse elements of a sentence to an infinite length (Pinker and Jackendoff 2005). Due to computers' comparably and seemingly infinite processing space as compared to the human mind, computers can recurse to depths far further than ever visible in human language. However, even though recursion is present in all human languages,[1] humans rarely can parse structures deeper than a depth of one at any given part of a sentence. This in fact works to actually limit the number of ambiguities and complex parses possible to encounter by a computer. As Ken described in his interview, "humans define what is 'correct.' Because humans are always right, computers will always play catch-up." Because of this, the field of NLP, though expansive and complex, has an eventual upper bound of knowledge to be fulfilled, due to the fact that although language is infinite in design, it is indeed finite in application.

In order for computers to be able to complete each of the variety of language tasks mentioned above and account for ambiguity present in language, a variety of machine learning techniques are commonly applied on large corpora (bodies of text) of data, both

[1] This universal recursion language faculty is debated by linguist Daniel Everett's controversial work regarding the Pirahã language in the Amazon region of South America (Everett 2009).

pre-processed and not, of a given language. This data can be either taken from written sources like newspapers or novels or from the internet itself, from websites and blogs. Using clever statistical algorithms on counts, frequencies, or n-grams (sets of words of n length) and running the results through machine learners provides a bulk of new research found in the field of computational linguistics in general.

When data is pre-processed, supervised learning techniques can be run that determine similarity between old and new data in order to appropriately classify new data into given categories. This type of machine learning is relatively easy to complete compared to unsupervised learning, to be described below. One downside of supervised learning techniques, however, is the requirement of pre-processed data. In general, for every hour of audio recorded a total of four to one hundred hours of manual work is necessary to process it, depending on what type of processing need be done. This severely limits the number of tasks that can be completed using this type of machine learning, especially in lesser-studied and lesser-documented languages.

The other type of machine learning algorithm commonly used is unsupervised learning, where correlations present in data are found and automatically clustered together based on similarity. This type of technique, while more crude and unfocused than supervised learning, does not require vast swaths of pre-processed data, alleviated major time constraints and available-data constraints that would otherwise limit research. However, because the computer finds its own patterns in this type of research, results may be unexpected or less accurate than those found through supervised learning techniques.

The NLP expert interviewee Ken Shan used to work on a variety of these problems in NLP, though for the last few years NLP has become less of a priority for Ken. Instead, as of 2014, Ken is currently working with the Defence Advanced Research Projects Agency (DARPA) to devise algorithms for probabilistic programming. This topic still remains ever-elusive to the group's full understanding; however, in his summary of this type of work and its possible impact on the future of NLP, Ken described that his eventual goal in this field is to devise an easy means for probabilistic distribution algorithms to be merged and dissected in order to expedite testing and work done using such algorithms. In other words, given that these types of algorithms could be easily modified as necessary as research advanced, more time could be spent on formulating and understanding data than on contriving new algorithms. If successful, this research would quickly and efficiently allow more varieties of tests to be run on data both found in NLP research and otherwise, providing a rapid influx of new results and possible findings through techniques such as machine learning, for example.

During the interview, Ken Shan elaborated about the future of NLP and what problems are currently being faced in the English language. The most important overarching points that he described included research on the analyzation of discourse structure in a text, building relations between sentences (like those found in examples of cause-and-effect, and implementing and updating databases automatically with world knowledge, as Watson is able to do. Discourse structure describes computers' parsing of

sentences to connect them together in an attempt to tie information gathered from previous sentences into what is currently being examined. This allows a general theme and flow of a sentence to be found that can alleviate issues like sentiment analysis where the change of sentiment across a text is important to fully understand.

The most important and perhaps most difficult task for the future of NLP is the obtaining of world-information from databases, allowing computers to build their own knowledge-banks from human data not designed for this purpose. This process is vastly unlike how current computers receive information: it is carefully selected and formatted by scientists to be manually inputted. The task of automatic information retrieval, however, requires a multitude of programs running together and cooperating simultaneously, while everything in the database must maintain a consistent structure so that information retrieval maintains efficiency. This task, when properly implemented, opens a literal world of knowledge to computers with near endless future applications.

Beatrice Santorini. 1990. Part-Of-Speech Tagging Guidelines for the Penn Treebank Project (3rd revision, 2nd printing).

Everett, D. L. "Pirahã Culture and Grammar: A Response to Some Criticisms." Language, vol. 85 no. 2, 2009, pp. 405-442.

Daniel Jurafsky and James H. Martin. 2000. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition (1st ed.). Prentice Hall PTR, Upper Saddle River, NJ, USA.

Hinrich Schütze. 1995. Distributional part-of-speech tagging. In Proceedings of the seventh conference on European chapter of the Association for Computational Linguistics (EACL '95). Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 141-148.

Kristina Toutanova, Dan Klein, Christopher D. Manning, and Yoram Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1 (NAACL '03), Vol. 1. Association for Computational Linguistics, Stroudsburg, PA, USA, 173-180.

Mobley, D. (n.d.). How Watson Works. Retrieved December 7, 2016, from http://www.cs.uky.edu/~raphael/grad/keepingCurrent/HowWatsonWorks.pdf

Papageorgiou Harris, Prokopidis Prokopis, Giouli Voula, and Piperidis Stelios. 2000. A unified pos tagging architecture and its application to Greek. In Proceedings of Second International Conference on Language Resources and Evaluation, Athens.

Pinker, Steven and Ray Jackendoff. 2005. The faculty of language: What's special about it? Cognition 95(2): 201–236.

Steven T. Piantadosi, Harry Tily, Edward Gibson. The communicative function of ambiguity in language. Cognition, 2012; 122 (3): 280

Stuart J. Russell and Peter Norvig. 2003. Artificial Intelligence: A Modern Approach (2 ed.). Pearson Education.