Enhanced Information Retrieval

Car Enhanced Demo

When Conversation does not scale

IBM Support

- IBM has software solutions in 27 capability areas
  - Each with many sub-categories

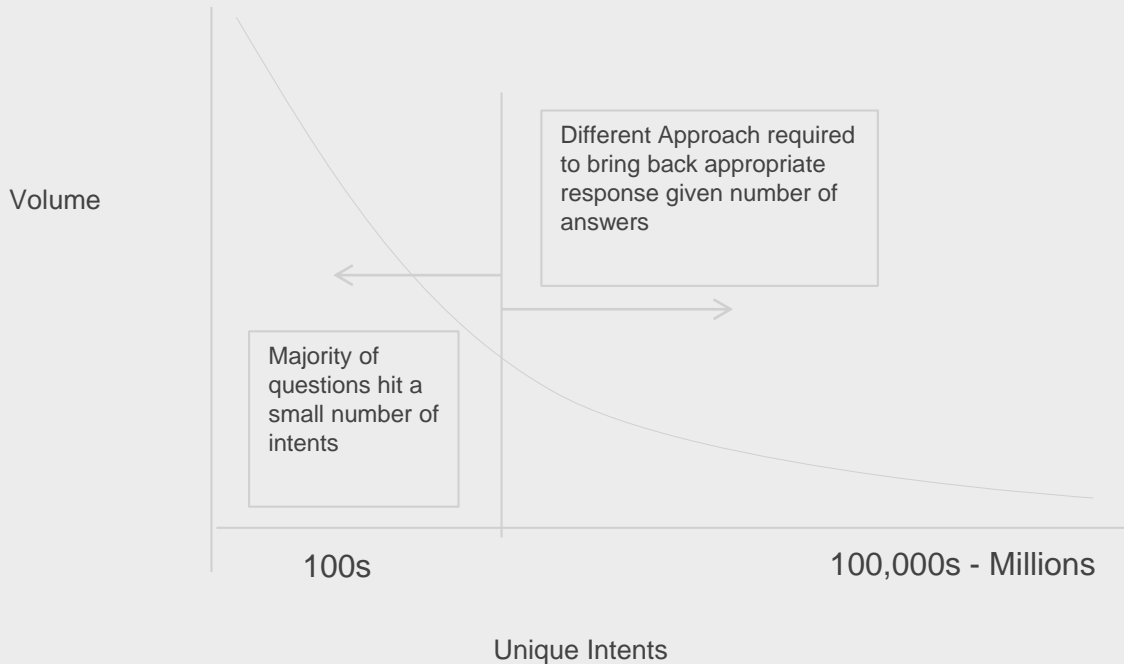Modeling this support space quickly becomes impossible.

IBM

Application infrastructure
Asset and facilities management
Big data
Business analytics
Cities management
Cognitive computing
Commerce and eCommerce
Connectivity and application integration
Continuous engineering
Data refinement, integration and governance
Data warehousing
Databases
DevOps
Digital experience
Enterprise content management
Enterprise modernization
Enterprise social and mail
IT service management
Intelligent business process management
Marketing and merchandising
Mobile development and integration
Mobile management and security
Operating Systems
Procurement
Security
Talent management
Unified endpoint management
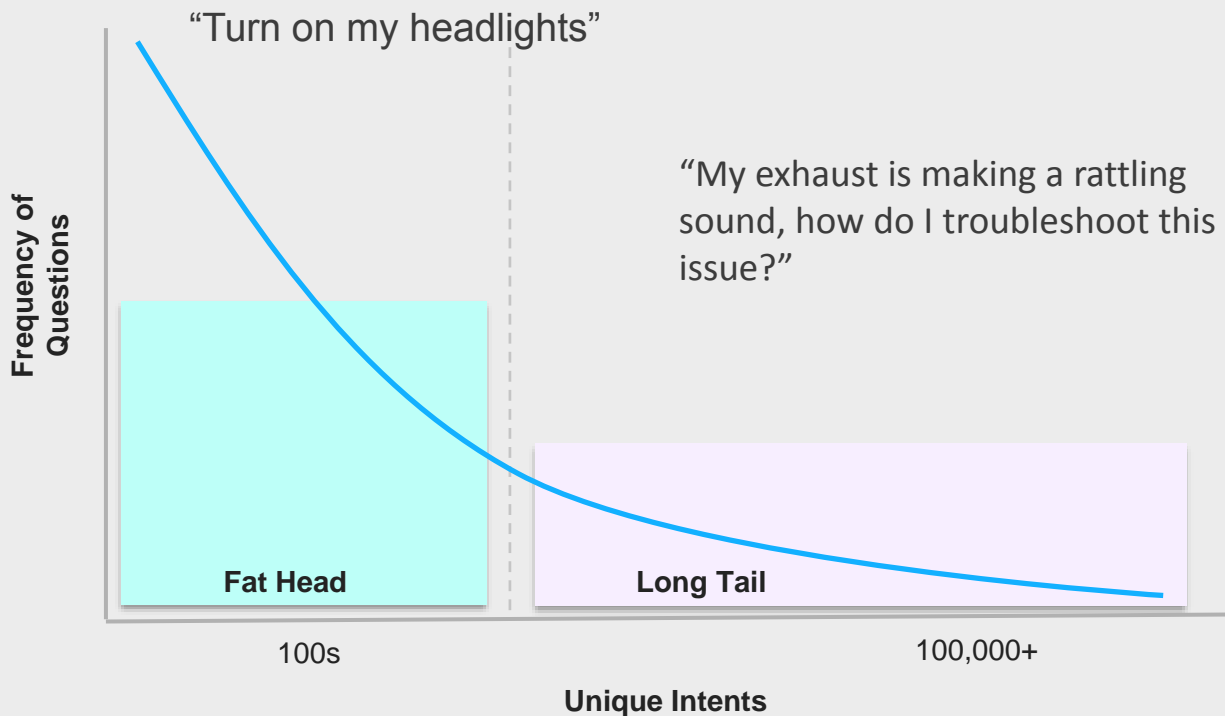
The Challenge of Scaling a Model

Some intents are
repeated over and over
and can be modeled

But there is a "long tail"
of infrequent intents

– Modeling the long tail
is impossible and
requires a different
approach

Volume

Different Approach required
to bring back appropriate
response given number of
answers

Majority of
questions hit a
small number of
intents

100s

100,000s - Millions

Unique Intents

Question Distribution

"Turn on my headlights"

"My exhaust is making a rattling sound, how do I troubleshoot this issue?"

Frequency of Questions

Fat Head

Long Tail

100s

100,000+

Unique Intents

Retrieve and Rank

Here Watson uses reasoning strategies that focus on identifying the most appropriate **answer**.

Watson Conversation

Here Watson uses reasoning strategies that focus on the language and context of the **question**.

Tying Watson Services Together

Most solutions combine multiple Service APIs

- Brought together with a use-case specific orchestration component
- Any number of services can be incorporated

"My exhaust is making a rattling sound, how do I troubleshoot this issue?"

I need to visit Zimbabwe as a British citizen with a criminal record—can I do that?

Orchestration

Conversation API

Conversation

1. Conversation is not confident it can answer the question

R&R API

Rank & Retrieve

2. R&R is used to search the corpus for the best answer

Fat Head
100s

Long Tail
100,000+

Unique Intents
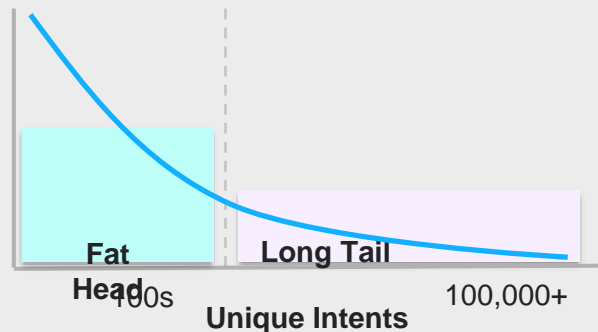
Retrieve and Rank as a Scaling Solution

Retrieve and Rank enables scaling to millions of entities and intents, ranking search results based on a trained model
- Scales at the cost of lower accuracy and precision of answer compared to Conversation

A customer application interacts with Conversation	If the application determines that its confidence is not high enough to answer the user's question, it will issue a query to R&R to search for documents which may be relevant	R&R returns a snippet of the document and some additional metadata (for example the Title, Author, and so on) to the calling app along with the URL where R&R crawled the document	The app will present this information to the user, and if the user clicks on the URL in their browser the original document will be retrieved from the web server (if the customer has configured this ability)

Enhanced Conversation Sample Demo

This app represents a fictional cognitive car. Commands issued to the car will be answered by the Conversation service. For example, **"Turn on my wipers"** or **"Turn up the radio"**. It is also possible to ask detailed questions about how to operate a car – details can be found in [this car manual](). These will be answered by the Retrieve and Rank service. For example, **"How do I check my tire pressure"** or **"How do I turn on cruise control?"** The app's responses are shown on the left while the right shows the detailed JSON responses from the server.

listen to rock music

Great choice! Playing some rock music for you.

how do i check my tire pressure

Great question. I found some ideas for you:

Collapse Results –

Tire Pressure Monitoring System

" check your tires as soon as possible, and inflate them to the proper pressure. Driving on a significantly under-inflated tire causes the tire to overheat"

Read the full document here

SYNC™ Applications and Services

" choices.

You do not need to say words that appear within square brackets. For example, for where (what can I say [available] commands) appears, you"

Read the full document here

Tire Care

Sent to Watson

```
1   {
2       "input": {
3           "text": "how do i check my tire pressure"
4       },
5       "context": {
6           "conversation_id": "10fe98fd-66da-420a-a749-98478a71!
7           "system": {
8               "dialog_stack": [
9                   "root"
10              ],
11              "dialog_turn_counter": 2,
12              "dialog_request_counter": 2
13          },
14          "default_counter": 0,
15          "call_retrieve_and_rank": false
16      }
17  }
```

Watson understands

```
1   {
2       "context": {
3           "conversation_id": "10fe98fd-66da-420a-a749-98478a71!
4           "system": {
5               "dialog_stack": [
6                   "root"
7               ],
8               "dialog_turn_counter": 3,
9               "dialog_request_counter": 3
10          },
11          "default_counter": 0,
```
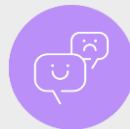
Other Notable Watson Services

Watson Speech to Text can be used anywhere there is a need to bridge the gap between the spoken word and its written form

Watson Tone Analyzer Service uses linguistic analysis to detect three types of tones from text: emotion, social tendencies, and language style

Watson Text to Speech provides a REST API to synthesize speech audio from an input of plain text

AlchemyLanguage offers several API functions as part of its text analysis service, each of which uses sophisticated natural language processing techniques to analyze your content.