Provided by the author(s) and University of Galway in accordance with publisher policies. Please cite the published version when available.

Title	Towards bootstrapping a chatbot on industrial heritage through term and relation extraction
Author(s)	Arcan, Mihael; O'Halloran, Rory; Robin, Cecile; Buitelaar, Paul
Publication Date	2022-11-20
Publication Information	Arcan, Mihael, O'Halloran, Rory, Robin, Cecile, & Buitelaar, Paul. (2022). Towards bootstrapping a chatbot on industrial heritage through term and relation extraction. Paper presented at the Workshop on Natural Language Processing for Digital Humanities, Online, 20 November.
Publisher	Association for Computational Linguistics (ACL)
Link to publisher's version	https://aclanthology.org/2022.nlp4dh-1.15.pdf
Item record	http://hdl.handle.net/10379/17779

Downloaded 2024-04-26T12:11:13Z

Some rights reserved. For more information, please see the item record link above.

Towards Bootstrapping a Chatbot on Industrial Heritage through Term and Relation Extraction

Mihael Arcan, Rory O'Halloran, Cécile Robin and Paul Buitelaar
Insight SFI Research Centre for Data Analytics,
Data Science Institute, University of Galway
`firstname.lastname@insight-centre.org`

Abstract

We describe initial work in developing a methodology for the automatic generation of a conversational agent or 'chatbot' through term and relation extraction from a relevant corpus of language data. We develop our approach in the domain of industrial heritage in the 18th and 19th centuries, and more specifically on the industrial history of canals and mills in Ireland. We collected a corpus of relevant newspaper reports and Wikipedia articles, which we deemed representative of a layman's understanding of this topic. We used the Saffron toolkit to extract relevant terms and relations between the terms from the corpus and leveraged the extracted knowledge to query the British Library Digital Collection and the Project Gutenberg library. We leveraged the extracted terms and relations in identifying possible answers for a constructed set of questions based on the extracted terms, by matching them with sentences in the British Library Digital Collection and the Project Gutenberg library. In a final step, we then took this data set of question-answer pairs to train a chatbot. We evaluate our approach by manually assessing the appropriateness of the generated answers for a random sample, each of which is judged by four annotators.

1 Introduction

Figure 1: Example of a "mill race" or "mill run" that was used to provide continuous water power to mills (image: author).

Conversational agents or 'chatbots' are a convenient way of making information available, as can be witnessed from the significant growth of chatbots used in all kinds of settings, from banks to public services.[1] Also in cultural heritage settings, chatbots are now being employed more and more to interact with visitors to websites and virtual exhibitions.[2] Although frameworks such as Rasa (Bocklisch et al., 2017) enable the development of sophisticated chatbots that allow for fluent dialogue, an important bottleneck is in collecting and defining the training data for such systems. Training data comes in the form of 'intent-question' pairs, for example: `order` - *Are you open*?; *Can I order*?; *Will you deliver*. The definition and collection of such training data for any given application domain are challenging and costly, in particular for more specific and content-rich topics such as in cultural heritage settings. The range of possible intents will be significantly larger and more varied than in typical commercial settings such as ordering products or services. We, therefore, explore the use of term and relation extraction from a relevant corpus of language data as a bootstrapping step in identifying relevant concepts that can serve as intents.

In this paper, we describe our work towards developing a methodology where we focus on term and relation extraction for end-to-end text gener-

---

[1] `https://www.marketsandmarkets.com/Market-Reports/smart-advisor-market-72302363.html`

[2] `https://pro.europeana.eu/page/culture-chatbot`

Galway Advertiser	The Journal	Buildings of Ireland	History of Ireland
Galway Educate Together	The Irish Times	Stormontfield Heritage	Galway Library
Irish Waterway History	Clifden Heritage	This is Galway	Galway.net
National Archives	Galway Museum	Wikipedia	

Table 1: Identified online resources related to the city of Galway.

ation. This allows us to be independent of the existing resources needed to train a conversational agent. We develop our approach in the domain of industrial heritage, and more specifically on the industrial history of canals and mills in Ireland.

2 Related Work

Term extraction for conversational agents was presented by Pereira et al. (2019) and applied for customer service dialogue in the FinTech domain. The authors provide insights on term extraction for automatic generation of taxonomies from customer service dialogue data, which can contribute to the conversational agent use case by taking advantage of existing dialogues between customers and agents as learning data about the domain. Similarly, Atzori et al. (2017) create a recommender system within the travel domain, leveraging lightweight access through chatbots to support travellers during their holiday stay.

Huang et al. (2007) presents an approach for extracting question-answer pairs as chat knowledge from online discussion forums to train a chatbot for a certain domain. The authors use an SVM classifier to identify and rank relevant question-answer pairs based on content similarity. Domain-specific terms have been identified as a key part of understanding user requests in chatbot interaction (Mohapatra et al., 2018).

Chao et al. (2021) study chatbot development using a systematic patent analytic approach. They apply text-mining techniques, including document term frequency analysis for key terminology extractions, a clustering method for identifying the subdomains, and Latent Dirichlet Allocation (LDA) for finding the key topics of a patent set.

Abu-Shawar and Atwell (2016) focus on transforming corpora to a specific chatbot format, which is used to retrain a chatbot system. For this task, the authors use different dialogue corpora, i.e., such as the British National Corpus of English (BNC) and the Quran, which is a monologue corpus where verse and following verse are turns. The main goal of this automation process is the ability to generate different chatbot prototypes that communicate in different languages based on the corpus.

In contrast to previous work, our approach does not leverage classification methods to align a question to a predefined intent or answer, respectively. Additionally, leveraging term and relation extraction on a relevant corpus of language data, our approach is not limited to existing resources, such as the Ubuntu Dialogue Corpus (Lowe et al., 2015), data sets drawn from Twitter (Ritter et al., 2010) or Reddit (Roller et al., 2021), needed to train a conversational agent.

3 Data

This section provides insights on the resources used to build a chatbot in the domain of industrial heritage in the 18th and 19th centuries, i.e. the Galway Data Set, British Library Digital Collection and the Project Gutenberg library.

3.1 Galway Data Set

For our work, we initially leveraged 14 online resources to extract the required data for term and relation extraction in the domain of industrial heritage (see Table 1). In addition to online resources, we also leverage Wikipedia,[3] a freely available encyclopaedia that is built by a collaborative effort of voluntary contributors, to further increase the data set for term and relation extraction.

3.2 British Library Digital Collection

The British Library Digital Collection (BLDC) includes a collection of digitised books created by the British Library. This is a collection of books that have been digitised and processed using Optical Character Recognition (OCR) software to make the text machine-readable. We used the Curatr online platform (cf. Section 5.2) to access BLDC to retrieve a corpus in the domain of industrial heritage.[4]

---

[3] `https://en.wikipedia.org/`

[4] See Appendix A.1 for the list of digitised books used in this work.

3.3 Project Gutenberg library

Project Gutenberg[5] is the oldest digital library founded in 1971 and aims to digitise and archive cultural works. Most of the items in its collection are the full texts of books or individual stories in the public domain. All files can be accessed for free under an open format layout, which stores more than 50,000 items in its collection. Most items are in the English language, but many non-English works are also available. There are multiple affiliated projects that provide additional content, including region- and language-specific works. We selected 100 items from the Project Gutenberg library that represent the 18th and the 19th century.[6]

4 Methodology

Within our work, we first leverage the Galway data set to extract most relevant terms and relations between them in the domain of industrial heritage. We use these terms and relations in the next step to extract sentences from the BLDC and the Project Gutenberg corpus containing these terms and relations.

4.1 Term Extraction

For our initial step in extracting the most relevant terms within the targeted domain, we leveraged the identified online resources and collected documents relating to the city of Galway. Once the documents were collected, we employ the Saffron framework[7] (see Section 5.1) to extract the 100 most relevant terms from the collected documents, with a maximum term length of four words. Candidate term retrieval is the first step in the term extraction process. Saffron extracts potential candidate terms using noun phrase extraction, which are filtered based on term length, as specified in the configuration.

After selecting candidate terms, Saffron evaluates their relevance to the domain and ranks them accordingly from the most relevant to the least relevant. Saffron uses a combination of scoring functions calculated for each of the candidate terms. It combines functions, such as `comboBasic`, `totalTfIdf`, `cValue` and `residualIdf` (Astrakhantsev, 2018), which are based on occurrence frequencies. More in detail, we leverage frequencies of candidate terms across the documents or occurrences as part of other candidate terms, and that are based on reference corpora, i.e., comparing occurrences in the data set versus a generic reference data set (”weirdness” function, with Wikipedia being used as reference corpus). Finally, a voting algorithm (Zhang et al., 2008) is then used to combine the functions.

The final set of terms is selected from the original list of candidate terms after ranking, by filtering the top 100 terms of the list.

4.2 Relation Extraction

In the next step, we first parsed the initial data set and extracted the dependencies between tokens by the usage of the Stanza[8] dependency parser (Qi et al., 2020). We retrieved the dependencies where the extracted terms were identified with their relation, e.g. `subj(flow, water)`. Finally, we identified triples, where two terms are linked through a relation. As an example, from the extracted dependencies `subj(leave, canal)` and `obj(leave, river)` we construct the triple `subj_obj(canal, leave, river)`.

4.3 Conversational Data Set Creation

In the final step, the extracted terms and the relations were used to query the BLDC corpus and the Project Gutenberg library to obtain a more relevant data to train the chatbot system. With this, we obtained four different data sets, i.e.:

- subject or object term data set: A subject or object term has to be present in the sentence from the BLDC and the Project Gutenberg corpus.
- subject and object term data set: The subject and the object term of the same triple have to be present in the sentence from the BLDC and the Project Gutenberg corpus.
- subject or object term and relation data set: The subject or object term with its relation within the same triple have to be present in the sentence from the BLDC Corpus or the Project Gutenberg corpus.
- concatenated corpus: A weighted corpus of the sub-corpora mentioned above is generated.

The final data set to train the chatbot is derived from the Galway data set, the BLDC and the

---

[5] https://www.gutenberg.org/

[6] See Appendix A.2 for the list of item used in this work.

[7] https://saffron.insight-centre.org/

[8] https://stanfordnlp.github.io/stanza/

		Questions		Answers	
	lines	tokens	types	tokens	types
`subject or object term`	659,433	3,062,197	120	29,964,660	123,381
`subject and object term`	70,584	570,689	114	4,088,592	41,944
`subject or object term and relation`	4,010	25,355	132	246,759	19,124
`concatenated corpus`	729,636	3,595,795	164	33,542,445	123,052

Table 2: Statistics on extracted question-sentence pairs based on the extracted terms and relations.

Question pattern	Term(s)	Relation	Term(s) embedded within a Question
What is a TERM?	canal	/	What is a canal?
Tell me about a TERM	staircase	/	Tell me about a staircase
I'm interested in TERM	steam engine	/	I'm interested in steam engine
Was TERM used in Galway?	gate used	/	Was lock gate used in Ireland?
What was TERM used for?	mill	/	What was mill used for?
What is the relation between SUBJECT-TERM and OBJECT-TERM?	chamber, gate	/	What is the relation between a chamber and a gate?
What does a SUBJECT-TERM PREDICATE?	bridge	cross	What does a bridge cross?

Table 3: Examples of question patterns and the extracted terms and relations embedded within a question.

Project Gutenberg corpus, which represents a broad overview of the industrial environment of late 18th and 19th century Ireland. As discussed, from the collected documents, key terms and relations between them were identified using the knowledge extraction framework Saffron and the dependency parser Stanza. The terms and relations serve as a means of extracting high-relevance sentences that inform the chatbot's proficiency. This resulted in 659,433 relevant sentences (Table 2), which contained at least one of the extracted terms. We used 90% of the sentences for training and 10% for validation (development set) purposes. We filter this corpus based on subject and object terms in combination with the relation that appeared in the sentence, resulting in four sub-corpora for chatbot generation. From the held-out evaluation set, 50 sentences were randomly selected for manual evaluation by the four annotators.

4.4 Question Generation

As end-to-end chatbots are trained based on question-answer pairs, we use the extracted terms and relations for the question part and embed them within manually defined questions. As an example, the extracted term *canal* would become *What is a canal?* Table 3 shows the patterns used to construct the questions needed to train the chatbot. Using the OpenNMT toolkit (Section 5.3), the chatbot learns to properly respond to a question through the identified sentences, which contains the relevant (extracted) terms and relations.

5 Experimental Setup

In this section, we give an overview of the Saffron framework used for term and relation extraction. We leverage these terms to query the BLDC corpus through the Curatr online platform. Furthermore, we provide information on OpenNMT and the architecture of the trained sequence-to-sequence neural network. Finally, we provide insights on the evaluation approach.

5.1 Term Extraction with Saffron

Term extraction was performed with the knowledge extraction framework Saffron. This open-source tool allows us to extract terms (i.e. multi-word expressions) of the domain of the corpus, i.e. here the industrial history of canals and mills. Several parameters can be specified, such as N, the number of terms extracted, which we set up to 100 in order to cover a range of various terms (Bordea et al., 2013). The minimum and the maximum length of the terms can be determined, which we set to one and four words, in order to obtain generic terms (e.g. *canal*) as well as more specific ones (e.g. *mill* *race*).

5.2 Curatr

Curatr[9] is an online platform providing access to the British Library Digital Collection. The platform hosts digitised versions of all English-language books from the British Library collection, corresponding to over thirty-five thousand unique titles, from 1700 to 1899. The data collection consists of over forty-six thousand unique volumes of text.

The system enables queries on the equivalent of over 12 million individual pages of text, which can be searched and sorted by author, title, year, and the actual full-text of the volumes themselves. This allows us to identify content relating to specific themes within little known or very long, unwieldy texts.

As Curatr supports the creation and export of smaller sub-corpora, we used it to filter the entire collection to produce a much smaller set of texts for closer inspection. We used the terms `mills` and `canals` to retrieve a corpus in the domain of industrial heritage in the 18th and 19th centuries.

5.3 Text Generation

The neural models for text generation were performed with the OpenNMT toolkit (Klein et al., 2017). We used the transform-based network with its default setting. The network used a six-layer encoder-decoder model with the attention mechanism enabled (Vaswani et al., 2017). To cover the entire vocabulary of the training set, we use sentencepiece to split the words into subword units. The training approach uses a batch size of 4,096, leveraging the ADAM optimiser (Kingma and Ba, 2015). We set the word embeddings' size to 500, and hidden layers to size 500, dropout = 0.1, respectively. We used a maximum sentence length of 50.

5.4 Evaluation Approach

The evaluation of responses of open-domain conversational agents, such as chatbots, is still an open question (Liu et al., 2016) since a variety of answers can be considered as correct. Therefore, we randomly selected 50 question-term pairs (out of the 100 pairs of the evaluation set) and evaluated manually the generated answers. Following the error classes by (Coughlin, 2003), four volunteers were assessing the chatbot's responses to the questions into three classes:

- Unacceptable = 1. Absolutely not comprehensible and/or little or no information generated accurately.
- Possibly Acceptable = 2. Possibly comprehensible (given enough context and/or time to work it out); some information generated accurately.
- Acceptable = 3. Not perfect (stylistically or grammatically odd), but definitely comprehensible, AND with all important information generated accurately.

In addition to the manual evaluation, we analysed the Inter Annotation Agreement (IAA) between the four annotators. For this, the Fleiss' Kappa (Fleiss et al., 1971) was calculated (Equation 1). P (actual agreement) and P_e (expected agreement) measure the reliability of agreement between a fixed number of annotators when assigning categorical ratings to several items or classifying items.

$$\kappa = \frac{P - P_e}{1 - P_e} \qquad (1)$$

6 Results and Discussion

In this section, we present the evaluation results of generated answers and provide some further insights into the challenges of generating accurate responses.

6.1 Evaluation Results

Table 4 illustrates the manual evaluation of the 50 automatically generated answers. All annotators marked each answer either as *unacceptable* (1), *possible acceptable* (2) or *acceptable* (3). The scores from the annotation campaign range from 1.30 to 2.54. As seen from the table, the annotators evaluated the responses generated from the `subject and object term` training set with the highest score, an average of 2.40. The answers generated from the training set `concatenated corpus` were annotated with the lowest scores. The chatbot trained on `subject and object term` benefits from various generated questions containing two relevant terms, while all other corpora contain more

---

[9] http://erdos.ucd.ie/curatr/about

	Annotator 1	Annotator 2	Annotator 3	Annotator 4	Average
`subject or object term`	2.20	2.00	1.92	1.64	1.94
`subject and object term`	2.54	2.46	2.24	2.34	2.40
`subject/object term and relation`	1.52	1.56	1.92	1.38	1.60
`concatenated corpus`	1.34	1.54	1.64	1.30	1.46

Table 4: The evaluation scores of each annotator, based on the 50 automatically generated answers with its average score.

	κ
`subject or object term`	0.21
`subject and object term`	0.33
`subject/object term and relation`	0.36
`concatenated corpus`	0.41

Table 5: Annotator agreement scores based on the quality of the generated answers.

general questions with only one term or the combination of a term and its relation, depending on if terms and relations were used to extract the sentences.

Inter Annotator Agreement Due to the annotation approach with four annotators, we calculated the Fleiss' κ score based on the evaluation of the quality of the generated answers. Table 5 shows the different scores for each of the different corpora the chatbot was trained on. The annotators achieved a fair agreement[10] evaluating the first three corpora, and moderate agreement[11] (Fleiss et al., 1971) evaluating the chatbot's answers trained on the `concatenated corpus`.

6.2 Discussion

As mentioned before, the evaluation of an open-domain chatbot is still an open challenge, as various answers can be considered as correct. In a manual analysis of the chatbot's responses, we observed meaningful answers across all different data sets. Nevertheless, the first two examples in Table 6, *Tell me about a cotton and a sawmill*) and *What is the relation between a gate and a river*, illustrate examples, where the system did not provide a meaningful answers. While the first example only gives answer related to one term, i.e., *sawmill*. Although both terms are present in the generated answer, the second example demonstrates an incomprehensible answer where additional information is missing in the sentence. The third example illustrates the lexical ambiguity of the word *lock*.[12] Whereas the desired answer relating to the *enclosure consisting of a section of canal*, the chatbot generates an answer with the meaning of *a fastener fitted to a door*.

Analysing the remaining examples in Table 6, i.e., *What is the relation between a log and water?* or *What is the relation between a lumber and a mill*, the chatbot provided acceptable responses. For the former example, the chatbot provides the information on switching the transport of *logs* from *water* to railroads, while for the latter, the system generates a meaningful explanation that *logs* were cut into *lumber* by a *mill*.

7 Conclusion and Future Work

We described an approach towards the bootstrapping of a chatbot in the cultural domain through the application of readily available tools for term and relation extraction and natural language generation. The evaluation shows that the end-to-end neural model produces acceptable results when asking questions about the industrial heritage of the 18th and 19th centuries. Nevertheless, the approach does require more extension and refinement to be useful for automatic chatbot development. The current limitations of the approach are in data collection, term and relation extraction and eval-

---

[10] fair agreement = 0.21 – 0.40

[11] moderate agreement = 0.41 – 0.60

[12] from Princeton WordNet, 1: a fastener fitted to a door or drawer to keep it firmly closed, 2: enclosure consisting of a section of canal that can be closed to control the water level; used to raise or lower vessels that pass through it. `http://wordnetweb.princeton.edu/perl/webwn?s=lock`

Question #1	Tell me about a cotton and a sawmill
Generated Answer	A sawmill or lumber mill is a facility where logs are cut into lumber.
Question #2	What is the relation between a gate and a river?
Generated Answer	On the clare side of the river, at the foot of thomond gate 587 limerick.
Question #3	Tell me about a city and a lock
Generated Answer	287 the gates of the city to be put in repair with locks and keys
Question #4	What is the relation between a log and water?
Generated Answer	Availability of railroad transportation for logs and lumber encouraged building of rail mills away from navigable water.
Question #5	What is the relation between a lumber and a paper?
Generated Answer	The larger pieces of wood that wo n't make lumber are chipped into wood chips and provide a source of supply for paper mills.
Question #6	What is the relation between a lumber and a mill?
Generated Answer	A sawmill or lumber mill is a facility where logs are cut into lumber.
Question #7	What is the relation between a boat and a wheel?
Generated Answer	[25] these floating mills had a wheel that was attached to a boat moored in a fast flowing river.
Question #8	Tell me about a machine and water
Generated Answer	The water power could be used for driving the splint and box-making machines.
Question #9	Tell me about a lock and water
Generated Answer	The lock is filled with water from upstream.

Table 6: Examples of questions and the generated answers from the best performing model.

uation. In particular, our current data set consists only of a small number of contemporary documents of general but not of specific relevance to the application under consideration, i.e. industrial heritage of Ireland in the 18th and 19th centuries. Instead, in future work, we want to focus on data collection in digital libraries on contemporary as well as historical documents specifically on topics that are of direct relevance to this application context. While we extracted relevant terms and relations in the targeted domain, these terms are mostly single-word terms. Therefore, we are planning to focus on extracting more multi-word terms, which will help us to identify relevant sentences for training the chatbot system. Further, the generation of questions based on the extracted terms and relations is currently limited to a template-based approach. We envision that the inclusion of neural models, such as `Text-To-Text Transfer Transformer` (T5) (Raffel et al., 2020) will generate better natural language questions. Furthermore, we plan to incorporate multi-modal approaches, i.e. incorporating images within the chatbot, for visual representation as well as for disambiguation approaches. Finally, we would like to include relevant historical expertise to better inform our approach from the use case perspective.

Acknowledgements

This publication has emanated from research conducted with the financial support of Science Foundation Ireland under Grant number SFI/12/RC/2289_P2. For the purpose of Open Access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

References

Bayan Abu-Shawar and Eric Atwell. 2016. Automatic extraction of chatbot training data from natural dialogue corpora. In *RE-WOCHAT: Workshop on Collecting and Generating Resources for Chatbots and Conversational Agents-Development and Evaluation, Portoroz, Slovenia.*

Nikita Astrakhantsev. 2018. Atr4s: toolkit with state-of-the-art automatic terms recognition methods in scala. *Language Resources and Evaluation*, 52(3):853–872.

Maurizio Atzori, Ludovico Boratto, and Lucio Davide Spano. 2017. Towards chatbots as recommendation interfaces. In *Proceedings of the Second Workshop on Engineering Computer-Human Interaction in Recommender Systems co-located with the 9th ACM SIGCHI Symposium on Engineering Interactive Computing Systems (EICS 2017), Lisbon, Portugal, June 26, 2017*, volume 1945 of *CEUR Workshop Proceedings*, pages 26–31. CEUR-WS.org.

Tom Bocklisch, Joey Faulkner, Nick Pawlowski, and Alan Nichol. 2017. Rasa: Open source language understanding and dialogue management.

Georgeta Bordea, Paul Buitelaar, and Tamara Polajnar. 2013. Domain-independent term extraction through domain modelling. In *Proceedings of the 10th International Conference on Terminology and Artificial Intelligence (TIA 2013)*.

Min-Hua Chao, Amy J. C. Trappey, Chun-Ting Wu, and Abd E.I.-Baset Hassanien. 2021. Emerging technologies of natural language-enabled chatbots: A review and trend forecast using intelligent ontology extraction and patent analytics. volume 2021, USA. John Wiley & Sons, Inc.

Deborah Coughlin. 2003. Correlating automated and human assessments of machine translation quality. In *In Proceedings of MT Summit IX*, pages 63–70.

J.L. Fleiss et al. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

Jizhou Huang, Ming Zhou, and Dan Yang. 2007. Extracting chatbot knowledge from online discussion forums. In *IJCAI*, volume 7, pages 423–428.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander M. Rush. 2017. OpenNMT: Open-Source Toolkit for Neural Machine Translation. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, System Demonstrations:67–72.

Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2122–2132, Austin, Texas. Association for Computational Linguistics.

Ryan Lowe, Nissan Pow, Iulian Serban, and Joelle Pineau. 2015. The Ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 285–294, Prague, Czech Republic. Association for Computational Linguistics.

Prateeti Mohapatra, Yu Deng, Abhirut Gupta, Gargi Dasgupta, Amit Paradkar, Ruchi Mahindru, Daniela Rosu, Shu Tao, and Pooja Aggarwal. 2018. Domain knowledge driven key term extraction for it services. In *International Conference on Service-Oriented Computing*, pages 489–504. Springer.

Bianca Pereira, Cecile Robin, Tobias Daudert, John P. McCrae, Pranab Mohanty, and Paul Buitelaar. 2019. Taxonomy extraction for customer service knowledge base construction. In *Semantic Systems. The Power of AI and Knowledge Graphs*, pages 175–190, Cham. Springer International Publishing.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Alan Ritter, Colin Cherry, and Bill Dolan. 2010. Unsupervised modeling of Twitter conversations. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 172–180, Los Angeles, California. Association for Computational Linguistics.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Eric Michael Smith, Y-Lan Boureau, and Jason Weston. 2021. Recipes for building an open-domain chatbot. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 300–325, Online. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.

Ziqi Zhang, Jose Iria, Christopher Brewster, and Fabio Ciravegna. 2008. A comparative evaluation of term recognition algorithms. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

A Appendices

A.1 British National Library Collection

	Title	Author(s)	Published
1	Kilkenny City and County Guide and Directory, etc. With a map.	Bassett, George Henry	1884
2	County Down Guide and Directory, including the borough of Newry, etc. With a map.	Bassett, George Henry	1886
3	The Book of Antrim. A manual and directory, etc. With a map.	Bassett, George Henry	1888
4	The Book of County Armagh. A manual and directory, etc. With a map.	Bassett, George Henry	1888
5	The history and antiquities of Tallaght, County Dublin	Handcock, William Domville	1877
6	The Post Chaise Companion ... By R. Lewis or William Wilson. The 3d edition, corrected and enlarged, with an entire new set of plates	Lewis, R., Wilson, William	1803
7	The Old Bridge of Athlone	Joly, John Swift	1881
8	The Post Chaise Companion ... By R. Lewis? or W. Wilson. The 4th edition, corrected and enlarged, with an entire new set of plates	Lewis, R., Wilson, William	1820
9	Ancient Naas. Extracted from the ”Journal of the County Kildare Archæological Society.”	De Burgh, Thomas J.	1895
10	The Ancient and Modern History of the Maritime Ports of Ireland	Marmion, Anthony	1855
11	The Ancient and Modern History of the Maritime Ports of Ireland.	Marmion, Anthony	1860
12	History of the Rise and Progress of Belfast, and Annals of the County Antrim, etc	Pilson, James Adair	1846
13	A Short Topographical and statistical account of the Bandon Union; with some observations on the trade, agriculture, manufactures and tideways of the district. With a map.	Spillar, William A.	1844
14	The Book of County Tipperary. A manual and directory, etc. With a map.	Bassett, George Henry	1889
15	History of Sligo, county and town, from the earliest ages to the close of the reign of Queen Elizabeth (to the present time). With illustrations, etc	Martin, William Gregory Wood	1882
16	Sights and Scenes in our Fatherland	Lacy, Thomas	1863
17	The Irish Commercial and Railway Gazetteer, showing every town and railway station in Ireland, alphabetically arranged, with county, distance from Dublin, etc	Leggatt, John E.	1879
18	Rigby’s Illustrated Guide to Belfast and the North of Ireland	Rigby, Thomas	1895
19	How to spend a month in Ireland, and what it will cost	Roney, Cusack Patrick	1861
20	The new Hand-Book of Ireland; an illustrated guide for tourists and travellers	Walker, John A., Godkin, James	1871
21	Corporation. The Council Book of the Corporation of Youghal ... Edited from the original, with annals and appendices ... by R. Caulfield. With plates, etc.	Caulfield, Richard	1878

22	The Council Book of the Corporation of Kinsale, from 1652 to 1800. Illustrated. ... Edited from the original, with annals and appendices compiled from public and private records, by R. Caulfield	Caulfield, Richard	1879
23	The History of the Two Ulster Manors of Finagh, in the County of Tyrone, and Coole, otherwise Manor Atkinson, in the County of Fermanagh, and of their owners	Corry, Somerset Richard Lowry	1881
24	Limerick: its history and antiquities, ecclesiastical, civil, and military, from the earliest ages, with copious historical, archæological, topographical, and genealogical notes and illustrations; maps, plates, and appendices, etc	Lenihan, Maurice	1866
25	Guide to the most picturesque tour in Western Europe. Embracing the ... Lakes ... of Killarney, Glengarriff, ... Blarney, etc	Unknown	1888
26	Descriptive particulars of portions of the estates of the late J. Sadlier, situate in the counties of Tipperary and Limerick, and comprising, amongst others, the ... the demesne of Kilcommon, Cahir Castle, etc	Sadleir, John	1857
27	The Irish Sketch-book ... Second edition. , The Irish Sketch-Book	Thackeray, William Makepeace	1857
28	The West of Ireland: its existing condition, and prospects. Letters originally published in "Saunders's News-Letter." With plates and maps.	Coulter, Henry	1862
29	The history of the County and City of Cork	Gibson, Charles Bernard	1861
30	History of Enniscorthy, etc	Flood, William Henry Grattan	1898
31	Limerick and its Sieges ... Illustrated	Dowd, James	1890
32	"Devia Hibernia." The road and route guide for Ireland of the Royal Irish Constabulary. Compiled and edited by G. A. de M. E. Dagg. With a map.	Dagg, George Amyrald De Montmorency Edwin	1893
33	Official Tourist Guide of the Dublin, Wicklow, and Wexford Railway. Illustrated. Compiled by A. T. Hennessy	Hennessy, A. T.	1895
34	The History of the Town of Belfast with ... a statistical survey of the parish of Belfast, etc. By George Benn. , Appendix	Benn, George	1823
35	The Irish Sketch-book ... Second edition. , The Irish Sketch-Book	Thackeray, William Makepeace	1886
36	Guy's South of Ireland Pictorial Guide, etc	Unknown	1890
37	A Narrative of an Excursion to Ireland, by the Deputy Governor, two members of the Court, and the Assistant Secretary, of the Honorable Irish Society, of London. 1825. By the Deputy Governor i.e. Gilpin Gorst , Appendix. II. Miscellaneous	Gorst, Gilpin	1825
38	Picturesque guide to the Lakes of Killarney	Unknown	1851
39	Ancient and modern Sketches of the County Westmeath: historical, traditional, and legendary	Woods, James	1890
40	Here and there through Ireland ... With illustrations ... Reprinted from the "Weekly Freeman."	Banim, Mary	1891
41	Irish Pictures drawn with pen and pencil, etc	Lovett, Richard	1888
42	The Handbook for Youghal, containing an account of St. Mary's Collegiate Church. With the Historical Annals of the Town. (Third Series.) By S. Hayman.	Unknown	1852

43	History of Sligo, county and town, from the earliest ages to the close of the reign of Queen Elizabeth (to the present time). With illustrations, etc	Martin, William Gregory Wood	1882
44	Ireland. Revised edition. With a map of Ireland.	Browning, Walter Westcott	1884
45	Reminiscences of my Irish Journey in 1849. With a preface by J. A. Froude. , Single Works. Reminiscences of my Irish Journey	Carlyle, Thomas	1882
46	A History of the City of Dublin	Gilbert, John Thomas	1854
47	Illustrated Handbook to Cork, the Lakes of Killarney, and the South of Ireland ... From the Irish Tourists' Handbook or rather "The Tourists' Illustrated Handbook for Ireland" , Appendix	Unknown	1859
48	The Irish Sketch-book ... Second edition, The Irish Sketch-Book	Thackeray, William Makepeace	1845
49	The history and antiquities of the county of the town of Carrickfergus, etc.	Macskimin, Samuel	1823
50	The County and City of Cork Remembrancer; or, Annals of the county and city of Cork, etc	Tuckey, Francis H.	1837
51	Historical Sketches of Monaghan, from the earliest records to the Fenian movement	Rushe, Denis Carolan	1895
52	The City of Cork, how it may be improved. Lecture, etc. With a view and plan.	Walker, Robert	1883
53	The Irish Tourist's illustrated handbook for visitors to Ireland in 1852. Second edition	Unknown	1852
54	Two trips to the Emerald Isle. By "Faed" ... Illustrated, etc	Unknown	1888
55	Guide to North of Ireland, Giant's Causeway and Belfast, with history of Belfast	Aickin, Robert	1890
56	Trim: its ecclesiastical ruins, its castle, etc. By Edward Evans.	Evans, Edward	1886
57	Guy's South of Ireland Pictorial Guide, etc	Guy	1891
58	An Improved Topographical and Historical Hibernian Gazetteer ... Scientifically arranged, with an appendix of ancient names. To which is added, an introduction to the ancient and modern History of Ireland	Hansbrow, G.	1835
59	Queenstown and the Places around Cork Harbour. A handy guide, etc	Unknown	1895
60	The new Hand-book for Youghal: containing notes and records of the ancient religious foundations, and the historical annals of the town. Fourth series	Hayman, Samuel	1858
61	A Tour in Ireland, in 1813 and 1814 ... By an Englishman i.e. J. G.?	Gough, John	1817
62	The Saxon in Ireland: or, the rambles of an Englishman in search of a settlement in the West of Ireland. By John Henry Ashworth.	Ashworth, John Hervey	1851
63	The Shannon and its Lakes; or, a short history of that noble stream from its source to Limerick	Harvey, R.	1896
64	Popular Traditions of Glasgow: historical, legendary and biographical	Wallace, Andrew	1889
65	Notes and Gleanings relating to the County of Wexford, in its past and present conditions	Doyle, Martin	1868

66	Mellifont Abbey, Co. Louth: its ruins and associations. A guide and popular history. With illustrations.	Unknown	1897
67	Guide to Ireland. With illustrations. , Appendix. II. Miscellaneous	Unknown	1898
68	The Irish Sketch-book ... Second edition, The Irish Sketch-Book	Thackeray, William Makepeace	1845
69	The History of Dundalk, and its environs; from the earliest period to the present time; with memoirs of its eminent men. With plates and maps.	O’flanagan, James Roderick, D’alton, John	1864
70	Ireland, as I saw it: the character, condition, and prospects of the people	Balch, William S.	1850
71	Here and there through Ireland ... With illustrations ... Reprinted from the ”Weekly Freeman.”	Banim, Mary	1891
72	Guide to Ireland. (Second edition.) With illustrations. , Appendix. II. Miscellaneous	Unknown	1899
73	Lough Corrib, its shores and islands: with notices of Lough Mask ... Illustrated, etc	Wilde, William Robert Wills	1867
74	Rambling Recollections of Old Glasgow. By ”Nestor.”	Unknown	1880
75	The Sunny Side of Ireland. How to see it by the Great Southern and Western Railway ... With seven maps and over 130 illustrations, etc	O’mahony, John	1898
76	John Bull and his Other Island	Bennett, Arthur	1890
77	Ierne; or, Anecdotes and incidents during a Life in Ireland, with notices of people and places. First series, Appendix. II. Miscellaneous	Unknown	1861
78	The History of Bandon, and the principal towns in the West Riding of County Cork. Enlarged edition, with ... illustrations and a portrait	Bennett, George	1869
79	Dignam’s Dublin Guide. With a handy map, etc	Dignam, James	1891
80	Ireland: its health resorts and watering places. With maps, etc	Flinn, David Edgar	1888
81	Three Months’ Tour in Ireland ... Translated and condensed by Mrs. Arthur Walter. With illustrations, Trois mois en Irlande	Walter, Arthur, Bovet, Marie Anne De	1891
82	Through the Green Isle; a gossiping guide to the districts traversed by the Waterford, Limerick and Western Railway system ... Illustrated, etc	Hurley, M. J.	1895
83	The Council Book of the Corporation of the City of Cork, from 1609 to 1643, and from 1690 to 1800. Edited ... with annals and appendices ... by Richard Caufield	Caulfield, Richard	1876
84	Picturesque Scenery in Ireland drawn by T. Creswick. ... With descriptive jottings by a Tourist	Creswick, Thomas	1873
85	Pococke’s Tour in Ireland in 1752. Edited, with an introduction and notes, by G. T. Stokes	Stokes, George Thomas, Pococke, Richard	1891
86	The South Isles of Aran, County Galway	Burke, Oliver Joseph	1887
87	The Council Book of the Corporation of the City of Cork, from 1609 to 1643, and from 1690 to 1800. Edited ... with annals and appendices ... by Richard Caufield	Caulfield, Richard	1876
88	A Fortnight in Ireland. Second edition	Head, Francis Bond	1852
89	A Fortnight in Ireland	Head, Francis Bond	1852

90	The Land of Eire. The Irish Land League. Its origin, progress and consequences. Preceded by a concise history of the various movements which have culminated in the last great agitation ... With a descriptive and historical account of Ireland from the earliest period to the present day. Illustrated by numerous fine engravings, etc	Devoy, John	1882
91	Historical Gleanings in Antrim and neighbourhood	Smith, William Sunderland	1888
92	An Ulster Parish: being a history of Donaghcloney, Waringstown. With plates.	Atkinson, Edward Dupré	1898
93	Irish Tourist Development. ”Visit Ireland:” a concise, descriptive, and illustrated guide to Ireland. Compiled by F. W. Crossley	Crossley, F. W.	1892
94	Ulster as it is; or, twenty-eight years' experience as an Irish Editor	Macknight, Thomas	1896
95	Miscellaneous Essays on Topography, Ethnology, Language, ... contributed to the Ulster Journal of Archæology. With maps. MS. letter by the author	Hume, Abraham	1859
96	Lough Erne, Enniskillen, Belleek, Ballyshannon, and Bundoran, with routes from Dublin to Enniskillen, etc	Wakeman, William Frederick	1870
97	The history of the County and City of Cork	Gibson, Charles Bernard	1861
98	A History of the City of Dublin	Gilbert, John Thomas	1854
99	Topographical Sketches of Armagh and Tyrone ... Second edition	Rogers, Edward	1874
100	History of Sligo, county and town, from the earliest ages to the close of the reign of Queen Elizabeth (to the present time). With illustrations, etc	Martin, William Gregory Wood	1882

A.2 Project Gutenberg

	Title	Author(s)	Year
1	A Child's Dream of a Star	Charles Dickens	2013 [EBook 42232]
2	A Child's History of England	Charles Dickens	1996 [eBook 699]
3	A Christmas Carol	Charles Dickens	1992 [eBook 46]
4	A Christmas Carol	Charles Dickens	2007 [EBook 24022]
5	A Tale of Two Cities	Charles Dickens	1994 [eBook 98]
6	Adventures of Huckleberry Finn	Mark Twain (Samuel Clemens)	1993 [eBook 76]
7	The Adventures of Tom Sawyer	Mark Twain	1993 [eBook 74]
8	Alice's Adventures in Wonderland	Lewis Carroll	1991 [eBook 11]
9	American Notes for General Circulation	Charles Dickens	2013 [eBook 675]
10	Anna Karenina	Leo Tolstoy	1998 [EBook 1399]
11	An Enemy of the People	Henrik Ibsen	2000 [EBook 2446]
12	Around the World in Eighty Days	Jules Verne	1994 [eBook 103]
13	Bardell v. Pickwick	Percy Fitzgerald	2008 [eBook 25985]
14	Barnaby Rudge	Charles Dickens	2006 [EBook 917]
15	The Importance of Being Earnest	Oscar Wilde	1997 [eBook 844]
16	Bleak House	Charles Dickens	1997 [eBook 1023]
17	Youth	Joseph Conrad	1996 [EBook 525]

18	The Confidence-Man	Herman Melville	2007 [eBook 21816]
19	A Connecticut Yankee in King Arthur’s Court	Mark Twain	1993 [eBook 86]
20	The Count of Monte Cristo	Alexandre Dumas	1998 [eBook 1184]
21	David Copperfield	Charles Dickens	1996 [Etext 766]
22	De Profundis	Oscar Wilde	2007 [eBook 921]
23	A Doll’s House	Henrik Ibsen	2001 [eBook 2542]
24	Dombey and Son	Charles Dickens	1997 [eBook 821]
25	The Picture of Dorian Gray	Oscar Wilde	1994 [eBook 174]
26	Dracula	Bram Stoker	1995 [eBook 345]
27	Dracula’s Guest	Bram Stoker	2003 [EBook 10150]
28	Emma	Jane Austen	1994 [eBook 158]
29	Frankenstein	Mary Wollstonecraft (Godwin) Shelley	1993 [eBook 84]
30	Ghosts	Henrik Ibsen	2005 [EBook 8121]
31	Great Expectations	Charles Dickens	1998 [eBook 1400]
32	Grimms’ Fairy Tales	Jacob Grimm and Wilhelm Grimm	2001 [eBook 2591]
33	The Happy Prince	Oscar Wilde	1997 [eBook 902]
34	Hard Times	Charles Dickens	2013 [eBook 786]
35	Heart of Darkness	Joseph Conrad	1995 [eBook 219]
36	Hedda Gabler	Henrik Ibsen	2003 [Etext 4093]
37	The Hound of the Baskervilles	Arthur Conan Doyle	2001 [eBook 2852]
38	Hunted Down	Charles Dickens	2014 [eBook 807]
39	An Ideal Husband	Oscar Wilde	1997 [eBook 885]
40	The Innocents Abroad	Mark Twain (Samuel Clemens)	2006 [EBook 3176]
41	John Gabriel Borkman	Henrik Ibsen,Translated William Archer	2006 [eBook 18792]
42	A Journey to the Centre of the Earth	Jules Verne	2006 [EBook 18857]
43	The Lady From The Sea	Henrik Ibsen	2008 [EBook 2765]
44	The Last Man	Mary Wollstonecraft Shelley	2006 [eBook 18247]
45	Life On The Mississippi	Complete MarkTwain (Samuel Clemens)	2006 [EBook 245]
46	Little Dorrit	Charles Dickens	2008 [EBook 963]
47	Through the Looking-Glass	Charles Dodgson, AKA Lewis Carroll	1991 [eBook 12]
48	Lord Jim	Joseph Conrad	2006 [EBook 5658]
49	The Man in the Iron Mask	Alexandre Dumas, Père	2001 [eBook 2759]
50	Master and Man	Leo Tolstoy	1997 [Etext 986]
51	The Master Builder	Henrik Ibsen	2003 [Etext 4070]
52	Metamorphosis	Franz Kafka	2002 [eBook 5200]
53	Moby-Dick; or The Whale	Herman Melville	2001 [eBook 2701]
54	Nostromo: A Tale of the Seaboard	Joseph Conrad	2006 [EBook 2021]
55	Oliver Twist, illustrated	Charles Dickens	2014 [eBook 46675]
56	Oliver Twist	Charles Dickens	1996 [eBook 730]
57	Our Mutual Friend	Charles Dickens	2006 [EBook 883]
58	Persuasion	Jane Austen	2008 [EBook 105]
59	Pictures from Italy	Charles Dickens, Illustrated Marcus Stone	2013 [eBook 650]
60	Pillars of Society	Henrik Ibsen	2010 [EBook 2296]

61	The Piazza Tales	Herman Melville	2005 [eBook 15859]
62	Pride and Prejudice	Jane Austen	1998 [eBook 1342]
63	The Prince and The Pauper	Mark Twain (Samuel Clemens)	2006 [EBook 1837]
64	Rosmerholm	Henrik Ibsen	2010 [EBook 2289]
65	Best Russian Short Stories	Various	2004 [EBook 13437]
66	Sevastopol	Lyof N. Tolstoï	2014 [EBook 47197]
67	The Adventures of Sherlock Holmes	Arthur Conan Doyle	2002 [eBook 1661]
68	Sketches by Boz	Charles Dickens	1997 [eBook 882]
69	Some Christmas Stories	Charles Dickens	2015 [eBook 1467]
70	Stories of Intellect	Various	2020 [EBook 61668]
71	Tales and Stories	Mary Wollstonecraft Shelley	2018 [EBook 56665]
72	Ten Years Later	Alexandre Dumas	2001 [eBook 2681]
73	The Footpath Way	Sidney Smith, William Hazlitt, Isaak Walton, Walter Scott, et al.	2019 [eBook 59813]
74	The Great English Short-Story Writers, Vol. 1	Various, et al	2003 [eBook 10135]
75	The Collected Works of Henrik Ibsen	Henrik Ibsen	2021 [eBook 66186]
76	The Letters of Charles Dickens	Charles Dickens	2008 [EBook 25852]
77	The Life And Adventures Of Nicholas Nickleby	Charles Dickens	2006 [EBook 967]
78	The Lock and Key Library	Hawthorne, Ed.	1999 [Etext 1831]
79	The Mystery of Edwin Drood	Charles Dickens	1996 [eBook 564]
80	The Old Curiosity Shop	Charles Dickens	1996 [eBook 700]
81	The Pickwick Papers	Charles Dickens	2009 [EBook 580]
82	The Posthumous Papers of the Pickwick Club,v. 2(of 2)	Charles Dickens	2014 [EBook 47535]
83	The Uncommercial Traveller	Charles Dickens	1997 [eBook 914]
84	Pierre; or The Ambiguities	Herman Melville	2011 [eBook 34970]
85	Bartleby The Scrivener	Herman Melville	2004 [eBook 11231]
86	The Secret Agent	Joseph Conrad	1997 [eBook 974]
87	The Secret Sharer	Joseph Conrad	1995 [EBook 220]
88	The Trial	Franz Kafka	2005 [EBook 7849]
89	Three Ghost Stories	Charles Dickens	2013 [eBook 1289]
90	The Three Musketeers	Alexandre Dumas, Père	1998 [eBook 1257]
91	Twenty Thousand Leagues under the Sea	Jules Verne	1994 [eBook 164]
92	Twenty Years After	Alexandre Dumas, Père	1998 [eBook 1259]
93	Typee	Herman Melville	1999 [eBook 1900]
94	Typhoon	Joseph Conrad	2006 [EBook 1142]
95	War and Peace	Leo Tolstoy	2001 [eBook 2600]
96	What Men Live By and Other Tales	Leo Tolstoy	2004 [EBook 6157]
97	When We Dead Awaken	Henrik Ibsen	2003 [EBook 4782]
98	White-Jacket	Herman Melville	2004 [eBook 10712]