CSI5180: Topics in AI Virtual Assistants

Assignment 3 Comparative studies

Submitted To:
Prof. Caroline Barrière

Team Member:
Abhisht Makarand Joshi: 300311048
Nidhi Kumari Chauhan: 300279256

Comparative Study 1 - Skill development

Virtual Assistants are part of our daily life, meaning that not only is there research going on to continue to enhance the platforms, but there is also a full ecosystem of vendors specialized in the development of skills for customers to add on their phones.

TO DO:

- Explore a few skill development companies, and choose 3 for your analysis.
- Describe each vendor, what it does, what is its target audience.
- Develop a set of 5 criteria for your comparative study. The criteria could be with respect to types of skills they suggest (e,g, home automation, games, etc), types of service offered (e.g. full programming, code adaptation, etc).
- Make a comparative table according to your set of criteria.

Solution 1:

Voiceflow: Voiceflow is a platform that allows developers and non-developers to build voice apps for virtual assistants like Alexa, Google Assistant, and others without writing any code. Voiceflow also offers templates and pre-built components that can be used to quickly create voice apps for specific use cases. The platform offers a drag-and-drop interface that makes it easy to create voice apps, and it supports a wide range of features like text-to-speech, speech-to-text, and natural language understanding. It was founded in 2018 by Braden Ream and Andrew Lawrence, and it is headquartered in Toronto, Canada.

Target audience: Voiceflow is aimed at both developers and non-developers who want to create voice apps for virtual assistants. Developers can use Voiceflow's platform to speed up the development process and create voice apps more efficiently. Non-developers can use Voiceflow to create voice apps without needing to write any code, which makes it more accessible for businesses and individuals who don't have programming experience. Voiceflow is also used by companies and organizations that want to create voice apps for their customers or employees.

Witlingo: Witlingo is a platform that offers tools and resources for building and publishing voice experiences for virtual assistants like Alexa and Google Assistant. It offers a voice-first design approach, which means that the user experience is optimized for voice interactions rather than visual interfaces. Witlingo supports features like personalized responses, integrations with third-party apps, and analytics. The platform also includes a content management system (CMS) that allows users to update and manage their voice experiences. Witlingo is aimed at businesses and organizations that want to create voice experiences for their customers or employees. Witlingo was founded in 2015 by Ahmed Bouzid, who previously worked on the team that built the first version of Alexa. The company is headquartered in McLean, Virginia, USA.

Target audience: Witlingo is aimed at businesses and organizations that want to create voice experiences for their customers or employees. This includes industries like healthcare, finance, and education, where voice interactions can provide added value and convenience. Witlingo's platform is designed to be flexible and customizable, so it can be tailored to meet the specific needs of each business or organization. Developers and non-developers can both use Witlingo's platform to create voice experiences, depending on their level of expertise.

Bespoken: Bespoken is a platform that offers tools and resources for testing, monitoring, and developing voice apps for virtual assistants like Alexa. The tools include a command-line interface (CLI) for testing and debugging voice apps, a virtual device emulator for simulating voice interactions, and a testing platform that lets developers make automated tests for their voice apps. Bespoken also offers consulting services to help businesses and organizations make and use voice apps. Bespoken is aimed at developers who want to ensure the quality of their voice apps before releasing them. Bespoken was founded in 2017 by John Kelvie and Emerson Sklar, and it is headquartered in Seattle, Washington, USA.
Target audience: Bespoken is aimed at developers who are building voice apps for virtual assistants. The tools provided by Bespoken are designed to simplify the development process and make it easier to create high-quality voice apps. The virtual device emulator, for example, allows developers to test their voice apps in a simulated environment without needing to use an actual virtual assistant device. Bespoken's testing platform also helps developers ensure that their voice apps are working properly and meeting the necessary quality standards. Additionally, Bespoken's consulting services can help businesses and organizations navigate the complexities of creating and deploying voice apps.

Comparative Study:

The table below represents the comparison between the 3 VA building and development platforms and compares them on different criteria.

Criteria	**Voiceflow**	**Witlingo**	**Bespoken**
Types of skills	Wide range of use cases, including home automation, games, productivity, education, and more.	Focused on business use cases, such as customer service, sales, education, and healthcare. Also supports consumer applications like home automation.	Can develop voice apps for any use case, with a focus on testing and debugging.
Types of services	User-friendly platform with drag-and-drop interface. Offers consulting services.	Content management system for managing and updating voice apps. Analytics and reporting tools. Professional services for design and development.	Provides suite of tools for developers, including command-line interface, virtual device emulator and testing platform. Also offers consulting services.
Integrations	Integrates with multiple platforms and services, including Amazon Alexa, Google Assistant, and Twilio.	Supports integration with popular CRM platforms and tools, including Salesforce, HubSpot, and Zendesk. Also supports custom integrations via APIs.	Integrates with multiple platforms and services, including Amazon Alexa, Google Assistant, and Twilio. Also supports custom integrations via APIs.

Developer community	Offers a robust community forum with resources and support for users.	Offers a community platform for users to connect, collaborate, and share resources.	Provides a developer community with access to resources, support, and best practices.
Pricing Model	Voiceflow offers a freemium pricing model with three tiers: Free, Plus, and Pro. Additionally, Voiceflow offers custom enterprise pricing for larger organizations with more complex needs.	Witlingo offers a subscription-based pricing model with different tiers (starter, pro, enterprise) based on the number of skills and level of support needed. It also offers a 30-day free trial for users to test out their platform before committing to a paid subscription.	Bespoken offers a usage-based pricing model, which means customers are charged based on their usage of the platform's services. It also offers custom enterprise pricing for larger organizations with more complex needs.

Comparison between Voiceflow, Witlingo and Bespokenon based on different criteria.

Result/discussion based on comparative criteria:

Overall, each of these companies has a unique focus and target audience. Voiceflow is geared towards users who want to create voice apps without needing to know how to code, and it supports a wide range of use cases. Witlingo is focused on providing voice solutions for businesses and organizations, with a voice-first design approach and professional services. Bespoken provides tools for developers to create high-quality voice apps with a strong focus on testing and debugging, and it can be used for any type of use case.
These three companies have different strengths and target audiences when it comes to pricing, integrations, developer communities, and natural language tools. Voiceflow offers a freemium model and integrations with many platforms and services, along with a robust developer community. Witlingo provides a subscription-based model with support for popular CRM platforms and a community platform for users to collaborate. Bespoken offers a subscription-based model with different tiers based on the number of requests and features needed, along with a developer community and a strong focus on testing and debugging.

References:

https://witlingo.com/
https://www.voiceflow.com/
https://bespoken.io/
https://en.wikipedia.org/wiki/Witlingo
https://techcrunch.com/2019/04/23/voiceflow-which-allows-anyone-to-make-voice-apps-without-coding-raises-3-5-million/

Comparative Study 2 - Open-Source Voice Assistants

This article, 7 Best Open Source Voice Assistants, describes some assistants not part of the large private companies. So there exists, outside of Apple, Amazon, Google, etc, a desire to develop virtual assistants that will have particular features. Often privacy is a main concern and something promoted by the open source developers.

TO DO:

- Explore a few of these assistants. Choose 3 assistants for your analysis.
- Describe each voice assistant, what it does, how it can be used, etc.
- Develop a set of 5 criteria for your comparative study. The criteria could be with respect to the adopting community, privacy issues, etc.
- Make a comparative table according to your set of criteria.

Solution 2:

An open-source virtual assistant is a free software that allows users to alter and personalise the code to meet their individual requirements. Open source VA provide users with a more flexible and transparent experience than proprietary VA, which are often closed source and managed by a single business. The benefit of open-source virtual assistants is that they are not linked to a specific hardware or software platform, thus they can be utilised on a variety of devices and OS. Although there are a number of open-source voice assistants that can be found online, we will only explore three of them here:

Mycroft - Free open-source voice assistant: Mycroft is the world's first open-source virtual assistant platform. It focuses on privacy and is made to be very customisable. It may be used on a range of devices and operating systems and understands and interprets user requests using NLP and ML. Mycroft offers a skills marketplace where users may download and configure skills to add new capabilities to the virtual assistant. Mycroft is capable of a variety of functions, like setting reminders, playing music, operating smart home appliances, and more. The fact that Mycroft may function totally locally implies that all processing and data storage takes place on the user's device rather than in the cloud, which is one of its distinctive features. Text-to-speech is also supported by Mycroft. Mycroft is distinctive in that it functions on a variety of electronics, such as desktop PCs, mobile phones, and embedded devices.

Rhasspy - Offline open-source voice assistant: Rhasspy is an open-source platform for creating customised voice assistants that is totally offline and offers voice assistant services for several human languages. It offers a range of functions, including voice command recognition, natural language understanding, and programmable talents, and it can be utilised on a variety of devices and operating systems and integration with Raspberry Pi, etc. With support for numerous languages and unique wake words, Rhasspy is made to be highly customised. It may be used to play music, create reminders, and more, as well as operate smart home appliances. Rhasspy's capacity to run entirely offline—without the need for an internet connection—means that all processing takes place locally on the user's device.

Jasper: An open-source platform called Jasper is used to create voice-activated software. It enables programmers to construct unique voice commands and use natural language to automate tasks on their computers. Jasper is based on Python and makes use of a number of libraries and Services, including the Google Voice Recognition API etc. Jasper can be used to control different elements of a machine, including

as opening and closing programmes, playing music, surfing the internet, and creating reminders. It can also be integrated with other hardware and software, allowing to send texts or manage smart home equipment. Jasper has a number of capabilities, such as customizable voice commands, unique integrations, and multilingual support.

Comparative Study:

The table below represents the comparison between the 3 open-source VAs and compares them on different criteria.

Criteria	**Mycroft**	**Rhasspy**	**Jasper**
No. of Languages supported	26 Languages	16 Languages	25 Languages
Integration with Hardware	Supports hardware integrations, like Mark I and Mark II devices. (Product of Mycroft).	Supports hardware integrations with Raspberry Pi and Matrix Voice etc.	Supports hardware integrations, with Raspberry Pi and Arduino.
Operation without the internet	Can be fully offline operation, including speech-to-text and text-to-speech.	Can fully offline operation, but needs some additional configuration	Cannot perform operations without the internet requires internet for voice recognition.
Wake Word	Built-in wake word is “Hey Mycroft”. To build a custom wake word for Precise (Mycroft's wake word Listner), you need to be acquainted with the Linux command line and have a fundamental understanding of machine learning. Training data and time are necessary.	Rhasspy supports Raven, Porcupine, Snowboy, and more wake-word listening systems as "Rhasspy Wake". Your device can have a custom wake word in several ways. One example: Record three WAV template files to teach Raven your wake word. If you want a unique wake word, Porcupine demands re-running their optimizer programme every 30 days. Make your wake word at the Picovoice Console.	Doesn’t have a custom wake word. But has built-in wake word - "Hello Jasper". The Jasper documentation doesn’t have any specific mention on a custom wake word.
Privacy	Strong emphasis on security and privacy. Eg. Mimic 3.	Strong emphasis on security and privacy.	There is no privacy emphasis, and voice recognition requires an internet connection.

Comparison between Mycroft, Raspy and Jasper based on different criteria.

Result/discussion based on comparative criteria:

In the following table, we compare three popular open-source virtual assistants: Mycroft, Rhasspy, and Jasper. The table compares a variety of features, both qualitative and quantitative, such as privacy, supported languages, custom wake word support, and more. Each platform in the table has its own set of advantages and disadvantages that make it better or worse suited to certain applications. Mycroft is highly customizable with a strong privacy focus with a maximum number of languages supported. Rhasspy is flexible and scalable with the offline operation, and Jasper is highly customizable with a RESTful API but is quite outdated and cannot function offline. Mycroft and Rhasspy support the most number of languages and provide custom wake word detection, privacy and operation without the internet.

References:

https://mycroft.ai/initiatives/
https://www.darkhackerworld.com/2021/07/open-source-voice-assistants.html
https://rhasspy.readthedocs.io/en/latest/
https://jasperproject.github.io/documentation/

Comparative Study 3a - QA datasets

This article, 10 Question-Answering Datasets To Build Robust Chatbot Systems lists different datasets for QA.

TO DO:

- Explore a few of these datasets. Choose 3 datasets for your analysis. You can choose some that are not mentioned in the article as well.
- Describe each dataset, what it contains, which International Evaluation (if any) it was used for, who was it developed by, etc. Show an example of its content.
- Develop a set of 5 criteria for your comparative study. The criteria could be with respect to the size, application, etc.
- Make a comparative table according to your set of criteria.

Solution 3a:

Chatbots, an artificial intelligence system that mimics human speech, are trained and evaluated using QA datasets. By providing a broad and diversified set of training examples that match the types of inquiries and responses users are likely to face, chatbot QA datasets can assist enhance the accuracy and efficacy of chatbots. They can also advise programmers on where the chatbot should use more training or refinement. We describe three datasets that we intend to investigate further.

SQuAD 2.0 (Stanford Question Answering Dataset): The Stanford Question Answering Dataset (SQuAD) is a reading comprehension dataset comprised of questions posed by crowd workers on a set of Wikipedia articles, where the answer to each question is a span of text taken directly from the corresponding reading passage, or the question may be unanswerable. SQuAD's name suggests that it is best suited for answering questions. The model is put to the test in a reading comprehension task, where it must read a passage and then respond to a series of questions about the text. In SQuAD2.0, the original 100,000

questions from SQuAD1.1 are combined with more than 50,000 hostile questions that were prepared by crowd workers to look like answerable ones but were not. SQuAD 2.0 contains more than 100,000 questions. In order to score highly on SQuAD2.0, systems need to be able to not only provide answers when applicable but also recognise when there is no appropriate response based on the paragraph's context.

Examples of SQuAD Dataset:

Reasoning	Description	Example	Percentage
Lexical variation (synonymy)	Major correspondences between the question and the answer sentence are synonyms.	Q: What is the Rankine cycle sometimes **called**? Sentence: The Rankine cycle is sometimes **referred** to as a practical Carnot cycle.	33.3%
Lexical variation (world knowledge)	Major correspondences between the question and the answer sentence require world knowledge to resolve.	Q: Which **governing bodies** have veto power? Sen.: **The European Parliament and the Council of the European Union** have powers of amendment and veto during the legislative process.	9.1%
Syntactic variation	After the question is paraphrased into declarative form, its syntactic dependency structure does not match that of the answer sentence even after local modifications.	Q: What Shakespeare scholar **is currently on the faculty**? Sen.: **Current faculty include** the anthropologist Marshall Sahlins, ..., Shakespeare scholar David Bevington.	64.1%
Multiple sentence reasoning	There is anaphora, or higher-level fusion of multiple sentences is required.	Q: What collection does **the V&A Theatre & Performance galleries** hold? Sen.: **The V&A Theatre & Performance galleries** opened in March 2009. ... **They** hold the UK's biggest national collection of material about live performance.	13.6%
Ambiguous	We don't agree with the crowdworkers' answer, or the question does not have a unique answer.	Q: What is the main goal of criminal punishment? Sen.: **Achieving crime control via incapacitation and deterrence** is a major goal of criminal punishment.	6.1%

Table 3: We manually labeled 192 examples into one or more of the above categories. Words relevant to the corresponding reasoning type are bolded, and the crowdsourced answer is underlined.

In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under gravity. The main forms of precipitation include drizzle, rain, sleet, snow, graupel and hail... Precipitation forms as smaller droplets coalesce via collision with other rain drops or ice crystals within a cloud. Short, intense periods of rain in scattered locations are called "showers".

What causes precipitation to fall?
gravity

What is another main form of precipitation besides drizzle, rain, snow, sleet and hail?
graupel

Where do water droplets collide with ice crystals to form precipitation?
within a cloud

Figure 1: Question-answer pairs for a sample passage in the SQuAD dataset. Each of the answers is a segment of text from the passage.

Image Source: https://arxiv.org/pdf/1907.06292.pdf

MS MARCO (Microsoft Machine Reading Comprehension Dataset): Microsoft Research's AI team built a massive dataset called MS MARCO, which stands for Human Generated MAchine Reading COmprehension Dataset. Sampled from Bing's search query logs, the dataset contains 1,010,916 questions, all of which have been anonymized, together with the corresponding human-made responses to those questions and another 182,669 completely rewritten generated answers. The primary goal of this dataset is to enhance the field of artificial intelligence and related disciplines through academic, not-for-profit research.

Example of MS MARCO:

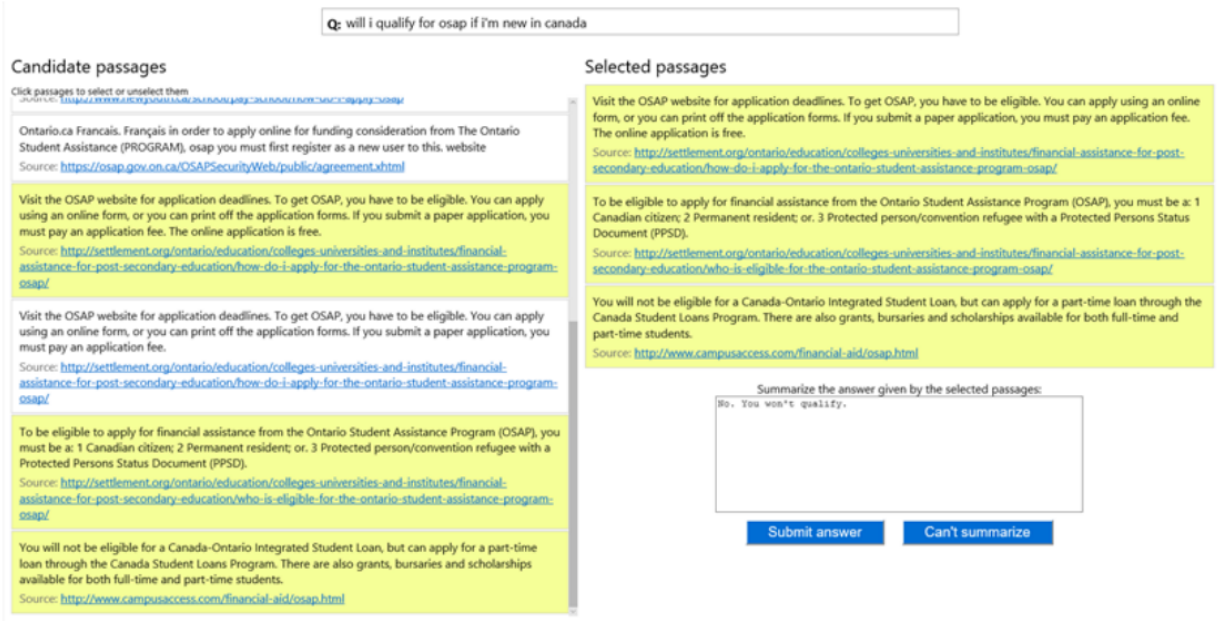


Figure 1: Simplified passage selection and answer summarization UI for human editors.

Image Source: https://arxiv.org/pdf/1611.09268.pdf

TweetQA: TWEETQA, a question-and-answer dataset focused on Twitter. The researchers at IBM and UC Santa Barbara have generated the first large-scale dataset for quality assurance (QA) on social media data. There are 10,898 articles, 17,794 tweets, and 13,757 question-answer pairs in the dataset. They proposed the first large-scale dataset for question-and-answering (QA) using data from social media platforms, whereas earlier efforts have focused on more formal material such as news and Wikipedia. They collected tweets used by journalists in order to make the tweets meaningful and intriguing. Human annotators are then requested to add further context to these tweets by adding their own queries and responses. Since the answers in SQuAD and similar QA datasets are extracted verbatim, they allow for more abstract reasoning. The model must process a question and brief tweet, then return a text phrase (which need not be in the tweet) as an answer.

Example ofTweetQA:

Passage: *Oh man just read about Paul Walkers death. So young. Ugggh makes me sick especially when it's caused by an accident. God bless his soul. – Jay Sean (@jaysean) December 1, 2013*

Q: *why is sean torn over the actor's death?*
A: *walker was young*

Image Source: https://tweetqa.github.io/

Comparative Study:

The table below represents the comparison between the 3 QA Datasets and compares them based on different criteria.

Criteria	SQuAD	MS MARCO	TweetQA
Size of the Dataset	Large dataset with more than 150000+ questions.	Large dataset with more than 1,010,916 questions.	Relatively smaller dataset with 13757 questions
What is the source of the dataset?	Wikipedia articles	Real Bing(Microsoft Search Engine) questions and human-generated answers.	Tweets (from Twitter)
Application of Dataset	Used in the SQuAD Challenge, which is a widely recognized benchmark for QA (Question-Answer) systems.	Non-commercial research, for benchmarking machine reading comprehension and question-answering models and QA.	QA systems targeting social media text. Focusing more on abstractive than extractive.
Length of Answers in the Dataset	Answer is mostly short in the dataset. The average length of Answers (#Words) = 3.35	Long as well as short answers in the dataset. The average length of Answers (#Words) = 3.52	Pretty short answers in the data set. The average length of Answers (#Words) = 2.45
Computational Requirements	Moderate computational requirements in terms of storage, memory utilisation, and processing time.	High computational requirements in terms of storage, memory utilisation, and processing time due to a large dataset.	Low computational requirements in terms of storage, memory utilisation, and processing time.

Comparison between SQuAD, MS MARCO and TweetQA datasets based on different criteria.

Result/discussion based on comparative criteria:

The study table contrasts the three QA datasets highlighting their primary shared characteristics and distinguishing features. Factors such as dataset size, question format, answer format, language, domain/topic, and other dimensions such as the number of unique sources, number of language support, and assessment metrics are all taken into account in the table. In comparison to TweetQA, which specialises in social media, SQuAD and MS MARCO are larger and more versatile. In addition, both SQuAD and MS MARCO provide both factual and general questions, with short and extended answers, while TweetQA is primarily tweet-based, with short answers. Depending on the expected performance and purpose of the chatbot, several considerations may affect which dataset is most suited for the task at hand.

References:

https://rajpurkar.github.io/SQuAD-explorer/
https://towardsdatascience.com/the-quick-guide-to-squad-cae08047ebee
https://analyticsindiamag.com/10-question-answering-datasets-to-build-robust-chatbot-systems/
https://microsoft.github.io/msmarco/
https://paperswithcode.com/dataset/ms-marco
https://tweetqa.github.io/
https://web.stanford.edu/class/archive/cs/cs224n/cs224n.1174/reports/2762006.pdf

Comparative Study 4 - Chatbot platforms

This article, from 2020, Building a Chatbot using Chatterbot in Python, by A. Navlani, lists a series of chatbot platforms. Another source, The 4 best chatbot builders in 2022, by Harry Guinnes, also gives a list for 2022. And again, another one: Best AI Chatbot Platforms for 2023, by Werner Geyser, gives a list for 2023. And these are just some articles providing lists, you can find even more… So you can see that it's a very active field and many start-ups or larger companies want to be a player in that chatbot development market.

TO DO:

- Explore a few platforms, and choose 3 for your analysis
- Describe each platform, what it does, what is its target audience.
- Develop a set of 5 criteria for your comparative study. The criteria could be with respect to availability, to ease of use, flexibility, etc.
- Make a comparative table according to your set of criteria.

Solution 4:

Dialogflow: Businesses can build chatbots for a variety of messaging services like Facebook Messenger, Slack, and Google Assistant using the Dialogflow chatbot platform. It uses natural language processing (NLP) to respond to what the user says, which makes conversations feel more natural. The platform has a graphical user interface (GUI) for building conversational flows. This lets developers design their own intents, entities, and responses. The chatbots from Dialogflow can also be used for lead generation, e-commerce, and customer care.

Target audience: Dialogflow targets developers and businesses who want to create conversational interfaces for their customers or internal use. The platform is made to make the development process easier, so it can be used by people who aren't experts in machine learning or natural language processing. But it also has advanced features and integrations that can be useful for experienced developers who want to build more complex conversational interfaces.

Dialogflow's key attributes are following:
Integration of NLP: Dialogflow uses NLP to make responses to user inputs. This makes conversations feel more natural and requires less knowledge of programming.
Pre-built agents: Dialogflow has pre-built agents for basic conversational situations, which makes it easier for organizations to use chatbots.
Integrations: Zapier, Shopify, and Salesforce are just a few of the third-party platforms that Dialogflow can be integrated with.
Customization: Dialogflow's chatbots can be highly customized, enabling organizations to build chatbots for specific use cases.
Free Plan: Dialogflow has a free plan that includes basic chatbot features. This makes it a good choice for small businesses that want to save money.

MobileMonkey: MobileMonkey is a platform for chatbots that lets businesses make chatbots for things like Facebook Messenger, Instagram, and web chat. It features a drag-and-drop interface that does not require any coding experience. The platform provides a visual interface for designing and building conversational flows, allowing users to create custom prompts, responses, and actions. MobileMonkey's chatbots can also be used for lead generation, customer service, and marketing campaigns.
Target audience: MobileMonkey focuses on companies of all sizes that want to build chatbots for their clients or internal use. The platform is made to make the development process simpler so that non-technical individuals with little or no experience in coding or chatbot development can use it. It does, however, also provide cutting-edge capabilities and integrations that may be appealing to professional developers looking to create more advanced chatbots.

Some key features of MobileMonkey include:-
Chatbot templates: To make it simple for businesses to get started with chatbots, MobileMonkey offers pre-built chatbot templates for a range of use cases.
Automatic lead generation: Chatbots can be utilized by companies to generate leads using MobileMonkey. For the purpose of facilitating lead generation, it has elements like lead magnets, forms, and scheduling.
Integrations: Salesforce, HubSpot, and MailChimp are just a few of the third-party platforms that MobileMonkey can be integrated with.
Free plan: MobileMonkey provides a free plan for basic chatbot capabilities, making it a cost-effective choice for small enterprises.

Imperson: Imperson is a platform that provides custom chatbot development services to businesses. The platform develops chatbots that can automate various business processes and increase consumer interaction using machine learning (ML) and natural language processing (NLP) techniques. The platform offers a visual design and construction interface for conversational flows, enabling users to design unique prompts, responses, and actions. Moreover, Imperson provides capabilities like analytics and platform connection

with different outside entities and and supports text, audio, video, AR, and VR.The bot is hosted and deployed by its full-service creative studio, which also offers a comprehensive analytics dashboard with real-time performance statistics. They regularly adjust your chatbot's performance based on user behaviour and monitor its success as well.

Target audience: Imperson targets businesses of all sizes that wish to automate their customer service, lead generating, and sales processes with chatbots. The platform's goal is to make chatbot development accessible to non-technical people without prior experience by streamlining the development process. It does, however, unlike other platforms, it also provide cutting-edge capabilities and integrations that may be useful to experienced developers to create more complex chatbots.

Comparative Study:

The table below represents the comparison between the 3 chatbot Platforms and compares them based on different criteria.

Criteria	**Dialogflow**	**Imperson**	**MobileMonkey**
Availability	Cloud-based platform	Cloud-based custom development services	Cloud-based platform
Ease of use	Requires some technical knowledge	Custom development service, less user-friendly for non-technical users	User-friendly, drag-and-drop interface with pre-built templates
Flexibility	Offers a wide range of integrations and customization options	Highly flexible and customizable according to user requirements	Offers some integrations and customizations options, but less than Dialogflow
Pricing	Offers a free tier with paid plans based on usage	Custom pricing based on project scope	Offers a free tier with paid plans based on features and usage
Natural language processing	Strong NLP capabilities with built-in machine learning models	Strong NLP capabilities with custom development	Good NLP capabilities with pre-built models and ability to train

Comparison between Dialogflow, Imperson and MobileMonkey on based different criteria.

Result/discussion based on comparative criteria:

Dialogflow, Imperson, and MobileMonkey are all cloud-based solutions. Imperson creates chatbots for particular clients, whereas Dialogflow and MobileMonkey are platforms for creating chatbots. Of the three, MobileMonkey has the easiest drag-and-drop interface and pre-built templates. To use Dialogflow and Imperson successfully, some technical knowledge is needed. Dialogflow is ideal for developers and

companies of all sizes because it has the most flexibility and integration options. Imperson's custom development services give it a great degree of adaptability and customization. Dialogflow provides more customization and integration options than MobileMonkey does. Strong natural language processing (NLP) capabilities are offered in Dialogflow and Imperson, with Dialogflow providing built-in machine learning models and Imperson providing custom development. With built-in models and the ability to train the bot, MobileMonkey has solid NLP skills.

References:

https://customers.ai/
https://influencermarketinghub.com/ai-chatbot-platforms/
https://zapier.com/blog/best-chatbot-builders/#MobileMonkey
https://www.chatbot.com/blog/conversational-design/
https://zapier.com/blog/best-chatbot-builders/#MobileMonkey
https://influencermarketinghub.com/ai-chatbot-platforms/
https://cloud.google.com/dialogflow
https://en.wikipedia.org/wiki/Dialogflow
https://www.imperson.com/

Comparative Study 6 - Data-to-Text Datasets and Evaluations

This Dataset Site lists 18 (as of March 2022) datasets when we search for "Data-to-Text Generation".

TO DO:

- Explore a few of these datasets.
- Choose 3 datasets for your analysis. Describe each dataset, what it contains, which International Evaluation (if any) it was used for, who was it developed by, etc. Show an example of its content.
- Develop a set of 5 criteria for your comparative study. The criteria could be with respect to the size, application, etc.
- Make a comparative table according to your set of criteria.

Solution 6:

WebNLG dataset

About Dataset (what it contains, which International Evaluation (if any) it was used for, who was it developed by, etc):

The WebNLG dataset is a benchmark dataset for natural language generation (NLG) from structured data. It consists of more than 18,000 records, each of which contains a set of triples that describe a particular entity (such as a person, place, or event), as well as corresponding natural language texts that describe the entity. The test set is split into two parts: seen, which contains inputs created for entities and relations belonging to DBpedia categories that were seen in the training data, and unseen, which contains inputs extracted for entities and relations belonging to 5 unseen categories.

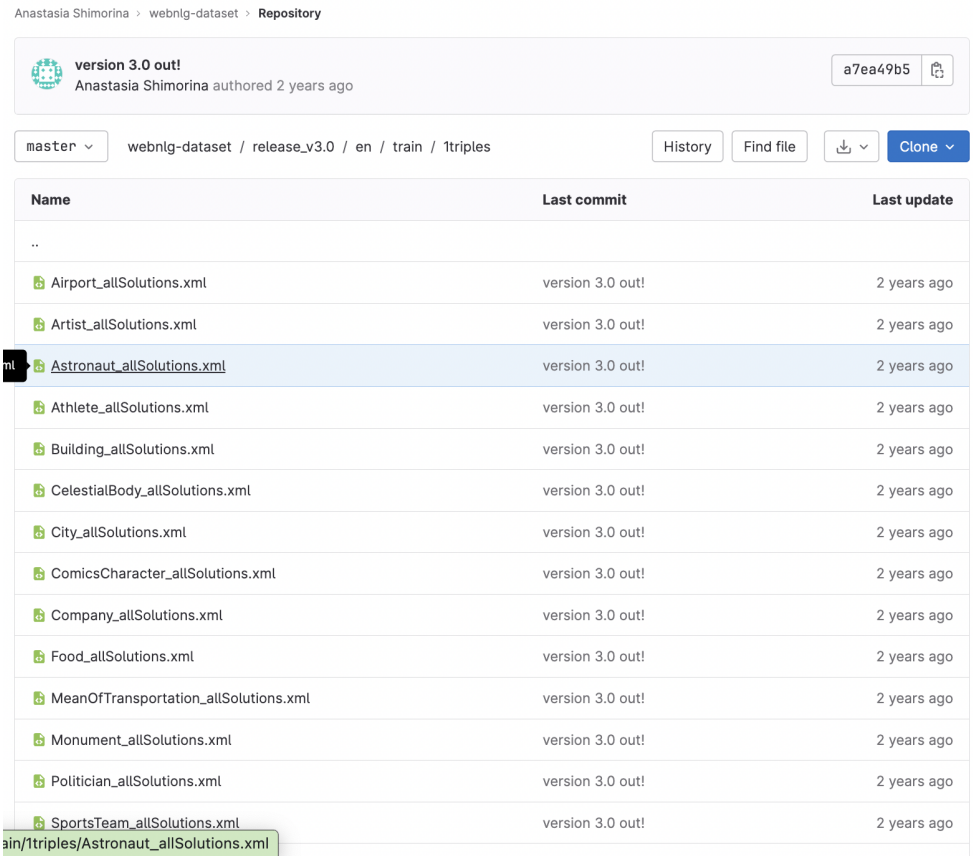
The dataset covers a wide range of domains, including geography, music, sports, and politics. The WebNLG dataset was developed to support research in data-to-text generation, particularly in the domain of web data. It has been used in several evaluations, including the Generation Challenges in the SemEval and INLG conferences, which aim to advance the state-of-the-art in data-to-text generation.

The WebNLG dataset was developed by researchers at the University of Sheffield, in collaboration with the Universitat Pompeu Fabra and the University of Edinburgh. It is freely available for research purposes, and it continues to be a valuable resource for researchers in the field of natural language generation.

Dataset used in (International Evaluation, research):

1. https://aclanthology.org/W17-3518/ **-** In the SemEval-2017 Generation Task, the WebNLG dataset was used for the task of data-to-text generation from RDF triples. The challenge involved generating textual descriptions of given entities and relations in the WebNLG dataset.
2. https://paperswithcode.com/paper/tplinker-single-stage-joint-extraction-of - The dataset has been used in this research as well.

Snapshot of dataset:



Snapshot of dataset

<u>PubMed Term, Abstract, Conclusion, Title Dataset</u>

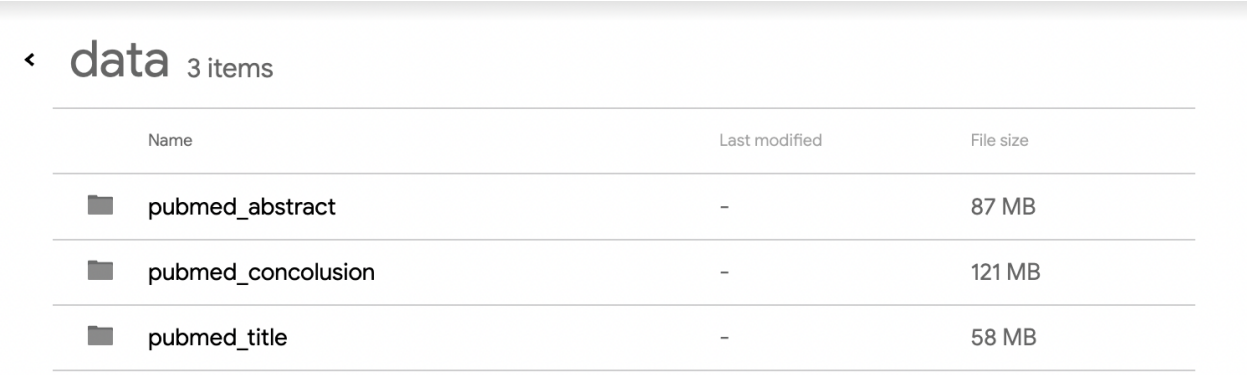
About Dataset (what it contains, which International Evaluation (if any) it was used for, who was it developed by, etc):

The PubMed Term, Abstract, Conclusion, Title Dataset (PACT) is a dataset for document classification and summarization tasks in the biomedical domain. It consists of more than 200,000 PubMed abstracts, each of which is labeled with one or more Medical Subject Headings (MeSH) terms that indicate the topic of the

article. The dataset also includes the title, abstract, and conclusion sections of each article, as well as information about the journal, publication date, and authors.
The PACT dataset was created to support research in text classification and summarization, particularly in the biomedical domain. It has been used in several evaluations, including the BioASQ challenge, which aims to advance the state-of-the-art in biomedical question answering.
The PACT dataset was developed by researchers at the University of Illinois at Urbana-Champaign and the National Library of Medicine. It is freely available for research purposes, and it continues to be a valuable resource for researchers in the field of biomedical text mining and natural language processing.

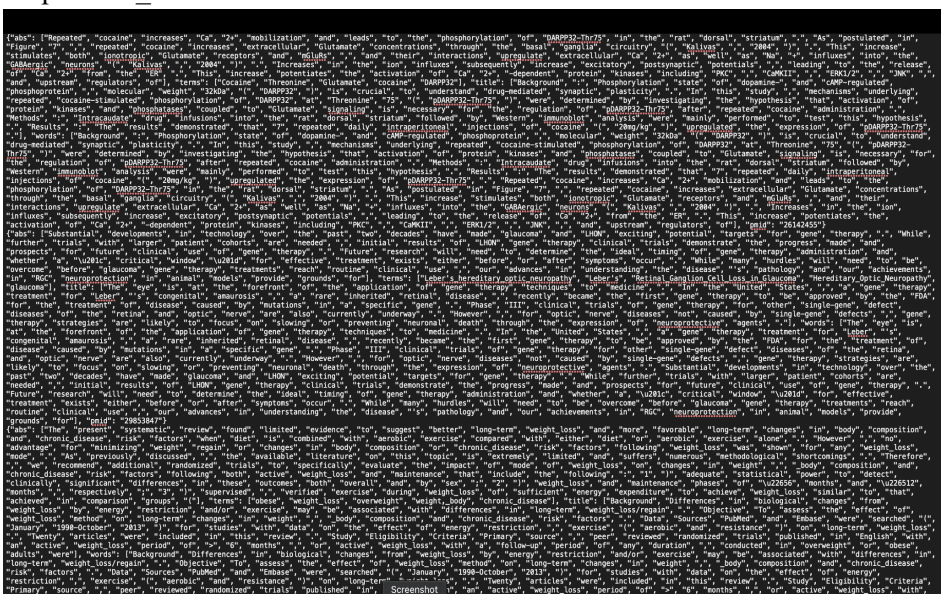
The dataset used in (International Evaluation, research):
It is used in the research work as mentioned below:
https://paperswithcode.com/paper/paperrobot-incremental-draft-generation-of

Snapshot of dataset:



Snapshot of dataset

A snapohot of a punmed_Conclusion - train.txt file.



Snapohot of a punmed_Conclusion - train.txt file.

E2E (End-to-End NLG Challenge) dataset

About Dataset (what it contains, which International Evaluation (if any) it was used for, who was it developed by, etc):

The E2E NLG Challenge dataset is a benchmark dataset for natural language generation (NLG) from structured data in the domain of restaurant information. It consists of approximately 50,000 records, each of which describes a restaurant using structured input data (such as restaurant name, location, cuisine, and price range) and a corresponding natural language text that describes the restaurant.

The E2E NLG Challenge dataset was developed to support research in data-to-text generation, particularly in the domain of restaurant information. It has been used in the yearly E2E NLG Challenge competition, which aims to advance the state-of-the-art in data-to-text generation for the restaurant domain.

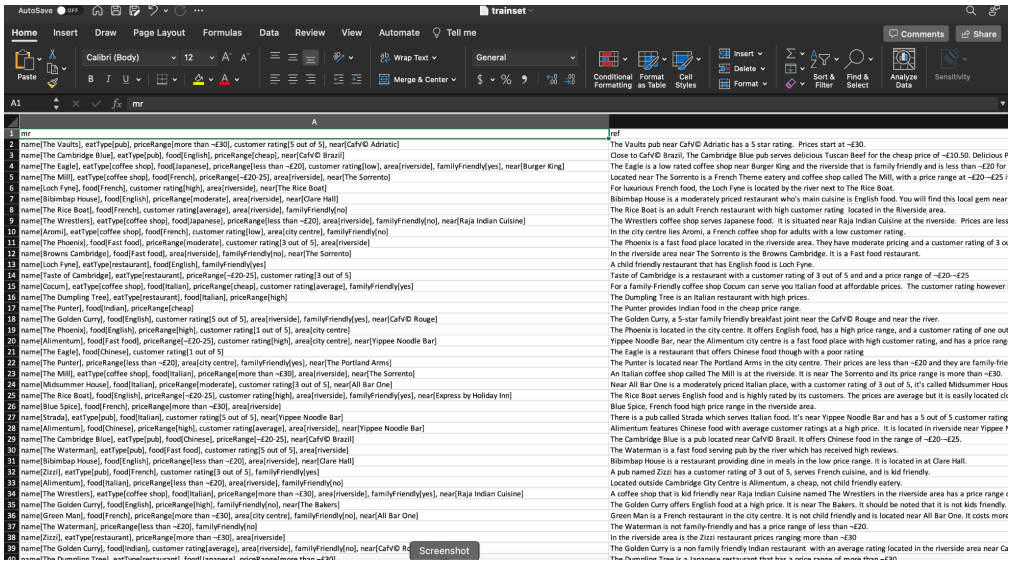
The E2E NLG Challenge dataset was developed by researchers at the University of Edinburgh and Heriot-Watt University. It is freely available for research purposes, and it continues to be a valuable resource for researchers in the field of natural language generation.

Dataset used in (International Evaluation, research):

The E2E (End-to-End) NLG Challenge dataset was developed specifically for the E2E Challenge, which was a competition that was part of the 11th International Natural Language Generation conference (INLG 2018). The challenge focused on the task of generating natural language text from structured data, such as database records or tables.

https://inlg2018.uvt.nl/

Snapshot of dataset:



Snapshot of dataset

Comparative Study:

The table below represents the comparison between the 3 Data-to-Text Datasets and compares them based on different criteria.

Criteria	WebNLG	E2E NLG Challenge	PubMed TACT
Purpose/Application	Natural Language Generation (NLG), Data-to-text generation	NLP and NLG, Data-to-text generation for restaurant reviews	Text Summarization and classification for biomedical and life sciences literature
Domain/Topic	Various domains such as music, food, sports, etc.	Restaurant Reviews	Biomedical and life sciences
Size	The WebNLG dataset consists of 25,298 (data,text) pairs and 9,674 distinct data units.	50,602 instances	300,000+ documents
Used in Evaluation	SemEval (2017), INLG Challenge (2017, 2018)	INLG Challenge (2018)	None
Dataset Attributes	- Contains 80k triples with annotations and surface realizations - Divided into 15 domains (e.g., book, film, music, etc.) - Includes 18,042 examples with inputs and multiple reference outputs	- Contains over 50,000 examples with input-output pairs - Includes multiple human reference outputs for each input. - Attributes of dataset are: Name, eat, Type, familyFriendly, priceRange, food, near, area, customerRating	- Contains over 1.4 million examples with titles, abstracts, and conclusions - Includes labels for article type, methodology, and clinical relevance

Comparison between WebNLG, E2E NLG Challenge and PubMed TACT based on different criteria.

Result/discussion based on comparative criteria:

The WebNLG dataset and the E2E NLG Challenge dataset were both created with natural language generation (NLG), whereas the PubMed TACT dataset was created with text categorization and summarization. While the E2E NLG Challenge dataset focuses specifically on restaurant reviews, the WebNLG dataset includes samples from a variety of domains, including geography, music, food, and sports.

The WebNLG dataset is primarily used for data-to-text generation tasks, where the goal is to generate natural language text based on structured data (e.g., triples). It includes multiple reference outputs for each

input, allowing for evaluation of different output variations. The dataset is diverse, covering a range of domains and topics.

The E2E NLG Challenge dataset is also used for data-to-text generation, but is focused specifically on generating restaurant reviews. It includes multiple human reference outputs for each input, allowing for evaluation of output quality. However, the dataset is limited in terms of domain and topic compared to the WebNLG dataset.

The PubMed TACT dataset is used for text classification and summarization tasks in the biomedical and life sciences fields. It includes over 1.4 million examples, making it the largest of the three datasets, and has labels for article type, methodology, and clinical relevance. However, the dataset is limited to a specific domain and does not include multiple reference outputs for each input like the WebNLG and E2E NLG Challenge datasets.

The E2E NLG Challenge dataset has 50,000 instances, the WebNLG dataset has about 18,000 instances, wherein it consists of 25,298 (data,text) pairs and 9,674 distinct data units, and the PubMed TACT dataset has approximately 300,000 text documents.

The WebNLG dataset has been used in international evaluations and challenges like SemEval (2017) and the INLG Challenge (2017, 2018). While the PubMed TACT dataset has not been used in any international evaluations or competitions but in research by PaperRobot, and the E2E NLG Challenge dataset was created specifically for the INLG Challenge (2018).

References:

https://eaglew.github.io/dataset/paperrobot_writing
https://aclanthology.org/W17-3518/
https://paperswithcode.com/dataset/webnlg
https://paperswithcode.com/dataset/e2e
https://paperswithcode.com/paper/evaluating-the-state-of-the-art-of-end-to-end
https://paperswithcode.com/paper/creating-training-corpora-for-nlg-micro
https://github.com/EagleW/PaperRobot