

# DDD (Digital Data Deception) Technology Watch Newsletter

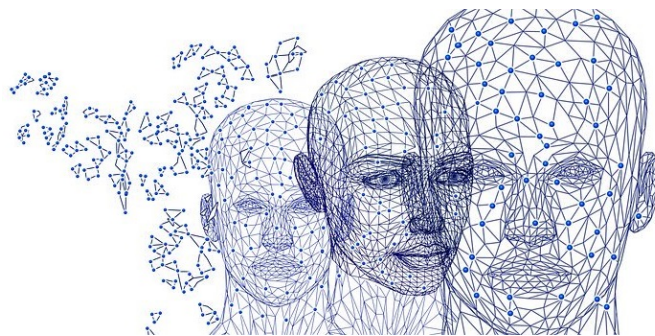
---

---

---

## Table of Contents

- Editorial
- List of Acronyms
- Definitions and Scope
- Stylised Text Generation
- Conversation
- Rewriting



*“All warfare is based on deception. Hence, when we are able to attack, we must seem unable; when using our forces, we must appear inactive; when we are near, we must make the enemy believe we are far away; when far away, we must make him believe we are near.”*

— Sun Tzu, *The Art of War*

**Editors:** Keenan Jones, Enes Altuncu, Virginia Franqueira, Yichao Wang, Sanjay Bhat-tacherjee and Shujun Li

**Affiliation:** Institute of Cyber Security for Society (iCSS), University of Kent, UK

**Contact Us:** [ddd-newsletter@kent.ac.uk](mailto:ddd-newsletter@kent.ac.uk)

---

---

## Editorial

The rapid growth in our abilities to generate artificial texts, leveraging new capacities for deep learning and powerful pre-trained models, such as OpenAI's GPT-3 [17], has meant that natural language generation (NLG) has never been more relevant as a tool for real world use.

Given the broad potential that comes with being able to generate convincing texts, NLG thus finds applications in a wide range of text-based tasks. In turn, NLG has been used in a wide range of fields, including chatbot development [87], story creation [1], and joke telling [3].

This breadth of application means that NLG has a powerful capacity for integrating itself into our everyday lives. Already, virtual assistants, like Amazon's Alexa and Apple's Siri, have found their way into the homes of hundreds of millions of users [65]. These devices rely on NLG-based modules responsible for generating dialogue in response to users' input in order to answer questions or complete tasks [65]. Beyond personal assistants, NLG systems have been used in a variety of other real-world applications. These include the medical field, in which dialogue and Q&A systems have been leveraged to aid with diagnoses and patient care [75, 96]. Education has also been a common focus of NLG applications, in which a wide variety of task-based systems including story-telling tools and Q&A chatbots have been proposed as a method of aiding student learning through virtual platforms and e-tutoring [3, 56].

With these real-world applications, however, comes the potential for deception and misuse [40]. Malicious chatbots could masquerade as genuine people in the hope of misleading them, or tricking them into divulging personal information [143]. NLG systems designed to persuade could be used by unscrupulous political leaders to profile and target vulnerable users with bespoke, automatically generated advertisements aimed at subtly skewing their perceptions of controversial issues [130]. Moreover,

anonymisation methods could be used to hide the authors of hateful or extremist online posts, granting them added protection from site moderators and law enforcement [40]. With every NLG system comes the potential for abuse, and as these systems improve in their abilities to produce text that passes as human, this potential for abuse becomes ever more relevant [118].

In this issue, we examine the wide range of tasks and subtasks that NLG can be used towards, highlighting the key ways in which these tasks can be used in the real-world for both good and ill. We focus on three core NLG tasks identified from the academic literature that have the potential for deception and misuse: **stylised text generation**, **conversation**, and **rewriting**.

For this purpose, we sourced a wide range of survey articles dedicated to providing general summaries of NLG methods, and to summarising the existing work conducted toward specific NLG tasks and their subtasks. The articles in this newsletter were taken from those sourced in the previous issue (NL-2022-3). To briefly summarise our approach, we examined a range of NLG-related academic venues, identifying literature reviews and other survey-like articles related to NLG that had been published since 2019. All papers identified were manually inspected and encoded for relevance to one of the core NLG tasks highlighted above, or a subtask thereof. These papers were then used to guide our conception of these NLG tasks, and provided us with the information necessary to summarise the common approaches to these tasks, and their various applications and potential uses in deception. For further reference, a more complete description of the paper collection process can be found in the Editorial section of NL-2022-3.

We hope you enjoy reading this issue. Feedback is always welcome, and should be directed to [ddd-newsletter@kent.ac.uk](mailto:ddd-newsletter@kent.ac.uk).



---

## List of Acronyms

- AI: Artificial Intelligence
- AP: Associated Press
- ATIS: Airline Travel Information System
- BERT: Bidirectional Encoder Representations from Transformers
- BLEU: Bilingual Evaluation Understudy
- BN: Broadcast News
- BNC: British National Corpus
- CCF: China Computer Federation
- CIDEr: Consensus-based Image Description Evaluation
- CSJ: Corpus of Spontaneous Japanese
- CTS: Conversational Telephone Speech
- CV: Computer Vision
- CoQA: Conversational Question Answering
- DA: Data Augmentation
- DNN: Deep Neural Network
- DRV: Dietary Reference Values
- DST: Dialogue State Tracking
- DSTC: Dialog State Tracking Challenges
- DUC: Document Understanding Conferences
- EMD: Earth Mover's Distance
- FAQ: Frequently Asked Question
- Q&A/QA: Question and Answer
- GAN: Generative Adversarial Network
- GDPR: General Data Protection Regulation
- GPT: Generative Pre-trained Transformer
- GYAFC: Grammarly Yahoo Answers Formality Corpus
- HCI: Human-Computer Interaction
- HEQ: Human Equivalence Score
- ICSI: International Computer Science Institute
- IMDb: Internet Movie Database
- IR: Information Retrieval
- ITAC: Informal Text Anonymisation Corpus
- JAPE: Joke Analysis and Production Engine
- L2R: Learning to Rank
- LCSTS: Large Scale Chinese Short Text Summarization Dataset
- LSTM: Long Short-Term Memory
- MATRICS: Multimodal (Task-oriented) Group Discussion
- MCNC: Multiple Choice Narrative Cloze
- METEOR: Metric for Evaluation for Translation with Explicit Ordering
- MIMIC-III: Medical Information Mart for Intensive Care
- ML: Machine Learning
- MTurk: Amazon Mechanical Turk
- NCT: Narrative Cloze Test
- NIST: National Institute of Standards and Technology
- NLG: Natural Language Generation/Generator
- NLP: Natural Language Processing
- NLPCC: Natural Language Processing & Chinese Computing
- NLU: Natural Language Understanding
- NPS: Naval Postgraduate School
- NYTAC: New York Times Annotated Corpus
- PHI: Protected Health Information
- PLM: Pre-trained Language Model
- PMI: Pointwise Mutual Information
- PoS: Part-of-Speech

- 
- PWKP: Parallel Wikipedia
  - RNN: Recurrent Neural Network
  - ROUGE: Recall-Oriented Understudy for Gisted Evaluation
  - SCT: Story Cloze Test
  - seqGAN: Sequence Generative Adversarial Nets
  - SQuAD: The Stanford Question Answering Dataset
  - ST: Speech Translation
  - STP: Simple Temporal Problem
  - QuAC: The Question Answering in Context Dataset
  - TAC: Text Analysis Conference
  - TDT2: Topic Detection and Tracking
  - TST: Text Style Transfer
  - VHA: The Veterans Health Administration
  - WMD: Word Mover's Distance
  - XAI: eXplainable Artificial Intelligence
-

---

# 1. Definitions and Scope

## 1.1. Applying NLG to the Real World

Natural language generation (NLG) focuses on the development of computational models capable of creating ‘convincing’ natural language text (that is, text that contains the fluency and coherency of human written text) from a typically minimal set of inputs [71].

In the previous issue of this newsletter, NL-2022-3, we provided a broad overview of the salient aspects of NLG. In NL-2022-3, we covered the core methods and technologies that have been leveraged to generate natural language, as well as the common datasets that have been used to build NLG models, and the typical evaluation approaches and protocols that are commonly leveraged to assess the performance of these NLG systems.

Given the breadth of this topic, however, we were only able to give a brief mention to the broad array of tasks, subtasks, and subsequent applications within the field of NLG [1, 3, 50, 55, 87]. As the powers of machine learning develop ever further, leveraging vast amounts of online data and new deep-learning approaches, NLG has very quickly found itself advancing from basic prototypes and proofs of concept to becoming an integral part of many of the current technologies used today [71]. In turn, any technology that relies on some form of dynamic interaction with a user, the automatic creation of text, or the processing of speech likely relies on some form of NLG.

In order to fully appreciate the dangers that can be posed by the misuse of NLG-based systems, it is crucial, therefore, to not only have a broad understanding of how NLG is conducted, but also to have a good appreciation of the many tasks that NLG can be applied to. By understanding this, we can then begin to build a more coherent picture as to how these various NLG tasks can be used maliciously.

In this issue, we thus identify the key high-level tasks that constitute NLG and their most common subtasks, examining how these tasks are conducted and evaluated, and how they are applied in the real world. We also consider ways in which each of these tasks and their common applications may allow for deception and misuse.

## 1.2. NLG Tasks and Subtasks

The common high-level NLG tasks considered in this newsletter are defined as follows.

- **Stylised Text Generation:** Stylised text generation refers to the creation of NLG systems aimed at generating *original* texts in a specific, user-designated style [88]. Examples of the desired style of writing are typically used as training or fine-tuning data, from which a given NLG model then attempts to automatically generate new texts that mimic the style, but not the content, of the examples. Style, in turn, is a broad term encompassing specific genres (e.g., fiction, non-fiction [1, 99]), text purposes (e.g., rhetoric, humour [3, 31]), and forms (e.g., academic papers, poems, novels [88]). We provide a complete overview of this task in Section 2.
- **Conversation:** Conversation refers to a series of subtasks in NLG in which the broad aim is the creation of a model that can dynamically generate responses to user inputs, thus facilitating conversation in some form [87]. This encompasses a number of subtasks including **task-oriented conversation**, in which the NLG system attempts to conduct a desired task through conversation with a user [145]; and **Q&A conversation**, in which the NLG system seeks to provide the desired answer to a given, user-specified question [95, 146]. This task is covered in Section 3.
- **Rewriting:** Rather than generating entirely new texts, rewriting tasks instead aim to leverage NLG systems that are able to reinterpret a given input text such that its underlying content and/or semantics are retained, whilst some user-specified attribute of its writing is changed [61]. Common subtasks include style-transfer [61], in which a given model attempts to retain the topic of a given text whilst changing some specified stylistic attribute (e.g., sentiment, toxicity, formality). Moreover, style-transfer can be adapted to preserve user privacy by removing or otherwise obfuscating aspects of authorial style [72, 74], thereby protecting the source of the original text. Other rewriting subtasks include summarisation [42], in which the summarising model attempts to generate a shortened version of a given input whilst retaining its overall content. We cover

---

the Rewriting task in Section 4.

In this issue, we examine each of these high level NLG tasks; identifying their key subtasks, discussing typical approaches to them, noting key relevant datasets, and examining the evaluation measures that are typically conducted. We also discuss the common applications in which these NLG tasks are used, and the manner in which these applications of NLG could lead to problems of deception or other forms of misuse. These issues of deception and misuse will then be carried on to the next issue, in which we will conduct a detailed examination of the risks of deception and misuse in NLG, and the key challenges and open questions that need considering in order to safeguard against these threats.

It is worth noting that whilst these tasks and subtasks are generally implemented and evaluated

distinctly from one another, they are not inherently discreet and can be utilised together in the development of a given real-world system. For instance, a conversation agent could leverage a humour module to provide it with some joke telling capabilities. Whilst stacking or otherwise combining NLG tasks is possible, the academic literature typically takes a task-centric view to NLG, focusing on tackling each task individually.

In order to best represent the current state-of-the-art research, we thus opt to leverage this task-based focus within this newsletter. In turn, the following sections are dedicated to covering each of the high level tasks above, with Section 2 focusing on stylised text generation, Section 3 covering conversation tasks, and Section 4 examining NLG rewriting.

---

## 2. Stylised Text Generation

### 2.1. Introduction

Stylised text generation, otherwise known as style-conditioned text generation, is an NLG task focused on the automatic creation of novel texts that contain a specific, desired writing style [88]. The broad nature of style encompasses a wide variety of subtasks including story generating [1], poem generation [112], humour generation [3], and the various sub-forms of each of these styles (e.g., different genres of story, different forms of poetry).

As the scope for stylised generation is only limited by the vast scope of different writing styles available, in this section we opt to focus on some of the most commonly studied forms of stylised generation in the current literature.

In turn, we examine the Creative Writing subtasks, including story, poem, and lyric generation, the humour generation subtask, which focuses on joke generation, and rhetoric generation, which is a subtask aimed at creating persuasive texts. We also examine a more unconventional form of stylised generation: text augmentation, which is typically applied to boosting supervised machine learning performance by using stylised generation to create new training samples, in the style of the existing training data [8].

### 2.2. Creative Writing

Creative writing is a type of stylised text generation where artificial intelligence (AI) and psychology intersect to teach computers how to mimic human creativity [1]. This includes generating stories, poems, lyrics and prose, either from scratch or based on some existing data, such as an incomplete story, or a painting. In this section, we focus on one of the most commonly studied form of creative writing: story generation.

Story generation is a sub-field of creative writing where the aim is to generate stories [1]. It is essentially the problem of mechanically selecting a sequence of events or actions that meet a set of criteria and integrating these together through prose writing to tell a story. The inputs of a typical story generation method include sets of events, an initial story, or a set of author goals. The model then leverages these inputs to automatically generate a coherent story.

Story generation approaches are classified into three groups:

**Structural Models:** These models employ schemas to generate structured stories by dividing the stories into slots. Then, similar fragments of previously collected and annotated stories are placed into the new story's slots. Structural models include graph-based and grammar-based approaches based on how the annotations are used to attach the fragments. Although structural models are easy-to-implement, they have a couple of drawbacks. Firstly, they only consider syntax of the story despite stories being semantic models in nature. Moreover, they are limited to producing stories that satisfy the predefined story structure. Lastly, they can suffer from the over-generation problem, meaning that they can generate non-story texts [1].

**Planning-based Models:** These models focus on the logical flow between the successive fragments rather than the overall structure of the story. The aim is to generate plots from the fragments, combining the fragments in a structured way to reach a story goal starting from an initial state. Planning-based approaches involve goal-directed, analogy-based and heuristic search approaches. An example output of a planning-based model is shown in Fig. 1.

**Machine Learning (ML) Models:** ML models, especially recurrent neural networks (RNN), are utilised by state-of-the-art story generation methods. They can learn the conditional probability distribution between story events from a story corpus to generate better stories. Other than generating new stories, ML models are also leveraged in other relevant tasks, including script learning and generation, and story completion.

Story generation evaluation is mostly based on assessment of quality rather than creativity. Human evaluation is the most commonly used approach although it is inflexible, time/effort consuming, and subjective.

A typical human evaluation approach is to ask human evaluators to rate the generated stories based on common quality criteria such as consistency, coherence, and interestingness. A discussion of the common quality criteria used in human evaluation can be found in NL-2022-3.

Another common approach is to ask evaluators to edit generated stories to make them more coherent,

---

### The Proud Knight 13

*It was the Spring of 1089, and a knight named Godwin returned to Camelot from elsewhere. A hermit named Bebe told Godwin that Bebe believed that if Godwin jousted then something bad would happen. Godwin was very proud. Because Godwin was very proud, Godwin wanted to impress his king. Godwin jousted. Godwin lost the joust. Godwin hated himself.*

*Moral: Pride goes before a fall.*

---

Figure 1: An example story generated with a planning-based method [1].

and calculate the story quality measure as the distance between the edit and the original story. The more edits that are needed to make the story coherent, the lower the quality of the generated story. Common edits include reordering, adding, deleting and changing events in the generated story.

Regarding machine evaluation, *Narrative Cloze Test (NCT)* is one of the most prominent approaches. It measures the system’s ability to predict a single event removed from a sequence of story events by generating a ranked list of guesses based on seen events. The system is then evaluated using average rank, recall@N (the recall rate within the top  $N$  guesses), and accuracy. *Story Cloze Test (SCT)* is a NCT-based approach designed for supervised learning approaches. It measures the system’s performance according to its ability to choose the correct ending for each story, labelled as “right ending” and “wrong ending”. *Multiple Choice Narrative Cloze (MCNC)* is another NCT-based approach where the system chooses the missing event from five randomly ordered events. Other than task-specific metrics, general NLG evaluation metrics, such as BiLingual Evaluation Understudy (BLEU), Metric for Evaluation of Translation with Explicit ORdering (METEOR), Consensus-based Image Description Evaluation (CIDEr), Recall-Oriented Understudy for Gisting Evaluation (ROUGE), perplexity, and Pointwise Mutual Information (PMI) are also commonly used for evaluating story generation methods. Whilst useful for efficient evaluation (particularly of a large number of outputs), these metrics have a number of drawbacks. Firstly, they require a gold standard corpus to compare the generated text against, which can conflict with the creative nature of story generation. In addition, they typically do not correlate with human judgements, raising questions as to the relevance of their scores [1].

For story generation evaluation, the following datasets are commonly used:

**Andrew Lang fairy tale corpus:** This dataset contains more than 400 stories from Andrew Lang’s Fairy Books (<http://www.mythfolklore.net/andrewlang>).

**ROCStories:** ROCStories is a story generation dataset containing over 98K everyday life stories, and it was constructed for use with SCT. (<https://cs.rochester.edu/nlp/rocstories/>).

**Children’s Book Test:** This dataset was built by Facebook with freely available children’s books taken from Project Gutenberg. The dataset contains over 600K stories. (<https://research.fb.com/downloads/babi/>).

**STORIUM:** STORIUM is a story generation dataset including 6K lengthy stories (125M tokens) with fine-grained natural language annotations, such as character goals and attributes. (<https://storium.cs.umass.edu/>).

Story generation can be used for a variety of different applications, including entertainment, education, and gaming. For instance, stories can be customised for each learner’s educational needs. Furthermore, interactive stories can be used to provide more interesting gaming experiences. From a deception perspective, it is possible that realistic and convincing stories can be generated to deceive and mislead people.

## 2.3. Humour Generation

Broadly speaking, humour generation is a stylised text generation subtask aimed at the creation of jokes [3]. Humour generation finds its roots in the wider field of computational humour, which focuses on the use of computational methods in the analysis and evaluation of humour [12]. This has taken the form of a variety of subtasks including humour recognition [14, 21], the development of systems capable of learning humour preferences of users [131], the automatic evaluation of humour [16],

as well as the development of relevant datasets and corpora [44].

Humour generation leverages the work conducted towards computational humour and seeks to merge it with NLG to develop systems capable of automatic joke telling. Specifically, humour generation typically revolves around one of three formulations [3]: (1) **Q&A joke telling**, in which a question is posed to instigate the joke, and the NLG system attempts to generate a punchline response; (2) **narrative joke generation**, in which longer form story-based jokes are generated; (3) **lexical replacement joke telling**, in which the system attempts to replace the words in an existing text to turn it into some form of pun or joke.

In general, humour generation has received far less study when compared to other forms of stylised text generation (such as creative writing), and there is less of a clear methodology or set of approaches used in the generation of jokes [3]. Moreover, most approaches fail to achieve quality humour generation that is sufficiently convincing or funny to human observers [3]. Thus, humour generation as it now stands is still very much in its infancy, requiring more study before effective humour generation can be achieved.

Currently, there stand two broad approaches to humour generation derived from the broader field of NLG: **Neural text generation**, which uses deep-learning approaches that have become commonplace in text generation more broadly [23]; and **Template-based generation**, which leverages more traditional approaches to text generation in which the system attempts to choose words to fill in missing slots in an existing text template to create new text [30].

Despite the omnipresence of neural methods in broader text generation [23], the application of neural text generation models to humour generation is less common [3]. Generally, most approaches have found neural text generation to be broadly unsuited to generating humour, finding that whilst neural methods are capable of creating jokes with a high level of originality and creativity, they are typically unable to add humour or joke telling within these texts. One of the first examples of the use of neural methods for humour generation was conducted in an undergraduate project by Yang and Sheng [142], who leverage a long short-term memory (LSTM) model to create jokes based on a user-specified topic. The authors trained their model on a large corpus of ap-

proximately 7,500 jokes alongside a corpus of news data to improve the model's knowledge of current affairs. In order to try and provoke a comedic response from the model, the authors attempted to promote incongruity in the generated text by having the model output words based on the probability they were assigned in the output layer, rather than the words with the highest overall probability [140]. Despite these efforts, the model was generally incapable of producing humorous text [3].

One of the only other notable examples of neural-based humour generation comes from Yu et al. [144], who focus specifically on pun generation [3]. To do this, the authors aimed to maximise incongruity, training a neural network using a Seq2Seq model and Wikipedia data. The model is given as input a polysemic word (a word with multiple possible meanings), and two of its definitions [144]. The model is then used to generate two sentences using this word – one for each of the two meanings provided as input. An encoder-decoder model is then trained to generate a single sentence, based on these two sentences, which uses the input word ambiguously to allude to both meanings – thereby creating a pun. Despite the novelty of the approach, however, this was still unable to consistently create humorous content [3]. As an example, the input *square*: 1) *a plane rectangle with four equal sides and four right angles, a four-sided regular polygon*; 2) *someone who does not understand what is going on*, yielded the resulting pun: *Little is known when he goes back to the square of the football club* [3].

Given the current inadequacies of neural methods in humour generation, most approaches have instead focused on the use of template-based generation systems [3]. Typically, a joke template is thus created, alongside a schema which encodes the relationships between the various template variables (the empty slots that the generation model attempts to fill) [3]. In joke generation, this schema typically encodes relationships that provoke incongruity and resolution (key aspects of joke construction). An example of one of the earliest template-based joke generation systems: the Joke Analysis and Production Engine (JAPE), can be found in Fig. 2.

Additionally, template systems rely on some form of knowledge base that provides information about the relationships between the various words that can be selected as candidates for each template variable [3]. Common approaches to constructing these

knowledge bases include ontology based systems, either using manually constructed lexicons mapping each word's relationship to each other, or pre-existing ontologies and databases such as WordNet, ConceptNet, and UniSyn [3]. Quantitative methods have also been proposed, which leverage probabilistic approaches such as the use of N-gram co-occurrence probabilities and vector similarity measures to select template variables that are best suited to humour [3]. Whilst template-based systems are more commonly used for joke construction, and are typically more successful at generating humour, these approaches are far more limited than neural methods due to the constraints posed by the template, and the lexicon from which template variable candidates are selected.

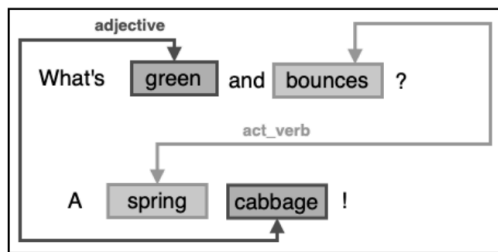


Figure 2: An example of one of the first template-based joke generating systems: JAPE. The system uses question and answer templates, leveraging word homophony to fill in the two template slots in the question and answer [13].

Beyond the generation process itself, efforts have also been made to develop adequate approaches to evaluating humour generating systems. Given the ambiguous and highly subjective nature of humour, however, this too has been hampered with difficulties [3]. Whilst more typical NLG-based metrics for examining the coherence of a given text have some usage in assessing the overall coherence of the generated joke, these are not particularly effective in measuring the humorousness of the output. Typically, studies have relied on human evaluation as the gold-standard, using some form of Likert-scale scoring systems to assess the success of a given output at being humorous [3]. Given the subjective nature of humour, however, this approach is by no means perfect – the fact an evaluator does not find a joke funny does not necessitate that that joke is not funny, this may be a matter of their personal preference and sense of humour.

To mitigate this, other approaches have been proposed. One method is to examine humour frequency. Rather than focusing on measuring the humorousness of individual outputs, this approach asks evaluators to score a set of outputs as being *funny* or *not funny* [3]. The percentage of funny jokes produced by the generator can then be measured. Whilst this method has benefits for measuring the overall performance of a system, it is still limited in its ability to assess the individual outputs. Other suggested approaches seek to leverage a modified version of the Turing test, in which evaluators are tasked with trying to differentiate between human and machine created jokes [3].

Although approaches to humour generation are still limited in their abilities, the broader field of computational humour has meant that a sizeable number of relevant humour datasets exist. These may become more relevant to humour generation in the near future, if state of the art neural methods that are commonplace in broader NLG are more successfully adapted to humour generation. Some examples of these humour datasets include:

**UR-FUNNY:** Created by Hasan et al. [44], UR-FUNNY is a multi-modal dataset containing the video, audio, and transcripts from 1,866 TED talks from 1,741 speakers across more than 400 topics. The transcriptions include markers for audience behaviour, which were used to identify snippets from the talks in which a joke's punchline was told using the audience laughter marker [44]. These 8,257 punchlines, and their preceding context were then extracted and annotated (including audio and video time points) as such. 8,257 negative samples were also extracted, where the last sentence/utterance of the snippet did not end in laughter. An example from the UR-FUNNY dataset can be found in Fig. 3. (<https://github.com/ROC-HCI/UR-FUNNY>).

**One-Liner Dataset:** The One-Liner dataset contains approximately 16,000 one-liner jokes collected using a web-based bootstrapping approach to automatically extract one-liners from a set of webpages [84]. (<https://www.kaggle.com/moradnejad/oneliners-datasets/version/1>).

**Pun of the Day Dataset:** The Pun of the Day dataset was collected by Yang et al. [141], and contains 2,423 puns and 2,403 not-punny sentences. All puns were extracted from the Pun of the Day website, and negative samples were extracted from a variety of news sources, including AP (associated

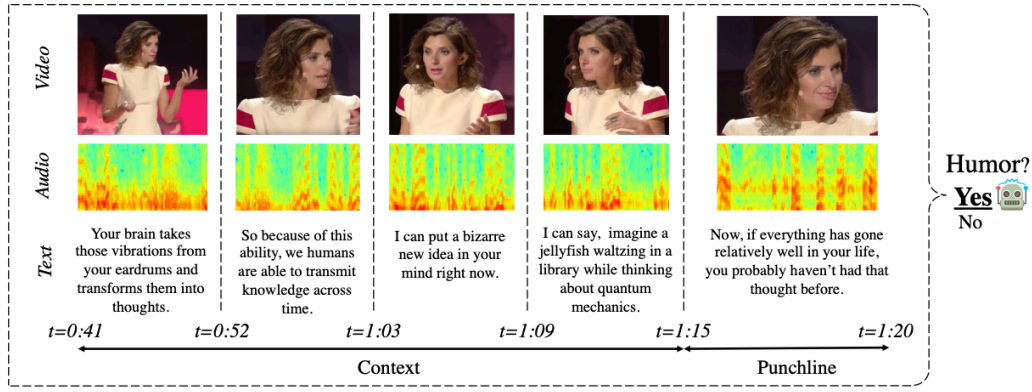


Figure 3: Example from the UR-FUNNY dataset, annotated with context and punchline. [44].

press) News, the New York Times, and Yahoo! Answer.

**Humicroedit:** A more recent dataset, the Humicroedit dataset is constructed of a series of 15,095 English news headlines posted on Reddit, paired with versions of each headline that have been edited, typically using single word replacement, to make them humorous [51]. The editing task, along with the humour evaluation of each edited headline were conducted using Mechanical Turk (MTurk) crowdworkers. Examples of the editing and evaluation tasks are presented in Fig. 4. (<https://www.cs.rochester.edu/u/nhossain/humicroedit.html>).

ORIGINAL:	Eric Trump: Those Who Oppose My Dad Are ' Not Even People '
EDITED:	Eric Trump: Those Who support My Dad Are ' Not Even People '
Substitute:	support

(a) The Headline Editing Task.

Orig:	EU says summit with Turkey provides no answers to concerns
Edit:	EU says gravy with Turkey provides no answers to concerns
	<input type="radio"/> 0 (Not Funny) <input type="radio"/> 1 (Slightly Funny) <input type="radio"/> 2 (Moderately Funny) <input type="radio"/> 3 (Funny)

(b) The Headline Grading Task.

Figure 4: Examples of the editing and evaluation tasks for the Humicroedit dataset [51].

Whilst humour generation is still in its infancy, there are clear applications that warrant its continued study. In turn, studies have also been dedicated to examining applications of computational humour, which have highlighted its potential value in a variety of aspects, including education – particularly in terms of aiding students with complex communication needs [106], and as part of conversational systems (covered in Section 3) as a means of im-

proving user experience [59]. However, the static nature of most humour generation approaches – and their reliance on template-based systems – means that current approaches are less suited to integration with downstream tasks such as education systems and chatbots where they may be of value, but where dynamic generation is needed.

The relatively poor performances of humour generators does mean, however, that their usage in deception or otherwise malicious purposes is currently quite limited. With this being said, the proposed integration of these systems with chatbots and other conversation forms of AI warrants further analysis and consideration, as this could lead to conversation systems far more capable of mimicking human dialogue [78]. This mimicry, in turn, could open the door for more dangerous forms of deception as people become less capable of distinguishing whether they are conversing with a human or a machine [41].

## 2.4. Rhetoric Generation

Rhetoric generation specifically refers to the use of NLG methods to create persuasive communications. [31]. Specifically, this involves the creation of a text output aimed at persuading an individual (or individuals) – the persuadee – to accept a given argument through the use of persuasive messaging embedded in the generated text [31].

Many studies in the field of rhetoric and persuasive generation have focused on the role of various psychosocial aspects that a given NLG system will need to leverage in order to create persuasive texts [31, 57]. In turn, Duerr and Gloor [31] identify four categories that underlie persuasive language generation. These categories are **Benevo-**

---

lence, **Linguistic Appropriacy**, **Logical Argumentation**, and **Trustworthiness**.

**Benevolence:** This category refers to aspects of language aimed at creating value for the persuadee. Identifying the absence or impact of factors that may alter the persuadee's benevolence towards the argument at hand is thus of value to successful persuasion [57]. Examples in this category include example giving, appeal to morality, and the use of social proof/expectations [31].

**Linguistic Appropriacy:** This category involves the profiling of the persuadee's lexical style in order that the persuading NLG system can leverage a style in its generation that achieves the highest degree of congruence between the generated persuading message, and the persuadee. Approaches in other NLG tasks, such as authorial style-transfer [61], have shown some degree of success in learning the latent styles of a target author in order to automatically rewrite texts in that author's style (see Section 4.2). Examples of linguistic appropriacy include the use of emphatics (pronouns like 'myself', 'yourself', etc.), specific word frequencies, and word familiarity [31].

**Logical Argumentation:** This category encompasses the ability of the persuading NLG model to present a text that contains arguments with consistent logic. Some attempts have been made to create systems capable of conducting or recognising logical reasoning and argumentation through the use of first order logic and semantic argumentation graphs [15, 86]. Logical argumentation includes the use of analogies, logical operators (e.g., if, then), and logical consistency [31].

**Trustworthiness:** The final category focuses on the capacity of the persuading system to establish trust with the persuadee. Attempts at psychological profiling as a means of identifying how this trust may be established have thus been suggested, using machine learning models to infer characteristics about individuals from their writing [127, 147]. This is particularly important to persuasive NLG, with previous studies having identified that users often display a lack of trust when dealing with chatbots and other dialogue systems [78]. Examples of this include the use of agreeableness, empathy, and emotionality [31].

Whilst some of the categories above have received a reasonable degree of focus in terms of technical implementations, particularly in regard to the automatic profiling of individuals from their writing, there are fewer works that have focused on creating

NLG systems capable of persuasive generation. Additionally, these works are often spread across a wide range of domains, likely owing to the breadth of applicability of persuasive generation. Currently, there is a lack in unified approaches to creating persuasive NLG systems.

Some examples of attempts towards persuasive NLG include that of Anselma and Mazzei [5], who developed an NLG system to encourage or discourage a user from eating a certain food based on their chosen diet, the foods they had previously eaten that day, and the nutritional value of the dish in question. The authors thus used a simple template-based generation system, which outputs one of five predetermined responses based on the system's evaluation of the input food. The reasoning module used to inform the template selection is built on a simple temporal problem (STP) framework, which is used to calculate whether a food is admissible based on a series of constraints derived from the user's macronutrient dietary recommended values (DRV). The degree to which a food is permitted based on these constraints and the degree to which it meets the user's DRV can then be used by the NLG module to select the appropriate template and template variables (see Fig. 5 for an example of the generated outputs).

Other attempts include the work by Munigala et al. [89], who created a system capable of generating persuasive sentences based on a fashion product specification. This approach used a series of modules to extract keywords from the input product description via Word2Vec embeddings, before leveraging these keywords and a domain specific knowledge base to identify the most relevant domain noun-phrases using these keywords. A neural language model was then used, leveraging as input the keywords and top phrases, alongside other domain-relevant verbs and adjectives and selected persuasive verbs, to generate persuasive summaries about the product [89].

Given the wide range of approaches used, there also exist little in the way of established metrics for evaluating persuasiveness. This lack of standardisation in this area is also likely in part due to the difficulty in measuring persuasion, which can be highly subjective [89]. In this space, common NLG-based metrics are still typically leveraged including BLEU, METEOR, and ROUGE [22], but these only provide a sense of the overall quality of the generated

C	D	Message Template	Translation
I.1	IPO	Questo piatto non va affatto bene, contiene davvero pochissime proteine!	This dish is not good at all, it's too poor in proteins!
I.2	IPO	Ora non puoi mangiare questo piatto perché è poco proteico. Ma se domenica mangi un bel piatto di fagioli allora lunedì potrai mangiarlo.	You cannot have this dish now because it doesn't provide enough proteins, but if you eat a nice dish of beans on Sunday, you can have it on Monday.
C.1	IPO	Va bene mangiare le patatine ma nei prossimi giorni dovrai mangiare più proteine.	It's OK to eat chips but in the next days you'll have to eat more proteins.
C.2	IPO	Questo piatto va bene, è solo un po' scarso di proteine. Nei prossimi giorni anche fagioli però! :)	This dish is OK, but it's a bit poor in proteins. In the next days you'll need beans too! :)
C.3	-	Ottima scelta! Questo piatto è perfetto per la tua dieta :)	Great choice! This dish is perfect for your diet :)

Figure 5: The five templates available in Anselma and Mazzei [5]’s diet management system. Column **C** indicates the classification of the food based on the STP reasoner (e.g., I.1 indicates the food is incompatible with the user’s DRV). Column **D** indicates whether the dish is too rich or too poor in the given macro-nutrient value.

text, not of its persuasive abilities [89]. In addition, other qualitative measures have been proposed, such as catchiness (i.e., is the text catchy or not), and relatedness (i.e., is the text related to the target/argument) domain, but even these fail to offer true measurements of persuasiveness [92]. Other studies have relied on the use of human evaluators participating as persuadees, using questionnaires to measure the degree to which the persuasive systems changed each evaluator’s mind on the target subject [57].

Given the breadth of the subject and its myriad applications, there are also a plethora of datasets that are potentially of use in the creation of persuasive NLG systems. Example of datasets focused on collecting and annotating arguments, debates, and persuasive text include:

**16k Persuasiveness Dataset:** Presented in [43], this dataset contains 16,000 argument pairs for 32 topics, where one argument is *for* the topic, and one *against*. All arguments were sampled from the debate portals *createdebate.com* and *procon.org*. MTurk annotators were then used to choose which argument was more convincing. An example of an argument pair from the dataset can be found in Fig. 6.

**Argument Annotated Essays:** This dataset is construed of 402 persuasive essays written by students on *essaysforum.com* [116]. The essays are written on a range of controversial topics, including “competition or cooperation—which is better?”. These essays were then annotated with their argument components, including any major claims and premises (<https://tudatalib.ulb.tu-darmstadt.de/handle/tudatalib/2422>). Extend-

ing this, Eger et al. [35] translated the dataset into a variety of languages (beyond the original English), including German, French, Spanish, and Chinese. All translations were conducted using Google translate, though the authors also used native German speakers to translate a subset of 402 essays. ([https://github.com/UKPLab/coling2018-xling\\_argument\\_mining](https://github.com/UKPLab/coling2018-xling_argument_mining)).

**Prompt:** Should physical education be mandatory in schools? **Stance:** Yes!

#### Argument 1

physical education should be mandatory cuz 112,000 people have died in the year 2011 so far and it's because of the lack of physical activity and people are becoming obese!!!!

#### Argument 2

YES, because some children don't understand anything except physical education especially rich children of rich parents.

Figure 6: Examples of an argument pair for a given topic in the 16K Persuasiveness dataset [43].

**Debate.org Corpus:** This dataset, curated by Durmus and Cardie [33], consists of 67,315 debates from *debate.org* across 23 different topic categories including politics, religion, health, science, and music. The dataset includes the debate texts themselves, alongside votes provided by users indicating their preference along a variety of dimensions, including the quality of the arguments and the debater conduct. Debates are staged across a series of rounds, with one debater being *for* the claim and the other *against*. For each round, each debater is able to put forth a single argument. An exam-

ple of a debate round is shown in Fig. 7. (<https://www.cs.cornell.edu/~esindurmus/ddo.html>).



Debate Rounds (3)	Comments (39)	Votes (89)
 [+14 words...], the year is better spent with a full time parent. Most children will learn very little at the preschool. [...+39 words]		
 [+11 words...], right school can be an excellent resource for a parent. A child needs to have a place to meet other children. [...+44 words]		

Figure 7: Example of a debate round on *debate.org* [33].

**Persuasion For Good Corpus:** Aimed at facilitating the development of persuasive systems that can be used for social good, this dataset contains 1,017 dialogues between MTurk workers [130]. Pairs of participants were assigned, with one participant the persuader and the other the persuadee. The persuader was then tasked with persuading the persuadee to donate to the charity *Save the Children*. All dialogues were multi-turn, with at least 10 conversational turns required. Alongside the dialogues, a number of annotations are included. These annotations record the various persuasion strategies used, e.g., logical appeal, emotional appeal, and personal stories. (<https://convokit.cornell.edu/documentation/persuasionforgood.html>).

Given the capacity for persuasive language generation to be integrated with a wide variety of other NLG tasks, this allows it to have a considerable range of potential applications. One of the more obvious roles is in advertising, where persuasive NLG systems could be used to generate advertising campaigns at scale, and potentially even to adapt them to individual user profiles [31]. This would offer the capacity for further scaling of existing personalised advertisement campaigns. Indeed, current work such as that of Munigala et al. [89] and their development of persuasive fashion product statements already shows indications of the potential in this area. Beyond business interests, this could also be leveraged for more positive campaigns aimed at achieving social good, such as in charity donation appeals as suggested in [130]. Additional suggestions have also been made for the use of persuasive generation for the advocacy of vaccinations and other medical and personal health areas [6, 31].

Beyond these broad range of applications, however, come the potential for considerable social risk. Whilst most NLG tasks have the capacity for deception and misuse, persuasive generation is particularly problematic in this regard. This is especially troubling as NLG produced text reaches the point of being indistinguishable from human text, which could lead to difficult ethical problems in which users are unwittingly duped by persuasive text generation systems [130]. The capacity of these systems to leverage personal profiling as part of their persuasion is particularly dangerous, as this could allow dishonest parties to identify those most vulnerable for persuasion, and use these powers to mislead them or convince them into doing things against their best interests.

Moreover, issues of a machine’s inability to tell “right” from “wrong” could lead to persuasive NLG systems inadvertently leveraging persuasive acts most would consider socially unacceptable in order to achieve its intended goal [118]. Given the automated nature of many of these systems, it may be difficult, even for well-intentioned designers, to create persuasive systems capable of ethical persuasion within the confines of social acceptability [118]. Additionally, the fact that individuals typically appear to have inherently negative views of systems capable of persuasion raise further ethical and moral questions in regards to their use – especially in cases where the exact behaviours of the NLG systems are hidden from the user [118].

## 2.5. Generative Text Augmentation

Generative text augmentation is a subset of the wider field of study, data augmentation (DA). DA, in sum, is the process of artificially creating new data samples through the modification of existing data samples, or the generation of new samples using existing data samples as training data [8]. This is typically used as a means of creating additional training data samples to help diversify a training set, thereby helping to reduce overfitting when training ML models – especially in cases where the amount of ‘genuine’ training data available is limited [36].

Initially, DA was almost exclusively applied to the field of computer vision (CV), using a variety of image transformation methods like cropping, flipping, and colour jittering existing images in the training dataset to create ‘new’ images to better

train CV models [114]. An example of CV DA using a colour augmentation approach can be found in Fig. 8.

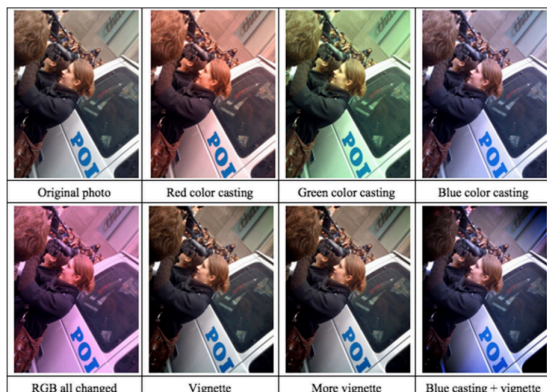


Figure 8: Examples of colour augmentation [133].

With the value of applying DA techniques being well established in the CV community [133], a natural extension to this is the application of DA to NLP problems, using augmentation methods to enlarge text-based datasets. The use of text augmentation poses a more difficult problem, however, as text data (characters, words, phrases, etc.) have far less granularity than image data (pixels, colours, etc.). This makes the text augmentation problem far more challenging than CV based augmentation.

A further challenge of text augmentation is the problem of *label preservation*. When making modifications or generating new data samples for use in training ML models, it is crucial that any modified or generated samples maintain the desired training labels. In the case of text augmentation, it can often be easy through the manipulation of text data to accidentally alter the original sample label. For instance, in the case of sentiment analysis (in which a text is classified as either *positive* or *negative* in sentiment), a DA method that randomly inserts words could accidentally alter the intended sentiment (i.e., by randomly inserting new negative words into a positive text) [8]. Despite these challenges, a wide range of methods have been proposed for text augmentation. A taxonomy of these approaches can be found in Fig. 9.

Typical approaches used for text augmentation are widely varied in nature. More straightforward methods, such as those at the character level, often focus on introducing noise artificially, through approaches such as random character swapping within text data samples [8]. Rule-based methods instead

leverage rules of grammar to make modifications to the text samples in the dataset, such as through the expanding or adding of common contractions, or through the addition of common spelling mistakes [8].

More sophisticated approaches have also been suggested, including the use of pre-trained language models (PLMs) to identify appropriate substitute replacement words in order to modify existing text samples [8]. Other approaches utilise translation tools at the document level, in which the text samples are translated to a given language, and then translated back into the original language. This has been shown to allow for reasonable degrees of paraphrasing and label preservation [8, 36].

With the demonstrated power of new neural techniques for NLG, which have shown increased capacities towards capturing specific styles [88], methods have also been proposed to leverage NLG as a form of generative augmentation [36]. A number of approaches have thus been suggested, leveraging a variety of methods including RNNs, seqGAN, and PLMs like GPT-2 [8, 99]. In turn, these models are typically trained or fine-tuned on existing text data samples, and then tasked with generating new text samples in their ‘style’ [4, 8, 129]. Through this, the generative models are able to artificially construct ‘new’ training texts that are still representative of the class label of the original data samples.

PLMs have shown particular promise in this area, with additional experiments being conducted to try to ensure the preservation of class labels during generation. Given the probabilistic nature of the generation process, this is essential to ensuring that the models retain the stylistic attributes of the original data sufficiently to preserve the desired class label. Examples of these approaches include Wang and Lillis [129], who opt to only use rarer instances to fine-tune their augmentation model; and Anaby-Tavor et al. [4], who use an intermediary classifier to identify text samples generated by their GPT model that retain the desired class label, before using the complete training set (the original data combined with the artificial, generated data) to train a final classifier.

Despite the promising role of generative text augmentation, and text augmentation more broadly, in developing better performing ML models – especially in situations where data is limited, or annotation is expensive – issues currently exist in regard to stan-

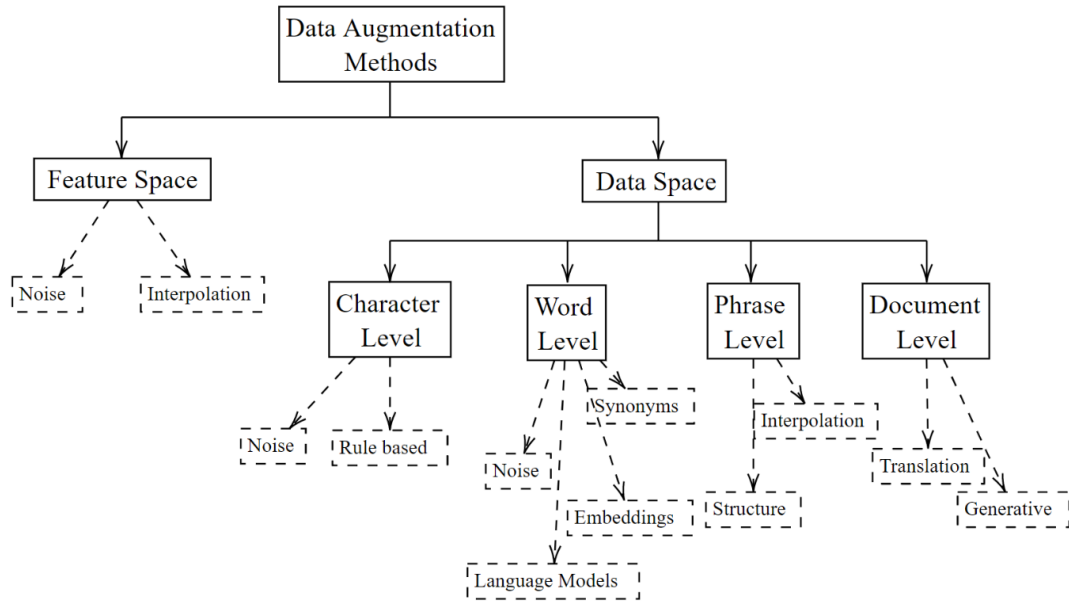


Figure 9: A taxonomy and grouping of data augmentation methods as presented in [8].

dardised approaches for DA evaluation. Currently, most approaches rely solely on the final ML model’s performance as the marker for successful augmentation. If the augmented data boosts model performance, then it is considered successful. Whilst this is certainly the key marker to evaluating text augmentation methods, critics have suggested that other metrics such as resource usage and language variety warrant consideration too [8]. This is especially relevant to NLG-based approaches, which can often require large amounts of computational resources in order to run effectively. Moreover, the lack of language variety in most PLMs, which are predominately trained on English-only datasets, may inhibit their utility in text augmentation of other languages [8, 36].

Given that the core role of generative text aug-

mentation is the boosting of ML model performance, it thus has a wide range of applications. These include a range of other NLG tasks, such as text summarisation and question answering [36]. As NLG tasks, and NLP tasks more broadly, generally require large amounts of training data, DA, and particularly generative text augmentation owing to its heightened abilities to introduce linguistic variety, have the potential to be of great value in improving model performance in a wide range of tasks [8]. Text DA may be of particular value in scenarios where sensitive or private data is needed to train a given model [8]. By using augmentation, ML engineers can reduce the amount of private data they need to gather, which could be of particular benefit to privacy preservation [36].

## 3. Conversation

### 3.1. Introduction

Arguably one of the most commonly studied tasks in NLG, and NLP more broadly, conversation tasks aim to build some form of model capable of automatically generating dynamic natural language messages in response to conversational inputs by a user (or set of users) [87].

Although often referred to by a variety of different terms, including *conversational agents*, *chatbots*, and *dialogue systems*, these terms all encapsulate the task of developing a model capable of simulating conversation in some manner [87].

Given its scope, the conversation task is generally split into three, broadly distinct (though some overlaps exist) subtasks: **task-oriented conversation**, **chat-oriented conversation**, and **Q&A conversation** [146]. Examples of these subtasks can be found in Fig. 10.

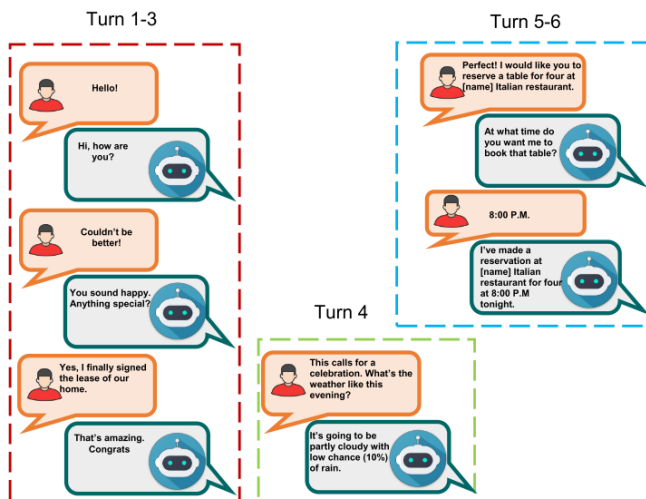


Figure 10: Examples of subtasks in conversational AI [146]. Turns 1–3 depicts a chat-oriented system, turn 4 a Q&A system, and turns 5–6 a task-oriented system.

Task-oriented conversation refers to the subtask of developing conversation systems capable of helping users complete a specified task, or set of tasks [145]. Chat-oriented conversation, instead focuses on more broadly simulating natural conversation, typically across a wider set of domains [145]. Finally, Q&A conversation focuses specifically on the creation of systems capable of answering user questions [2].

In this section, we begin by examining general approaches to the conversation task, looking at the broadly applicable design choices, modules, evaluation procedures, and datasets that are typically used in this space. We also highlight the key general applications of conversation systems, whilst also considering the role that deception could play in these use-cases. From there, we then examine the specific formulations and common approaches taken for task-oriented and Q&A based dialogue systems, alongside any subtask specific evaluation methods, datasets, and applications. We do not include a specific focus on chat-oriented conversations as this has been less well studied, with current approaches generally relying on expanding task-oriented methods of developing conversational systems.

### 3.2. General Approaches

#### 3.2.1. System Goals

Broadly speaking, all conversational models can be centred around a series of general, common goals. The model's ability to achieve these goals (or a subset of them) will, in turn, allow it to more or less successfully simulate conversation. These are:

**User Support:** Probably the most common goal of conversation agents, this goal focuses on the ability of the conversation system to support the user in whatever application it is being used in [87]. This support could come through assisting the user in achieving a set task, or by retrieving information that the user requests [9].

**Information Request:** Moving beyond user support, information request refers specifically to the goal of retrieving specific information that a user desires. Whilst clearly relevant to the Q&A conversation subtasks, some degree of information request ability is still typically needed in the construction of other forms of conversational system. Given the central nature of knowledge sharing and question asking in more natural forms of conversation, most conversational systems require some ability towards identifying and returning specific information. This is typically leveraged through the curation of specific knowledge databases, either more broadly defined or specific to a set of relevant domains [87].

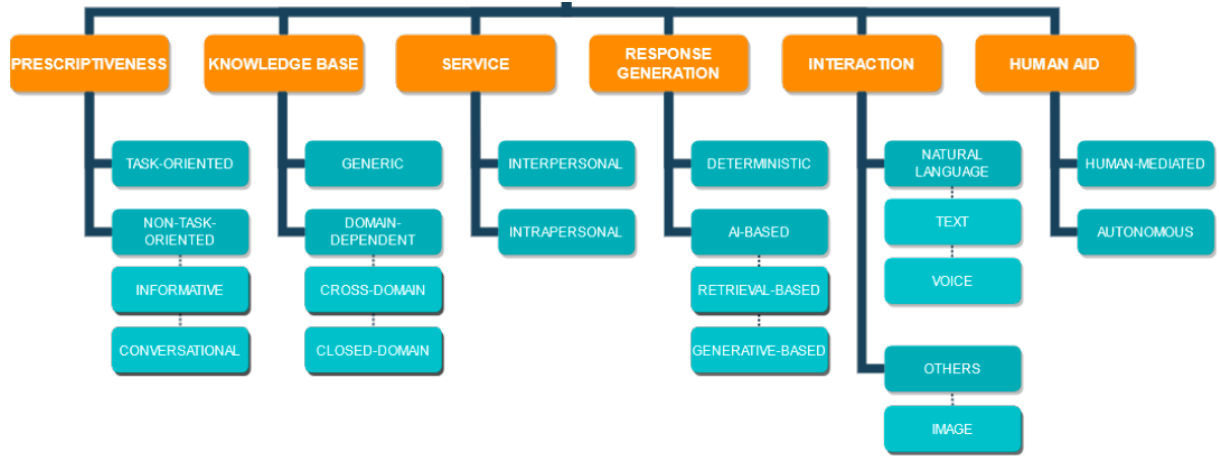


Figure 11: Breakdown of the various design dimensions used when developing conversational agents [87].

**User Engagement:** Also central to most conversation systems, particularly those that facilitate conversation across multiple conversational turns, is the ability of the conversational agent to adequately interest and engage the user it is conversing with [87]. This overlaps with the common human-computer interaction (HCI) aspects of developing chatbots, and is crucial in developing systems that are capable of ‘natural’ conversation [24].

**Information Collection:** A final key goal of most conversational systems is information collection. Whilst this is often specifically formulated towards the collection of key pieces of information about a user (e.g., in a task-oriented systems aimed at providing medical diagnoses [39]), some degree of information collection is essential in most conversational systems [7]. Through information collection, the conversational system is better able to build a profile of the user it is chatting with, allowing for the creation of more personal and relevant responses and utterances [87].

Regardless of the specific subtasks, the development of quality chatbots involves the implementation of many modules [145, 146], all of which need integrating in order to create an effective conversational agent. This moves beyond basic NLG, to consider natural language understanding (NLU) to comprehend user inputs [76], information retrieval (IR) to extract and generate relevant responses [145], intent classification to understand user intent [76], and myriad other tasks besides [87].

### 3.2.2. System Design

In order to design a given conversational system,

considerations have to be given to a series of design dimensions; namely, the prescriptiveness of the system, the knowledge base used, its intended service, the form of response generation used, the mode of interaction, and the degree of human-aid to be leveraged. Each of these design dimensions will require implementation and integration into the overall architecture of the chatbot itself. A breakdown of the most common design dimensions can be found in Fig. 11.

**Prescriptiveness:** This design element refers to the intended subtasks(s) that the chatbot aims to achieve. As described in Section 3.1, these can be divided into task-oriented systems, chat-oriented conversational systems, and informative Q&A systems [87, 146].

**Knowledge Base:** This dimension describes the manner and form by which the knowledge base used by the conversational agent is implemented. The knowledge base is a key element, as it will impact the model’s ability to respond to inputs regarding different topics, and the choice of knowledge base typically reflects the intended use of the conversational system [146]. Generic knowledge bases contain data that is non-specific to a certain topic or set of topics, and aim to encapsulate a wide range of (typically general) knowledge [87]. Domain-specific knowledge bases, instead, contain data relevant to a specific set of subject areas only [87]. Cross-domain knowledge bases contain information from a small set of topics, and closed domain a single subject area [87].

**Service:** This design element defines how the conversational agent interacts with the user. Interpersonal systems are more generic, and do not build

any specific relationship with each user it interacts with [87]. Intrapersonal systems, instead, aim to build context and user dependent profiles that inform its response generation [87].

**Response Generation:** This design choice informs how the system will go about building the desired response to a given conversational input. Deterministic systems use prescribed structures to link the input data to a relevant (often pre-written) response [2, 146]. This includes template approaches, in which the system selects the relevant key terms to fill in blank spaces in a pre-written response. Retrieval-based systems instead leverage machine learning methods to predict relevant responses from a set of pre-defined responses – this is often formulated as a typical machine learning classification problem, where each response is the equivalent of a given class [87]. Generative approaches, instead, leverage NLG – typically through the use of deep learning models – to dynamically generate responses based on the user input [145].

**Interaction:** This element refers to the manner in which the chatbot system interacts with users. Typically, this will be via natural language through text, but voice responses are also possible – as seen in the recent rise in virtual assistants such as Amazon Alexa [87, 122].

**Human Aid:** Whilst generally less considered, this refers to the level of autonomy given to the chatbot. Whilst most research has focused on the development of autonomous conversational agents, it is also possible to develop human-in-the-loop systems, in which human assistance is leveraged to aid the conversational system [87].

In turn, these design considerations will determine the intended task that the conversational agent is developed toward, the manner in which it will be implemented to achieve the task (including the means by which it generates response), and the way in which its interactions will be developed.

### 3.2.3. System Evaluation

Beyond the design of the system itself, it is also essential that consideration is given to the creation of the evaluation approaches that will be used to assess the performance of the conversational system in question. Whilst many of the evaluation procedures used are specific to each subtask (measuring its ability to achieve that subtask specifically) [145, 146], there are also more general evaluative measures that

are often implemented to measure chatbot quality. In turn, these evaluation strategies often mirror the broader approaches taken to evaluate NLG systems (as discussed in NL-2022-3).

A common means of evaluating the abilities of a conversational system is through the quality criterion approach [53]. Through this, the conversational system is scored based on how well it meets a defined quality criterion or set of quality criteria [87]. These are typically focused on the specific text output, measuring how well it meets human standards of natural language. Common criteria include *fluency*: how well the intended language is mimicked [22]; *factuality*: how logically coherent and ‘true’ the response is [22]; and *typicality*: how likely it is you’d expect to see a response of this nature from a human author [53]. These criteria are typically assessed through human evaluators, leveraging some form of scale-based scoring system (e.g., Likert Scales) [22].

In the case of conversational systems, these quality criteria are typically characterised broadly into **functional suitability, efficiency, usability and security** [87].

**Functional Suitability:** This set of criteria encapsulates how correct and appropriate the system’s outputs are. Correctness overlaps most strongly with broad NLG quality criteria, measuring the overall ability of the system to create convincing responses of a reasonable quality. Appropriateness, then measures the degree to which the content of these responses is appropriate, given the user’s input and the desired tasks to which the conversational agent is being used towards. [87]

**Efficiency:** Less relevant to the model’s response generation, this examines how effectively the conversation system manages its resources, and how quickly it can generate responses to user input [87].

**Usability:** This set of quality criteria refers to the ease with which users can interact with the systems. This and efficiency (above) are key HCI considerations when developing conversational systems, though are less related to the model’s performance in terms of generating ‘correct’ responses [24].

**Security:** These quality criteria measure the degree to which the conversational agent is capable of protecting user privacy, and is resistant to malicious interaction. This includes data management considerations such as how personal data is stored, but also encompasses how ‘trustworthy’ the conversational agent is. As these chatbots typically act au-

tonomously, it is crucial that users can trust that the responses they generate will be appropriate and reliable, and not mislead them in some manner.

As some of these quality criteria are more widely conceived than the more general NLG-based criteria, often requiring broader considerations of the overall behaviour and/or performance of the chatbot systems in question, more dynamic evaluation approaches are often favoured. These almost always necessitate a reliance on human evaluation (as opposed to automated metrics), with interviews, broader questionnaires and focus groups all being typical in evaluating a given conversational agent [87].

It is worth noting that quantitative and more automated metrics are still commonly used, but these are more often leveraged in measuring a system's performance in a given subtask (e.g., accuracy in a Q&A setting [146]), or to measure the performance of individual modules within the conversational system's hierarchy [7]. For instance, accuracy might be leveraged to measure the quality of the dialogue state tracking module, a common element in many chatbot systems that monitors the conversation history to help inform the chatbot's response [7].

### 3.2.4. Datasets

Given the scope of conversational AI, including its broad potential for real-world application, and the variety of design dimensions needed to create fully-fledged conversational agents, there exists a wide array of datasets developed for both training and evaluating dialog systems. We present a few of the most popularly used below:

**The Multi-Domain Wizard-of-Oz (MultiWoz) Dataset:** The MultiWoz dataset includes a set of human conversations across a wide range of domains and topics [18]. These domains include hotel, restaurant, police, and hospital. The dataset includes multiple conversational turns, and multi-domain conversations. The dataset also includes annotated dialogue acts indicating the intents and slot-value pairs of a given piece of dialogue. For instance, the dialogue act *INFORM(domain=restaurant,price=cheap)* indicates an intent to inform, with slots for the 'domain' and the 'price', and values of 'restaurant', and 'cheap' [18]. The annotations can then be used to evaluate the performance of the intent classification and slot-value prediction of a given conversational agent. This ability is often essential to

the agent's ability to generate a relevant response or accomplish a given task. (<https://github.com/budzianowski/multiwoz>).

**Ubuntu Dialogue Corpus:** The Ubuntu Dialogue Corpus is an unlabelled dataset containing almost 1 million two person conversations in English extracted from Ubuntu's chat-logs, with more than 7 million utterances being recorded [77]. This corpus is generally leveraged by language model based conversational systems capable of utilising vast amounts of unlabelled text data. (<https://github.com/rkadlec/ubuntu-ranking-dataset-creator>).

**The Naval Postgraduate School (NPS) Chat Corpus:** The NPS Chat Corpus contains more than 10,000 English posts extracted from a variety of online chat services [38]. The dataset has also been annotated with part-of-speech (PoS) tags and dialogue speech acts [38]. (<http://faculty.nps.edu/cmartell/npschat.htm>).

**OpenSubtitles:** The OpenSubtitles dataset is a multilingual dataset containing the subtitles for a wide range of movies and TV programs in more than 60 different languages [73]. The dataset contains more than 300 million pieces of dialogue, and also captures multi-person dialogue – with an average of 2 to 6 speakers per script [81]. (<https://opus.nlpl.eu/OpenSubtitles-v2018.php>).

**British National Corpus (BNC):** The BNC is an multi-party English corpus created by Oxford University Press in the 1980s, making it one of the oldest corpora of its kind [70]. The dataset contains more than 800 pieces of dialogue across a wide range of genres and domains, including transcribed speech, fiction, magazines, and newspapers. The dataset also includes basic PoS tagging [81]. (<http://www.natcorp.ox.ac.uk/>).

### 3.2.5. Applications

Due to their ability to be leveraged in both open-domain contexts and subject-specific areas, and the value of automating dynamic user interactions in a wide range of fields, chatbots have been proposed for a vast array of general applications. An overview of some of the common application domains for conversational agents can be found in Fig. 12.

This has led to the proposal of chatbots of various kinds in specific industries, including tourism (e.g., flight booking, holiday planning) [87, 146], and the restaurant industry (e.g., restaurant finding, restaurant booking) [87, 145]. Chatbots have

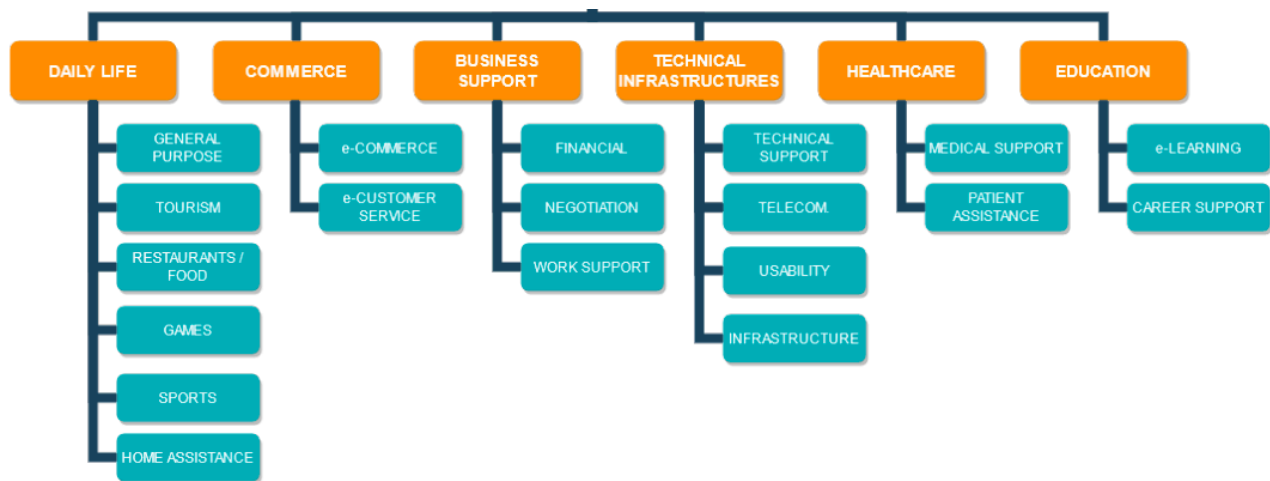


Figure 12: Breakdown of the common application domains relevant to the development of conversational systems [87]

also been more generally proposed as facilitator of customer service and technical support in commerce more broadly [28, 134], with different subtask based conversational systems having the potential to be applied here [145, 146].

Extending this, chatbots have also been proposed as facilitators of workplace support, offering technical support and/or assistance with financial tasks, negotiation and teamwork [87]. By automating these support tasks through the use of a chatbot-based system, businesses can aim to make these interactions more natural for employees, increasing the ease with which they can leverage these forms of assistance [87].

Healthcare is another area of application in which chatbot based systems have been proposed [63, 66]. Beyond previously mentioned applications of general customer support, chatbots have also been suggested as support tools for healthcare professionals, assisting with prescriptions and diagnoses. Chatbots have also been suggested as being of value in aiding patients, with therapy-based conversation systems and conversational symptom-checker systems being proposed [63, 87].

A final, common application domain is that of education [49, 97]. Chatbots have thus been proposed as a means of answering student's FAQs, providing e-tutoring to students, as well as being suggested as a form of automated careers advice counselling and as assistants for accessing and searching institution libraries [87, 97].

Due to its broad capability for application to the real-world, its direct interaction with users, and the

typical need for the chatbot to profile each user in some manner, there are many fears in regard to the potential capabilities toward deception and misuse that may come with developing real-world chatbot systems. Moreover, given that many modules and sub-systems need to be integrated together to develop the conversational agent, this opens up the potential for weaknesses to be exploited throughout a given chatbot's architecture.

Additionally, the dynamic nature of the system, in which it needs to respond to a range of (sometimes hard to predict) user inputs, means that it can be hard to gauge how chatbot systems might react in certain scenarios. This can make it difficult for users to truly place their trust in these systems, whilst also raising difficult questions in regard to the protection of user privacy when interacting with these systems. This is particularly problematic, given the proposed use of chatbots in a variety of sensitive applications, including education and healthcare.

Examples of these potential issues include out of domain problems, in which the chatbot is asked to engage with a user about a subject that is outside its knowledge base [143]. In this case, the chatbot may respond with information that is misleading or otherwise false, which could inadvertently deceive or trick the user in some manner [143]. If we consider the case of a healthcare symptom checker system being asked about an illness it lacks requisite knowledge of, we can see how this could be potentially dangerous. These out of domain issues could also be leveraged by malicious users, who could utilise the lack of knowledge of a chatbot in a given area to

<b>Utterance</b>	I	want	to	listen	to	Hey	Jude	by	The	Beatles
<b>Slot</b>	O	O	O	O	O	B-SONG	I-SONG	O	B-ARTIST	I-ARTIST
<b>Intent</b>	PLAY_SONG									

Figure 13: Output of the slot filling and intent classification tasks conducted by an NLU module, for a given input dialogue. Slot labels are in BIO format, where B indicates the start of a slot span, I the inside of a span, and O a word that does not belong to a slot. The NLU attempts to predict the correct intent label and slot labels for the given utterance [76].

fool users into unknowingly divulging private information [143].

Other dangers include the potential for chatbots systems to produce undesired responses that are derogatory or offensive, when presented with certain inputs [143]. This has been noted as particularly problematic for generative chatbot systems that leverage powerful PLMs. Since these language models leverage vast amounts of (typically web-based) data, distinct biases and tendencies towards extreme language have been noted [11]. It is thus possible that in certain situations, chatbots may provide responses to users that reflect the biases in its training data and/or knowledge base, which could disrupt user trust or potentially even lead to the user being misled or biased in some manner [143].

Additionally, issues exist in regard to user privacy concerns, particularly in regard to divulging personal data [110]. The use of chatbots has led to the potential removal of user agency over their data, with users typically having limited understanding over how personal information divulged to a chatbot might be leveraged, with users typically having limited recourse to delete this data in future [110]. Again, given the proposed use of chatbots in sensitive applications where personal data disclosure is likely necessary, such as in healthcare, this is a pertinent problem [63]. Moreover, given the often generative nature of most chatbots, it can be hard to ensure that the conversational agents do not inadvertently request information from the user that they do not wish to provide – an issue that could be particularly problematic in out of domain scenarios [143]. It is also difficult, given the lack of control over user inputs, to ensure that users do not inadvertently surrender more private data than is intended [109]. Given that some chatbots may integrate user inputs into their training data or knowledge base to improve their performance, this could lead to issues of chatbots unintentionally leaking user data in future conversations [109, 143].

### 3.3. Task-Oriented Conversation Systems

Task-oriented conversation systems are a conversation subtask aimed at creating conversational agents aimed at performing a set of desired tasks on behalf of a user [146]. These systems therefore rely on conversation with the user in order to gain a sense of their desires and preferences towards a given task, before using this knowledge to perform the task in the way the user wishes [146].

At their core, most task-oriented conversation systems are constructed around three separate modules: the Natural Language Understanding (NLU) module, the Policy Learning module, and the Response Generation module [7, 76, 145].

The NLU module is responsible for the initial processing of a given utterance or input text from the user [76]. In turn, the core function of the NLU is to provide intent classification and slot-value prediction for the given piece of dialogue (See Fig. 13 for an example). This, in turn, can be integrated with a dialogue state tracking (DST) module (typically as a joint model), which allows the NLU to leverage previous utterances from the user in its prediction of the overall intent of the current dialogue presented to the conversational systems. The range of possible intents and slot-values are generally encoded in a pre-defined ontology which is typically closed-domain – using domains relevant to the desired task the conversational system is used to conduct.

From this, the Policy Learning module is used to decide what action is to be taken by the conversational system, in order to guide the user to a specific task, leveraging the predicted dialogue states achieved by the NLU module [58, 76].

Finally, the generation module is used to create a response based on the predicted dialogue state and the chosen policy. This is generally implemented either through a template-based approach, in which the system identifies an appropriate response template and template values, or (less frequently) using

a probabilistic generative method (such as through the use of powerful generative language models like GPT-2) [76].

In order to train and evaluate task-oriented conversational systems, many of the datasets mentioned in Section 3.2 have been popularly used. This is especially true of the MultiWoz dataset, as it comes annotated with dialogue acts that are particularly useful for evaluating the NLU components of the conversational system [18]. As a large proportion of the work on task-oriented systems has been dedicated to its NLU aspects, there also exist a range of datasets aimed at specifically evaluating task-oriented intent classification and slot-value prediction:

#### **Airline Travel Information System (ATIS)**

**dataset:** The ATIS is a highly popular, single-turn dataset used for benchmarking NLU modules [46, 76]. This dataset is composed of approximately 5,000 utterances focused on airline travel, e.g., queries focused on flight searching. Each utterance is annotated with the appropriate slot and intent labels needed to evaluate the accuracy of a given NLU system in this domain. (<https://www.kaggle.com/hassanamin/atis-airlinetravelinformationsystem>).

**MEDIA:** The MEDIA dataset is focused around hotel booking scenarios, containing a series of simulated conversations in French between a tourist and a hotel concierge [83]. The dataset contains approximately 18,000 utterances, and is labelled with slots (intent labels are not provided) including the number of people, the date, and the hotel facility [76]. (<https://catalogue.elra.info/en-us/repository/browse/ELRA-S0272/>).

**Snips Dataset:** This dataset was curated by crowdsourcing spoken conversation using the Snips voice platform [27]. The conversational data was generated by using Amazon Mechanical Turks (MTurk) and other crowdsourcing platforms to create artificial utterances based on a provided set of intents and slots. A variety of domains were used for the intent-slot sets provided, including restaurant bookings, movie schedule requests, and song playing requests. (<https://github.com/sonos/nlu-benchmark>).

**Facebook Multilingual Dataset:** This dataset attempts to address the lack of non-English data by curating a multilingual dataset of task-specific dialogues [111]. This dataset thus contains 57,000 dialogues, including 8,600 Spanish utterances and 5,000 Thai utterances across a vari-

ety of domains including weather, alarm, and reminder. All dialogues are annotated with both intents and slots. (<https://ai.facebook.com/blog/democratizing-conversational-ai-systems-through-new-data-sets-and-research/>).

#### **Dialog State Tracking Challenges (DSTC)**

**2 & 3:** DSTC 2 & 3 are two English datasets aimed at specifically evaluating DSTs [47, 48]. These datasets are composed of human-machine conversation related to both restaurants and tourism. DSTC2 contains over 3,000 dialogues (1,612 training dialogues, 506 development dialogues, and 1,117 testing dialogues), and is labelled with the turn-level semantics of each dialogue, which the given DST attempts to predict via the current dialogue and dialogue history to that point. DSTC3, on the other hand, is aimed at assessing the ability of DSTs to predict slot-values in unseen, out of domain situations, and thus only contains 2,265 dialogues for testing. (<https://github.com/matthen/dstc>).

As is the case with most datasets in NLP, English is the predominant language in dataset curation for task-oriented conversation systems. However, there have been attempts in recent years to develop datasets for additional languages. Beyond the MEDIA and Facebook datasets mentioned above, the popular ATIS dataset has been translated into a variety of different languages, including Hindi, Turkish, and Indonesian [120, 123], as well as, through the MultiAtis++ dataset [135], into Spanish, Portuguese, German, French, Chinese, and Japanese. The SNIPS dataset has also been adapted to Italian by Bellomaria et al. [10].

Due to its task-oriented nature, these type of conversational system have been popularly leveraged in a range of applications, especially as a means of improving booking systems for a range of businesses. In turn, task-oriented conversational systems have been proposed as a means of aiding in hotel booking, the reservation of restaurants, and for holiday bookings. Given the very specific domain of these tasks, it has been more straightforward to implement systems capable of extracting the relevant intents and slot-values from these forms of conversation in order to extract enough information from a customer to facilitate the booking. Other applications have also been proposed, including the usage of task-oriented systems in online shopping, in which the system can be leveraged as a means of finding relevant products, as well as an interface in which purchases can

be made [140].

Task-oriented nature is also inherently related to the development of virtual assistants (such as Alexa, Siri, etc.), as these systems are inherently tasks oriented by design. Through this, a variety of task-oriented applications are encompassed, including (beyond the above) tasks such as alarm and reminder setting, and weather checking.

Given these applications, task-oriented systems are thus vulnerable to the wide range of deceptive or otherwise malicious actions discussed in Section 3.2. Out of domain issues pose a particular problem here, as task-oriented systems are typically narrowly focused on one, or a small set of domain-specific tasks. This could lead to the task-oriented system attempting to perform actions it is not capable of doing, or of performing tasks in an undesired manner. Privacy issues are also problematic in this case, especially when the applications of the systems involve the parsing of sensitive user data. Applications such as holiday booking or product purchases bring risks as this may involve the task-oriented conversation system having to handle user payment data. Thus, issues of adequate storage of this data by the conversational systems also exist, as poor security here could result in the leaking of this private data.

### 3.4. Q&A Systems

Q&A systems are centred around the development of dialogue systems capable of answering user questions. These systems can be formulated either as closed-domain, in which the questions are related to one, or a small subset, of topics; or open-domain, in which questions can be drawn from a wide range of topics [145, 146]. To some extent, Q&A systems can be viewed as a highly specific version of task-oriented systems (indeed, question answering is often part of a task-oriented system's overall task), but the approaches taken to developing Q&A systems typically differ from those of task oriented system, often relying more heavily on comprehension and information retrieval rather than response generation [2].

Beyond the classes of open-domain and closed-domain, Q&A systems can also be either single-turn, or multi-turn [146]. Single-turn systems are required to provide an answer based on a single dialogue input from a user, whereas multi-turn (also called conversational Q&A) systems allow for multiple dialogue turns and questions – where the Q&A system will

typically have to rely, in some part, on the dialogue history to answer any questions posed.

For single-turn Q&A, the most common approach is to try and identify the answer to the question within some form of knowledge base. For closed-domain tasks, this will typically be a dataset, or datasets, specifically relevant to these specific topics, whereas in an open-domain setting, these will typically be large knowledge bases such as Wikipedia. The common modules for a single-turn Q&A system are the **Question Analysis** module, the **Document Retrieval** module, and the **Answer Extraction Module** [148]. Alongside the question, the system will often also be provided with some form of context to aid in answering the question [148]. Figure 14 provides an overview of a common single-turn Q&A architecture.

The Question Analysis module is generally tasked with predicting the correct question class in order to inform the structure of the answer to be generated, and formulating the optimum query by which the knowledge base's documents can be searched for the answer. For question classification, common question classes include factoid questions, e.g., how, why, where; confirmation questions, e.g., yes, no; and listing questions, i.e., listing items in a given order [146].

The Document Retrieval module is then tasked with using the query created by the Question Analysis module to identify the document or subset of documents within the knowledge base that are most likely to contain the answer to the question [148]. This is typically treated as a form of information retrieval task (IR). Common approaches include Boolean models, in which Boolean expressions are used to match questions to documents and vector models, in which vector representations and similarity metrics are leveraged to find the most relevant documents [148]. Language model based approaches have also been suggested, which rank documents based on the probability of the model generating the question given the document – which is used to fine-tune the model [148].

After the most relevant document has been identified, paragraph ranking is then conducted to find the paragraph within the candidate document that is most likely to contain the answer to the question. One of the most popular approaches to achieve this is through learning to rank (L2R) [56], which utilises supervised learning approaches to identify the opti-

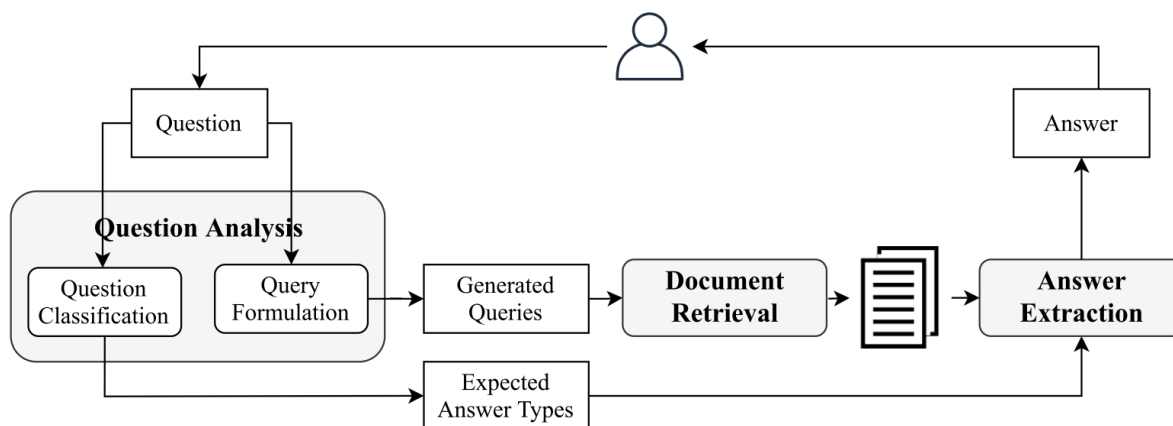


Figure 14: Typical architecture of a single-turn Q&A system [148].

mum ranking of candidate paragraphs relative to the question.

The Answer Extraction module is then leveraged to identify the answer, within the candidate paragraph, to the user’s question. This is often approached via span prediction, in which the text span in the candidate paragraph that contains the answer is identified [56]. Many approaches exist to do this, though the use of PLMs such as BERT have become especially popular in recent years [56]. NLG-based approaches for selecting the answer using text generation rather than span selection have also been proposed, those these are less common and are typically unable to achieve the performances recorded by span-selection based models [148].

For multi-turn systems, approaches still typically conceive of the Q&A problem as one of answer span selection from a knowledge base of documents [148]. The key challenge presented then is one of dialogue history modelling in order to best leverage any historical information in past utterances that can be used to identify the most relevant answer-span [146]. To do this, a history selection module is typically incorporated to select the past utterances most relevant to answering the question. This is generally either done through a  $k$ -turn based approach, in which the past  $k$  number of utterances are used, or dynamic selection where a trained model is used to dynamically assess the contributions of each utterance to answering the question [146]. Some form of history encoding and modelling is then used to integrate the historical dialogue with the current question. Commonly used approaches include conventional word embeddings, and contextualised word embeddings

leveraging PLMs [146].

Evaluation of Q&A based system is generally conducted via the use of some form of common accuracy metric such as F1-score to measure the degree to which the Q&A systems correctly identifies the answer to the question [148]. For generative answering approaches, common automated metrics such as BLEU have also been leveraged to measure the quality of the answer generated [148]. A further metric that has been proposed is Human equivalence score (HEQ), which measures a system’s performance relative to that of the average human [26].

The metrics used are typically tied into the specific dataset used to train and evaluate the Q&A system. Some of the most commonly used datasets in this subtask are:

**SQuAD:** The Stanford Question Answering Dataset (SQuAD) is a dataset containing more than 100,000 questions posed by crowd-workers on a specific set of Wikipedia articles [101]. This takes a span-finding approach to Q&A, where the annotated answer to each question is a specific segment of text in one of the Wikipedia articles. Alongside each question, the relevant passage is also provided to the system as context. Extensions to this dataset include SQuAD 2.0, which expands the dataset to include unanswerable questions (which the system must identify as such) [100], and SQuAD<sub>open</sub> [25], which expands SQuAD to leverage the entirety of Wikipedia [148]. (<https://rajpurkar.github.io/SQuAD-explorer/>).

**QuAC:** The Question Answering in Context dataset (QuAC) also leveraged Wikipedia articles for span-prediction based Q&A systems [26]. Un-

Characteristics	CoQA	QuAC
<b>Data Source</b>	Passages collected from 7 diverse domains e.g. children stories from MCTest, news articles from CNN, Wikipedia articles etc.	Sections from Wikipedia articles filtered in the “people” category associated with subcategories like culture, animal, geography, etc.
<b>Conversational Setup</b>	Questioner-answerer setting where both have access to the entire context.	Teacher-Student setting where the teacher has access to the full context for answering, while the student has only the title and summary of the article
<b>Requires External Knowledge?</b>	Yes	No
<b>Question Type</b>	Factoid	Open-ended, highly contextual
<b>Answer Type</b>	Free-form with an extractive rationale.	Extractive span which can be yes/no or ‘No Answer’.
<b>Dialog Acts</b>	No	Yes
<b>Max Turns per Dialog</b>	15	11
<b>Unanswerable Questions</b>	Yes	Yes
<b>Total Number of Questions</b>	126K	100K
<b>Total Number of Dialogs</b>	8K	14K

Figure 15: Comparison between the CoQA and QuAC datasets [146].

like SQuAD, however, QuAC examines multi-turn Q&A using a student teacher scenario, in which the student keeps asking for further clarifications on a specific topic. This dataset contains more than 100,000 conversation turns across 14,000 conversations, where each turn consists of a question and an answer, with each question reliant on knowledge of past dialogue to answer correctly. The dataset also leans more heavily on open-ended style questions (why, how). Evaluation is done through macro-averaged F1-score for word overlap of the system’s answer and the correct span is used for evaluation, alongside HEQ (as described above). (<https://quac.ai/>).

**CoQA:** Similar to QuAC, the Conversational Question Answering (CoQA) dataset examines conversational question answering using span prediction of Wikipedia articles [103]. CoQA contains more than 127,000 conversation turns across 8,000 conversations, where each turn consists of a factoid question and an answer, and each question is reliant on knowledge of past dialogue to be answered correctly. Unlike QuAC, CoQA includes unanswerable ques-

tions. As seen in QuAC, macro-average F1 score is again used for evaluation on CoQA. A comparison between QuAC and CoQA can be found in Fig. 15. (<https://stanfordnlp.github.io/coqa/>).

**SearchQA:** SearchQA utilises a span-prediction approach to Q&A using question and answer pairs selected from the television programme *Jeopardy!* [32]. Alongside each question answer pair is a series of relevant Google snippets within which the answer span can be identified. The system is thus challenged with identifying the correct answer span within the Google snippets. SearchQA contains more than 140,000 questions, with each question coupled with an average of 50 snippets. As with CoQA and QuAC, F1-score is most typically used to evaluate Q&A systems on SearchQA. (<https://github.com/nyu-dl/dl4ir-searchQA>).

These example datasets present just a few of the vast number of Q&A training and benchmark datasets that have been released in recent years [148]. This is due to the problem that Q&A datasets are often quickly “solved” [107]. This solving, however, is typically not a sign of the quality of

---

the Q&A system so much as its ability to leverage annotation artefacts and lexical cues to artificially achieve high performances. Moreover, the current approach to training and evaluating Q&A means that systems are unable to generalise away from the dataset in which they are trained. Even in cases where the questions being asked are of the same domain as the training set, most systems are unable to generalise to different datasets. This severely limits the current applicability of most Q&A approaches.

These weaknesses also mean that many Q&A systems are vulnerable to adversarial attack. In [128], for example, the authors identify the presence of universal adversarial triggers that can be used to prompt specific offensive outputs from a Q&A based system. Examining Q&A systems trained on the

SQuAD dataset, the authors proposed a method of identifying specific triggers that could be used to prompt the answer “to kill American people” for 72% of all “why” questions posed. Other studies have noticed similar weaknesses, in which adversarial question framing can provide incorrect responses and a loss in performance from the Q&A model [60, 108]. These vulnerabilities emphasise the current inability of Q&A systems to genuinely perform language comprehension, finding that they all too often rely on superficial cues to generate answers [60]. This, in turn, raises questions in regard to their current suitability to real-world applications, such as in education systems, FAQs and other customer-service roles [87, 107].

---

## 4. Rewriting

### 4.1. Introduction

We define rewriting tasks as tasks in which a given text is rewritten such that its original meaning is preserved, but some additional attributes are changed. This section covers two main tasks involving text rewriting: text style transfer and summarisation.

### 4.2. Text Style Transfer

Text Style Transfer (TST) is an NLG task which aims to rewrite a text according to a specific style “property” or “attribute”. Therefore, the TST task “aims to change the stylistic properties of any given text while preserving its style-independent content” [55]. This is a data-driven approach [61] focused on changing the syntax aspect of a text with respect to a given style attribute, whilst keeping the semantics intact, thereby fulfilling linguistic variation [55].

#### 4.2.1. TST Applications

TST has applications in four main domains. These are:

**Writing assistance:** TST functionalities can be incorporated into tools to help users tailor or improve a written text according to a specified attribute. For example, it may be applied to the text of business emails or reports to make them look more professional, therefore improving formality (involving attributes informal  $\rightarrow$  formal) or politeness (involving attributes impolite  $\rightarrow$  polite).

**Persuasive communication:** TST is a powerful mechanism for:

- Better engaging with an intended audience such as consumers (e.g., the style of a generic marketing text can be personalised according to a user profile), readers (e.g., image captions or headlines, which can be adapted to become more attractive according to attributes like humour, romance and clickbait [62]) and laymen or experts (e.g., a layman style can be used to make an expert text more readable while an expert style can be used to make a layman text appear more accurate and professional [55]).
- Reaching a target community effectively using gender (e.g., male  $\longleftrightarrow$  female), political ideology (e.g., Democrats  $\longleftrightarrow$  Republicans), or through shared views and interests (e.g., emotions, topics).
- Improving accessibility such as by using text simplification (complex  $\rightarrow$  simple).
- Adapting to user preferences and circumstances such as changing the sentiment conveyed in a text (negative  $\longleftrightarrow$  positive). Another example of this are chatbots which can adapt their script style according to the type of interaction (e.g., a casual style for suggesting products to customers and a formal style for handling customers complaints [55]).

**Authorship imitation:** TST can be used to mimic the style of a given author by changing the authorial style of the input text to that of a target author, e.g., through adapting the “word choice, syntactic structures, figurative language, and sentence arrangement” [55] of the input text. This is, therefore, a more holistic and complex application of style transfer.

**Re-styling for Social Good:** TST can also be used to improve text, such as social media posts and tweets, in terms of biasness (biased  $\rightarrow$  neutral) and toxicity (offensive  $\rightarrow$  non-offensive).

#### 4.2.2. TST System Design

Traditionally, TST has been achieved using *parallel data* where there are matching text pairs for different styles,, e.g., informal and formal, positive and negative, modern English and Shakespearean English. In this case, sequence-to-sequence models (as covered in NL-2022-3) and variations of those are often applied [61]. Despite the existence of a number of such datasets for specific use cases, there are several TST cases where there is a lack of parallel data available. Therefore, non-parallel datasets and the methods based on explicit or implicit *disentanglement* of style and content [55], i.e., methods that “disentangle text into its content and [style] attribute in the latent space” [61], have started to emerge and remain an active research area.

Explicit disentanglement follows three types of action [61]: (1) encode the given text  $t$  with the



Figure 16: An example of explicit disentanglement for Text Style Transfer (Delete-Retrieve-Generate framework) retrieved from Hu et al. [55]. The source style attribute is negative sentiment and the target style attribute is positive sentiment.

source style attribute  $a$  in a latent representation; (2) manipulate the latent representation to remove  $a$ ; and (3) decode into text  $t'$  with target style attribute  $a'$ . An example technique in this category is illustrated in Fig. 16 where a) corresponds to action (1) and b) includes actions (2) and (3).

Implicit disentanglement involves [55]: (1) learning the latent representations of content  $c$  and of style attribute  $a$  for the given text  $t$ ; and (2) combining content  $c$  with the latent representation of the target style attribute  $a'$  to generate text  $t'$ . An example technique in this category is Adversarial Learning where two models are typically used – one adversarial network model for (1) and a style-embedding model for (2).

Another stream of solutions for TST is to not rely on disentanglement but rather build a pseudo-parallel dataset and then apply more traditional methods such as sequence-to-sequence models. There are two main approaches to achieve this [61]: retrieval-based, using existing datasets to identify pairs of sentences semantically similar using a metric; and generation-based, using an iterative process to generate the dataset.

Another recent unsupervised trend for TST solutions, moving away from disentanglement, is to use Transformer-based models (as covered in NL-2022-3) such as in the work by Dai et al. [29].

#### 4.2.3. TST Evaluation

In terms of automated evaluation, the quality of TST solutions are typically determined by the following three main criteria [55].

- **Transferred style strength:**

The *Style Transfer Accuracy* metric measures “whether each sample generated by the model

conforms to the target [style] attribute” [61] and is calculated as

$$\frac{\# \text{ test samples correctly classified}}{\# \text{ all test samples}}$$

An alternative metric is the *Earth Mover’s Distance* (EMD), which measures the minimum cost to turn the style distribution of the given text  $t$  into the generated text  $t'$ . It can also be regarded as a measure of intensity of the style transfer [85].

- **Semantic preservation:**

The goal here is to measure the similarity of content between the given text  $t$  and the generated text  $t'$ . *BLEU* is the most widely used metric for TST solutions using parallel datasets, although others such as ROUGE and METEOR are also used [61]. On the other hand, *sBLEU* and *Cosine Similarity* are the mostly used for non-parallel settings [55]. Please refer to NL-2022-3 for an overview of metrics used for NLG systems.

The *Part-of-Speech Distance* (POS) is a metric specifically used to evaluate unsupervised TST solutions [121]. It relies on two vectors of tags, where tags can be nouns or verbs depending on the relevance of those for the style attributes in question – one vector from the given text  $t$  and the other vector from the generated text  $t'$ . The distance between those two vectors can then be calculated based on cosine similarity. The POS score will be high if  $t'$  does not contain semantically similar tags of  $t$ .

The *Word Mover’s Distance* (WMD), proposed by Kusner et al. [67], measures the “dissimilarity between two text documents as the min-

imum amount of distance that the embedded words of one document need to travel to reach the embedded words of another document” [67], in this case between the given text  $t$  and the generated text  $t'$ . The metric is illustrated in Fig. 17, where words in bold from both documents are embedded into the word2vec space. Then, the cumulative distance for all words from  $t$  in the space to travel in order to match the words from  $t'$  is calculated. WMD has been shown to correlate better with human evaluation than BLEU [85].

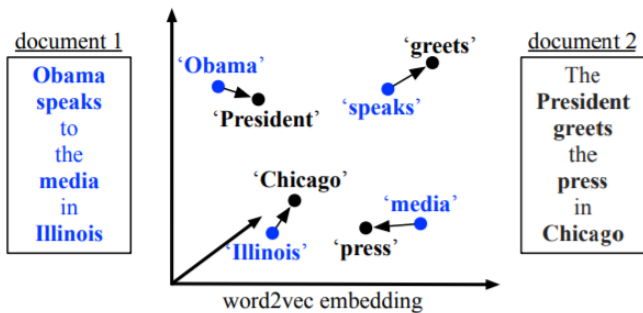


Figure 17: Illustration of the Word Mover’s Distance (WMD) metric by [67]. Words in bold from both documents are embedded into the Word2Vec space; the metric calculates the travel distance for words from document 1 to match words of document 2 in the space.

- **Text fluency (or naturalness):**

The Perplexity score, calculated using a PLM for all style attributes on the training data, is the most commonly used metric to evaluate fluency of TST output, although its correlation with human evaluation remains a subject of debate [61]. The lower the perplexity score of a generated sentence (from text  $t'$ ), the more aligned it is with the training dataset (from text  $t$ ) [55].

Human evaluation, although hampered by issues such as subjectivity [55] and irreproducibility [61], is also often used to provide insights about transferred style strength, semantic preservation, and text fluency given pairs of sentences from text  $t$  and from the generated text  $t'$ . According to Hu et al. [55], best practices indicate the need to “use 100 outputs for each style transfer direction (e.g., 100 outputs for formal  $\rightarrow$  informal, and 100 outputs for informal  $\rightarrow$  formal), and two human annotators for each task”.

#### 4.2.4. TST Subtasks and Datasets

All subtasks and datasets listed in this section relate to transferring style between two attributes.

*Sentiment Subtask:* This is a very popular TST subtask; it involves the styles “positive” and “negative”. There are 3 main datasets related to it mentioned below. These datasets were all pre-processed to exclude neutral reviews.

**Yelp:** This is a non-parallel dataset containing positive and negative real-world restaurant reviews [113]. All reviews have up to 10 sentences, with 250K negative sentences and 350K positive sentences in total. (<https://www.yelp.com/dataset>).

**Amazon:** This is a non-parallel dataset containing positive (approximately 278K) and negative (approximately 279K) real-world Amazon users’ reviews of products [45]. Please note that these numbers come from Hu et al. [55]. (<https://s3.amazonaws.com/amazon-reviews-pds/readme.html>).

**IMDb:** This is a non-parallel dataset containing real-world movie reviews collected from the Internet Movie Database [79]. It contains 50K reviews with no more than 30 reviews for the same movie, and an equal split between positive and negative reviews. (<https://ai.stanford.edu/~amaas/data/sentiment/>).

*Formality Subtask:* This TST subtask involves the contrasting styles “formal” and “informal”. Figure 18 shows an example of an informal sentence (input) and four corresponding formal sentences (Ref-0 to Ref-3) with variations of length and punctuation. Hu et al. [55] pointed out that formality is more complex than sentiment style transfer because it is more subjective – different individuals may have very different perceptions of what is formal. The Grammarly Yahoo Answers Formality Corpus (GYAFC) is the most popular dataset used for this subtask.

	Informal Formality $\longleftrightarrow$ Formal Formality
input	He loves you, too, girl...Time will tell.
Ref-0	He loves you as well, but only time can tell what will happen.
Ref-1	He loves you too, lady...time will tell.
Ref-2	He loves you, as well. Time will tell.
Ref-3	He loves you too and time will tell.

Figure 18: An illustration of formality text style transfer with an informal sentence and 4 corresponding formal versions of it (Ref-0 to Ref-3) [55].

**GYAFC:** The GYAFC dataset contains parallel data, i.e., 110K pairs of formal-informal sen-

tences [102]. It was built with sentences collected from a Yahoo Answers corpus for Entertainment & Music and Family & Relationship, and processed to remove sentences that were too long or too short. The sentences deemed informal were then translated into formal ones via crowd-sourcing. (<https://github.com/raosudha89/GY AFC-corporus>).

*Politeness Subtask:* This TST subtask relates to the text styles “polite” and “impolite”. It is interesting to notice that written expressions of politeness are culture-dependent even for the same language, and are also affected by social structures [80].

**Politeness:** This is a non-parallel dataset containing 1.39 million sentences that were automatically labelled [80]. The sentences were collected from the Enron corpus, and therefore reflect politeness in the context of email exchanges in an American corporation. (<https://github.com/tag-and-generate/politeness-dataset>).

*Authorship Subtask:* This TST subtask aims to target a particular writing style from a linguistic point-of-view, therefore, it is more artistic compared to all other subtasks [61]. It is a form of “paraphrasing” [139]. There are two datasets that are commonly used for research in this space.

**Shakespeare:** This parallel dataset contains sentences in Shakespearean English style and corresponding sentences in modern English style [139], totalling around 21K pairs. (<https://github.com/cocoxu/Shakespeare>).

**Bible:** This parallel dataset contains over 1.5 million pairs of sentences aligned by verse numbers from “the eight publicly available versions” of the Bible [20]. (<https://github.com/keithcarlson/StyleTransferBibleData>).

*Simplicity Subtask:* This TST subtask relates to text styles “complex” and “simple”. It aims to simplify text to make it more accessible for laymen, e.g., removing lexical or syntactic complexity. The target audience of text simplification also involves people with low literacy levels, such as children and non-native speakers, and people suffering from different kinds of reading comprehension, e.g., autism, aphasia, dyslexia [1]. Some techniques useful to achieve simplification are [149]: splitting (e.g., transforming long sentences into shorter ones), dropping (e.g., making sentences more concise and sharper), re-

ordering (e.g., rearranging sentences to make them easier to understanding), and substitution (e.g., replacing jargon and difficult terms with simpler synonyms).

For evaluating text simplification, common automated metrics such as BLEU are used. However, there also exist various task-specific metrics proposed for simplification evaluation. For instance, FKBLEU combines a paraphrase generation metric, iBLEU, with a readability metric, Flesch-Kincaid Index, to measure how adequate and readable a simplified text is [138]. Another metric is SARI, which measures the goodness of words that are added, deleted and kept by the simplification system [138]. Furthermore, readability indices are also commonly used to estimate how difficult a simplified text is to read [1, 115].

Several datasets are available for TST simplification where they aim to foster research and development for more effective communication between healthcare professionals and healthcare consumers, i.e., for medical text simplification.

**PWKP:** The Parallel Wikipedia (PWKP) dataset contains over 108K pairs of complex sentences from English Wikipedia and simple sentences from the Simple English Wikipedia, which targets children and adults learning English [149]. (<https://huggingface.co/datasets/turk>).

**van den Bercken et al.’s EXPERT datasets:** These are three separate datasets proposed by van den Bercken et al. [125]: EXPERT-FULLY, EXPERT-PARTIAL and AUTOMATED-FULLY. The datasets contain pairs of complex medical sentences and corresponding simple sentences, also drawing from Wikipedia and Simple Wikipedia. The EXPERT-FULLY dataset has 2,267 fully aligned medical sentences, the EXPERT-PARTIAL dataset has 3,148 partially aligned sentences, and the AUTOMATED-FULLY dataset has 3,797 fully aligned medical sentences. (<https://github.com/myTomorrows-research/public/tree/main/WWW2019>).

**MIMIC-III:** This is a non-parallel dataset containing real-world clinical sentences written in professional (medical) style and in consumer (layman patient) style. It has 443K sentences in professional language, and 73K sentences in consumer language [132]. (<https://physionet.org/content/mimiciii/1.4/>).

**BenchLS:** BenchLS is a combination of two

non-parallel lexical simplification datasets, LexMTurk and LSeval, containing 929 instances in total. Each instance consists of a sentence, a target complex word, and several (7.37 on average) candidate substitutions ranked according to their simplicity [93]. ([https://zenodo.org/record/2552393#.YYkZLdmP0\\_8](https://zenodo.org/record/2552393#.YYkZLdmP0_8)).

**NNSeval:** NNSeval is a non-parallel dataset that covers complex words for non-native speakers. All sentences in the dataset were taken from Wikipedia, LexMTurk and LSeval, and 400 non-native speakers identified the complex target words. The resulting dataset contains 239 instances [94]. ([https://zenodo.org/record/2552381#.YYkZRNnPO\\_8](https://zenodo.org/record/2552381#.YYkZRNnPO_8)).

**SS Corpus:** This is a parallel corpus containing 492,993 aligned sentences extracted by pairing *Simple English Wikipedia* with *English Wikipedia* [64]. (<https://github.com/tmu-nlp/sscorpus>).

**MSD:** MSD is a parallel dataset motivated by the example shown in Fig. 19 where expert sentences (upper sentences) are simplified into layman sentences (lower sentences) [19]. It contains 130K expert-style sentences and 114K layman-style sentences. (<https://srhthu.github.io/expertise-style-transfer/#disclaimer>).

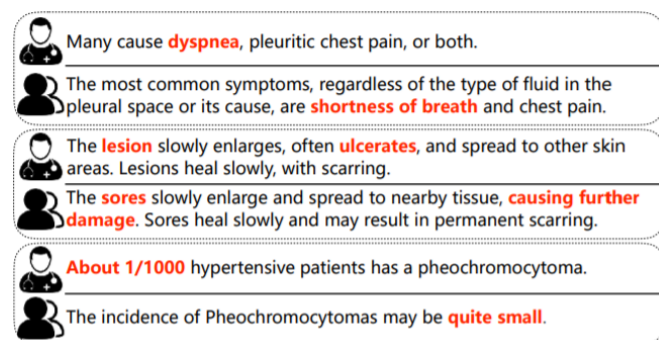


Figure 19: An illustration of simplification text style transfer with 3 pairs of sentences in expert style (upper sentences) and in layman style (lower sentences), adapted from figure by Cao et al. [19].

**Newsela:** This is a parallel corpus containing 1,130 news articles with four simplified versions each. The simplified versions were written by professional editors at *Newsela*, a company that produces reading materials for children [137]. (<https://newsela.com/data/>).

*Gender Subtask:* This TST subtask draws from

socio-linguistics research showing that gender is associated with language choices [90]. Assuming gender as a biological binary attribute of an individual, it typically involves text styles “female” and “male”.

**Yelp Gender:** This non-parallel dataset has been made available by Prabhumoye et al. [98]. It builds on a previous private dataset by Reddy and Knight [104] compiled from the Yelp Dataset Challenge 2016 and annotated with “male” and “female” labels for reviews in gender-neutral domains. The reviews were divided into more than 2.5 million sentences, where “only sentences that are strongly indicative of a gender” were kept [98]. ([http://tts.speech.cs.cmu.edu/style\\_models/gender\\_data.tar](http://tts.speech.cs.cmu.edu/style_models/gender_data.tar))

**RtGender:** This is a non-parallel dataset containing over 2.5 million sentences compiled from responses to online posts or videos where the gender of the (source) author and the gender of the responder were clear. The source-responder sentences have been extracted from comments on Facebook (US politicians and public figures), TED talks, Fitocracy (fitness), and Reddit [126]. (<https://nlp.stanford.edu/robvoigt/rtgender/>).

*Toxicity Subtask:* This TST subtask relates to text styles “offensive” and “non-offensive”. Changing style from the former to the latter contrasts with the approach of simply filtering and removing such content online, especially in relation to social media posts.

**Nogueira dos Santos et al.’s Twitter and Reddit datasets:** These are non-parallel datasets by Nogueira dos Santos et al. [91]. The authors used “sentences/tweets with size between 2 and 15 words and removed repeated entries”. Their Twitter dataset contains just under 2 million entries, while their Reddit dataset contains over 7 million entries.

#### 4.2.5. Multiple-attribute TST

Research has also examined the potential of transfer text style across multiple attributes. The idea is illustrated by Lample et al. [69] in Fig. 20 where, given an input sentence and a style (among a range of supported styles), the system generates a sentence in the opposite matching style. Such a system has to be trained using a dataset for each style and, in terms of evaluation of the transferred style strength, it has to be provided for each pair of styles separately [68].

Relaxed ↔ Annoyed	
Relaxed	Sitting by the Christmas tree and watching Star Wars after cooking dinner. What a nice night 🍷🌲💎
Annoyed	Sitting by the computer and watching The Voice for the second time tonight. What a horrible way to start the weekend 😡😡😡
Annoyed	Getting a speeding ticket 50 feet in front of work is not how I wanted to start this month 😞
Relaxed	Getting a haircut followed by a cold foot massage in the morning is how I wanted to start this month 😊
Male ↔ Female	
Male	Gotta say that beard makes you look like a Viking...
Female	Gotta say that hair makes you look like a Mermaid...
Female	Awww he's so gorgeous 😍 can't wait for a cuddle. Well done 🍷xxx
Male	Bro he's so f***ing dope can't wait for a cuddle. Well done bro
Age 18-24 ↔ 65+	
18-24	You cheated on me but now I know nothing about loyalty 😏 ok
65+	You cheated on America but now I know nothing about patriotism. So ok.
65+	Ah! Sweet photo of the sisters. So happy to see them together today .
18-24	Ah 😊 Thankyou 🍷 #sisters 🍷 happy to see them together today

Figure 20: Illustration of a solution which supports multiple-attribute style transfer by Lample et al. [69]; the first line of each box contains a given sentence and style, and the following line contains the given target style and the automatically generated sentence.

Multiple-attribute style transfer, as opposed to binary style transfer, is regarded as the research direction for future research and development since it has the potential to “explore richer and more dynamic tasks” Hu et al. [55].

#### 4.2.6. Anonymisation

According to the GDPR legislation, “anonymisation is the complete and irreversible process of removing personal identifiers, both direct and indirect, that may lead to an individual being identified” [72]. It has been approached from a privacy preserving perspective and from a NLP perspective. The former focuses on risks of disclosure of personal identifiers by an adversary, while the latter focuses on linguistic patterns helpful to infer personal identifiers, such as gender, age, race, geographical location and affiliations (called quasi-identifiers [72]), that can be used, e.g., for user profiling [90] or discrimination [104].

In the domain of NLP, anonymisation is a TST task, akin to style transfer (non-anonymised → anonymised), that has been predominately approached in two ways: *de-identification* and *obfuscation*. De-identification aims to detect and remove personal identifiers, and is often applied to the medical domain, i.e. to Protected Health Information (PHI). Whereas obfuscation aims to detect and rewrite text to “reduce the leakage of sensitive information” [136] while retaining text semantics and fluency. To illustrate obfuscation, let’s consider the

following examples from [136]:

Original: *I am a software engineer with 18 years of working experience.*  
 Rewritten: *I am a software engineer with more than 10 years of working experience.*  
 Original: *I went with my girlfriend and another couple.*  
 Rewritten: *I went with my friend and another couple.*

The obfuscation subtask can use the following tailored evaluation metrics [136]: average entropy (applied to all predictions of the classifier; higher entropy means less sensitive data leakage); predicted accuracy and modified accuracy (rate of accepted sentence modification).

Datasets used for anonymisation have all gone through a “surrogate process” to replace real-world personal identifiers with fictitious but realistic ones.

**2010 i2b2 NLP challenge corpus:** This dataset contains over 800 diabetic patient medical records [124], manually “annotated for an extended set of PHI categories” [72].

**VHA:** The Veterans Health Administration dataset contains 800 clinical notes also manually annotated with PHI categories such as “Social Security Numbers, Patient Names, and Dates” [37]. ([Download.](#))

**2016 CEGS N-GRID:** This dataset contains 1K psychiatric intake records and “more than 34,000

PHI phrases, with an average of 34 PHI phrases per record” [119]. It is annotated with PHI such as doctor name, patient name and hospital location.

**ITAC:** The Informal Text Anonymisation Corpus contains 2.5K personal emails with 31,926 personal identifiers including [82] with direct identifiers (e.g., name of individuals) and quasi identifiers (e.g., name of organisations) [72].

Efforts similar to the ITAC resulted in a number of datasets in languages other than English. One such dataset is:

**Code Alltag 2.0:** This dataset contains over 240K emails in German [34]. (<https://github.com/codealltag>)

### 4.3. Summarisation

Text summarisation is a rewriting subtask, which aims to generate a short, coherent version of a text that contains the main ideas, topics, and/or concepts of the original text [52]. Considering the vast amounts of digital textual data available, text summarisation can decrease the time needed for text processing in multiple contexts.

In the literature, two major approaches for text summarisation exist: **extractive summarisation** and **abstractive summarisation**. The former tries to identify the most relevant utterances or sentences from input text which describe the main theme. The latter aims to generate a fluent and concise summary, paraphrasing the intent of the input text in a shortened form [52].

Human evaluation is commonly used for evaluating text summarisation methods. Human evaluation of text summarisation mostly focuses on the ability of the generated text to capture the key contents of the input text. The focus of human evaluators is therefore generally directed towards measuring the informativeness, coverage, focus, and relevance of the summary text. Furthermore, more general measures of text fluency, readability, coherence and repetition are also often considered to evaluate linguistic quality. Human evaluation methods typically leverage common scoring methods, including Likert-type scales, rank-based annotations, and pairwise comparisons. Other proposed methods are Best-worst scaling (BWS), which is a specific type of ranking-oriented evaluation that requires annotators to specify only the first and last rank, and question-answering (QA) [117].

Some of the popular datasets used for evaluating text summarisation are as follows:

**Document Understanding Conferences (DUC):** DUC is a series of conferences run by the National Institute of Standards and Technology (NIST) focusing on the area of text summarisation. From 2001 to 2007, text summarisation datasets have been provided in the scope of the conferences. The datasets contain news articles from AQUAINT, TIPSTER and TREC corpora, and are available upon request. (<https://duc.nist.gov/data.html>).

**Text Analysis Conference (TAC):** TAC is another conference series run by NIST between 2008-2011. Each conference had a summarisation track where a relevant dataset has been distributed. The datasets contain news articles, and are available upon request. (<https://tac.nist.gov/data>).

**New York Times Annotated Corpus (NYTAC):** The NYTAC dataset contains over 1.8M news articles published by New York Times between 1987-2007, as well as 650K article summaries. The articles were manually summarised by library scientists. (<https://catalog.ldc.upenn.edu/LDC2008T19>)

**Large Scale Chinese Short Text Summarization Dataset (LCSTS):** The LCSTS dataset is a Chinese text summarisation dataset constructed from a Chinese micro-blogging website, Sina Weibo [54]. It contains over 2M real Chinese short texts with short summaries provided by the author of each text. (<http://icrc.hitsz.edu.cn/Article/show/139.html>).

**CCF Conference on Natural Language Processing & Chinese Computing (NLPCC):** NLPCC is a series of conferences on NLP organised annually in China. The conferences had a track for text summarisation in 2015, 2017 and 2018, where a text summarisation dataset was distributed for each year. (<http://tcci.ccf.org.cn/>).

Text summarisation has various applications. For example, medical conversation summarisation aims to summarise conversations between doctors, nurses, and patients about the proposed diagnoses and treatments so that patients can review them later without having to deal with a full record or transcript [75]. Text summarisation can also be used to increase the efficiency in processing long documents such as scientific papers [42] and long speeches [105].

---

#### 4.3.1. Speech Summarisation

Beyond more conventional text summarisation, which aims to summarise text inputs, speech summarisation has also been the focus of a considerable degree of study.

Speech summarisation aims to identify the most important content within human speech and then generate a condensed form of text suitable for the needs of a given task [105]. Unlike standard text summarisation approaches which mostly generate text from another text, speech summarisation methods take audio data as the input, and utilise speech recognition to process it. Nonetheless, speech summarisation approaches are similar to the standard text summarisation approaches, using the major approaches of **extractive summarisation** and **abstractive summarisation** discussed earlier.

For evaluating speech summarisation, qualitative and quantitative metrics have been used. Human evaluation typically leverages similar quality criteria to that of standard text summarisation, including readability, coherence, usefulness and completeness. Beyond human evaluation, automated evaluation using common NLG metrics, such as ROUGE (and its variants, ROUGE-1, ROUGE-2, ROUGE-3, ROUGE-L, ROUGE-SU4 and ROUGE-W), are common. Other performance measures, including precision, recall, F-measure, word accuracy and Pyramid, are also often used to measure how well a generated summary's content matches the content of the reference summary.

For evaluation of speech summarisation, the following datasets have been used:

**AMI:** The AMI meeting corpus is a multi-modal data set consisting of 100 hours of meeting recordings in English. Furthermore, it contains manually produced orthographic transcriptions for each individual speaker, as well as a wide range of other annotations, including extractive and abstractive summaries. (<https://groups.inf.ed.ac.uk/ami/corpus/>).

**International Computer Science Institute (ICSI) Meeting Corpus:** The ICSI Meeting Corpus is an English audio dataset consisting of approximately 70 hours of meeting recordings. It also contains orthographic transcriptions, and manual annotations of dialog acts and speech quality. (<https://groups.inf.ed.ac.uk/ami/icsi/>)

**Multimodal (Task-oriented) gRoup dIs-CuSsion (MATRICS):** The MATRICS corpus

contains discussions in English among four native Japanese speakers on three different topics. It involves 9 hours of group meeting recordings consisting of 29 dialogues by 10 conversation groups. (<https://github.com/IUI-Lab/MATRICS-Corpus>).

**Corpus of Spontaneous Japanese (CSJ):** CSJ is a dataset containing Japanese spoken language data and information for use in linguistic research. The dataset consists of 958 hours of lectures and task-oriented dialogues in Japanese. (<https://ccd.ninjal.ac.jp/cs/en/>).

**Topic Detection and Tracking (TDT2):** TDT2 contains news data collected daily from nine news sources in two languages (American English and Mandarin Chinese) over a period of six months. (<https://catalog.ldc.upenn.edu/LDC2001T57>).

**RT-03 MDE:** This dataset contains English Conversational Telephone Speech (CTS) and Broadcast News (BN) transcripts and annotations covering 40 hours of CTS and 20 hours of BN data. Annotations include fillers (e.g., “um”, “err”), discourse markers (e.g., “you know”), and semantic units (e.g., statements, questions). (<https://catalog.ldc.upenn.edu/LDC2004T12>).

**Mandarin Speech Data Across Taiwan Broadcast News (MATBN):** The MATBN Mandarin Chinese broadcast news corpus contains a total of 198 hours of broadcast news from the Public Television Service Foundation (Taiwan) with corresponding transcripts and annotations. (<http://slam.iis.sinica.edu.tw/corpus/MATBN-corpus.htm>).

**Switchboard-1:** This dataset contains a collection of approximately 2,400 two-sided telephone conversations among 543 speakers (302 male, 241 female) from the US, containing 260 hours of speech. (<https://catalog.ldc.upenn.edu/LDC97S62>).

**Fisher:** Fisher is another telephone speech dataset, containing 2,000 hours of conversational speech data in English. It has been built for the DARPA Effective, Affordable Reusable Speech-to-text (EARS) program.

**Spoken Language Data base (BEA):** BEA is a Hungarian spontaneous speech dataset that consists of 250 hours of speech data.

Speech summarisation has several bespoke applications, including improved efficiency and cost reduction in telephone contact centres (e.g., by identifying call topics, automatic user satisfaction evaluation, and efficiency monitoring of agents), more

---

efficient progress tracking in project meetings, the facilitation of learning using online courses, digital scribes, and conversational agents. Speech summarisation can be used for deception, with the main idea of a speech being taken out of context in its textual summary to mislead readers.

---

---

## References

- [1] Arwa I. Alhussain and Aqil M. Azmi. 2021. Automatic Story Generation: A Survey of Approaches. *Comput. Surveys* 54, 5, Article 103 (2021), 38 pages. <https://doi.org/10.1145/3453156>
- [2] Reem Alqifari. 2019. Question Answering Systems Approaches and Challenges. In *Proceedings of the Student Research Workshop Associated with RANLP 2019*. INCOMA, 69–75. [https://doi.org/10.26615/issn.2603-2821.2019\\_011](https://doi.org/10.26615/issn.2603-2821.2019_011)
- [3] Miriam Amin and Manuel Burghardt. 2020. A Survey on Approaches to Computational Humor Generation. In *Proceedings of the 4th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*. ICCL, 29–41. <https://aclanthology.org/2020.latechclfl-1.4>
- [4] Ateret Anaby-Tavor, Boaz Carmeli, Esther Goldbraich, Amir Kantor, George Kour, Segev Shlomov, Naama Tepper, and Naama Zwerdling. 2020. Do Not Have Enough Data? Deep Learning to the Rescue!. In *Proceedings of the 2020 AAAI Conference on Artificial Intelligence (AAAI'20)*, Vol. 34. AAAI, 7383–7390. <https://doi.org/10.1609/aaai.v34i05.6233>
- [5] Luca Anselma and Alessandro Mazzei. 2015. Towards Diet Management with Automatic Reasoning and Persuasive Natural Language Generation. In *Portuguese Conference on Artificial Intelligence*. Springer, 79–90. [https://doi.org/10.1007/978-3-319-23485-4\\_8](https://doi.org/10.1007/978-3-319-23485-4_8)
- [6] Luca Anselma and Alessandro Mazzei. 2020. Building a Persuasive Virtual Dietitian. In *Informatics*, Vol. 7. MDPI, Article 27, 26 pages. <https://doi.org/10.3390/informatics7030027>
- [7] Vevake Balaraman, Seyedmostafa Sheikhalishahi, and Bernardo Magnini. 2021. Recent Neural Methods on Dialogue State Tracking for Task-Oriented Dialogue Systems: A Survey. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL'21)*. Association for Computational Linguistics, 239–251. <https://aclanthology.org/2021.sigdial-1.25>
- [8] Markus Bayer, Marc-André Kaufhold, and Christian Reuter. 2021. A Survey on Data Augmentation for Text Classification. arXiv:2107.03158 [cs.CL] <https://arxiv.org/abs/2107.03158>
- [9] Samuel Bell, Clara Wood, and Advait Sarkar. 2019. Perceptions of Chatbots in Therapy. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, Article LBW1712, 6 pages. <https://doi.org/10.1145/3290607.3313072>
- [10] Valentina Bellomaria, Giuseppe Castellucci, Andrea Favalli, and Raniero Romagnoli. 2019. Almayave-SLU: A New Dataset for SLU in Italian. arXiv:1907.07526 [cs.CL] <https://arxiv.org/abs/1907.07526>
- [11] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT'21)*. ACM, 610–623. <https://doi.org/10.1145/3442188.3445922>
- [12] Kim Binsted, Anton Nijholt, Oliviero Stock, Carlo Strapparava, G. Ritchie, R. Manurung, H. Pain, Annalu Waller, and D. O'Mara. 2006. Computational Humor. *IEEE Intelligent Systems* 21, 2 (2006), 59–69. <https://doi.org/10.1109/MIS.2006.22>
- [13] Kim Binsted and Graeme Ritchie. 1994. An Implemented Model of Punning Riddles. In *Proceedings of the 12th National Conference on Artificial Intelligence (AAAI'94)*. AAAI, 633–638. <https://www.aaai.org/Papers/AAAI/1994/AAAI94-096.pdf>

- 
- [14] Vladislav Blinov, Valeria Bolotova-Baranova, and Pavel Braslavski. 2019. Large Dataset and Language Model Fun-Tuning for Humor Recognition. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL'19)*. Association for Computational Linguistics, 4027–4032. <https://doi.org/10.18653/v1/P19-1394>
- [15] Karsten Block, Simon Trumm, Premtim Sahitaj, Stefan Ollinger, and Ralph Bergmann. 2019. Clustering of Argument Graphs using Semantic Similarity Measures. In *KI 2019: Advances in Artificial Intelligence – 42nd German Conference on AI, Kassel, Germany, September 23–26, 2019, Proceedings*. Springer, 101–114. [https://doi.org/10.1007/978-3-030-30179-8\\_8](https://doi.org/10.1007/978-3-030-30179-8_8)
- [16] Pavel Braslavski, Vladislav Blinov, Valeria Bolotova, and Katya Pertsova. 2018. How to Evaluate Humorous Response Generation, Seriously?. In *Proceedings of the 2018 Conference on Human Information Interaction & Retrieval (CHIIR'18)*. ACM, 225–228. <https://doi.org/10.1145/3176349.3176879>
- [17] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. arXiv:2005.14165 [cs.CL] <https://arxiv.org/abs/2005.14165>
- [18] Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Inigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 2018. MultiWOZ—A Large-Scale Multi-Domain Wizard-of-Oz Dataset for Task-Oriented Dialogue Modelling. arXiv:1810.00278 [cs.CL] <https://arxiv.org/abs/1810.00278>
- [19] Yixin Cao, Ruihao Shui, Liangming Pan, Min-Yen Kan, Zhiyuan Liu, and Tat-Seng Chua. 2020. Expertise Style Transfer: A New Task Towards Better Communication between Experts and Laymen. arXiv:2005.00701 [cs.CL] <https://arxiv.org/abs/2005.00701>
- [20] Keith Carlson, Allen Riddell, and Daniel Rockmore. 2018. Evaluating Prose Style Transfer with the Bible. *Royal Society Open Science* 5, 19, Article 171920 (2018), 13 pages. <https://doi.org/10.1098/rsos.171920>
- [21] Andrew Cattle and Xiaojuan Ma. 2018. Recognizing Humour Using Word Associations and Humour Anchor Extraction. In *Proceedings of the 27th International Conference on Computational Linguistics (COLING'18)*. Association for Computational Linguistics, 1849–1858. <https://aclanthology.org/C18-1157/>
- [22] Asli Celikyilmaz, Elizabeth Clark, and Jianfeng Gao. 2020. Evaluation of Text Generation: A Survey. arXiv:2006.14799 [cs.CL] <https://arxiv.org/abs/2006.14799>
- [23] Khyathi Raghavi Chandu and Alan W. Black. 2020. Positioning Yourself in the Maze of Neural Text Generation: A Task-Agnostic Survey. arXiv:2010.07279 [cs.CL] <https://arxiv.org/abs/2010.07279>
- [24] Ana Paula Chaves and Marco Aurélio Gerosa. 2019. How Should My Chatbot Interact? A Survey on Human-Chatbot Interaction Design. arXiv:1904.02743 [cs.CL] <https://arxiv.org/abs/1904.02743>
- [25] Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading Wikipedia to Answer Open-Domain Questions. arXiv:1704.00051 [cs.CL] <https://arxiv.org/abs/1704.00051>
-

- 
- [26] Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen-tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. QuAC: Question Answering in Context. arXiv:1808.07036 [cs.CL] <https://arxiv.org/abs/1808.07036>
- [27] Alice Coucke, Alaa Saade, Adrien Ball, Théodore Bluche, Alexandre Caulier, David Leroy, Clément Doumouro, Thibault Gisselbrecht, Francesco Caltagirone, Thibaut Lavril, Maël Primet, and Joseph Dureau. 2018. Snips Voice Platform: An Embedded Spoken Language Understanding System for Private-by-Design Voice Interfaces. arXiv:1805.10190 [cs.CL] <https://arxiv.org/abs/1805.10190>
- [28] Lei Cui, Shaohan Huang, Furu Wei, Chuanqi Tan, Chaoqun Duan, and Ming Zhou. 2017. SuperAgent: A Customer Service Chatbot For E-Commerce Websites. In *Proceedings of the 2017 Annual Meeting of the Association for Computational Linguistics (ACL'17) – System Demonstrations*. Association for Computational Linguistics, 97–102. <https://aclanthology.org/P17-4017/>
- [29] Ning Dai, Jianze Liang, Xipeng Qiu, and Xuanjing Huang. 2019. Style Transformer: Unpaired Text Style Transfer without Disentangled Latent Representation. arXiv:1905.05621 [cs.CL] <https://arxiv.org/abs/1905.05621>
- [30] Kees van Deemter, Mariët Theune, and Emiel Krahmer. 2005. Real Versus Template-Based Natural Language Generation: A False Opposition? *Computational Linguistics* 31, 1 (2005), 15–24. <https://doi.org/10.1162/0891201053630291>
- [31] Sebastian Duerr and Peter A. Gloor. 2021. Persuasive Natural Language Generation – A Literature Review. arXiv:2101.05786 [cs.CL] <https://arxiv.org/abs/2101.05786>
- [32] Matthew Dunn, Levent Sagun, Mike Higgins, V. Ugur Guney, Volkan Cirik, and Kyunghyun Cho. 2017. SearchQA: A New Q&A Dataset Augmented with Context From a Search Engine. arXiv:1704.05179 [cs.CL] <https://arxiv.org/abs/1704.05179>
- [33] Esin Durmus and Claire Cardie. 2019. Exploring the Role of Prior Beliefs For Argument Persuasion. arXiv:1906.11301 [cs.CL] <https://arxiv.org/abs/1906.11301>
- [34] Elisabeth Eder, Ulrike Krieg-Holz, and Udo Hahn. 2020. CodE Alltag 2.0 — A Pseudonymized German-Language Email Corpus. In *Proceedings of the 12th Language Resources and Evaluation Conference (LREC'20)*. European Language Resources Association, 4466–4477. <https://aclanthology.org/2020.lrec-1.550>
- [35] Steffen Eger, Johannes Daxenberger, Christian Stab, and Iryna Gurevych. 2018. Cross-Lingual Argumentation Mining: Machine Translation (and a Bit of Projection) is All You Need! arXiv:1807.08998 [cs.CL] <https://arxiv.org/abs/1807.08998>
- [36] Steven Y. Feng, Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura, and Eduard Hovy. 2021. A Survey of Data Augmentation Approaches for NLP. arXiv:2105.03075 [cs.CL] <https://arxiv.org/abs/2105.03075>
- [37] Oscar Ferrandez, Brett R. South, Shuying Shen, F. Jeffrey Friedlin, Matthew H. Samore, and Stephane M. Meystre. 2013. BoB, A Best-of-breed Automated Text De-identification System for VHA Clinical Documents. *Journal of the American Medical Informatics Association* 20 (2013), 77–83. Issue 1. <https://doi.org/10.1136/amiajnl-2012-001020>
- [38] Eric N. Forsythand and Craig H. Martell. 2007. Lexical and Discourse Analysis of Online Chat Dialog. In *Proceedings of the 2007 International Conference on Semantic Computing (ICSC'07)*. IEEE, 19–26. <https://doi.org/10.1109/ICSC.2007.55>
-

- 
- [39] Tobias Gentner, Timon Neitzel, Jacob Schulze, and Ricardo Buettner. 2020. A Systematic Literature Review of Medical Chatbot Research From a Behavior Change Perspective. In *Proceedings of the 2020 IEEE 44th Annual Computers, Software, and Applications Conference (COMPSAC'20)*. IEEE, 735–740. <https://doi.org/10.1109/COMPSAC48688.2020.0-172>
- [40] Tommi Gröndahl and N. Asokan. 2019. Text Analysis in Adversarial Settings: Does Deception Leave a Stylistic Trace? *Comput. Surveys* 52, 3, Article 45 (2019), 36 pages. <https://doi.org/10.1145/3310331>
- [41] David Gros, Yu Li, and Zhou Yu. 2021. The R-U-A-Robot Dataset: Helping Avoid Chatbot Deception by Detecting User Questions About Human or Non-Human Identity. arXiv:2106.02692 [cs.CL] <https://arxiv.org/abs/2106.02692>
- [42] Yash Gupta, Pawan Sasanka Ammanamanchi, Shikha Bordia, Arjun Manoharan, Deepak Mittal, Ramakanth Pasunuru, Manish Shrivastava, Maneesh Singh, Mohit Bansal, and Preethi Jyothi. 2021. The Effect of Pretraining on Extractive Summarization for Scientific Documents. In *Proceedings of the Second Workshop on Scholarly Document Processing (SDP'21)*. Association for Computational Linguistics, 73–82. <https://doi.org/10.18653/v1/2021.sdp-1.9>
- [43] Ivan Habernal and Iryna Gurevych. 2016. Which Argument is More Convincing? Analyzing and Predicting Convincingness of Web Arguments Using Bidirectional LSTM. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 1589–1599. <https://doi.org/10.18653/v1/P16-1150>
- [44] Md. Kamrul Hasan, Wasifur Rahman, Amir Ali Bagher Zadeh, Jianyuan Zhong, Md. Iftexhar Tanveer, Louis-Philippe Morency, and Mohammed Ehsan Hoque. 2019. UR-FUNNY: A Multimodal Language Dataset for Understanding Humor. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP'19)*. Association for Computational Linguistics, 2046–2056. <https://doi.org/10.18653/v1/D19-1211>
- [45] Ruining He and Julian McAuley. 2016. Ups and Downs: Modeling the Visual Evolution of Fashion Trends with One-Class Collaborative Filtering. In *Proceedings of the 2016 International World Wide Web Conference (WWW'16)*. ACM, 507–517. <https://doi.org/10.1145/2872427.2883037>
- [46] Charles T. Hemphill, John J. Godfrey, and George R. Doddington. 1990. The ATIS Spoken Language Systems Pilot Corpus. In *Speech and Natural Language: Proceedings of a Workshop Held at Hidden Valley*. Association for Computational Linguistics, 96–101. <https://aclanthology.org/H90-1021/>
- [47] Matthew Henderson, Blaise Thomson, and Jason D. Williams. 2014. The Second Dialog State Tracking Challenge. In *Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL'14)*. Association for Computational Linguistics, 263–272. <https://doi.org/10.3115/v1/W14-4337>
- [48] Matthew Henderson, Blaise Thomson, and Jason D. Williams. 2014. The Third Dialog State Tracking Challenge. In *Proceedings of the 2014 IEEE Spoken Language Technology Workshop (SLT'14)*. IEEE, 324–329. <https://doi.org/10.1109/SLT.2014.7078595>
- [49] Sebastian Hobert. 2019. How Are You, Chatbot? Evaluating Chatbots in Educational Settings – Results of a Literature Review. *DELFI* (2019), 259–270. [https://doi.org/10.18420/delfi2019\\_289](https://doi.org/10.18420/delfi2019_289)
- [50] M.D. Zakir Hossain, Ferdous Sohel, Mohd Fairuz Shiratuddin, and Hamid Laga. 2019. A Comprehensive Survey of Deep Learning for Image Captioning. *ACM CSUR* 51, 6, Article 118 (2019), 36 pages. <https://doi.org/10.1145/3295748>
-

- 
- [51] Nabil Hossain, John Krumm, and Michael Gamon. 2019. “President Vows to Cut <Taxes> Hair”: Dataset and Analysis of Creative Text Editing for Humorous Headlines. arXiv:1906.00274 [cs.CL] <https://arxiv.org/abs/1906.00274>
- [52] Shengluan Hou (侯圣峦), Shuhan Zhang (张书涵), and Chaoqun Fei (费超群). 2019. A Survey to Text Summarization: Popular Datasets and Methods / 文本摘要常用数据集和方法研究综述. *Journal of Chinese Information Processing / 《中文信息学报》* 33, 5 (2019), 1–16. <http://jcip.cipsc.org.cn/CN/Y2019/V33/I5/1>
- [53] David M. Howcroft, Anya Belz, Miruna-Adriana Clinciu, Dimitra Gkatzia, Sadid A. Hasan, Saad Mahamood, Simon Mille, Emiel van Miltenburg, Sashank Santhanam, and Verena Rieser. 2020. Twenty Years of Confusion in Human Evaluation: NLG Needs Evaluation Sheets and Standardised Definitions. In *Proceedings of the 13th International Conference on Natural Language Generation (INLG’20)*. Association for Computational Linguistics, 169–182. <https://aclanthology.org/2020.inlg-1.23>
- [54] Baotian Hu, Qingcai Chen, and Fangze Zhu. 2016. LCSTS: A Large Scale Chinese Short Text Summarization Dataset. arXiv:1506.05865 [cs.CL] <https://arxiv.org/abs/1506.05865>
- [55] Zhiqiang Hu, Roy Ka-Wei Lee, Charu C. Aggarwal, and Aston Zhang. 2020. Text Style Transfer: A Review and Experimental Evaluation. arXiv:2010.12742 [cs.CL] <https://arxiv.org/abs/2010.12742>
- [56] Zhen Huang, Shiyi Xu, Minghao Hu, Xinyi Wang, Jinyan Qiu, Yongquan Fu, Yuncai Zhao, Yuxing Peng, and Changjian Wang. 2020. Recent Trends in Deep Learning Based Open-Domain Textual Question Answering Systems. *IEEE Access* 8 (2020), 94341–94356. <https://doi.org/10.1109/ACCESS.2020.2988903>
- [57] Anthony Hunter, Lisa Chalaguine, Tomasz Czernuszenko, Emmanuel Hadoux, and Sylwia Polberg. 2019. Towards Computational Persuasion via Natural Language Argumentation Dialogues. In *KI 2019: Advances in Artificial Intelligence – 42nd German Conference on AI, Kassel, Germany, September 23–26, 2019, Proceedings*. Springer, 18–33. [https://doi.org/10.1007/978-3-030-30179-8\\_2](https://doi.org/10.1007/978-3-030-30179-8_2)
- [58] Vladimir Ilievski, Claudiu Musat, Andreea Hossmann, and Michael Baeriswyl. 2018. Goal-Oriented Chatbot Dialog Management Bootstrapping with Transfer Learning. arXiv:1802.00500 [cs.CL] <https://arxiv.org/abs/1802.00500>
- [59] Ryosuke Iwakura, Tomohiro Yoshikawa, and Takeshi Furuhashi. 2018. A Basic Study on Generating Back-Channel Humor Phrases for Chat Dialogue Systems. In *Proceedings of the 2018 Joint 10th International Conference on Soft Computing and Intelligent Systems (SCIS) and 19th International Symposium on Advanced Intelligent Systems (ISIS’18)*. IEEE, 1263–1266. <https://doi.org/10.1109/SCIS-ISIS.2018.00198>
- [60] Robin Jia and Percy Liang. 2017. Adversarial Examples for Evaluating Reading Comprehension Systems. arXiv:1707.07328 [cs.CL] <https://arxiv.org/abs/1707.07328>
- [61] Di Jin, Zhijing Jin, Zhiting Hu, Olga Vechtomova, and Rada Mihalcea. 2020. Deep Learning for Text Style Transfer: A Survey. arXiv:2011.00416 [cs.CL] <https://arxiv.org/abs/2011.00416>
- [62] Di Jin, Zhijing Jin, Joey Tianyi Zhou, Lisa Orri, and Peter Szolovits. 2020. Hooks in the Headline: Learning to Generate Headlines with Controlled Styles. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL’20)*. Association for Computational Linguistics, 5082–5093. <https://doi.org/10.18653/v1/2020.acl-main.456>
-

- 
- [63] Mladjan Jovanovic, Marcos Baez, and Fabio Casati. 2020. Chatbots as Conversational Healthcare Services. *IEEE Internet Computing* 1 (2020), 44–51. <https://doi.org/10.1109/MIC.2020.3037151>
- [64] Tomoyuki Kajiwaru and Mamoru Komachi. 2016. Building a Monolingual Parallel Corpus for Text Simplification Using Sentence Similarity Based on Alignment Between Word Embeddings. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. Association for Computational Linguistics, 1147–1158. <https://aclanthology.org/C16-1109/>
- [65] Veton Kepuska and Gamal Bohouta. 2018. Next-Generation of Virtual Personal Assistants (Microsoft Cortana, Apple Siri, Amazon Alexa and Google Home). In *Proceedings of the 2018 IEEE 8th Annual Computing and Communication Workshop and Conference (CCWC'18)*. IEEE, 99–103. <https://doi.org/10.1109/CCWC.2018.8301638>
- [66] Tobias Kowatsch, Marcia Nißen, Chen-Hsuan Iris Shih, Dominik Rüegger, Dirk Volland, Andreas Filler, Florian Künzler, Filipe Barata, Dirk Büchter, Björn Brogle, et al. 2017. Text-based Healthcare Chatbots Supporting Patient and Health Professional Teams: Preliminary Results of a Randomized Controlled Trial on Childhood Obesity. In *Persuasive Embodied Agents for Behavior Change (PEACH'17) Workshop, co-located with the 17th International Conference on Intelligent Virtual Agents (IVA'17)*. ACM, 10. <https://www.alexandria.unisg.ch/252944/>
- [67] Matt J. Kusner, Yu Sun, Nicholas I. Kolkin, and Kilian Q. Weinberger. 2015. From Word Embeddings to Document Distances. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning (ICML'15)*. PMLR, 957–966. <https://doi.org/10.5555/3045118.3045221>
- [68] Chih-Te Lai, Yi-Te Hong, Hong-You Chen, Chi-Jen Lu, and Shou-De Lin. 2019. Multiple Text Style Transfer by using Word-level Conditional Generative Adversarial Network with Two-Phase Training. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP'19)*. Association for Computational Linguistics, 3579–3584. <https://aclanthology.org/D19-1366/>
- [69] Guillaume Lample, Sandeep Subramanian, Eric Smith, Ludovic Denoyer, Marc'Aurelio Ranzato, and Y-Lan Boureau. 2019. Multiple-Attribute Text Rewriting. In *Proceedings of the 2019 International Conference on Learning Representations (ICLR'19)*. ICLR 2019 Conference Committee, 20. <https://openreview.net/pdf?id=H1g2NhC5KQ>
- [70] Geoffrey Neil Leech. 1992. 100 Million Words of English: The British National Corpus (BNC). *Language Research / 语学研究* 28, 1 (1992), 1–13. <https://s-space.snu.ac.kr/handle/10371/85926>
- [71] Junyi Li, Tianyi Tang, Wayne Xin Zhao, and Ji-Rong Wen. 2021. Pretrained Language Models for Text Generation: A Survey. arXiv:2105.10311 [cs.CL] <https://arxiv.org/abs/2105.10311>
- [72] Pierre Lison, Ildikó Pilán, David Sanchez, Montserrat Batet, and Lilja Øvrelid. 2021. Anonymisation Models for Text Data: State of the art, Challenges and Future Directions. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJNLP'21)*. Association for Computational Linguistics, 4188–4203. <https://aclanthology.org/2021.acl-long.323>
- [73] Pierre Lison and Jörg Tiedemann. 2016. OpenSubtitles2016: Extracting Large Parallel Corpora from Movie and TV Subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. Association for Computational Linguistics, 923–929. <https://aclanthology.org/L16-1147/>
-

- 
- [74] Robert Lockwood and Kevin Curran. 2017. Text Based Steganography. *International Journal of Information Privacy, Security and Integrity* 3, 2 (2017), 134–153. <https://doi.org/10.1504/IJIPSI.2017.088700>
- [75] Jessica López Espejel. 2019. Automatic Summarization of Medical Conversations, a Review. In *Actes de la Conférence sur le Traitement Automatique des Langues Naturelles (TALN) PFIA 2019. Volume III : RECITAL*. ATALA, 487–498. <https://aclanthology.org/2019.jeptalnrecital-recital.3>
- [76] Samuel Louvan and Bernardo Magnini. 2020. Recent Neural Methods on Slot Filling and Intent Classification for Task-Oriented Dialogue Systems: A Survey. In *Proceedings of the 28th International Conference on Computational Linguistics (COLING’20)*. ICCL, 480–496. <https://doi.org/10.18653/v1/2020.coling-main.42>
- [77] 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. arXiv:1506.08909 [cs.CL] <https://arxiv.org/abs/1506.08909>
- [78] Xueming Luo, Siliang Tong, Zheng Fang, and Zhe Qu. 2019. Frontiers: Machines vs. Humans: The Impact of Artificial Intelligence Chatbot Disclosure on Customer Purchases. *Marketing Science* 38, 6 (2019), 937–947. <https://doi.org/10.1287/mksc.2019.1192>
- [79] Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning Word Vectors for Sentiment Analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, 142–150. <https://aclanthology.org/P11-1015/>
- [80] Aman Madaan, Amrith Setlur, Tanmay Parekh, Barnabas Poczós, Graham Neubig, Yiming Yang, Ruslan Salakhutdinov, Alan W. Black, and Shrimai Prabhumoye. 2020. Politeness Transfer: A Tag and Generate Approach. arXiv:2004.14257 [cs.CL] <https://arxiv.org/abs/2004.14257>
- [81] Khyati Mahajan and Samira Shaikh. 2021. On the Need for Thoughtful Data Collection for Multi-Party Dialogue: A Survey of Available Corpora and Collection Methods. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL’21)*. Association for Computational Linguistics, 338–352. <https://aclanthology.org/2021.sigdial-1.36>
- [82] Ben Medlock. 2006. An Introduction to NLP-based Textual Anonymisation. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC’06)*. European Language Resources Association (ELRA), 1051–1056. [http://www.lrec-conf.org/proceedings/lrec2006/pdf/200\\_pdf.pdf](http://www.lrec-conf.org/proceedings/lrec2006/pdf/200_pdf.pdf)
- [83] Marie-Jean Meurs, Frédéric Duvert, Frédéric Béchet, Fabrice Lefevre, and Renato De Mori. 2008. Semantic Frame Annotation on the French MEDIA Corpus. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC’08)*. Association for Computational Linguistics, 1014–1018. <https://aclanthology.org/L08-1491/>
- [84] Rada Mihalcea and Carlo Strapparava. 2005. Making Computers Laugh: Investigations in Automatic Humor Recognition. In *Proceedings of 2005 Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (EMNLP’05)*. Association for Computational Linguistics, 531–538. <https://doi.org/10.18653/v1/P19-1394>
- [85] Remi Mir, Bjarke Felbo, Nick Obradovich, and Iyad Rahwan. 2019. Evaluating Style Transfer for Text. arXiv:1904.02295 [cs.CL] <https://arxiv.org/abs/1904.02295>
- [86] Marie-Francine Moens. 2018. Argumentation Mining: How Can a Machine Acquire Common Sense and World Knowledge? *Argument & Computation* 9, 1 (2018), 1–14. <https://doi.org/10.3233/AAC-170025>
-

- 
- [87] Quim Motger, Xavier Franch, and Jordi Marco. 2021. Conversational Agents in Software Engineering: Survey, Taxonomy and Challenges. arXiv:2106.10901 [cs.CL] <https://arxiv.org/abs/2106.10901>
- [88] Lili Mou and Olga Vechtomova. 2020. Stylized Text Generation: Approaches and Applications. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*. Association for Computational Linguistics, 19–22. <https://doi.org/10.18653/v1/2020.acl-tutorials.5>
- [89] Vitobha Munigala, Abhijit Mishra, Srikanth G. Tamilselvam, Shreya Khare, Riddhiman Dasgupta, and Anush Sankaran. 2018. PersuAIDE! An Adaptive Persuasive Text Generation System for Fashion Domain. In *Companion Proceedings of the Web Conference 2018*. ACM, 335–342. <https://doi.org/10.1145/3184558.3186345>
- [90] Dong Nguyen, A. Seza Drogruoz, Carolyn P. Rose, and Franciska de Jong. 2016. Computational Sociolinguistics: A Survey. *Computational Linguistics* 42 (2016), 537–593. Issue 3. [https://doi.org/10.1162/COLI\\_a\\_00258](https://doi.org/10.1162/COLI_a_00258)
- [91] Cicero Nogueira dos Santos, Igor Melnyk, and Inkit Padhi. 2018. Fighting Offensive Language on Social Media with Unsupervised Text Style Transfer. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL’18)*. Association for Computational Linguistics, 189–194. <https://aclanthology.org/P18-2031/>
- [92] Gözde Özbal, Daniele Pighin, and Carlo Strapparava. 2013. BRAINSUP: Brainstorming Support for Creative Sentence Generation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 1446–1455. <https://aclanthology.org/P13-1142>
- [93] Gustavo Paetzold and Lucia Specia. 2016. Benchmarking Lexical Simplification Systems. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC’16)*. Association for Computational Linguistics, 3074–3080. <https://aclanthology.org/L16-1491/>
- [94] Gustavo Paetzold and Lucia Specia. 2016. Unsupervised Lexical Simplification for Non-Native Speakers. *Proceedings of the 2016 AAAI Conference on Artificial Intelligence (AAAI’16)* 30, 1 (2016), 3761–3767. <https://ojs.aaai.org/index.php/AAAI/article/view/9885>
- [95] Charulata Patil and Manasi Patwardhan. 2020. Visual Question Generation: The State of the Art. *Comput. Surveys* 53, 3, Article 47 (2020), 22 pages. <https://doi.org/10.1145/3383465>
- [96] John Pavlopoulos, Vasiliki Kougia, and Ion Androutsopoulos. 2019. A Survey on Biomedical Image Captioning. In *Proceedings of the Second Workshop on Shortcomings in Vision and Language (SiVL’19)*. Association for Computational Linguistics, 26–36. <https://doi.org/10.18653/v1/W19-1803>
- [97] José Quiroga Pérez, Thanasis Daradoumis, and Joan Manuel Marquès Puig. 2020. Rediscovering the Use Of Chatbots in Education: A Systematic Literature Review. *Computer Applications in Engineering Education* 28, 6 (2020), 1549–1565. <https://doi.org/10.1002/cae.22326>
- [98] Shrimai Prabhumoye, Yulia Tsvetkov, Ruslan Salakhutdinov, and Alan W. Black. 2018. Style Transfer Through Back-Translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL’18)*. Association for Computational Linguistics, 866–876. <https://aclanthology.org/P18-1080/>
- [99] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. *Language Models Are Unsupervised Multitask Learners*. Technical Report. OpenAI. <https://d4mucfpxsywv.cloudfront.net/better-language-models/language-models.pdf>
-

- 
- [100] Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know What You Don't Know: Unanswerable Questions for SQuAD. arXiv:1806.03822 [cs.CL] <https://arxiv.org/abs/1806.03822>
- [101] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ Questions for Machine Comprehension of Text. arXiv:1606.05250 [cs.CL] <https://arxiv.org/abs/1606.05250>
- [102] Sudha Rao and Joel Tetreault. 2018. Dear Sir or Madam, May I introduce the GYAFC Dataset: Corpus, Benchmarks and Metrics for Formality Style Transfer. arXiv:1803.06535 [cs.CL] <https://arxiv.org/abs/1803.06535>
- [103] Siva Reddy, Danqi Chen, and Christopher D. Manning. 2019. CoQA: A Conversational Question Answering Challenge. *Transactions of the Association for Computational Linguistics* 7 (2019), 249–266. [https://doi.org/10.1162/tacl\\_a\\_00266](https://doi.org/10.1162/tacl_a_00266)
- [104] Sravana Reddy and Kevin Knight. 2016. Obfuscating Gender in Social Media Writing. In *Proceedings of the 2016 EMNLP Workshop on Natural Language Processing and Computational Social Science*. Association for Computational Linguistics, 17–26. <https://aclanthology.org/W16-5603/>
- [105] Dana Rezazadegan, Shlomo Berkovsky, Juan C. Quiroz, A. Baki Kocaballi, Ying Wang, Lilliana Laranjo, and Enrico Coiera. 2020. Automatic Speech Summarisation: A Scoping Review. arXiv:2008.11897 [cs.CL] <https://arxiv.org/abs/2008.11897>
- [106] Graeme Ritchie, Ruli Manurung, Helen Pain, Annalu Waller, Rolf Black, and Dave O'Mara. 2007. A Practical Application of Computational Humour. In *Proceedings of the 4th International Joint Workshop on Computational Creativity (IJWCC'07)*. ACC, 91–98. <https://www.doc.gold.ac.uk/isms/CC07/CC07Proceedings.pdf#page=97>
- [107] Anna Rogers and Anna Rumshisky. 2020. A Guide to the Dataset Explosion in QA, NLI, and Commonsense Reasoning. In *Proceedings of the 28th International Conference on Computational Linguistics: Tutorial Abstracts*. ICCL, 27–32. <https://doi.org/10.18653/v1/2020.coling-tutorials.5>
- [108] Sara Rosenthal, Mihaela Bornea, and Avirup Sil. 2021. Are Multilingual BERT Models Robust? A Case Study on Adversarial Attacks for Multilingual Question Answering. arXiv:2104.07646 [cs.CL] <https://arxiv.org/abs/2104.07646>
- [109] Rahime Belen Sağlam and Jason R.C. Nurse. 2020. Is Your Chatbot GDPR Compliant? Open Issues in Agent Design. In *Proceedings of the 2nd Conference on Conversational User Interfaces (CUI'20)*. ACM, Article 16, 3 pages. <https://doi.org/10.1145/3405755.3406131>
- [110] Rahime Belen Sağlam, Jason R.C. Nurse, and Duncan Hodges. 2021. Privacy Concerns in Chatbot Interactions: When to Trust and When to Worry. In *HCI International 2021 - Posters: 23rd HCI International Conference, HCII 2021, Virtual Event, July 24–29, 2021, Proceedings, Part II*. Springer, 391–399. [https://doi.org/10.1007/978-3-030-78642-7\\_53](https://doi.org/10.1007/978-3-030-78642-7_53)
- [111] Sebastian Schuster, Sonal Gupta, Rushin Shah, and Mike Lewis. 2018. Cross-Lingual Transfer Learning for Multilingual Task Oriented Dialog. arXiv:1810.13327 [cs.CL] <https://arxiv.org/abs/1810.13327>
- [112] Sakib Shahriar. 2021. GAN Computers Generate Arts? A Survey on Visual Arts, Music, and Literary Text Generation using Generative Adversarial Network. arXiv:2108.03857 [cs.AI] <https://arxiv.org/abs/2108.03857>
- [113] Tianxiao Shen, Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2017. Style Transfer from Non-Parallel Text by Cross-Alignment. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*. Curran Associates Inc., 6833–6844. [10.5555/3295222.3295427](https://arxiv.org/abs/1704.05021)
-

- 
- [114] Connor Shorten and Taghi M. Khoshgoftaar. 2019. A Survey on Image Data Augmentation for Deep Learning. *Journal of Big Data* 6, 1, Article 60 (2019), 48 pages. <https://doi.org/10.1186/s40537-019-0197-0>
- [115] Punardeep Sikka and Vijay Mago. 2020. A Survey on Text Simplification. arXiv:2008.08612 [cs.CL] <https://arxiv.org/abs/2008.08612>
- [116] Christian Stab and Iryna Gurevych. 2017. Parsing Argumentation Structures in Persuasive Essays. *Computational Linguistics* 43, 3 (2017), 619–659. [https://doi.org/10.1162/COLI\\_a\\_00295](https://doi.org/10.1162/COLI_a_00295)
- [117] Julius Steen and Katja Markert. 2021. How to Evaluate a Summarizer: Study Design and Statistical Analysis for Manual Linguistic Quality Evaluation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Association for Computational Linguistics, 1861–1875. <https://aclanthology.org/2021.eacl-main.160>
- [118] Oliviero Stock, Marco Guerini, and Fabio Pianesi. 2016. Ethical Dilemmas for Adaptive Persuasion Systems. In *Proceedings of the 30th AAAI Conference on Artificial Intelligence (AAAI'16)*. AAAI, 4157–4161. <https://www.aaai.org/ocs/index.php/AAAI/AAAI16/paper/viewPaper/12349>
- [119] Amber Stubbs, Michele Filannino, and Ozlem Uzuner. 2017. De-identification of Psychiatric Intake Records: Overview of 2016 CEGS N-GRID Shared Tasks Track 1. *Journal of Biomedical Informatics* 75 (2017), S4–S18. <https://doi.org/10.1016/j.jbi.2017.06.011>
- [120] Raymond Hendy Susanto and Wei Lu. 2017. Neural Architectures for Multilingual Semantic Parsing. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, 38–44. <https://doi.org/10.18653/v1/P17-2007>
- [121] Youzhi Tian, Zhiting Hu, and Zhou Yu. 2018. Structured Content Preservation for Unsupervised Text Style Transfer. arXiv:1810.06526 [cs.CL] <https://arxiv.org/abs/1810.06526>
- [122] Varun Tiwari, Mohammad Farukh Hashmi, Avinash Keskar, and N.C. Shivaprakash. 2020. Virtual Home Assistant For Voice Based Controlling and Scheduling With Short Speech Speaker Identification. *Multimedia Tools and Applications* 79, 7 (2020), 5243–5268. <https://doi.org/10.1007/s11042-018-6358-x>
- [123] Shyam Upadhyay, Manaal Faruqui, Gokhan Tür, Hakkani-Tür Dilek, and Larry Heck. 2018. (Almost) Zero-Shot Cross-Lingual Spoken Language Understanding. In *Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'18)*. IEEE, 6034–6038. <https://doi.org/10.1109/ICASSP.2018.8461905>
- [124] Ozlem Uzuner, Brett R. South, Shuying Shen, and Scott L. DuVall. 2011. 2010 i2b2/VA Challenge on Concepts, Assertions, and Relations in Clinical Text. *Journal of the American Medical Informatics Association* 18 (2011), 552–556. Issue 5. <https://doi.org/10.1136/amiajnl-2011-000203>
- [125] Laurens van den Bercken, Robert-Jan Sips, and Christoph Lof. 2019. Evaluating Neural Text Simplification in the Medical Domain. In *Proceedings of the 2019 World Wide Web Conference (WWW'19)*. ACM, 3286–3292. <https://doi.org/10.1145/3308558.3313630>
- [126] Rob Voigt, David Jurgens, Vinodkumar Prabhakaran, Dan Jurafsky, and Yulia Tsvetkov. 2018. Rt-Gender: A Corpus for Studying Differential Responses to Gender. In *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC'18)*. European Language Resources Association (ELRA), 2814–2820. <https://aclanthology.org/L18-1445>
-

- 
- [127] Helen J. Wall, Claire C. Campbell, Linda K. Kaye, Andy Levy, and Navjot Bhullar. 2019. Personality Profiles and Persuasion: An Exploratory Study Investigating the Role of the Big-5, Type D Personality and the Dark Triad on Susceptibility to Persuasion. *Personality and Individual Differences* 139 (2019), 69–76. <https://doi.org/10.1016/j.paid.2018.11.003>
- [128] Eric Wallace, Shi Feng, Nikhil Kandpal, Matt Gardner, and Sameer Singh. 2019. Universal Adversarial Triggers for Attacking and Analyzing NLP. arXiv:1908.07125 [cs.CL] <https://arxiv.org/abs/1908.07125>
- [129] Congcong Wang and David Lillis. 2019. Classification for Crisis-Related Tweets Leveraging Word Embeddings and Data Augmentation. In *Proceedings of the TREC 2019-B Incident Streams Track*. 1–8. <https://trec.nist.gov/pubs/trec28/papers/CS-UCD.IS.pdf>
- [130] Xuewei Wang, Weiyan Shi, Richard Kim, Yoojung Oh, Sijia Yang, Jingwen Zhang, and Zhou Yu. 2019. Persuasion For Good: Towards a Personalized Persuasive Dialogue System for Social Good. arXiv:2010.03538 [cs.CL] <https://arxiv.org/abs/1906.06725>
- [131] Klaus Weber, Hannes Ritschel, Ilhan Aslan, Florian Lingenfelser, and Elisabeth André. 2018. How to Shape the Humor of a Robot-Social Behavior Adaptation Based on Reinforcement Learning. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction (ICMI'18)*. ACM, 154–162. <https://doi.org/10.1145/3242969.3242976>
- [132] Wei-Hung Weng, Yu-An Chung, and Peter Szolovits. 2019. Unsupervised Clinical Language Translation. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD'19)*. ACM, 3121–3131. <https://doi.org/10.1145/3292500.3330710>
- [133] Ren Wu, Shengen Yan, Yi Shan, Qingqing Dang, and Gang Sun. 2015. Deep Image: Scaling Up Image Recognition. arXiv:1501.02876 [cs.CV] <https://arxiv.org/abs/1501.02876>
- [134] Anbang Xu, Zhe Liu, Yufan Guo, Vibha Sinha, and Rama Akkiraju. 2017. A New Chatbot for Customer Service on Social Media. In *Proceedings of the 2017 CHI conference on Human Factors in Computing Systems (CHI'17)*. ACM, 3506–3510. <https://doi.org/10.1145/3025453.3025496>
- [135] Puyang Xu and Ruhi Sarikaya. 2013. Convolutional Neural Network Based Triangular CRF for Joint Intent Detection and Slot Filling. In *Proceedings of the 2013 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU'13)*. IEEE, 78–83. <https://doi.org/10.1109/ASRU.2013.6707709>
- [136] Qiongkai Xu, Lizhen Qu, Chenchen Xu, and Ran Cui. 2019. Privacy-Aware Text Rewriting. In *Proceedings of the 12th International Conference on Natural Language Generation (INLG'19)*. Association for Computational Linguistics, 247–257. <https://aclanthology.org/W19-8633/>
- [137] Wei Xu, Chris Callison-Burch, and Courtney Napoles. 2015. Problems in Current Text Simplification Research: New Data Can Help. *Transactions of the Association for Computational Linguistics* 3 (2015), 283–297. [https://doi.org/10.1162/tacl\\_a\\_00139](https://doi.org/10.1162/tacl_a_00139)
- [138] Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. 2016. Optimizing Statistical Machine Translation for Text Simplification. *Transactions of the Association for Computational Linguistics* 4 (2016), 401–415. [https://doi.org/10.1162/tacl\\_a\\_00107](https://doi.org/10.1162/tacl_a_00107)
- [139] Wei Xu, Alan Ritter, Bill Dolan, Ralph Grishman, and Colin Cherry. 2012. Paraphrasing for Style. In *Proceedings of 24th International Conference on Computational Linguistics (COLING'12)*. The COLING 2012 Organizing Committee, 2899–2914. <https://aclanthology.org/C12-1177/>
-

- 
- [140] Zhao Yan, Nan Duan, Peng Chen, Ming Zhou, Jianshe Zhou, and Zhoujun Li. 2017. Building Task-Oriented Dialogue Systems for Online Shopping. In *Proceedings of the 31st AAAI Conference on Artificial Intelligence (AAAI'17)*. AAAI, 4618–4625. <https://www.aaai.org/ocs/index.php/AAAI/AAAI17/paper/viewPaper/14261>
- [141] Diyi Yang, Alon Lavie, Chris Dyer, and Eduard Hovy. 2015. Humor Recognition and Humor Anchor Extraction. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP'15)*. Association for Computational Linguistics, 2367–2376. <https://aclanthology.org/D15-1284/>
- [142] Ren Er Yang and Quan Sheng. 2017. Neural Joke Generation. Final Project Reports of Course CS224n. <https://web.stanford.edu/class/archive/cs/cs224n/cs224n.1174/reports/2760332.pdf>
- [143] Winson Ye and Qun Li. 2020. Chatbot Security and Privacy in the Age of Personal Assistants. In *Proceedings of the 2020 IEEE/ACM Symposium on Edge Computing (SEC'20)*. IEEE, 388–393. <https://doi.org/10.1109/SEC50012.2020.00057>
- [144] Zhiwei Yu, Jiwei Tan, and Xiaojun Wan. 2018. A Neural Approach to Pun Generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 1650–1660. <https://doi.org/10.18653/v1/P18-1153>
- [145] Munazza Zaib, Quan Z. Sheng, and Wei Emma Zhang. 2020. A Short Survey of Pre-Trained Language Models For Conversational AI – A New Age in NLP. arXiv:2104.10810 [cs.CL] <https://arxiv.org/abs/2104.10810>
- [146] Munazza Zaib, Wei Emma Zhang, Quan Z. Sheng, Adnan Mahmood, and Yang Zhang. 2021. Conversational Question Answering: A Survey. arXiv:2106.00874 [cs.CL] <https://arxiv.org/abs/2106.00874>
- [147] Brahim Zarouali, Tom Dobber, Guy De Pauw, and Claes de Vreese. 2020. Using a Personality-Profiling Algorithm to Investigate Political Microtargeting: Assessing the Persuasion Effects of Personality-Tailored Ads on Social Media. *Communication Research* (2020), 1–26. <https://doi.org/10.1177/0093650220961965>
- [148] Fengbin Zhu, Wenqiang Lei, Chao Wang, Jianming Zheng, Soujanya Poria, and Tat-Seng Chua. 2021. Retrieving and Reading: A Comprehensive Survey on Open-Domain Question Answering. arXiv:2101.00774 [cs.CL] <https://arxiv.org/abs/2101.00774>
- [149] Zhemin Zhu, Delphine Bernhard, and Iryna Gurevych. 2010. A Monolingual Tree-based Translation Model for Sentence Simplification. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING'10)*. The COLING 2010 Organizing Committee, 1353–1361. <https://aclanthology.org/C10-1152/>