Transparency
Note: QnA Maker

Updated 05/06/2021
# Table of Contents

What is a Transparency Note? ................................................................................................................................. 3
Introduction to QnA Maker ....................................................................................................................................... 3
Characteristics and limitations of QnA Maker ........................................................................................................... 6
  Enhance system performance .................................................................................................................................. 6
  QnA extraction from the source ............................................................................................................................ 6
Learn more about responsible AI ............................................................................................................................. 8
Learn more about QnA Maker ................................................................................................................................... 8
Contact us .............................................................................................................................................................. 8
About this document .................................................................................................................................................. 8
What is a Transparency Note?

An AI system includes not only the technology, but also the people who will use it, the people who will be affected by it, and the environment in which it is deployed. Creating a system that is fit for its intended purpose requires an understanding of how the technology works, what its capabilities and limitations are, and how to achieve the best performance. Microsoft’s Transparency Notes are intended to help you understand how our AI technology works, the choices system owners can make that influence system performance and behavior, and the importance of thinking about the whole system, including the technology, the people, and the environment. You can use Transparency Notes when developing or deploying your own system, or share them with the people who will use or be affected by your system.

Microsoft’s Transparency Notes are part of a broader effort at Microsoft to put our AI Principles into practice. To find out more, see the Microsoft AI principles.

Introduction to QnA Maker

QnA Maker is a cloud-based, natural language processing service that easily creates a natural conversational layer over your data. It can be used to find the most appropriate answer for a specified natural language input, from your custom knowledge base of information. See the list of supported languages here.

QnA Maker is commonly used to build conversational client applications, which include social media applications, chat bots, and speech-enabled desktop applications. A client application based on QnA Maker can be any conversational application that communicates with a user in natural language to answer a question.

QnA Maker uses several Azure resources, each for a different purpose: Azure Cognitive Search, App Server and App Service plan and Application Insights. All customer data (question answers and chatlogs) is stored in the region where the customer deploys the dependent service instances. For more details on dependent services see here.

The basics of QnA Maker

The first step in using QnA Maker is training and preparing the QnA service to recognize the questions and answers that may be developed from your content. QnA Maker imports your content into a knowledge base of question and answer pairs. The import process extracts information about the relationship between the parts of your structured and semi-structured content to infer relationships between the question and answer pairs.

The extracted QnA pairs are displayed in the following way:
You can edit these question and answer pairs, and add new pairs yourself. When you’re satisfied with the content of your knowledge base, you can publish it, which will make it ready to be used to respond to questions sent to your client applications. At the second step, your client application sends the user’s question to your QnA Maker service API. Your QnA Maker service processes the question and responds with the best answer.

For more details, see the QnA Maker documentation.

Terms and definitions

<table>
<thead>
<tr>
<th>Term</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Knowledge base</td>
<td>A collection of questions, answers, and metadata that have been extracted from content sources or added manually. The collection is then used to develop question and answer pairs. Queries to the QnA service are matched against the contents of the knowledge base.</td>
</tr>
<tr>
<td>Active learning</td>
<td>Consumes the feedback from use of the system to provide suggestions (in the form of new questions) to the knowledge base owner to improve the contents of their knowledge base. Learn more here.</td>
</tr>
<tr>
<td>Multi-turn</td>
<td>Sometimes additional information is needed for QnA Maker to determine the best answer to a user question. QnA Maker asks a follow-up question to the user.</td>
</tr>
<tr>
<td>Metadata</td>
<td>Additional information in the form of a name and value that you can associate with each QnA pair in your knowledge base. Metadata can be used to pass context and filter results.</td>
</tr>
<tr>
<td>Synonyms</td>
<td>Alternate terms that can be used interchangeably in the knowledge base.</td>
</tr>
</tbody>
</table>

Example use cases

You can use QnA Maker in multiple scenarios and across a variety of industries. Typically information retrieval use cases are best suited for QnA Maker where there are usually one or only a few correct responses to a user question. Scenarios or topics that have a wide variety of viewpoints, worldviews, geopolitical views, controversial content, etc. will be more difficult to answer correctly. Customers should be aware that providing this type of content via QnA Maker can create negative sentiment and reactions, and result in negative publicity. If you do provide this type of content, consider adding source attribution to allow your users to evaluate the answers for themselves.

Some typical scenarios where QnA Maker is recommended are:
- **Customer support:** In most customer support scenarios, common questions get asked frequently. QnA Maker lets you instantly create a chat bot from existing support content, and this bot can act as the front line system for handling customer queries. If the questions can’t be answered by the bot, then additional components can help identify and flag the question for human intervention.

- **Enterprise FAQ bot:** Information retrieval is a challenge for enterprise employees. Internal FAQ bots are a great tool for helping employees get answers to their common questions. QnA Maker enables various departments, such as human resources or payroll, to build FAQ chat bots to help employees.

- **Instant answers over search:** Many search systems augment their search results with instant answers, which provide the user with immediate access to information relevant to their query. Answers from QnA Maker can be combined with the results from document search to offer an instant answer experience to the end user.

Considerations when choosing other use cases

- **Avoid high-risk scenarios:** The machine learnt algorithm used by QnA Maker optimizes the performance based on the data it is trained on, however there will always be edge cases where the correct answer isn’t returned for a user query which the system doesn't understand well. When you design your scenarios with QnA Maker, be aware of the possibility of false positive results. It is advisable to create a dataset of the top queries asked in your scenario and the corresponding expected answers, and periodically test the service for the correctness of the responses. For example:

  - **Healthcare:** This often requires high precision, and wrong information can have life-threatening consequences. Consider the example of a Doctor Assistant bot that uses QnA Maker to understand the patient’s symptoms and match it to common illnesses. Likewise, any bots that are designed to converse with patients with mental health issues, such as depression or anxiety, must be very careful of the responses returned. QnA Maker can be helpful in parsing through clinical terminology and deriving useful question and answer pairs, but is not designed, intended or made available to create medical devices, and is not designed or intended and should not be used as a substitute for professional medical advice, diagnosis, treatment, or judgment. Customer is solely responsible for displaying and/or obtaining appropriate consents, warnings, disclaimers and acknowledgements to end users of their implementation.

- **Avoid open domain scenarios:** QnA Maker is meant to answer questions from a particular domain knowledge base, not open-ended questions, or out-of-domain questions. Using out-of-domain questions with QnA Maker runs the risk of returning incorrect responses. For example:

  - **Social bots:** Bots that are meant for generic chit-chat, not related to a particular domain, are difficult to design with QnA Maker. In these scenarios, the user intents and viewpoints can range widely (for example, sports, fashion, politics, and religion). Building a QnA Maker knowledge base is best used for facts and/or discovery of content. Using QnA Maker for diverse worldview topics may be challenging and we recommend customers consider more careful review or curating of such content.

  - **Handling inappropriate conversations:** It’s possible that users will initiate inappropriate conversations with the bot, including expletives or hate speech. The bot designer must be very careful about how to handle these conversations, and make sure that these intents are detected with high accuracy and the appropriate response given. It’s difficult to build a comprehensive knowledge base in QnA Maker containing every variation of inappropriate utterances possible. It
is therefore better to handle such cases with a rule based system, for example the user utterances can be quickly checked for the presence of any words from a pre-processed blocklist of inappropriate keywords. This is not part of the QnA Maker service and would need to be developed on top of the QnA Maker service.

Characteristics and limitations of QnA Maker

Depending on your scenario and the input data, you might experience different levels of performance from QnA Maker. A common way to evaluate the quality of responses is to create a set of commonly asked queries in your scenario, and check whether the QnA Maker response matches the expected response. The batch testing tool will help you in evaluating your query set. The following sections discuss key concepts and best practices to improve the performance.

Enhance system performance

There are two main technology pieces of QnA Maker: QnA extraction from the source, and returning the best response for a query.

QnA extraction from the source

QnA Maker is able to extract question and answer pairs from semi-structured content. The algorithm looks for a repeating pattern in the source documents, or for a particular layout of the content, to determine which sections constitute a question and answer. You can extract from content such as FAQ URLs, product manuals, and support documentation. When extracting, QnA Maker focuses on using content that is suitable for a chat bot, which typically has a small surface area.

Note the following limitations and best practices:

- The extraction is primarily focused on text and images (from URLs and documents). Other elements, like tables and images in files are ignored in the extraction. This might lead to incomplete extractions.
- The extraction relies heavily on the structure of each document and looks for a pattern of question and answer pairs or recurring topics and sections. Content that doesn't have a dominant structure, or that has different formats in the same document, makes it more difficult for QnA Maker to extract questions and answers (for example, news articles and blogs).
- When you’re deciding on the sources for your knowledge base, choose content that has some structure to it. For more information, see Importing from data sources. Also keep in mind that the extracted answers need to be displayed in a chat bot, which usually has limitations in rendering content.
- Sources that are in a secure storage, such as OneDrive or DropBox, are currently unsupported in QnA Maker. You can only extract files stored in SharePoint.

Return the best response for a query

QnA Maker uses sophisticated natural language processing and ranking technology to match an incoming user query with the best QnA match in its index. To optimize performance, these models are trained on several open source and custom datasets. However, depending on the contents of your knowledge base, the relevance of responses may vary. Learn more about the ranking process here. Each response is associated with a confidence score, which usually denotes the quality of the response. The higher the confidence scores the higher the likelihood that the answer is correct for the asked question. False positive results occur when a wrong answer has a high confidence score, or the correct answer is not the top result.
Note the following limitations and best practices:

- Repeating the same word set within different questions in question and answer pairs will reduce the likelihood that the right answer is chosen for a particular user query with those words. For example, you might have two separate QnAs with the following questions: "where is the parking location", "where is the ATM location". Because these two QnAs use such similar words, this similarity might cause very similar scores for many user queries that are phrased like "where is the <x> location". Instead, you might clearly differentiate with queries like "where is the parking lot" and "where is the ATM".

- Add as many alternate questions as you need, but keep the alternate questions simple. Adding more words or phrasings that aren't part of the main goal of the question doesn't help QnA Maker find a match. Your user might enter questions with either a conversational style of text, ("How do I add a toner cartridge to my printer?"), or a keyword search ("toner cartridge"). The knowledge base should have both styles of questions in order to correctly return the best answer. If you aren't sure what keywords a customer is entering, use telemetry data to analyze queries.

- If a question in the knowledge base is ambiguous or could have multiple possible responses to it, ask the user for additional information. For more details, see Use follow-up prompts to create multiple turns of a conversation.

- Metadata helps narrow down the results of a user query based on metadata tags. The knowledge base answer can differ based on the metadata tag, even if the query is the same. For example, "where is restaurant parking located" can have a different answer depending on the location of the restaurant branch.

- QnA Maker supports some basic synonyms in the English language like I am, I'm, etc. To add synonyms to keywords that take different forms, use the Alterations API to add case-insensitive word alterations. For example, if the original word is Buy, you might create the following synonyms: purchase, net-banking, and net banking. All the knowledge bases created in a particular QnA Maker service will share the synonym list. Synonyms are not shared across QnA Maker services.