From Eliza to Siri and Beyond

Luísa Coheur^{1,2}

¹ INESC-ID Lisboa, Lisbon, Portugal

luisa.coheur@inesc-id.pt

² Instituto Superior Técnico/Universidade de Lisboa, Lisbon, Portugal

Abstract. Since Eliza, the first chatbot ever, developed in the 60s, researchers try to make machines understand (or mimic the understanding) of Natural Language input. Some conversational agents target small talk, while others are more task-oriented. However, from the earliest rule-based systems to the recent data-driven approaches, although many paths were explored with more or less success, we are not there yet. Rule-based systems require much manual work; data-driven systems require a lot of data. Domain adaptation is (again) a current hot-topic. The possibility to add emotions to the conversational agents' responses, or to make their answers capture their “persona”, are some popular research topics. This paper explains why the task of Natural Language Understanding is so complicated, detailing the linguistic phenomena that lead to the main challenges. Then, the long walk in this field is surveyed, from the earlier systems to the current trends.

Keywords: Natural language processing · Natural Language Understanding · Chatbots

1 Introduction

Since the first chatbots from the 60s (such as ELIZA [42]) to the current virtual assistants (such as SIRI¹), many things have changed. However, and despite the incredible achievements done in Natural Language Processing (NLP), we are far from creating a machine capable of **understanding natural language**, a long-standing goal of **Artificial Intelligence (AI)** and probably the ultimate goal of NLP.

Understanding natural language is an extremely complex task. Researchers in NLP have been struggling with it since the early days. Even the concept of “understanding” is not consensual. Some authors consider that whatever the implemented methods are, if a system is capable of providing a correct answer to some natural language input, then we can say that it was able to correctly interpret that input. Other authors assume that the mapping of the input sentence into some **semantic representation**, which captures its meaning, is necessary.

¹ <https://www.apple.com/siri/>.

The process of mapping the natural language input into the semantic representation is called **semantic parsing**.

Deep Learning brought a new way of doing things and also an extra verve to the field. However, independently of the approach followed, and even if some interesting and accurate dialogues with machines can be achieved in some limited domains, we can quickly find out inconsistencies in conversational agents responses. In fact, no current system genuinely understands language.

This paper is organized as follows: Sect. 2 surveys some of the main challenges when dealing with natural language understanding, and Sect. 3 give some historical perspective of the main advances in the area. Then, Sect. 4 discuss current trends and Sect. 5 concludes and points to some future work.

2 Challenges

Most researchers do not realize that language is this complex until embracing a NLP task involving understanding natural language. Indeed, we manage to communicate with some success. Consequently, we are not aware that the sequences of words that we produce and interpret within our dialogues are extremely variable (despite obeying to certain syntax rules), often ambiguous (several interpretations are often possible) and that, sometimes, sophisticated reasoning (sometimes considering features that go beyond natural language) needs to be applied so that a fully understanding can be achieved.

In fact, a factor that makes the computational processing of our language terribly complex is **language variability**, that is, our ability to say the same thing in so many different ways (e.g., *yes*, *right*, *Ok*, *Okay*, *Okie dokie*, *looks good*, *absolutely*, *of course* are just some ways of expressing agreement). If a bot operates in a strictly closed domain it is possible to gather the semantics of the most common questions that will be posed to it². Still, the main problem is not the semantics of the most common questions, but their form, which can vary a lot. For instance, in some closed domains we can build a list of FAQs representing what will be asked to a virtual agent. However, we will hardly have a list of all the paraphrases (sentences with the same meaning) of those questions. As an example, consider the following sentences (from [12]):

1. *Symptoms of influenza include fever and nasal congestion;*
2. *Fever and nasal congestion are symptoms of influenza;*
3. *A stuffy nose and elevated temperature are signs you may have the flu.*

Sentences (1) and (2) can be easily identified as paraphrases, by simply considering the lexical units in common of both sentences. However, sentence (3) will only be identified as a paraphrase of (1) and (2) if we know that *fever* and *elevated temperature* are similar or equal concepts and the same between *nasal congestion* and *stuffy nose*. WordNet [20] and current word embeddings do, indeed, solve some of these problems, but not all.

² Nevertheless, we can hardly predict the flow of the dialogue, unless the machine takes the initiative and do not leave much room for innovation to the human.

Another factor that makes language so complicated is the fact that it is inherently **ambiguous**. Sentences with different meanings emerge from **ambiguity at the lexical level**. For instance, some meanings of the word *light* are³:

1. *comparatively little physical weight or density*
2. *visual effect of illumination on objects or scenes as created in pictures.*

This leads to ambiguous sentences like *I will take the light suit*.

There is a well known NLP task called **Word Sense Disambiguation** that targets the specific problem of lexical ambiguity. Nonetheless, lexical entities are not the only source of ambiguity. A good example of **syntactic ambiguity** is the classical sentence *I saw a man in the hill with a telescope*. Who had the telescope? Who was on the hill? Many interpretations are possible. Another good example of ambiguity is the sentence *João and Maria got married*. We do not even notice that this sentence is ambiguous. However, did they marry each other or with other people?

In most cases we can find the correct interpretation of a sentence by considering the **context** in which it is uttered. Many works (including recent ones, such as the work described in [37]) propose different ways of dealing with context, which is still a popular research topic. Yet, to complicate things further, although context can help to dismantle some productions, it can also lead to more interpretations of a sentence. For instance, *I found it hilarious* can be detected by a *Sentiment Analyser* as a positive comment, which is correct if we are evaluating a comic film, but probably not if we refer to a horror movie or a drama. Another example: *I'll be waiting for you at 4.p.m. outside school* can be a typical line of a dialogue between mother and son, but can also be a bullying threat. Context is everything.

A further difficulty is that it is impossible to predict all the different sentences that will be posed to a virtual agente, unless, as previously mentioned, it operates in a really strictly closed domain. Therefore, a bottleneck of conversational agents are Out-of-Domain (OOD) requests. For instance, as reported in [2], Edgar Smith [11], a virtual butler capable of answering questions about Monserrate Palace in Sintra, was reasonably effective when the user was asking **In-Domain questions**. However, people kept asking it **OOD questions**, such as *Do you like Cristiano Ronaldo?*, *Are you married?*, *Who is your favourite actress?*. Although it might be argued that, in light of their assistive nature, such systems should be focused in their domain-specific functions, the fact is that people become more engaged with these applications if OOD requests are addressed [5, 24].

Several other linguistica phenomena make dialogues even more difficult to follow by a machine. For instance, for interpreting the sentences *Rebello de Sousa and Costa went to Spain*. *The president went to Madrid and the prime-minister to Barcelona.*, we need to know that Marcelo is the president and that Costa is the prime-minister. The NLP task that deals with this phenomenon is **Coreference Resolution**. Also, **elliptical constructions** – omission of one or more

³ According with <https://muse.dillfrog.com/meaning/word/light>.

words in a sentence that can be inferred – make things extremely difficult to machines (e.g. *Is Gulbenkian's museum open on Sundays? And on Saturdays?*). In addition, we use **idioms** (e.g., *it's raining cats and dogs*), **colocations** and many multi-word expressions whose meaning is not necessarily related with the meaning of its parts. For instance, colocations are sequences up to three words that we learn to use since ever, but that we do not really know why we say it that way (e.g., in Portuguese we say *perdi o avião* (literally *I lost the plane*), to say *I missed my flight*). Why the verb *perdi*? (by the same token, why *miss*?). If for a foreign language learner the production (and understanding) of these expressions is a real challenge, it certainly is too for a computer. Moreover, we cannot ignore **humor** and **sarcasm**, which further complicate the machine tasks of understanding language (e.g., *Those who believe in telekinetics, raise my hand.* – Kurt Vonnegut).

Furthermore, the simple fact that we might be using speech when interacting with the machine (not to mention sign language) adds an extra layer of problems. For instance, a noisy environment or a strong accent can be enough to unable our understanding of what is being said (the same for a virtual agent). Also, the way we say something influences the interpretation of a sentence. For instance, *Good morning* can be said, in an unpleasant tone, to those who arrive late at class, meaning *Finally!*.

To conclude, the way we express ourselves has to do with our mastery of language, which, in turn, is influenced by the time in which we live, by our age, social condition, occupation, region where we were born and/or inhabit, emotional state, among others. All those features make each one of us a special case (and computers prefer archetypal patterns). A truly robust application should be able to handle all of our productions. Therefore, it must be equipped with some reasoning ability, which must also take into account all non-verbal elements involved in a conversation (e.g., the interlocutor's facial expression, the scenario, etc.). And the truth is that we are still far from being able to integrate all these variables in the (natural language) understanding process.

3 Historical Perspective

As previously said, ELIZA is considered to be the first chatbot ever, developed in the 1960s, with the aim of simulating a psychotherapist. Although it was able to establish a conversation, simulating it was a human being, its virtual model was based in rephrasing the user input, whenever it matched a set of hand-crafted rules and also in providing content-free remarks (such as *Please go on.*) in the absence of a matching. For instance, ELIZA could have a rule like the following one, in which * is the wildcard and matches every sequence of words; the (2) means that the sequence of words captured by the second wildcard would be returned in ELIZA's answer:

Rule: * *you are* */ *What makes you think I am (2)?*

If the user input was, for instance, *I think you are bright*, it would answer *What makes you think I am bright?*. At that time, many people believed they

were talking with another human (the “ELIZA effect”). Having no intention of modelling the human cognitive process and despite its simplicity, ELIZA showed how a software program can cause a huge impact by the mere illusion of understanding. Nowadays, ELIZA is still one of the most widely known applications in AI. ELIZA and subsequent bots (such as the paranoid mental patient PARRY [9] or JABBERWACKY [8]), definitely provided the seeds to many different directions to explore. For instance, the idea of having a virtual agent with a “persona” that explains the flow (and the flaws) of the conversation continues to be widely used. Another idea that is due to the first chatbots and that is still being explored today is that of “learning by talking”. For example, if a user asks *What do you think of Imagine Dragons?* and the bot does not know how to answer, it will record the question. The next time it interacts with a human, that same question will be asked by the bot and (hopefully) a possible answer will be gathered. Of course, this approach can go wrong and a very (recent) popular case was that of a chatbot that was willing to learn by interacting with humans. Within hours of being in use, the chatbot was racist, nazi and vulgar, and had to be turned off. A survey on the early chatbots and their contributions can be found in [25].

Another important line of research emerged in the beginning of the 70s, due to three seminal papers of Richard Montague: *English as a Formal Language* [21], *The Proper Treatment of Quantification in Ordinary English* [22], and *Universal Grammar* [23]. Montague proposed a formal framework to map syntactic structures into semantic representations, providing a systematic and **compositional** way of doing it. Although Montague’s work was limited to a very small subset of the English language and difficult to extend, the idea of taking advantage of syntax to build semantic forms in a compositional way was used in the many **Natural Language Interfaces to Databases (NLIDB)** that popped-up by that time (mostly during the 80s), and that are still being developed nowadays, although with different techniques ([3] surveys classical NLIDBs and [1] more recent ones).

Then, **Question/Answering (QA) (and Dialogue)** systems started to come out. Contrary to NLIDB, QA systems knowledge sources are not (necessarily) ground in databases; they are quite similar otherwise. On the subject of Dialogue systems, they are designed to engage in a dialogue with the user to get the information they need to complete a task (for instance), and, thus, there is more than just a question and an answer involved. The strong development of the former in the late 90s and beginning of the XXI century was partially due evaluation *fora*, such as TREC⁴ (since 99) and CLEF⁵ (since 2003), which provided tasks entirely dedicated to QA, such as QA@CLEF⁶. This allowed researchers to straightforwardly compare their systems, as everybody was evaluated with the same test sets. A side effect of these competitions was the release of data that become usually available to the whole community. This certainly also consolidated the rise of Machine Learning techniques against traditional rule-based

⁴ <https://trec.nist.gov>.

⁵ <http://www.clef-initiative.eu/home>.

⁶ <http://www.clef-initiative.eu/track/qaclef>.

approaches. Nowadays, an important task in NLP is **Machine Reading Comprehension** that targets to understand unstructured text in order to answer questions about it, but new related challenges are still emerging, such as the Conversational Question Answering Challenge (CoQA) [30].

In what concerns Dialogue Systems, several domains were explored from the early days. Particularly prolific were the conversational agents targeting the concept of Edutainment, that is, education through entertainment. Following this strategy, several bots have animated museums all over the world: the 3D animated Hans Christian Andersen [4] established multimodal conversations about the writer’s life and tales, Max [27] was employed as guide in the Heinz Nixdorf Museums Forum, Sergeant Blackwell [31], installed in the Cooper-Hewitt National Design Museum in New York, was used by the U.S. Army Recruiting Command as a hi-tech attraction and information source, and the previously mentioned **Edgar Smith** (Fig. 1).

Fig. 1. Edgar Smith in Monserrate.

In some of these systems the agent’s knowledge base was constituted of pairs of sentences (S_1 , S_2), where S_2 (the answer) is a response to S_1 (the trigger). Their “understanding” process was based on a **retrieval approach**: if the user says something that is “close” to some trigger that the agent has in its knowledge base (that is, if the user input matches or rephrases a trigger), then it will return the correspondent answer. Others based their approach in information extraction techniques, capable of detecting the user intentions and extract relevant entities from the dialogues in order to capture the “meaning” of the sentence. A typical general architecture of the latter systems combined natural language understanding and generation modules (sometimes template-based). A dialogue manager was present in most approaches. Nowadays, **end-to-end**

data-driven systems – that is systems trained directly from virtual data to optimize an objective function [17] –, have replaced or at least try to replace these architectures.

Then, things started to happen fast: Watson wins Jeopardy! in 2011 (a big victory to NLP), Apple releases Siri, also in 2011, Google Now appears in 2012, and the world faces a new level of conversational agents: the virtual assistants built by these colossal companies⁷. In the meantime, Deep Learning starts to win in all fronts, and Sequence to Sequence (seq2seq) models start to be successfully applied to Machine Translation [36]. Considering that these models take as input a sequence of words in one language and output a sequence of words in another language, the first generative dialogue systems based on these models did not take long to appear, and the already mentioned **end-to-end (dialogue) systems** came to light (v.g. [18, 35, 37, 39, 44, 45, 47]). These systems differ from retrieval-based systems as in the latter pre-defined responses are given to the user; in the generative-based approaches responses are generated in run-time. The majority of these systems are trained to engage in general conversations (chit-chat agents) and, therefore, make use of movie subtitles (or Reddit⁸ data).

Besides seq2seq architectures, two concepts are responsible for many of the latest achievements: **Neural Word Embeddings** and **Pre-trained Language Models**. Machine learning algorithms cannot usually directly deal with plain text. Therefore, the idea of converting words into vectors has been explored for a long time. Word Embeddings are functions that map words (or characters, paragraphs or even documents) into vectors (and neural networks can learn these mappings). More recently deep learning led to the creation of several embedding types, from context-free to contextual models, from character- to word-based, etc. Considering context-free embeddings, each word form is associated with a single embedding, and, thus, the word *light* will have a single embedding associated. In contextual models, the embedding captures a specific meaning of the word, and, therefore, *light* will have several vectors associated, according with the context in which it occurs. Examples of context-free embeddings are the ones created with Word2Vec models [19]; examples of the latter are ELMo [26] and BERT [10]. A simple way to use these models in dialogue systems is, for instance, in a retrieval-base approach, calculate the embedding of the agent’s knowledge base sentences and the embedding associated with the user given sentence. Then, find the cosine similarity between the embedding of the latter and the ones from the knowledge base, and return the one with the highest value. In what concerns language models, these can be seen as models that are trained to predict the next word (or character) in a sequence. For instance, by counting n-grams (sequences of n tokens) we can build a language model. These have many application scenarios. For instance, in a translation setup, decide which, from possible translations of a source sentence, is more probable, considering the target language (and the n-grams observed in that target language). An example of a language model is

⁷ Microsoft’s Cortana and Alexa were born in 2014.

⁸ <https://www.reddit.com>.

GPT-2 [29], trained in 40GB of text data, and developed by OPenAI⁹. It should be said that behind some of the most successful models (as BERT and GPT-2), there is a Transformer-base model. Transformers [38] were introduced in 2017 and are enjoying great success in NLP.

Several problems are still under research. The next section presents some of the current trends.

4 Current Trends

Several challenges lie ahead of the current end-to-end conversational agents. Just to name a few, some systems are unable to track the topic of the conversation, or are prone to generate trivial (and universal) responses such as “ok” or “I don’t know” (the “universal answer” drawback). Some authors propose neural models that enable dialogue state tracking (v.g. [48]), or new methods that inject diversity in the generated responses (v.g. [33]).

Current systems try to take into account important features of the user request, as for instance, their sentiment [6, 14, 34]. This is particularly important for support bots, as, for instance, a very unhappy customer (negative polarity) should not receive an answer starting with *Hello, we are so happy to hear from you!!!*. Researchers also explore how to add declarative knowledge to neural network architectures. As an example, in [16], neural networks are augmented with First-Order Logic. Domain adaptation is also, as previously said, a hot-topic (v.g [28]). Some current research follows in the transfer-learning paradigm: pre-trained Language Models are fine-tuned with specific data. For instance, in [7], GPT-2 is used as a pre-trained model and fine-tuned in task-oriented dialogue systems; in [43] a model is pre-trained on the BookCorpus dataset [49], and, then, tuned on the PERSONA-CHAT corpus [46]. The latter corpus was created to allow the building of chit-chat conversational agents, with a configurable, but persistent persona [46]. This idea of creating a bot with a consistent persona is also the topic of research of many current works [13, 15].

Many more proposals, not necessarily following in the end-to-end paradigm, are also worth to be mentioned. For instance, in [41] the authors propose the building of a semantic parser overnight, in which crowdsourcing is used to paraphrase canonical utterances (automatically built), into natural language questions; in [40] the computer learns, from scratch, the language used by people playing a game in the blocks world. Users can use whatever language they want, and they can even invent one.

5 Main Conclusions and Future Work

Since its early days, the NLP community has embraced the task of building conversational agents. However, and despite all the recent achievements (mainly due to Deep Learning), a short conversation with these systems quickly exposes their

⁹ <https://openai.com>.

weaknesses [32], including the lack of a consistent personality. The community is pretty aware of these limitations, and recent work is focusing on boosting the conversational agents' capabilities. Pre-trained models will certainly continue to be explored, as well as ways to enrich the model training with different types of knowledge.

The adaptation of a bot to a specific user is also something to explore, as the given answer will probably differ regardless of whether we interact with the bot once (for instance, when buying tickets for visiting a specific monument) or regularly (for instance, to reserve a hotel in a particular platform). In the latter case we want our assistant to remember some information about the user.

Acknowledgements. I would like to express my gratitude to Vânia Mendonça, who gave me very detailed comments about this document. However, the responsibility for any imprecision lies with me.

References

1. Affolter, K., Stockinger, K., Bernstein, A.: A comparative survey of recent natural language interfaces for databases. *VLDB J.* **28**(5), 793–819 (2019). <https://doi.org/10.1007/s00778-019-00567-8>
2. Ameixa, D., Coheur, L., Fialho, P., Quaresma, P.: Luke, I am your father: dealing with out-of-domain requests by using movies subtitles. In: Bickmore, T., Marsella, S., Sidner, C. (eds.) *IVA 2014. LNCS (LNAI)*, vol. 8637, pp. 13–21. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-09767-1_2
3. Androutsopoulos, I., Ritchie, G., Thanisch, P.: Natural language interfaces to databases - an introduction. *Nat. Lang. Eng.* **1**(1), 29–81 (1995). <https://doi.org/10.1017/S135132490000005X>
4. Bernsen, N.O., Dybkjær, L.: Meet Hans Christian Andersen. In: *Proceedings of Sixth SIGdial Workshop on Discourse and Dialogue*, pp. 237–241 (2005)
5. Bickmore, T., Cassell, J.: How about this weather? Social dialogue with embodied conversational agents. In: *Socially Intelligent Agents: The Human in the Loop*, pp. 4–8. AAAI Press, Menlo Park (2000)
6. Bothe, C., Magg, S., Weber, C., Wermter, S.: Dialogue-based neural learning to estimate the sentiment of a next upcoming utterance. In: Lintas, A., Rovetta, S., Verschure, P.F.M.J., Villa, A.E.P. (eds.) *ICANN 2017. LNCS*, vol. 10614, pp. 477–485. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-68612-7_54
7. Budzianowski, P., Vulic, I.: Hello, it's GPT-2 - how can i help you? Towards the use of pretrained language models for task-oriented dialogue systems (2019). <https://doi.org/10.17863/CAM.47097>
8. Carpenter, R., Freeman, J.: *Computing Machinery and the Individual: The Personal Turing Test* (2005). <http://www.jabberwacky.com>
9. Colby, K.M.: Ten criticisms of PARRY. *SIGART Newslett.* 5–9 (1974)
10. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Minneapolis, Minnesota, Volume 1 (Long and Short Papers), pp. 4171–4186. Association for Computational Linguistics, June 2019. <https://doi.org/10.18653/v1/N19-1423>. <https://www.aclweb.org/anthology/N19-1423>

11. Fialho, P., et al.: Meet EDGAR, a tutoring agent at MONSERRATE. In: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations, Sofia, Bulgaria, pp. 61–66. Association for Computational Linguistics, August 2013. <https://www.aclweb.org/anthology/P13-4011>
12. Fialho, P., Coheur, L., Quaresma, P.: From lexical to semantic features in paraphrase identification. In: Rodrigues, R., Janousek, J., Ferreira, L., Coheur, L., Batista, F., Oliveira, H.G. (eds.) 8th Symposium on Languages, Applications and Technologies (SLATE 2019). OpenAccess Series in Informatics (OASICS), Dagstuhl, Germany, vol. 74, pp. 9:1–9:11. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik (2019). <https://doi.org/10.4230/OASICS.SLATE.2019.9>. <http://drops.dagstuhl.de/opus/volltexte/2019/10876>
13. Herzig, J., Shmueli-Scheuer, M., Sandbank, T., Konopnicki, D.: Neural response generation for customer service based on personality traits. In: Proceedings of the 10th International Conference on Natural Language Generation, Santiago de Compostela, Spain, pp. 252–256. Association for Computational Linguistics, September 2017. <https://doi.org/10.18653/v1/W17-3541>. <https://www.aclweb.org/anthology/W17-3541>
14. Kong, X., Li, B., Neubig, G., Hovy, E., Yang, Y.: An adversarial approach to high-quality, sentiment-controlled neural dialogue generation. In: AAAI 2019 Workshop on Reasoning and Learning for Human-Machine Dialogues (DEEP-DIAL 2019), Honolulu, Hawaii, January 2019. <https://arxiv.org/abs/1901.07129>
15. Li, J., Galley, M., Brockett, C., Spithourakis, G., Gao, J., Dolan, B.: A persona-based neural conversation model. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, Berlin, Germany, (Volume 1: Long Papers), pp. 994–1003. Association for Computational Linguistics, August 2016. <https://doi.org/10.18653/v1/P16-1094>. <https://www.aclweb.org/anthology/P16-1094>
16. Li, T., Srikumar, V.: Augmenting neural networks with first-order logic. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, pp. 292–302. Association for Computational Linguistics, July 2019. <https://doi.org/10.18653/v1/P19-1028>. <https://www.aclweb.org/anthology/P19-1028>
17. Lowe, R.T., Pow, N., Serban, I.V., Charlin, L., Liu, C., Pineau, J.: Training end-to-end dialogue systems with the ubuntu dialogue corpus. *D&D* **8**(1), 31–65 (2017). <http://dad.uni-bielefeld.de/index.php/dad/article/view/3698>
18. Lu, Y., Keung, P., Zhang, S., Sun, J., Bhardwaj, V.: A practical approach to dialogue response generation in closed domains. *CoRR* abs/1703.09439 (2017)
19. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. In: Bengio, Y., LeCun, Y. (eds.) 1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, 2–4 May 2013, Workshop Track Proceedings (2013). <http://arxiv.org/abs/1301.3781>
20. Miller, G.A.: WordNet: a lexical database for English. *Commun. ACM* **38**, 39–41 (1995)
21. Montague, R.: English as a formal language. In: Thomason, R. (ed.) *Formal Philosophy. Selected papers of Richard Montague*, pp. 188–221. Yale University Press, New Haven (1974)
22. Montague, R.: The proper treatment of quantification in ordinary English. In: Thomason, R. (ed.) *Formal Philosophy. Selected papers of Richard Montague*, pp. 247–270. Yale University Press, New Haven (1974)

23. Montague, R.: Universal grammar. In: Thomason, R. (ed.) *Formal Philosophy. Selected papers of Richard Montague*, pp. 223–246. Yale University Press, New Haven (1974)
24. Patel, R., Leuski, A., Traum, D.: Dealing with out of domain questions in virtual characters. In: Gratch, J., Young, M., Aylett, R., Ballin, D., Olivier, P. (eds.) *IVA 2006. LNCS (LNAI)*, vol. 4133, pp. 121–131. Springer, Heidelberg (2006). https://doi.org/10.1007/11821830_10
25. Pereira, M.J., Coheur, L., Fialho, P., Ribeiro, R.: Chatbots’ greetings to human-computer communication. *CoRR* abs/1609.06479 (2016). <http://arxiv.org/abs/1609.06479>
26. Peters, M., et al.: Deep contextualized word representations. In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, New Orleans, Louisiana, Volume 1 (Long Papers), pp. 2227–2237. Association for Computational Linguistics, June 2018. <https://doi.org/10.18653/v1/N18-1202>. <https://www.aclweb.org/anthology/N18-1202>
27. Pfeiffer, T., Liguda, C., Wachsmuth, I., Stein, S.: Living with a virtual agent: seven years with an embodied conversational agent at the heinz nixdorf museumsforum. In: *Proceedings of the International Conference Re-thinking Technology in Museums 2011 - Emerging Experiences*, pp. 121–131. Thinkk Creative & The University of Limerick (2011)
28. Qian, K., Yu, Z.: Domain adaptive dialog generation via meta learning. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy, pp. 2639–2649. Association for Computational Linguistics, July 2019. <https://doi.org/10.18653/v1/P19-1253>. <https://www.aclweb.org/anthology/P19-1253>
29. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I.: *Language models are unsupervised multitask learners* (2019)
30. Reddy, S., Chen, D., Manning, C.D.: CoQA: a conversational question answering challenge. *Trans. Assoc. Comput. Linguist.* **7**, 249–266 (2019). <https://www.aclweb.org/anthology/Q19-1016>
31. Robinson, S., Traum, D., Ittycheriah, M., Henderer, J.: What would you ask a conversational agent? Observations of human-agent dialogues in a museum setting. In: *International Conference on Language Resources and Evaluation (LREC)*, Marrakech, Morocco (2008)
32. Serban, I.V., Lowe, R., Henderson, P., Charlin, L., Pineau, J.: A survey of available corpora for building data-driven dialogue systems: the journal version. *Dialogue Discourse* **9**(1) (2018)
33. Shao, Y., Gouws, S., Britz, D., Goldie, A., Strophe, B., Kurzweil, R.: Generating high-quality and informative conversation responses with sequence-to-sequence models. In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Copenhagen, Denmark, pp. 2210–2219 (2017)
34. Shi, W., Yu, Z.: Sentiment adaptive end-to-end dialog systems. In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, Melbourne, Australia, (Volume 1: Long Papers), pp. 1509–1519. Association for Computational Linguistics, July 2018. <https://doi.org/10.18653/v1/P18-1140>. <https://www.aclweb.org/anthology/P18-1140>
35. Song, Y., Yan, R., Feng, Y., Zhang, Y., Zhao, D., Zhang, M.: Towards a neural conversation model with diversity net using determinantal point processes. In: *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, New Orleans, Louisiana, USA, 2–7 February 2018 (2018)

36. Sutskever, I., Vinyals, O., Le, Q.V.: Sequence to sequence learning with neural networks. In: Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N.D., Weinberger, K.Q. (eds.) *Advances in Neural Information Processing Systems* 27, pp. 3104–3112. Curran Associates, Inc. (2014). <http://papers.nips.cc/paper/5346-sequence-to-sequence-learning-with-neural-networks.pdf>
37. Tian, Z., Yan, R., Mou, L., Song, Y., Feng, Y., Zhao, D.: How to make context more useful? An empirical study on context-aware neural conversational models. In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, Vancouver, Canada, (Volume 2: Short Papers), pp. 231–236. Association for Computational Linguistics, July 2017
38. Vaswani, A., et al.: Attention is all you need. In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS 2017, pp. 6000–6010. Curran Associates Inc., Red Hook (2017)
39. Vinyals, O., Le, Q.V.: A neural conversational model. CoRR abs/1506.05869 (2015)
40. Wang, S.I., Liang, P., Manning, C.D.: Learning language games through interaction. In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, Berlin, Germany, (Volume 1: Long Papers), pp. 2368–2378. Association for Computational Linguistics, August 2016. <https://doi.org/10.18653/v1/P16-1224>. <https://www.aclweb.org/anthology/P16-1224>
41. Wang, Y., Berant, J., Liang, P.: Building a semantic parser overnight. In: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, Beijing, China, (Volume 1: Long Papers), pp. 1332–1342. Association for Computational Linguistics, July 2015. <https://doi.org/10.3115/v1/P15-1129>. <https://www.aclweb.org/anthology/P15-1129>
42. Weizenbaum, J.: ELIZA - a computer program for the study of natural language communication between man and machine. *Commun. ACM* **9**, 36–45 (1966)
43. Wolf, T., Sanh, V., Chaumond, J., Delangue, C.: TransferTransfo: a transfer learning approach for neural network based conversational agents. CoRR abs/1901.08149 (2019). <http://arxiv.org/abs/1901.08149>
44. Yao, L., Zhang, Y., Feng, Y., Zhao, D., Yan, R.: Towards implicit content-introducing for generative short-text conversation systems. In: *Empirical Methods in Natural Language Processing (EMNLP)*, pp. 2190–2199. Association for Computational Linguistics, Copenhagen (2017)
45. Yin, J., Jiang, X., Lu, Z., Shang, L., Li, H., Li, X.: Neural generative question answering. In: *International Joint Conference on Artificial Intelligence IJCAI*, pp. 2972–2978. IJCAI/AAAI Press, New York (2016)
46. Zhang, S., Dinan, E., Urbanek, J., Szlam, A., Kiela, D., Weston, J.: Personalizing dialogue agents: I have a dog, do you have pets too? In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, Melbourne, Australia, (Volume 1: Long Papers), pp. 2204–2213. Association for Computational Linguistics, July 2018. <https://doi.org/10.18653/v1/P18-1205>. <https://www.aclweb.org/anthology/P18-1205>
47. Zhao, T., Eskenazi, M.: Towards end-to-end learning for dialog state tracking and management using deep reinforcement learning. In: *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, Los Angeles, pp. 1–10. Association for Computational Linguistics, September 2016. <https://doi.org/10.18653/v1/W16-3601>. <https://www.aclweb.org/anthology/W16-3601>

48. Zhao, T., Zhao, R., Eskenazi, M.: Learning discourse-level diversity for neural dialog models using conditional variational autoencoders. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, Vancouver, Canada, (Volume 1: Long Papers), pp. 654–664. Association for Computational Linguistics, July 2017. <https://doi.org/10.18653/v1/P17-1061>
49. Zhu, Y., et al.: Aligning books and movies: towards story-like visual explanations by watching movies and reading books. In: Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), ICCV 2015, pp. 19–27. IEEE Computer Society, USA (2015). <https://doi.org/10.1109/ICCV.2015.11>. <https://doi.org/10.1109/ICCV.2015.11>