Making NLP Work: Data Extraction Holds the Key

Abstract

Leading communications and media businesses are increasingly leveraging AI to improve operations, customer experience, and monetization. There are a number of areas where AI can be applied, such as personal voice assistants, customer feedback apps, developing customer insights, processing unstructured data, and so on. Natural Language Processing (NLP) is a key component of AI and plays a significant role in the effectiveness of such applications.

NLP is a way for computers to analyse, understand, and derive meaning from human language in a smart and useful way. It can be used to organize and structure knowledge to perform tasks such as automatic summarization, translation, named entity recognition, relationship extraction, sentiment analysis, speech recognition, and topic segmentation.

The key to effective deployment of NLP is data extraction. This paper outlines an approach for data extraction, while highlighting how contextualizing the extraction process to a specific business domain leads to a significant advantage in the speed and accuracy of data extraction, in turn resulting in superior user experiences.

The Opportunity

Plummeting computing power costs and the shift to digital are widening the application of Big Data technologies. For instance, meaningful insights for various business applications can be generated from various data sources such as text, images, audios and videos, leveraging NLP. For CSPs, NLP can help reduce costs without compromising customer service—think chatbots. Google Home and Amazon Alexa, both make extensive use of NLP.

NLP driven AI is helping companies lower the cost of customer service. Spectrum Communications' 'Ask the Spectrum Virtual Assistant'[1] uses NLP driven AI to address a wide variety customer queries. These range from troubleshooting and identifying service outages to ordering Pay-Per-View events. The virtual assistant even automatically directs users to human customer service agents when it is unable to resolve a query. Such assistants and chatbots come with the added advantage of round-the-clock availability.

Companies are using NLP for sentiment analysis to offer innovative solutions for brand marketers   to drive effective campaigns and grow revenues. The technology is also being deployed to increase the effectiveness of advertising solutions by aligning the advertising being served and content being consumed. It has a role in cost reduction as well, by enabling automatic development of content. Both the Associated Press and the BBC use Wordsmith, an NLP automated database, to create huge volumes of stories within seconds.[2]

Making Sense of it All

The amount of unstructured data collected by enterprise applications is increasing day by day. Quite often, the 'text' that must be understood by NLP applications is highly unstructured—it is written in a non-standard manner, may be grammatically incorrect, and could have spelling or vocabulary errors.

Consider these examples:

- Say a customer writes, 'The bill for my account A12345 is very high'. In this case the account number A12345 is labeled by the word account in the text. But this may not be the case always, as the text can alternatively be constructed as, "The bill for A12345 is very high'. The customer may assume that the one who is processing this text will understand that A12345 refers to her account number.
- At times, there can be a dependency among the data. For example, a customer identifier type and the actual identifier value are interdependent and these would typically be provided in natural language text in the form 'my <ID type> is <ID value>', like, my PAN is AXXPX1111N'. Processing this text must correctly infer that the ID type is PAN and the ID value is AXXPX1111N.
- The parameters, if they are names, can also sometimes be qualified and the qualification may or may not be informative. For instance a customer may say, 'My April bill amount is X, but he may also write, 'My unfathomable bill amount is X.' The latter does indicate the frustrated sentiment but does not necessarily qualify the label 'bill' with any additional information. So this might get interpreted as the latest bill.

Thus, the challenge lies in identifying a meaningful label or name for the information. The label should be understood identically across systems after analyzing dependency among the provided data, non-informative qualification of parameters and their variation across domains. However, it's easier said than done.

The 4+1 Solution

Effectively uncovering the intent of a written or oral expression has a significant impact on the quality of the applications using NLP. While there are many ways to extract data, we propose a four-step approach that maximizes the probability of identifying the intent accurately.

Step 1: The first step is 'extraction of possible name-value pairs.' There could be zero to many name-value pairs in an interaction. An example of a name-value pair would be a name of 'Account Id' and value of 'A12345' in an interaction, 'My Account Id' is 'A12345'. To extract the right name value pairs, adopt a combination of Parts of Speech and Symantic

Seperators Approach. The effectiveness lies in the implementation of the above approaches. To illustrate, implementing the Parts of Speech approach itself should use a set of techniques, typically in a range of eight to 10. Adopting the right techniques ensures your deduction of name-value pairs is accurate.

Step 2: The second step is to 'match the extracted names against the defined parameters and derive the context specific data'. The most critical aspect in this step is building a comprehensive domain dictionary to ensure that the lookup is accurate. For instance, an 'Account ID' has a specific meaning in the context.

Step 3: The third step is to 'extract possible values for the defined parameters by pattern match and type match.' While one user could use the text 'My Account ID is A12345' in an interaction, it is equally possible that another user could just use 'A12345'. The ability to identify the parameter in context of an interaction and extracting the values using the combination of domain dictionary and pre-defined metadata is critical.

Step 4: The fourth step is to 'extract values for dependent parameters'. To illustrate, let us use an example:

Bot: "Could you provide an identify proof for further processing?"

User: "My Aadhar / SSN ID is S1234."

This step implicitly illustrates that there are two parameters: One is the 'Identify Proof Type' whose value is 'Aadhar/SSN ID' and the second parameter is 'Aadhar/SSN ID' whose value is 'S1234'. Again the domain dictionary plays a critical role in identifying such dependent parameters.

Effective NLP: The Game Changer in AI Adoption

This above approach was tested in a live environment for one of our clients at the HOBS Experience page on Facebook.com. The simulation involved defining different 'intents' and associated parameters. These texts were simulated such that the data for the values was provided as part of the text. Using the algorithm discussed above, the extraction of values for the parameters was computed.

The results showed that 'values' were extracted completely for 88% of parameters, and partially for another 11%. This was under the condition that the values are given in the text but the overall structure could vary. The analysis of this showed that the probability of extracting the value for the parameter using the above algorithm is close to 100%. If the number of samples used for testing were increased, then this can be better inferred, and the algorithm can be tuned for the most optimal execution path.

The effectiveness of NLP data extraction lies in deploying a structured approach in tandem with a comprehensive domain dictionary. As the adoption of NLP and AI increases across several business applications in the communications industry, their ability to influence customer experience has grown multifold. To make the most of these promising technologies, businesses must adopt solutions built on the best practices drawn from both technology and domain areas.

References

1. Spectrum Communications' 'Ask the Spectrum Virtual Assistant , https://www.spectrum.net/support/general/ask-charter-virtual-assistant-reference/
2. BBC News, "Robo-journalism: How a computer describes a sports match" (September 2015), accessed Jan 2018,   http://www.bbc.com/news/technology-34204052

About The Author

Sangeetha Ramakrishnan

Sangeetha Ramakrishnan is part of TCS' Communications, Media, and Information Services business unit. With over 14 years of experience in the telecom domain, she is currently the Lead Technical Architect - AI for TCS' HOBS Platform. Ms. Ramakrishnan specializes in Natural Language Processing, bots, and UI/UX. She is the architect for key framework components of TCS' HOBS Platform including the chassis, authorization, UI style guides, rest APIs for security, and multi-tenancy meta data configurations. She holds a Bachelor's degree in Engineering (Electronics and Communications) from the Guindy Engineering College, Anna University, Chennai, India.

The author acknowledges the valuable inputs provided by Ramachandran Selvarajan and Raghavan Venkatasubramanian without which this paper would not have been possible.

Contact

Visit the Communications, Media & Technology page on www.tcs.com

Email: global.cmi@tcs.com

Blog: Next Gen CMI

Subscribe to TCS White Papers

TCS.com RSS: http://www.tcs.com/rss_feeds/Pages/feed.aspx?f=w
Feedburner: http://feeds2.feedburner.com/tcswhitepapers

About Tata Consultancy Services Ltd (TCS)

Tata Consultancy Services is an IT services, consulting and business solutions organization that delivers real results to global business, ensuring a level of certainty no other firm can match. TCS offers a consulting-led, integrated portfolio of IT and IT-enabled, infrastructure, engineering and assurance services. This is delivered through its unique Global Network Delivery Model™, recognized as the benchmark of excellence in software development. A part of the Tata Group, India's largest industrial conglomerate, TCS has a global footprint and is listed on the National Stock Exchange and Bombay Stock Exchange in India.

For more information, visit us at www.tcs.com

Experience certainty.

IT Services
Business Solutions
Consulting

TCS Design Services I M I 02 I 18