Linguistic vs Machine Learning? Why the real winner is a mixed approach.

Zachary Wilkins
Senior Computational Linguist
Artificial Solutions
I. The typical customers who want to build a new Virtual Assistant

II. How we use linguistic rules

III. How we use machine learning

IV. Pros & cons of ML-only vs rules-only VAs

V. Our mixed ML-rules approach

VI. How we expose our VAs
The typical customers who want to build a new Virtual Assistant

- Lack Data
- Want Control
- Unclear Regarding Solutions
How we use ML - Part of Speech Tagger

Try out

What's the time?

My watch says it's 8:13 PM GMT+1. If you live in a different time zone, I advise you to check the current time elsewhere, for example on your computer or mobile phone.

Response Info

11:13:20 AM 😊

- Input Summary

[whats] [the] [time]: %S,UQuestion | %S,Quote | %S,Classification | %S,Others | %S,Classification,Conf,Low
[whats]: %S,PRON,POS | %S,INTERROG,POS | %S,VBP,POS | %S,PREP,POS | %S,3RPERS,POS
[the]: %S,DET,POS
[time]: %S,NN,POS | %S,SG,POS

- Answer Details

1

- Output Parameters

1

- Watched Variables

2
How we use linguistic rules
Language objects – our intent building blocks (“Teneo Language Resources”)

<table>
<thead>
<tr>
<th>HOW.DO.I.PHR</th>
<th>GOOD.ADJ.PHR</th>
<th>VACATION.NN.SYN</th>
<th>CITIES.NN.LIST</th>
</tr>
</thead>
<tbody>
<tr>
<td>How can I</td>
<td>Good</td>
<td>Holiday</td>
<td>LA/Los Angeles</td>
</tr>
<tr>
<td>Can you help me to</td>
<td>Great</td>
<td>Vacay</td>
<td>Seattle</td>
</tr>
<tr>
<td>How could I</td>
<td>Excellent</td>
<td>Vacation</td>
<td>Toronto</td>
</tr>
<tr>
<td>Tell me how to</td>
<td>Wonderful</td>
<td></td>
<td>New York/NY/NYC</td>
</tr>
<tr>
<td>How does one</td>
<td>Nice</td>
<td></td>
<td>Austin</td>
</tr>
<tr>
<td>How would one</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I want to be able to</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I want to know how to</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Teneo Language Resources
Teneo Languages
# 35 Teneo Languages

## Advanced Level

- **ENGLISH** *
- **GERMAN** *
- **DUTCH** *
- **SWEDISH** *
- **DANISH** *
- **NORWEGIAN**

## Standard Level

- **BASQUE**
- **BELARUSIAN**
- **BOSNIAN**
- **BULGARIAN**
- **CATALAN**
- **CROATIAN**
- **CZECH**
- **ESPERANTO**
- **FINNISH**
- **GALICIAN**
- **GREEK**

**ADVANCED LANGUAGE SUPPORT:**
- Lexical Resources
- Autocoding
- Dialogue Resources
- *Pos Tagger / Morphological Analyzer

**STANDARD LANGUAGE SUPPORT:**
- Teneo Input Processing
- Language Tokenization
- Sentence Identification
- Basic Number Recognizer
Teneo Language Resources: Annotations, LRs, DRs

I. Input Processors (Annotations)  
   Chain of modules used to pre-process user inputs (sentence split, tokenize, spell correct)

II. Lexical Resources  
   Re-usable Linguistic building blocks to compose NLU rules (Understanding conditions)

III. Dialogue Resources  
   Re-usable flows with generic dialogue capabilities
Other linguistic rules-based VA functionality

- Contextual Understanding
- Interruptibility
- Intent Granularity
- Follow-up Skipping
How we use machine learning
How we use ML - Teneo Discovery

- Text mining tool
- Based on unsupervised machine learning
- Powerful data visualization provides easy-to-digest understanding of data to business side
- Crucial for project scoping, solution optimization
- First step for incorporating ML into a solution
How we use ML for intent recognition

- First, intent and/or topic are classified and stored in memory along with the confidence score.

- Part of speech tagging model trained on a dictionary built using our ‘language object’ data annotates the user input.

- We have used our data to optimize models for Chinese (Mandarin), Dutch, English, French, German, Italian, Japanese, Spanish and Swedish.
How we use ML – depends on the amount of data

Intent Recognition vs. Topic Modeling

“I want to know my Account Balance”

Topic = “Account”
Linguistic vs Machine Learning Conversational Systems
## MACHINE LEARNING vs LINGUISTIC APPROACH

### Advantages

**MACHINE LEARNING**
- No linguists required
- Higher recall
- More scalable
- Lower engineering effort

**LINGUISTIC APPROACH**
- Complete control over Virtual Assistant behavior
- Higher precision
- Requires little or no data
- Less rework if virtual assistant is built in a new language
- Small behavior tweaks are easy
- Context easier to maintain

### Disadvantages

**MACHINE LEARNING**
- Requires a high amount of data, and again when a new language is added
- Update requests face machine learning 'black box' issue
- Even with very high amounts of data, undesired behavior is more likely than rule-based approach

**LINGUISTIC APPROACH**
- Less scalable / more engineering efforts
- More prone to human error
- Higher maintenance costs later
Why a ‘Mixed’ approach?

Project can begin development with very little data

Rules-based VA can be released to small group of customers for testing and data collection

Behaviors and responses can be optimized while data is collected
Baby travel (Read only)

Classification: Baby travel
- Classification: High
- Intent: Baby Travel

Refine Understanding

1. Age
   - If your child is not yet 2 years old on both the outbound and the inbound journey, it will travel sitting on your lap. Every adult may travel with 2 babies, but only 1 baby may sit on his/her...

2. Luggage
   - For babies younger than 2 years of age you can bring one small, folding pram for free. The pram must fit in the overhead compartment or under your seat. You may also bring piece(s)...

3. Stroller or car seat
   - Per child, a collapsible stroller or pram can be taken free of charge. If you booked a seat for your child, we recommend that you bring your own car/child seat on board.

4. Otherwise...
   - Want to bring baby food with you on board? The restrictions on bringing liquids do not apply to baby food. Although we have a limited amount of special baby food available on board...

5. Baby food
   - Get generic answer
How we expose our VAs
Our 15 years of experience building VAs tells us that people expect the solution to:

I. Go off-topic/out of domain
II. Get social
III. Get personal
IV. Be rude/polite
V. Be clever enough to handle that
Public VAs
We’re front-end agnostic
Successful Intelligent Assistant platforms require ingesting massive amounts of both structured and unstructured data to build linguistic-based conversational systems that understand user intents to help complete a task successfully. But the hype surrounding Deep Neural Networking, Artificial Intelligence, Big Data and Analytics has led many decision makers to believe that Machine Learning, alone, can recognize intent and provide proper responses without human supervision. Real-world experience emphatically shows otherwise. Artificial Solutions’ patented, hybrid approach is flexible and allows enterprise personnel to use linguistic tools to train an Intelligent Assistant -- even without data -- then use inputs of data from multiple sources as context to optimize the IA’s ability to carry on natural language dialogs that embody the tone, personality and other attributes that are consistent with a company’s brand.

- Dan Miller, Lead Analyst & Founder, Opus Research, Opus Research