Chapter 1: See Other File

Chapter 2: See Other File

Chapter 3: Automation and Crowd-Sourcing for Data

The fastest method of creating large neural-scale datasets is through automatic generation of synthetic data. This chapter discusses a large audio dataset created with Text-To-Speech technology, and the limitations thereof (Section 3.1).[1] We reach similar conclusions from another automatically-created dataset, intended for evaluating co-reference (Chapter **??**). Both datasets, while large, are not real-

[1] Denis Peskov, Joe Barrow, Pedro Rodriguez, Graham Neubig, and Jordan Boyd-Graber. 2019. Mitigating noisy inputs for question answering. In Conference of the International Speech Communication Association. Peskov is responsible for the data creation, the gathering of recordings from users, running the neural models, figure and table design, and paper writing.

istic and motivate using humans for **generation** (Chapter **??**). Furthermore, both datasets ultimately depend on **experts** for validation (Chapter **??**).

3.1 Automated Data Creation for Question Answering

Progress on question answering (QA) has claimed human-level accuracy. However, most factoid QA models are trained and evaluated on clean text input, which becomes noisy when questions are spoken due to Automatic Speech Recognition (ASR) errors. This consideration is disregarded in trivia match-ups between machines and humans: IBM Watson (Ferrucci, 2010) on Jeopardy! and QB matches between machines and trivia masters (Boyd-Graber et al., 2018) provide text data for machines while humans listen. A fair assessment of an Artificial Intelligence's ability to answer human trivia questions would subject the machine to speech input, akin to how typical human would process sound.[2]

Hence, computers should be provided with the same audio input that a human would hear. In order to process the audio and answer the question, the computer needs a model to decode the audio into textual format. Unfortunately, there are no large *spoken* corpora of factoid questions with which to train models; text-to-speech software can be used as a method for generating training data at scale for question answering models (Section 3.2). Although synthetic data is less realistic than true human-spoken questions it easier and cheaper to collect at scale, which is important for training. These synthetic data are still useful;, models trained

[2] An audibly impaired person would be delivered questions in a non-audio medium, but would still experience a cognitive delay, unlike a machine.

on synthetic data are applied to human spoken data from QB tournaments and Jeopardy! (Section 3.4.1).

Noisy ASR is particularly challenging for QA systems (Figure 3.1). While humans and computers might know the title of a "revenge novel centering on Edmund Dantes by Alexandre Dumas", transcription errors may mean deciphering "novel centering on edmond dance by alexander <*unk*>" instead. Dantes and Dumas are low-frequency words in the English language and hence likely to be misinterpreted by a generic ASR model; however, they are particularly important for answering the question. Additionally, the introduction of distracting words (e.g., "dance") causes QA models to make errors (Jia and Liang, 2017). Key terms like named entities are often missing, which is detrimental for QA (Section 3.2.1).

Previous approaches to mitigate ASR noise for answering mobile queries (Mishra and Bangalore, 2010) or building bots (Leuski et al., 2009) typically use unsupervised methods, such as term-based information retrieval. Our datasets for training and evaluation can produce *supervised* systems that directly answer spoken questions. Machine translation (Sperber et al., 2017) also uses ASR confidences; we evaluate similar methods on QA.

Specifically, some accuracy loss from noisy inputs can be mitigated through a combination of forcing unknown words to be decoded as the closest option (Section 3.3.2), and incorporating the uncertainties of the ASR model directly in neural models (Section 3.3.3). The forced decoding method reconstructs missing terms by using terms audibly similar to the transcribed input. Word-level confidence scores incorporate uncertainty from the ASR system into a Deep Averaging Network, introduced earlier (Background Section ??). These methods are compared against baseline methods on our synthetic and human speech datasets for Jeopardy! and QB (Section 3.4).

3.2 Automatically Generating a Speech Dataset

Neural networks require a large training corpus, but recording hundreds of thousands of questions is not feasible. Methods for collecting large scale audio data include Generative Adversarial Networks (Donahue et al., 2018) and manual recording (Lee et al., 2018). For manual recording, crowd-sourcing with the required quality control (speakers who say "cyclohexane" correctly) is prohibitively expensive. As an alternative, we generate a data-set with Google Text-to-Speech on 96,000 factoid questions from a trivia game called Quizbowl (Boyd-Graber et al., 2018), each with 4–6 sentences for a total of over 500,000 sentences.[3] We then decode these utterances using the Kaldi chain model (Peddinti et al., 2015), trained on the Fischer-English dataset (Cieri et al., 2004) for consistency with past results on mitigating ASR errors in MT (Sperber et al., 2017). This model decodes enough noise into our data to test mitigation strategies.[4]

[3] `http://cloud.google.com/text-to-speech`

[4] This model has a Word Error Rate (WER) of 15.60% on the eval2000 test set. The WER increases to 51.76% on our QB data, which contains out of domain vocabulary. Since there is no existing work in question answering, we use machine translation as proxy for determining an appropriate Word Error Rate, as intentional noise has been added to this subdomain (Michel and Neubig, 2018; Belinkov and Bisk, 2018). The most BLEU improvement in machine translation under noisy conditions could be found in this middle WER range, rather than in values below 20%

3.2.1 Why Question Answering is challenging for ASR

Question Answering (QA) requires the system to provide a correct answer out of many candidates based on the question's wording. ASR changes the features of the recognized text in several important ways: the overall vocabulary is quite different and important words are corrupted. First, it reduces the overall vocabulary. In our dataset, the vocab drops from 263,271 in the original data to a mere 33,333. This is expected, as our ASR system only has 42,000 words in its vocab, so the long tail of the Zipf's curve is lost. Second, unique words—which may be central to answering the question—are lost or misinterpreted; over 100,000 of the words in the original data occur only once. Finally, ASR systems tend to delete unintentionally delete words, which makes the sentences shorter. In our QB data, the average number of words decreases from 21.62 to 18.85 per sentence.

The decoding system is able to express uncertainty by predicting <*unk*>. These account for slightly less than 10% of all our word tokens, but is a top-2 prediction for 30% of the 260,000 original words. For QA, words with a high TF-IDF measure are valuable. While some words are lost, others can likely be recovered: "hellblazer' becoming "blazer", "clarendon" becoming "claritin". We evaluate this by fitting a TF-IDF model on the Wikipedia dataset and then comparing the average TF-IDF per sentence between the original and the ASR data. The average TF-IDF score,

or above 80% (Sperber et al., 2017). Retraining the model on the QB domain would mitigate this noise; however, in practice one is often at the mercy of a pre-trained recognition model due to changes in vocabularies or speakers.

the most popular metric for evaluating how important a word is for a document, drops from 3.52 to 2.77 per sentence. Examples of this change can be seen in Figure 3.1.

For generalization, we test the effect of noise on two types of distinct questions. QB questions, which are generally four to six sentences long, test a user's depth of knowledge; early clues are challenging and obscure but they progressively become easy and well-known. Competitors can answer these types of questions at any point. Computer QA is competitive with the top players (Yamada et al., 2018). Jeopardy! questions are single sentences and can only be answered after the question ends. To test this alternate syntax, we use the same method of data generation on a dataset of over 200,000 Jeopardy questions (Dunn et al., 2017b).

3.3 Mitigating noise

This section discusses two approaches to mitigating the effects of missing and corrupted information caused by ASR systems. The first approach—forced decoding—exploits systematic errors to arrive at the correct answer. The second uses confidence information from the ASR system to down-weight the influence of low-confidence terms. Both approaches improve accuracy over a baseline DAN model and show promise for short single-sentence questions. However, a IR approach is more effective on long questions since noisy words are completely avoided during the answer selection process.

3.3.1 IR baseline

The IR baseline reframes Jeopardy! and QB QA tasks as document retrieval tasks with an inverted search index. We create one document per distinct answer; each document has a text field formed by concatenating all questions with that answer together. At test time new, unseen questions are treated as queries, and documents are scored using BM25 (Ramos, 2003; Robertson et al., 2009). We implement this baseline with Elastic Search and Apache Lucene.

3.3.2 Forced decoding

We have systematically lost information due to ASR decoding. We could predict the answer if we had access to certain words in the original question and further postulate that wrong guesses are better than knowing that a word is unknown.

As a first step, we explored commercial solutions—Bing, Google, IBM, Wit—with low transcription errors. However, their APIs ensure that an end-user often cannot extract anything more than one-best transcriptions, along with an aggregate confidence for the sentence. Additionally, the proprietary systems are moving targets, harming reproducibility.

Therefore, we use Kaldi (Povey et al., 2011) for all experiments. Kaldi is a commonly-used, open-source tool for ASR; its maximal transparency enables approaches that incorporate uncertainty into downstream models. Kaldi provides not only top-1 predictions, but also confidences of words, entire lattices, and phones (Table 3.1). Each item in the sequence represents a word and has a confidence in

Clean	For 10 points, name this revenge novel centering on Edmond Dantes, written by Alexandre Dumas
1-Best	for$^{0.935}$ ten$^{0.935}$ points$^{0.871}$ same$^{0.617}$ this^{1} . . . revenge novel centering on <*unk*> written by alexander <*unk*> . . .
“Lattice”	for$^{0.935}$ [eps]$^{0.064}$ pretend$^{0.001}$ ten$^{0.935}$ . . . pretend point points point name same named name names this revenge novel . . .
Phones	f_B$^{0.935}$ er_E$^{0.935}$ t_B$^{0.935}$ eh_I^{1} n_E$^{0.935}$ . . . p_B oy_I n_I t_I s_E sil s_B ey_I m_E dh_B ih_I s_E r_B iy_I v_I eh_I n_I jh_E n_B aa_I v_I ah_I l_I . . .

Table 3.1: As original data are translated through ASR, it degrades in quality. One-best output captures per-word confidence. Full lattices provide additional words and phone data captures the raw ASR sounds. Our confidence model and forced decoding approach could be used for such data in future work.

range [0, 1] correspond to the respective.

The typical end-use of an ASR system wants to know when when a word is not recognized. By default, a graph will have a token that represents an unknown; in Kaldi, this becomes <*unk*>. At a human-level, one would want to know that an out of context word happened.

However, when the end-user is a downstream model, a systematically wrong prediction may be better than a generic statement of uncertainty. So by removing all reference to <*unk*> in the model, we force the system to decode “Louis Vampas” as

"Louisiana" rather than <*unk*>.[5] The risk we run with this method is introducing words not present in the original data. For example, "count" and "mount" are similar in sound but not in context embeddings. Hence, we need a method to downweight incorrect decoding.

3.3.3 Confidence augmented DAN

The errors introduced by ASR can hinder sequence neural models as key phrases are potentially corrupted. We modify the original DAN model, introduced in Background Section **??**, to use word-level confidences from the ASR system as a feature and be robust to corrupted phrases. In increasing order of complexity, the variations are: a Confidence Informed Softmax DAN, a Confidence Weighted Average DAN, and a Word-Level Confidence DAN. We represent the confidences as a vector $\mathbf{c}$, where each cell c_{i} contains the ASR confidence of word w_{i}.

The simplest model averages the confidence across the whole sentence and adds it as a feature to the final output classifier. For example in Table 3.1, "for ten points" averages to 0.914. We introduce an additional weight in the output $\mathbf{W^c}$, which adjusts our prediction based on the average confidence of each word in the question.

However, most words have high confidence, and thus the average confidence of a sentence or question level is high. To focus on *which* words are uncertain we weight the word embeddings by their confidence attenuating uncertain words before

[5] More specifically, <*unk*> is removed from the Finite State Transducer, which sets the input/output for the ASR system.

calculating the DAN average. In the previous example—“for ten points”—“for” and “ten” are frequently occuring words and have a confidence of .935, while “points” has a lower confidence of .871. The next word—“same”—should actually be “name” and hence the embedding referenced is incorrect. But, the lower confidence of .617 for this prediction decreases the overall weight of the embedding in the model.

Weighting by the confidence directly removes uncertain words, but this is too blunt an instrument, and could end up erasing useful information contained in low-confidence words, so we instead learn a function based on the raw confidence from our ASR system. Thus, we recalibrate the confidence through a learned function f:

$$f(\mathbf{c}) = \mathbf{W^{(c)}}\mathbf{c} + \mathbf{b^{(c)}} \tag{3.1}$$

and then use that scalar in the weighted mean of the DAN representation layer:

$$\mathbf{r}^{**} = \frac{\sum_i^N \mathbf{E}[w_{\mathrm{i}}] * f(c_{\mathrm{i}})}{N}. \tag{3.2}$$

In this model, we replace the original encoder $\mathbf{r}$ with the new version $\mathbf{r}^{**}$ to learn a transformation of the ASR confidence that down-weights uncertain words and up-weights certain words. This final model is referred to as our “Confidence Model”.

Architectural decisions are determined by hyperparameter sweeps. They include: having a single hidden layer of 1000 dimensionality for the DAN, multiple drop-out, batch-norm layers, and a scheduled ADAM optimizer. Our DAN models train until convergence, as determined by early-stopping. Code is implemented in

PyTorch (Paszke et al., 2017), with TorchText for batching.[6]

3.4 Results

Achieving 100% accuracy on this dataset is not a realistic goal, as not all test questions are answerable (specifically, some answers do not occur in the training data and hence cannot be learned by a machine learning system). Baselines for the DAN (Table 3.2) establish realistic goals: a DAN trained and evaluated on the *same train and dev set*, only in the original non-ASR form, correctly predicts 54% of the answers. Noise drops this to 44% with the best IR model and down to ≈ 30% with neural approaches.

Since the noisy data quality makes full recovery unlikely, we view any improvement over the neural model baselines as recovering valuable information. At the question-level, strong IR outperforms the DAN by around 10%.

Since IR can avoid all the noise while benefiting from additional independent data points, it scales as the length of data increases. There is additional motivation to investigate this task at the sentence-level. Computers can beat humans at the game by knowing certain questions immediately; the first sentence of the QB question serves as a proxy for this threshold. Our proposed combination of forced decoding with a neural model led to the highest test accuracy results and outperforms the IR one at the sentence level.

A strong TF-IDF IR model can top the best neural model at the multi-sentence

[6] Code, data, and additional analysis available at `https://github.com/DenisPeskov/QBASR`

question level in QB; multiple sentences are important because they progressively become easier to answer in competitions. However, our models improve accuracy on the shorter first-sentence level of the question. This behavior is expected since IR methods are explicitly designed to disregard noise and can pinpoint the handful of unique words in a long paragraph; conversely they are less accurate when they extract words from a single sentence.

3.4.1 Qualitative Analysis & Human Data

The synthetic dataset facilitates large-scale machine learning, but ultimately we care about performance on human data. For QB we record questions read by domain experts at a competition. To account for variation in speech, we record five questions across ten different speakers, varying in gender and age; this set of fifty questions is used as the human test data. Table 3.3 provides examples of variations. For Jeopardy! we manually parsed a complete episode by question.

The predictions of the regular DAN and the confidence version can differ. As one example, input about The House on Mango Street, which contains words like “novel”, “character”, and “childhood” alongside a corrupted name of the author, the regular DAN predicts The Prime of Miss Jean Brodie, while our version predicts the correct answer. As another example the model in Table 3.3 predicts “London” if “beaumont” and “john” are preserved, but “Baghdad” if the proper nouns, but not “palace” and “city”, are lost.

3.5 Implications of Automation

The advantages of this method are cost and scalability, which is demanded by the current paradigm of neural models. This however comes at the expense of quality. A limitation of our past work in **automation** is generalization: text-to-speech only has female voices and is consistently decoded, while the voices of real humans are decoded with large variations. Unseen data points are likely to confound a model trained on unnatural data. Additionally, automated data creation still depends on having quality source data, that often has to come from expert users. In this project, we are able to record **found** questions that were already written by Quizbowl **experts**. Writing hundreds of thousands of our questions would not have been tractable. Hence, expert design is necessary for automation, as implemented in our other project (Chapter **??**).

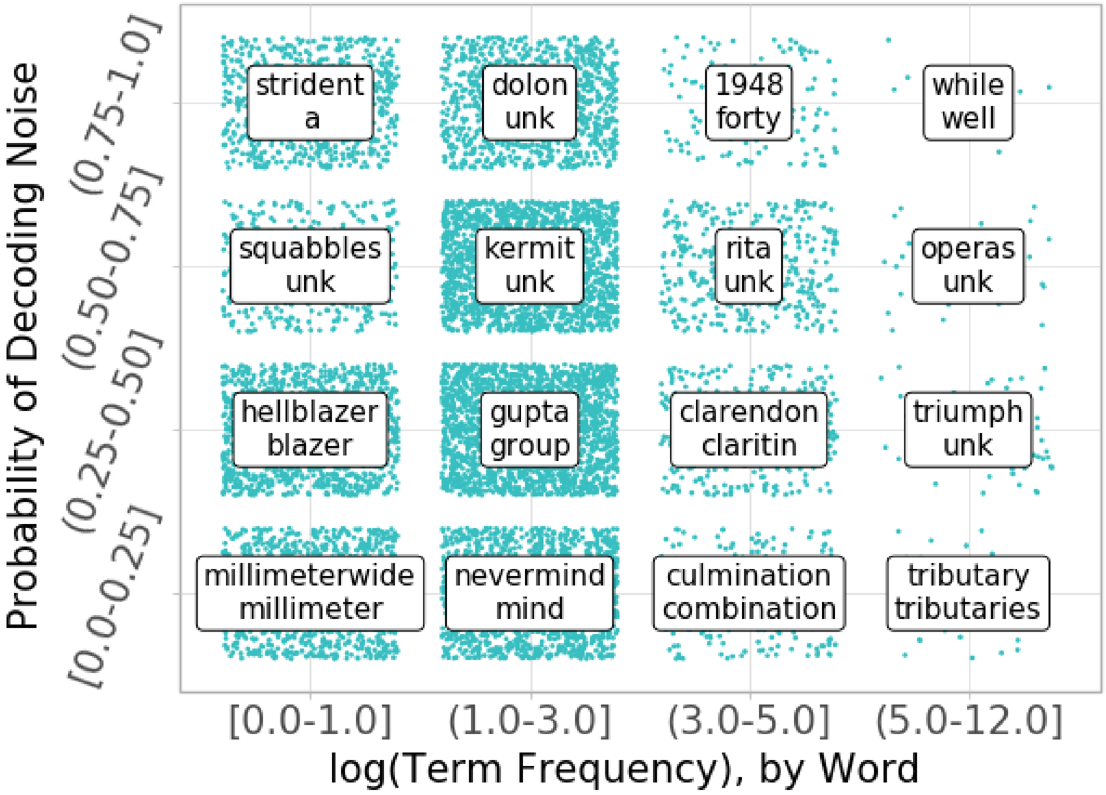

Figure 3.1: ASR errors on QA data: original spoken words (top of box) are garbled (bottom). While many words become into "noise"—frequent words or the unknown token—consistent errors (e.g., "clarendon" to "clarintin") can help downstream systems. Additionally, words reduced to <*unk*> (e.g., "kermit") can be useful through forced decoding into the closest incorrect word (e.g., "hermit" or even "car").

	QB				Jeopardy!	
	Synth		Human		Synth	Human
Model	Start	End	Start	End		
Methods Tested on Clean Data						
IR	0.064	0.544	0.400	1.000	0.190	0.050
DAN	0.080	0.540	0.200	1.000	0.236	0.033
Methods Tested on Corrupted Data						
IR base	0.021	0.442	0.180	0.560	0.079	0.050
DAN	0.035	0.335	0.120	0.440	0.097	0.017
FD	0.032	0.354	0.120	0.440	0.102	0.033
Confidence	0.036	0.374	0.120	0.460	0.095	0.033
FD+Conf	0.041	0.371	0.160	0.440	0.109	0.033

Table 3.2: Both forced decoding (FD) and the best confidence model improve accuracy. Jeopardy only has an At-End-of-Sentence metric, as questions are one sentence in length. Combining the two methods leads to a further joint improvement in certain cases. IR and DAN models trained and evaluated on clean data are provided as a reference point for the ASR data.

Speaker	Text
Base	John Deydras, an insane man who claimed to be Edward II, stirred up trouble when he seized this city's Beaumont Palace.
S1	unk an insane man who claimed to be the second unk trouble when he sees unk beaumont → Richard_I_of_England
S2	john dangerous insane man who claims to be the second stirring up trouble when he sees the city's beaumont → London
S3	unk dangerous insane man who claim to be unk second third of trouble when he sees the city's unk palace → Baghdad

Table 3.3: Variation in different speakers causes different transcriptions of a question on Oxford. The omission or corruption of certain named entities leads to different answer predictions, which are indicated with an arrow.

Bibliography

Abejide Olu Ade-Ibijola, Ibiba Wakama, and Juliet Chioma Amadi. 2012. An expert system for automated essay scoring (aes) in computing using shallow nlp techniques for inferencing. *International Journal of Computer Applications*, 51(10).

Amazon. 2021. Amazon Mechanical Turk. `http://www.mturk.com/`. [Online; accessed 03-January-2021].

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2019. On the cross-lingual transferability of monolingual representations. *CoRR*, abs/1910.11856.

Sue Atkins, Jeremy Clear, and Nicholas Ostler. 1992. Corpus design criteria. *Literary and linguistic computing*, 7(1):1–16.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, et al. 2016. Ms marco: A human generated machine reading comprehension dataset. *arXiv preprint arXiv:1611.09268*.

Juan M Banda, Ramya Tekumalla, Guanyu Wang, Jingyuan Yu, Tuo Liu, Yuning Ding, and Gerardo Chowell. 2020. A twitter dataset of 150+ million tweets related to covid-19 for open research. *Type: dataset*.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Yonatan Belinkov and Yonatan Bisk. 2018. Synthetic and natural noise both break neural machine translation. In *Proceedings of the International Conference on Learning Representations*.

Adam Berger, Peter F Brown, Stephen A Della Pietra, Vincent J Della Pietra, John R Gillett, John Lafferty, Robert L Mercer, Harry Printz, and Lubos Ures. 1994. The candide system for machine translation. In *HUMAN LANGUAGE TECHNOLOGY: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994*.

Jean Berko. 1958. The child's learning of english morphology. *Word*, 14(2-3):150–177.

Léon Bottou. 2010. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010*, pages 177–186. Springer.

L. E. Bourne, J. Kole, and A. Healy. 2014. Expertise: defined, described, explained. *Frontiers in Psychology*, 5.

Jordan Boyd-Graber. 2020. What question answering can learn from trivia nerds. In *Proceedings of the Association for Computational Linguistics*.

Jordan Boyd-Graber, Shi Feng, and Pedro Rodriguez. 2018. *Human-Computer Question Answering: The Case for Quizbowl*. Springer Verlag.

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. In *Proceedings of Advances in Neural Information Processing Systems*.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 2018. Multiwoz-a large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling. In *Proceedings of Empirical Methods in Natural Language Processing*.

Michael Buhrmester, Tracy Kwang, and Samuel D Gosling. 2011. Amazon's mechanical Turk: A new source of inexpensive, yet high-quality data? *Perspectives on psychological science: a journal of the Association for Psychological Science*, 6 1:3–5.

Bill Byrne, Karthik Krishnamoorthi, Chinnadhurai Sankar, Arvind Neelakantan, Daniel Duckworth, Semih Yavuz, Ben Goodrich, Amit Dubey, Kyu-Young Kim, and Andy Cedilnik. 2019. Taskmaster-1: Toward a realistic and diverse dialog dataset. In *Proceedings of Empirical Methods in Natural Language Processing*.

Chris Callison-Burch, Lyle Ungar, and Ellie Pavlick. 2015. Crowdsourcing for nlp. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorial Abstracts*, pages 2–3.

Jesse J Chandler and Gabriele Paolacci. 2017. Lie for a dime: When most prescreening responses are honest but most study participants are impostors. *Social Psychological and Personality Science*, 8(5):500–508.

Wendy W Chapman, Prakash M Nadkarni, Lynette Hirschman, Leonard W D'avolio, Guergana K Savova, and Ozlem Uzuner. 2011. Overcoming barriers to nlp for clinical text: the role of shared tasks and the need for additional creative solutions.

James Cheng, Monisha Manoharan, Yan Zhang, and Matthew Lease. 2015. Is there a doctor in the crowd? diagnosis needed! (for less than $5). *iConference 2015 Proceedings.*

Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen-tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. Quac: Question answering in context. In *Proceedings of Empirical Methods in Natural Language Processing*.

Christopher Cieri, David Miller, and Kevin Walker. 2004. The fisher corpus: a resource for the next generations of speech-to-text. In *Proceedings of the Language Resources and Evaluation Conference*.

Jonathan H. Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. 2020a. Tydi qa: A benchmark for information-seeking question answering in typologically diverse languages. In *Transactions of the Association for Computational Linguistics*.

Jonathan H Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. 2020b. Tydi qa: A benchmark for information-seeking question answering in typologically diverse languages. In *Transactions of the Association for Computational Linguistics*.

Louise Deléger, Magnus Merkel, and Pierre Zweigenbaum. 2009. Translating medical terminologies through word alignment in parallel text corpora. *Journal of Biomedical Informatics*, 42(4):692–701.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Djellel Difallah, Elena Filatova, and Panos Ipeirotis. 2018. Demographics and dynamics of mechanical turk workers. In *Proceedings of the eleventh ACM international conference on web search and data mining*, pages 135–143.

Chris Donahue, Bo Li, and Rohit Prabhavalkar. 2018. Exploring speech enhancement with generative adversarial networks for robust speech recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing*.

Matthew Dunn, Levent Sagun, Mike Higgins, V. Ugur Güney, Volkan Cirik, and Kyunghyun Cho. 2017a. SearchQA: A new Q&A dataset augmented with context from a search engine. *CoRR*, abs/1704.05179.

Matthew Dunn, Levent Sagun, Mike Higgins, V. Ugur Güney, Volkan Cirik, and Kyunghyun Cho. 2017b. Searchqa: A new Q&A dataset augmented with context from a search engine. *CoRR*, abs/1704.05179.

Alfred Dürr. 2005. *The cantatas of JS Bach: with their librettos in German-English parallel text*. OUP Oxford.

Jeffrey L Elman. 1990. Finding structure in time. *Cognitive science*, 14(2):179–211.

David A. Ferrucci. 2010. Build Watson: an overview of DeepQA for the Jeopardy! challenge. In *19th International Conference on Parallel Architecture and Compilation Techniques*, pages 1–2.

Elena Filatova, Vasileios Hatzivassiloglou, and Kathleen McKeown. 2006. Automatic creation of domain templates. In *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, pages 207–214.

Timothy W. Finin, William Murnane, Anand Karandikar, Nicholas Keller, Justin Martineau, and Mark Dredze. 2010. Annotating named entities in twitter data with crowdsourcing. In *Conference of the North American Chapter of the Association for Computational Linguistics*.

Ailbhe Finnerty, Pavel Kucherbaev, Stefano Tranquillini, and Gregorio Convertino. 2013. Keep it simple: Reward and task design in crowdsourcing. In *Proceedings of the Biannual Conference of the Italian Chapter of SIGCHI*, pages 1–4.

John R Firth. 1957. A synopsis of linguistic theory, 1930-1955. *Studies in linguistic analysis*.

Felix A Gers, Jürgen Schmidhuber, and Fred Cummins. 1999. Learning to forget: Continual prediction with lstm.

Yoav Goldberg. 2017. Neural network methods for natural language processing. *Synthesis Lectures on Human Language Technologies*, 10(1):1–309.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. *Advances in neural information processing systems*, 27:2672–2680.

Rob van der Goot. 2021. We need to talk about train-dev-test splits. In *Proceedings of Empirical Methods in Natural Language Processing*, pages 4485–4494, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Maharshi Gor, Kellie Webster, and Jordan Boyd-Graber. 2021. Toward deconfounding the influence of subject's demographic characteristics in question answering. In *Emperical Methods in Natural Language Processing*, page 6.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R. Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *NAACL-HLT*.

Matthew Henderson, Pawel Budzianowski, Iñigo Casanueva, Sam Coope, Daniela Gerz, Girish Kumar, Nikola Mrksic, Georgios Spithourakis, Pei-Hao Su, Ivan Vulic, and Tsung-Hsien Wen. 2019. A repository of conversational datasets. *CoRR*, abs/1904.06472.

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)*, pages 263–272, Philadelphia, PA, U.S.A. Association for Computational Linguistics.

Na Hong, Andrew Wen, Majid Rastegar Mojarad, Sunghwan Sohn, Hongfang Liu, and Guoqian Jiang. 2018. Standardizing heterogeneous annotation corpora using hl7 fhir for facilitating their reuse and integration in clinical nlp. In *AMIA Annual Symposium Proceedings*, volume 2018, page 574. American Medical Informatics Association.

W John Hutchins. 2004. The georgetown-ibm experiment demonstrated in january 1954. In *Conference of the Association for Machine Translation in the Americas*, pages 102–114. Springer.

Mohit Iyyer, Anupam Guha, Snigdha Chaturvedi, Jordan Boyd-Graber, and Hal Daumé III. 2016. Feuding families and former friends: Unsupervised learning for dynamic fictional relationships. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1534–1544.

Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. In *Proceedings of Empirical Methods in Natural Language Processing*.

Qiqi Jiang, Chuan-Hoo Tan, Chee Wei Phang, Juliana Sutanto, and Kwok-Kee Wei. 2013. Understanding chinese online users and their visits to websites: Application of zipf's law. *International journal of information management*, 33(5):752–763.

Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke S. Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the Association for Computational Linguistics*.

Daniel Jurafsky and James H Martin. 2000. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice Hall PTR.

Kushal Kafle, Mohammed Yousefhussien, and Christopher Kanan. 2017. Data augmentation for visual question answering. In *Proceedings of the 10th International Conference on Natural Language Generation*, pages 198–202.

Marzena Karpinska, Nader Akoury, and Mohit Iyyer. 2021. The perils of using mechanical turk to evaluate open-ended text generation. In *Proceedings of Empirical Methods in Natural Language Processing.*

Bryan Klimt and Yiming Yang. 2004. The enron corpus: A new dataset for email classification research. In *European Conference on Machine Learning*, pages 217–226. Springer.

Shailesh Kochhar, Stefano Mazzocchi, and Praveen Paritosh. 2010. The anatomy of a large-scale human computation engine. In *Proceedings of the acm sigkdd workshop on human computation*, pages 10–17.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5, pages 79–86. Citeseer.

Philipp Koehn and Christof Monz. 2006. Manual and automatic evaluation of machine translation between european languages. In *Proceedings on the Workshop on Statistical Machine Translation*, pages 102–121.

Ana Kozomara and Sam Griffiths-Jones. 2014. mirbase: annotating high confidence micrornas using deep sequencing data. *Nucleic acids research*, 42(D1):D68–D73.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105.

Abhimanu Kumar and Matthew Lease. 2011. Learning to rank from a noisy crowd. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, pages 1221–1222.

Jonathan K. Kummerfeld. 2021. Quantifying and avoiding unfair qualification labour in crowdsourcing. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 343–349, Online. Association for Computational Linguistics.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

Chia-Hsuan Lee, Shang-Ming Wang, Huan-Cheng Chang, and Hung-Yi Lee. 2018. Odsqa: Open-domain spoken question answering dataset. In *2018 IEEE Spoken Language Technology Workshop (SLT).*

Gondy Leroy and James E Endicott. 2012. Combining nlp with evidence-based methods to find text metrics related to perceived and actual text difficulty. In *Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium*, pages 749–754.

Anton Leuski, Ronakkumar Patel, David Traum, and Brandon Kennedy. 2009. Building effective question answering characters. In *Proceedings of the Annual SIGDIAL Meeting on Discourse and Dialogue*.

David D Lewis, Yiming Yang, Tony G Rose, and Fan Li. 2004. Rcv1: A new benchmark collection for text categorization research. *Journal of machine learning research*, 5(Apr):361–397.

Patrick Lewis, Barlas Oğuz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. 2019. Mlqa: Evaluating cross-lingual extractive question answering. *arXiv preprint arXiv:1910.07475*.

Tal Linzen. 2020. How can we accelerate progress towards human-like linguistic generalization? In *Proceedings of the Association for Computational Linguistics*.

Ryan Lowe, Nissan Pow, Iulian Serban, and Joelle Pineau. 2015. The ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. *arXiv preprint arXiv:1506.08909*.

Mitchell P Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. 1993. Building a large annotated corpus of english: The penn treebank. *Computational linguistics*.

Andrey Andreyevich Markov. 1906. Extension of the law of large numbers to dependent quantities. *Izv. Fiz.-Matem. Obsch. Kazan Univ.(2nd Ser)*, 15:135–156.

Winter Mason and Siddharth Suri. 2012. Conducting behavioral research on amazon's mechanical turk. *Behavior research methods*, 44(1):1–23.

Julian McAuley, Christopher Targett, Qinfeng Shi, and Anton Van Den Hengel. 2015. Image-based recommendations on styles and substitutes. In *Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval*, pages 43–52.

Warren S McCulloch and Walter Pitts. 1943. A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5(4):115–133.

Merriam-Webster. Crowdsourcing. In *Merriam-Webster.com dictionary*.

Paul Michel and Graham Neubig. 2018. Mtnt: A testbed for machine translation of noisy text. In *Proceedings of Empirical Methods in Natural Language Processing*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

George A. Miller. 1995. Wordnet: A lexical database for english. *COMMUNICATIONS OF THE ACM*, 38:39–41.

Taniya Mishra and Srinivas Bangalore. 2010. Qme!: A speech-based question-answering system on mobile devices. In *Conference of the North American Chapter of the Association for Computational Linguistics*.

Tom Mitchell. 1997. Introduction to machine learning. *Machine Learning*, 7:2–5.

Ethan Mollick and Ramana Nanda. 2016. Wisdom or madness? comparing crowds with expert evaluation in funding the arts. *Manag. Sci.*, 62:1533–1553.

Barzan Mozafari, Purnamrita Sarkar, Michael J. Franklin, Michael I. Jordan, and Samuel Madden. 2014. Scaling up crowd-sourcing to very large datasets: A case for active learning. *Proc. VLDB Endow.*, 8:125–136.

Mathias Müller, Annette Rios, Elena Voita, and Rico Sennrich. 2018. A large-scale test set for the evaluation of context-aware pronoun translation in neural machine translation. *arXiv preprint arXiv:1810.02268*.

Andrew Y Ng and Michael I Jordan. 2002. On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. In *Advances in neural information processing systems*, pages 841–848.

An T Nguyen, Matthew Lease, and Byron C Wallace. 2019. Explainable modeling of annotations in crowdsourcing. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*, pages 575–579.

Stefanie Nowak and Stefan Rüger. 2010. How reliable are annotations via crowdsourcing: a study about inter-annotator agreement for multi-label image annotation. In *Proceedings of the international conference on Multimedia information retrieval*, pages 557–566.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the Association for Computational Linguistics*, pages 311–318.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in pytorch. In *Conference on Neural Information Processing Systems: Autodiff Workshop: The Future of Gradient-based Machine Learning Software and Techniques*.

Vijayaditya Peddinti, Guoguo Chen, Vimal Manohar, Tom Ko, Daniel Povey, and Sanjeev Khudanpur. 2015. Jhu aspire system: Robust lvcsr with tdnns, ivector adaptation and rnn-lms. In *Automatic Speech Recognition and Understanding (ASRU), IEEE Workshop on*, pages 539–546.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of Empirical Methods in Natural Language Processing*.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proc. of NAACL*.

Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Nagendra Goel, Mirko Hannemann, Yanmin Qian, Petr Schwarz, and Georg Stemmer. 2011. The Kaldi speech recognition toolkit. In *IEEE Workshop on Automatic Speech Recognition and Understanding*.

David MW Powers. 1998. Applications and explanations of zipf's law. In *New methods in language processing and computational natural language learning*.

Raimon H. R. Pruim, Maarten Mennes, Daan van Rooij, Alberto Llera, Jan K. Buitelaar, and Christian F. Beckmann. 2015. Ica-aroma: A robust ica-based strategy for removing motion artifacts from fmri data. *NeuroImage*, 112:267–277.

Junfei Qiu, Qihui Wu, Guoru Ding, Yuhua Xu, and Shuo Feng. 2016. A survey of machine learning for big data processing. *EURASIP Journal on Advances in Signal Processing*, 2016(1):67.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the Association for Computational Linguistics*.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of Empirical Methods in Natural Language Processing*.

Juan Ramos. 2003. Using tf-idf to determine word relevance in document queries. In *Proceedings of the International Conference of Machine Learning*.

Siva Reddy, Danqi Chen, and Christopher D Manning. 2019. Coqa: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266.

Philip Resnik, Mari Broman Olsen, and Mona Diab. 1999. The bible as a parallel corpus: Annotating the 'book of 2000 tongues'. *Computers and the Humanities*, 33(1-2):129–153.

Philip Resnik and Noah A Smith. 2003. The web as a parallel corpus. *Computational Linguistics*, 29(3):349–380.

Stephen Robertson, Hugo Zaragoza, et al. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389.

Pedro Rodriguez, Joe Barrow, Alexander Miserlis Hoyle, John P. Lalor, Robin Jia, and Jordan Boyd-Graber. 2021. Evaluation examples are not equally informative: How should that change NLP leaderboards? In *Proceedings of the 59th*

Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 4486–4503, Online. Association for Computational Linguistics.

Pedro Rodriguez, Shi Feng, Mohit Iyyer, He He, and Jordan Boyd-Graber. 2019. Quizbowl: The case for incremental question answering. *arXiv preprint arXiv:1904.04792*.

Frank Rosenblatt. 1958. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386.

Ronald Rosenfeld. 2000. Two decades of statistical language modeling: Where do we go from here? *Proceedings of the IEEE*, 88(8):1270–1278.

Sebastian Ruder. 2016. An overview of gradient descent optimization algorithms. *arXiv preprint arXiv:1609.04747*.

Claude Sammut and Geoffrey I. Webb, editors. 2010. *Mean Squared Error*, pages 653–653. Springer US, Boston, MA.

Claude Elwood Shannon. 1948. A mathematical theory of communication. *Bell system technical journal*.

Claude Elwood Shannon, Warren Weaver, et al. 1949. mathematical theory of communication.

Kathryn Sharpe Wessling, Joel Huber, and Oded Netzer. 2017. MTurk Character Misrepresentation: Assessment and Solutions. *Journal of Consumer Research*, 44(1):211–230.

Elben Shira and Matthew Lease. 2010. Expert search on code repositories.

Jason Smith, Herve Saint-Amand, Magdalena Plamadă, Philipp Koehn, Chris Callison-Burch, and Adam Lopez. 2013. Dirt cheap web-scale parallel text from the common crawl. In *Proceedings of the Association for Computational Linguistics*, pages 1374–1383.

Matthew G Snover, Nitin Madnani, Bonnie Dorr, and Richard Schwartz. 2009. Terplus: paraphrase, semantic, and alignment enhancements to translation edit rate. *Machine Translation*, 23(2):117–127.

Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y Ng. 2008. Cheap and fast—but is it good?: Evaluating non-expert annotations for natural language tasks. In *Proceedings of Empirical Methods in Natural Language Processing*.

Wee Meng Soon, Hwee Tou Ng, and Daniel Chung Yong Lim. 2001. A machine learning approach to coreference resolution of noun phrases. *Computational linguistics*, 27(4):521–544.

Matthias Sperber, Graham Neubig, Jan Niehues, and Alex Waibel. 2017. Neural lattice-to-sequence models for uncertain inputs. In *Proceedings of the Association for Computational Linguistics*.

Siddharth Suri, Daniel G. Goldstein, and Winter A. Mason. 2011. Honesty in an online labor market. In *Proceedings of the 11th AAAI Conference on Human Computation*, AAAIWS'11-11, page 61–66. AAAI Press.

Jennifer EF Teitcher, Walter O Bockting, José A Bauermeister, Chris J Hoefer, Michael H Miner, and Robert L Klitzman. 2015. Detecting, preventing, and responding to "fraudsters" in internet research: ethics and tradeoffs. *Journal of Law, Medicine & Ethics*, 43(1):116–133.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. Fever: a large-scale dataset for fact extraction and verification. *arXiv preprint arXiv:1803.05355*.

Robert Tibshirani. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288.

AM Turing. 1950. Computing machinery and intelligence.

Donna Vakharia and Matthew Lease. Beyond mechanical turk: An analysis of paid crowd work platforms.

Dániel Varga, András Kornai, Viktor Nagy, László Németh, and Viktor Trón. 2007. Parallel corpora for medium density languages.

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

Ellen M Voorhees et al. 1999. The trec-8 question answering track report. In *Trec*, volume 99, pages 77–82.

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85.

Maja Vukovic and Claudio Bartolini. 2010. Towards a research agenda for enterprise crowdsourcing. In *International Symposium On Leveraging Applications of Formal Methods, Verification and Validation*, pages 425–434. Springer.

Eric Wallace, Shi Feng, Nikhil Kandpal, Matt Gardner, and Sameer Singh. 2019a. Universal adversarial triggers for attacking and analyzing nlp. In *Proceedings of Empirical Methods in Natural Language Processing*.

Eric Wallace, Pedro Rodriguez, Shi Feng, Ikuya Yamada, and Jordan Boyd-Graber. 2019b. Trick me if you can: Human-in-the-loop generation of adversarial examples for question answering. *Transactions of the Association for Computational Linguistics*, 7:387–401.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. In *Proceedings of Advances in Neural Information Processing Systems.*

Zeerak Waseem. 2016. Are you a racist or am i seeing things? annotator influence on hate speech detection on twitter. In *NLP+CSS@EMNLP.*

Mark E Whiting, Grant Hugh, and Michael S Bernstein. 2019. Fair work: Crowd work minimum wage with one line of code. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 7, pages 197–206.

Jake Ryland Williams, Paul R Lessard, Suma Desu, Eric M Clark, James P Bagrow, Christopher M Danforth, and Peter Sheridan Dodds. 2015. Zipf's law holds for phrases, not words. *Scientific reports*, 5:12209.

Ludwig Wittgenstein. 1953. *Philosophical Investigations.*

Marty J Wolf, Keith W Miller, and Frances S Grodzinsky. 2017. Why we should have seen that coming: comments on microsoft's tay "experiment," and wider implications. *The ORBIT Journal*, 1(2):1–12.

Stephen M Wolfson and Matthew Lease. 2011. Look before you leap: Legal pitfalls of crowdsourcing. *Proceedings of the American Society for Information Science and Technology*, 48(1):1–10.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144.*

Ikuya Yamada, Ryuji Tamaki, Hiroyuki Shindo, and Yoshiyasu Takefuji. 2018. Studio Ousia's quiz bowl question answering system. In *NIPS Competition: Building Intelligent Systems*, pages 181–194.

Kaiyu Yang, Klint Qinami, Li Fei-Fei, Jia Deng, and Olga Russakovsky. 2020. Towards fairer datasets: Filtering and balancing the distribution of the people subtree in the imagenet hierarchy. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 547–558.

Dani Yogatama, Cyprien de Masson d'Autume, Jerome Connor, Tomás Kociský, Mike Chrzanowski, Lingpeng Kong, Angeliki Lazaridou, Wang Ling, Lei Yu, Chris Dyer, and Phil Blunsom. 2019. Learning and evaluating general linguistic intelligence. *arXiv preprint arXiv:1901.11373.*

Omar Zaidan and Chris Callison-Burch. 2011. Crowdsourcing translation: Professional quality from non-professionals. In *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*, pages 1220–1229.

Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. 2018. Texygen: A benchmarking platform for text generation models. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 1097–1100.

George Kingsley Zipf. 1935. *The psycho-biology of language: An introduction to dynamic philology*, volume 21. Psychology Press.