# Literature Review

In the past decade, conversational chatbots have seen a surge in popularity. The virtual assistant, such as offerings from Google Assistant and Amazon’s Alexa, are now entering our homes with Internet of Things devices. In 2017, Google Assistant was installed on over 400 million devices [1]. Furthermore, specialized chatbots have seen an influx within banking, retail, and healthcare. Chatbots represent a trend towards using natural language in the realm of human-computer interaction (HCI). This literature review will explore how chatbots are implemented, their benefits, and how this project can innovate within the chatbot space.

## Chatbots

When discussing chatbots, the beginning of their history is usually cited as Alan Turing’s 1950 article “Computing Machinery and Intelligence” [2], wherein Turing describes a test to determine whether a human evaluator can distinguish between a human and a machine during a natural language conversation. This test became known as the Turing test, and asks the question ‘Can machines think?’. However, the goal of many chatbots is not to create true artificial intelligence, but rather to using pattern matching and conversational responses to mimic the responses of a human.

One of the first programs to attempt the Turing test was ELIZA, created by Joseph Weizenbaum between 1964 and 1966 [3]. ELIZA consisted of a language analyser and a set of rules by which the ‘chatterbot’ followed. ELIZA used a script called DOCTOR was designed to simulate responses of a psychotherapist during a psychiatric interview – predominantly achieved by the therapist mirroring the responses of the patient[].

* Current uses

Chatbots are ubiquitous in modern life, with most modern smartphones pre-equipped with a personal assistant such as Google Assistant or Apple’s Siri.

Specific uses

## Natural Language Processing

## Machine Learning

## Datasets

Typically, chatbots are divided into two groups, open-domain and closed-domain [4]. In an open-domain system, the conversation can go in any direction, and the user can talk to the chatbot about any topic. A closed-domain system is restricted to a narrower topic area or set of function – these are the chatbots we see most in real-world applications such as customer service and banking. For this project, the focus will be on a closed-domain system as the goal is to create a chatbot that can achieve a goal – these are often called Goal-Oriented (GO) Chatbots [4]. However, to create a GO chatbot, one must have a goal the chatbot should achieve, and a dataset from which to learn.

The Ubuntu Dialogue Corpus (UDC), is one of the largest public dialog datasets available [5], consisting of 1 million multi-turn dialogues from users receiving technical support for Ubuntu-related problems [6]. **continue**

In terms of knowledge bases that lend themselves to the question and answer format, Wikipedia is the world’s largest collaboratively edited source of encyclopaedic knowledge [7]. In terms of size, it eclipses the size of the Encyclopaedia Britannica, its nearest rival, by a factor of ten [8] – as of 12 November 2019, there are over 5.9 million articles in English, and over 51 million articles in the 306 languages officially covered by the Wikimedia Foundation [9]. However, Wikipedia’s content is only fit for human reading [7], and is hard to process computationally. Many attempts have been made to formalise and structure this data [7], [8], [10], but this review will focus one of these being DBpedia.

DBpedia is a crowd-sourced effort to extract structured content from various Wikimedia projects [11], including Wikipedia. The English version of the DBpedia knowledge base describes 4.58 million things, out of which 4.22 million are classified in a consistent ontology [], consisting of 320 classes described by 1,650 different properties [12]. This structure enables programs to process this data effectively, including a chatbot application. The size of the DBpedia Ontology are shown in Figure 1, which demonstrates the scale of the project, and how this might be effective for the project.

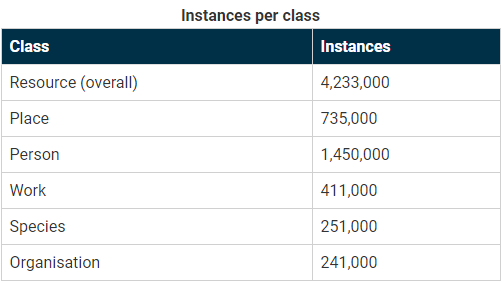


Figure 1: DBpedia Ontology instances per class [13]

The DBpedia extraction framework is responsible for extracting data from Wikipedia into a structured knowledge base, an overview of which is shown in Figure 2. This extraction is structured into four phases, as described in [12]:

**Input:** Wikipedia pages are read from an external source, either from a Wikipedia dump, or using the MediaWiki API.

**Parsing:** Each Wikipedia page is parsed, which transforms the source code of the Wikipedia page into an Abstract Syntax Tree (AST). An AST is a tree representation of the syntactic structure of the source code.

**Extraction:** The Abstract Syntax Tree of each page is forwarded to the extractors. There are many types of extractors, which will later be described, which extract data such as labels, images and infoboxes. Each extractor takes an AST as an input, and yields a set of Resource Description Framework (RDF) statements. These are XML statements which describe properties and values of resources.

**Output:** These RDF statements are written into sinks, which receive the data.

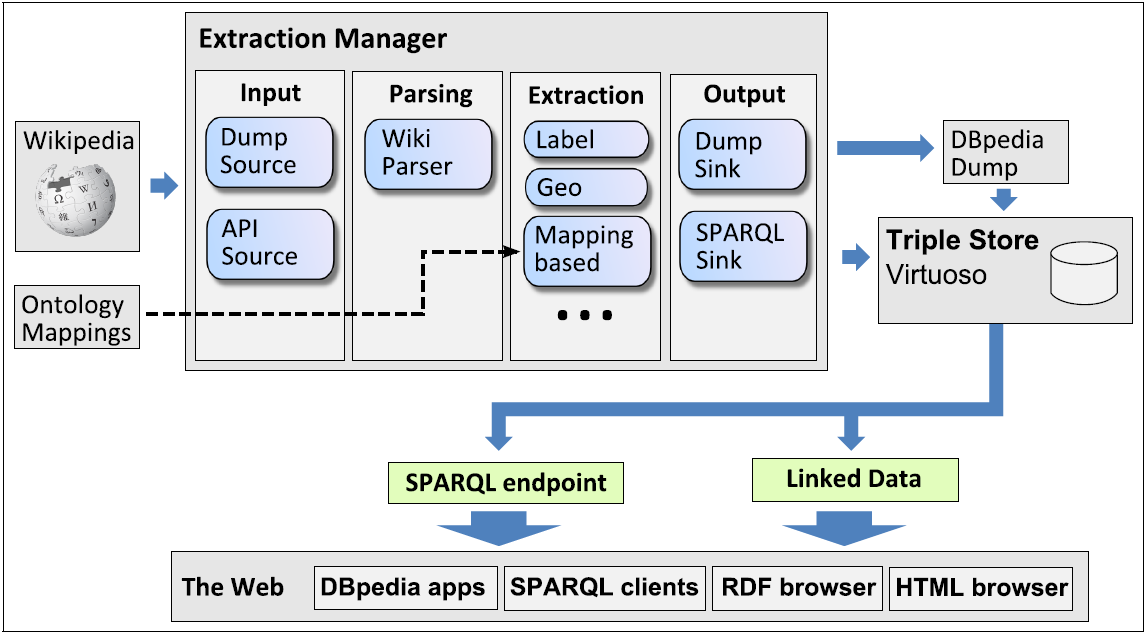


Figure 2: DBpedia extraction framework [12]

* Ontology
* Endpoint

## Programming Languages

* AIML
* Java
* Python
* Libraries

## Existing Solutions

* Google Search/Assistant
* DBPedia chatbot

## Conclusion

# References

|  |  |
| --- | --- |
| [1] | R. Chandra, “How Google Home and the Google Assistant helped you get more done in 2017,” 5 Jan 2018. [Online]. Available: https://blog.google/products/assistant/how-google-home-and-google-assistant-helped-you-get-more-done-in-2017/. [Accessed 5 Nov 2019]. |
| [2] | A. M. Turing, “Computing machinery and intelligence,” *Mind,* vol. 59, no. 236, pp. 433-60, 1950. |
| [3] | J. Weizenbaum, Computer Power and Human Reason: From Judgment to Calculation, New York: W.H. Freeman and Company, 1976. |
| [4] | V. Ilievski, “Building Advanced Dialogue Managers for Goal-Oriented Dialogue Systems,” 16 Mar 2018. [Online]. Available: https://arxiv.org/pdf/1806.00780.pdf. [Accessed 13 Nov 2019]. |
| [5] | S. Kojouharov, “Ultimate Guide to Leveraging NLP & Machine Learning for your Chatbot,” Chatbots Life, 18 Sep 2016. [Online]. Available: https://chatbotslife.com/ultimate-guide-to-leveraging-nlp-machine-learning-for-you-chatbot-531ff2dd870c. [Accessed 13 Nov 2019]. |
| [6] | R. Lowe, N. Pow, J. Pineau and I. V. Serban, “The Ubuntu Dialogue Corpus: A Large Dataset for Research in Unstructed Multi-Turn Dialog Systems,” 4 Feb 2016. [Online]. Available: https://arxiv.org/pdf/1506.08909.pdf. [Accessed 13 Nov 2019]. |
| [7] | M. Völkel, M. Krötzch, D. Vrandecic, H. Haller and R. Studer, “Semantic Wikipedia,” in *Proceedings of the 15th international conference on World Wide Web*, Edinburgh, 2006. |
| [8] | O. Medelyan, D. Milne, C. Legg and I. H. Witten, “Mining meaning from Wikipedia,” *International Journal of Human-Computer Studies,* vol. 67, no. 9, pp. 716-754, 2009. |
| [9] | Wikimedia Foundation, “List of Wikipedias,” 17 Oct 2019. [Online]. Available: https://meta.wikimedia.org/wiki/List\_of\_Wikipedias. [Accessed 13 Nov 2019]. |
| [10] | F. Wu and D. S. Weld, “Autonomously semantifying wikipedia,” in *CIKM '07 Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, Lisbon, 2007. |
| [11] | DBpedia, “About | DBpedia,” 2019. [Online]. Available: https://wiki.dbpedia.org/about. [Accessed 13 Nov 2019]. |
| [12] | J. Lehmann, et al., “DBpedia – A Large-scale, Multilingual Knowledge Base Extracted from Wikipedia,” *Semantic Web,* vol. 6, no. 2, pp. 167-195, 2015. |
| [13] | DBpedia, “Ontology | DBpedia,” 2019. [Online]. Available: https://wiki.dbpedia.org/services-resources/ontology. [Accessed 13 Nov 2019]. |