Blow the Dog Whistle: A Chinese Dataset for Cant Understanding with Common Sense and World Knowledge

Canwen Xu[1]*, **Wangchunshu Zhou**[2]*, **Tao Ge**[3], **Ke Xu**[2], **Julian McAuley**[1], **Furu Wei**[3]
[1] University of California, San Diego [2] Beihang University [3] Microsoft Research Asia
[1]{cxu,jmcauley}@ucsd.edu [3]{tage,fuwei}@microsoft.com
[2]zhouwangchunshu@buaa.edu.cn,kexu@nlsde.buaa.edu.cn

Abstract

Cant is important for understanding advertising, comedies and dog-whistle politics. However, computational research on cant is hindered by a lack of available datasets. In this paper, we propose a large and diverse Chinese dataset for creating and understanding cant from a computational linguistics perspective. We formulate a task for cant understanding and provide both quantitative and qualitative analysis for tested word embedding similarity and pretrained language models. Experiments suggest that such a task requires deep language understanding, common sense, and world knowledge and thus can be a good testbed for pretrained language models and help models perform better on other tasks.[1]

1 Introduction

A cant[2] (also known as doublespeak, cryptolect, argot, anti-language or secret language) is the jargon or language of a group, often employed to exclude or mislead people outside the group (McArthur et al., 2018). Cant is crucial for understanding advertising (Dieterich, 1974) and both ancient and modern comedy (Sommerstein, 1999; Prasetyo, 2019). Also, it is the cornerstone for infamous dog-whistle politics (López, 2015; Albertson, 2015).

Here, we summarize the key elements for cant: (1) Both a cant and its reference (i.e., *hidden word*) should be in the form of common natural text (not another symbol system, e.g., Morse code). (2) There is some shared information between the cant users (i.e., *the insiders*) that is not provided to the people outside the group. (3) A cant should be deceptive and remain undetected to avoid being decrypted by people outside the group (i.e., *the outsiders*). These elements make the creation and understanding of cant subtle and hard to observe (Taylor, 1974). To the best of our knowledge, currently there are very few resources available for the research of cant.

In this paper, we create a dataset for studying cant, `DogWhistle`, centered around the aforementioned key elements (examples shown in Figure 1). We collect the data with a well-designed online game under a player-versus-player setting (see Section 3.1). The dataset includes abundant and diverse cant for a wide spectrum of hidden words. We find that cant understanding requires a deep understanding of language, common sense and world knowledge, making it a good testbed for next-generation pretrained language models. Our dataset also serves as a timely and complex language resource that can help models perform better on other tasks through Intermediate Task Transfer (Pruksachatkun et al., 2020).

2 Related Work

The use of cant has long been studied in linguistics research (Pei, 1973; Pulley, 1994; Albertson, 2006; Squires, 2010; Henderson and McCready, 2017, 2019b,a; Bhat and Klein, 2020). However, due to a lack of language resources, there are few studies in computational linguistics research. Henderson and McCready (2020) attempted to model the dog-whistle communications with a functional, agent-based method.

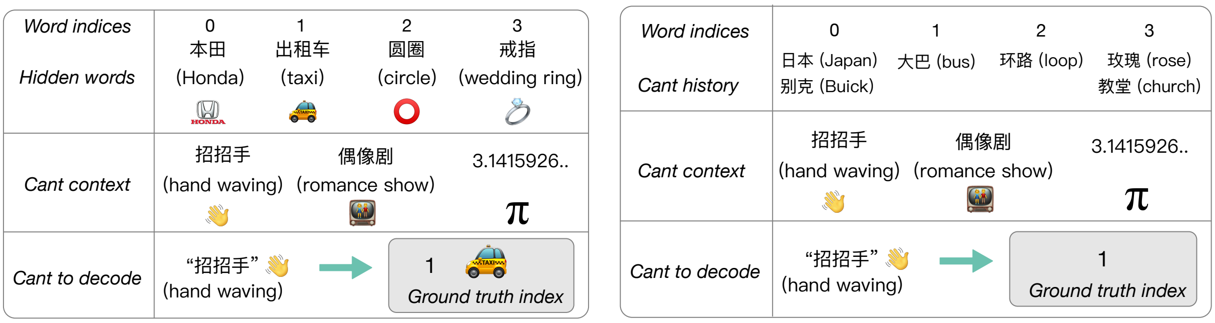
As a related topic in computational linguistics, some previous studies investigate coded names in human language. Zhang et al. (2014) analyzed and generated coded names of public figures. Zhang et al. (2015) designed an automatic system to decode the coded names. Huang et al. (2017) exploited a knowledge graph to identify coded names. Huang et al. (2019) leveraged multi-modal information to align coded names with their references.

*Equal Contribution. Work done at Microsoft Research Asia.

[1] The code is available at https://github.com/JetRunner/dogwhistle. The data and leaderboard are available at https://competitions.codalab.org/competitions/30451.

[2] https://en.wikipedia.org/wiki/Cant_(language)

Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 2139–2145
June 6–11, 2021. ©2021 Association for Computational Linguistics



Word indices Hidden words	0 本田 (Honda)	1 出租车 (taxi)	2 圆圈 (circle)	3 戒指 (wedding ring)
Cant context	招招手 (hand waving) 👋	偶像剧 (romance show) 📺		3.1415926.. π
Cant to decode	“招招手” 👋 (hand waving)	→	1 🚕 Ground truth index	

(a) *Insider* subtask. In this subtask, we mimic communication between insiders. The input (white background) is hidden words, cant context and a cant to decode. The model should output the index of the predicted hidden word (gray background). The hidden words are visible in this subtask.

Word indices Cant history	0 日本 (Japan) 别克 (Buick)	1 大巴 (bus)	2 环路 (loop)	3 玫瑰 (rose) 教堂 (church)
Cant context	招招手 (hand waving) 👋	偶像剧 (romance show) 📺		3.1415926.. π
Cant to decode	“招招手” 👋 (hand waving)	→	1 Ground truth index	

(b) *Outsider* subtask. In this subtask, an outsider tries to decrypt the communication by reading the cant history from previous rounds. The input is cant histories, cant context and a cant to decode (white background). The model should output the index of the predicted cant history (gray background). The hidden words are not visible in this subtask.

Figure 1: Input and output examples of the two subtasks of `DogWhistle`. See Appendix A for more examples.

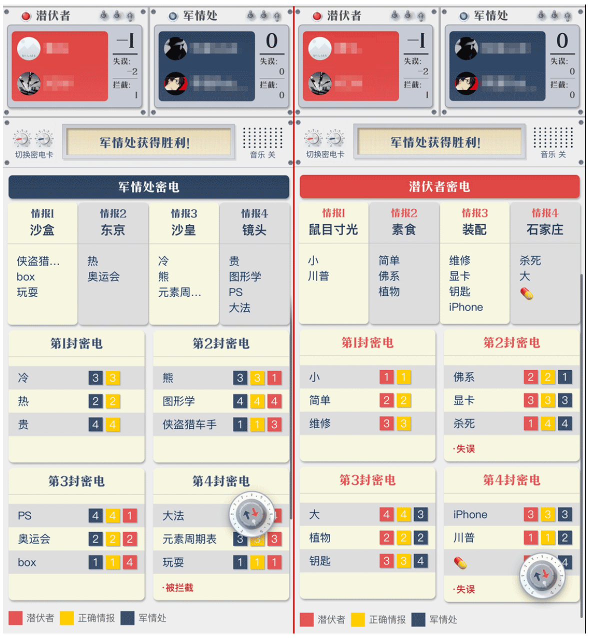
Our work differs from the above in the following ways: (1) Previous studies focused on coded names for public figures; the source and variety of these coded names is limited. The hidden words in our dataset are sampled from a common dictionary and are of high diversity. (2) The coded names in previous studies are used by users to bypass a censor (mostly a rule-based automatic text matching system). Conversely, our data are collected under an adversarial setting, pressuring users to mislead human adversaries. Thus, our work is ideal for evaluating recent progress on Natural Language Understanding (NLU) (Devlin et al., 2019; Lan et al., 2020; Liu et al., 2019; Sun et al., 2019b; Xu et al., 2020c; Zhou et al., 2020; Xu et al., 2020a).

3 Data Collection

Previous studies (Dergousoff and Mandryk, 2015; van Berkel et al., 2017) reveal that gamification can often improve the quality of collected data. Instead of collecting data from the wild like most datasets (Zhang et al., 2014, 2015; Xu et al., 2020b), we collect the data from historical game records of *Decrypto Online*, a well-designed online board game. The screenshot of the user interface is shown in Figure 2.

3.1 Game Design

The game design is adapted from the board game *Decrypto*.[3] Four players (e.g., A, B, C and D) are divided into two teams (e.g., A and B vs. C and D), with each trying to correctly interpret the cant presented to them by their teammates while cracking the codes they intercept from the opposing team.



潜伏者 [redacted] [redacted] −1 失误: −2 拦截: 1
军情处 [redacted] [redacted] 0 失误: 0 拦截: 0
切换密电卡 军情处获得胜利! 音乐 关

军情处密电

情报1 沙盒	情报2 东京	情报3 沙皇	情报4 镜头
侠盗猎… box 玩耍	热 奥运会	冷 熊 元素周…	贵 图形学 PS 大法

第1封密电
冷 3 3
热 2 2
贵 4 4

第2封密电
熊 3 3 1
图形学 4 4 4
侠盗猎车手 1 1 3

第3封密电
PS 4 4 1
奥运会 2 2 2
box 1 1 4

第4封密电
大法
元素周期表 3 3
玩耍 1 1 1
·被拦截

🟥 潜伏者 🟨 正确情报 ■ 军情处

潜伏者 [redacted] [redacted] −1 失误: −2 拦截: 1
军情处 [redacted] [redacted] 0 失误: 0 拦截: 0
切换密电卡 军情处获得胜利! 音乐 关

潜伏者密电

情报1 鼠目寸光	情报2 素食	情报3 装配	情报4 石家庄
小 川普	简单 佛系 植物	维修 显卡 钥匙 iPhone	杀死 大 💊

第1封密电
小 1 1
简单 2 2
维修 3 3

第2封密电
佛系 2 2 1
显卡 3 3 3
杀死 1 4 4
·失误

第3封密电
大 4 4 3
植物 2 2 2
钥匙 3 3 4

第4封密电
iPhone 3 3 3
川普 1 1 2
💊 4
·失误

🟥 潜伏者 🟨 正确情报 ■ 军情处

Figure 2: Screenshot of the user interface. The left and right halves of the screenshot are the screens for the two teams, respectively. The top section is the teams' scores. The middle section contains the hidden words and cant history. The bottom section is the cant to decode for each round.

In more detail, each team has their own screen, and in this screen there are four words numbered 0-3. Both players on the same team can see their own words while hiding the words from the opposing team. In the first round, each team does the following: One team member receives a randomly generated message that shows three of the digits 0-3 in some order, e.g., 3-1-0. They then give cant

[3] We recommend this video showing how to play the game: https://youtu.be/2DBg7Z2-pQ4

	train	dev	test
# games	9,817	1,161	1,143
# rounds	76,740	9,593	9,592
# word comb.	18,832	2,243	2,220
# uniq. words	1,878	1,809	1,820
# cant	230,220	28,779	28,776
avg. word len.	2.11	2.12	2.13
avg. cant len.	2.10	2.10	2.09

Table 1: Statistics of our collected `DogWhistle` dataset.

that their teammates must use to guess this message. For example, if A and B's four words are “本田” (Honda), “出租车” (taxi), “圆圈” (circle), and “戒指” (wedding ring), then A might say “招招手-偶像剧-3.14" (“hand waving”-“romance show”-3.14) and hope that their teammate B can correctly map those cant to 0-2-1. If B guesses incorrectly, the team would receive one “failure mark”.

Starting in the second round, a member of each team must again give a clue about their words to match a given three-digit message. One member from the other team (e.g., C) then attempts to guess the message. Taking Figure 1b as an example, based on the cant histories from previous rounds, C can roughly guess the code is 0-2-1. If C is correct, C and D would receive one “success mark”. After every round, the real messages that both teams were trying to pass will be revealed.

The rounds continue until a team collects either its second success mark (to win the game) or its second failure mark (to lose the game).

3.2 Additional Rules and Restrictions

The participants are explicitly asked not to create a cant based on its position, length, and abbreviation. That is to say, to mimic the creation of cant, we emphasize the importance of semantics instead of the morphology. To enforce this, all input that contains the same character as in one of the four words will be automatically rejected. As emojis have been playing an important role in online communications nowadays (Chen et al., 2019), emojis are allowed as valid input.

3.3 Data Cleaning and Split

For data cleaning, we remove all rounds with an empty cant. We also exclude rounds where the player fails to write a cant within the given time limit (one minute). We randomly split the data into training, development and test sets with an 8:1:1 ratio, such that all rounds of a game are in the same split. We also ensure there is no overlapping combination of hidden words between splits. We show the statistics of the training, development and test sets in Table 1. In contrast to 288k cant phrases for 1.9k hidden words in our dataset, data collected by previous studies (Zhang et al., 2014, 2015; Huang et al., 2017) are quite small, often containing hundreds of coded names for a small set of entities.

4 Experiments and Analysis

4.1 Task Formulation

As shown in Figure 1, we have subtasks named *insider* and *outsider*, respectively. For the *insider* subtask, we try to decode the cant to one of the hidden words. For the *outsider* subtask, the hidden words are invisible and the goal is to decrypt the messages based on the communication history. We formulate the task of decoding the cant in a similar format to multi-choice reading comprehension tasks (Lai et al., 2017; Zellers et al., 2018; Clark et al., 2018). We consider the cant context and the cant to decode as the “context” and “question” (respectively) as in multi-choice reading comprehension tasks. For the candidate answers, we use the hidden words and the set of cant histories for the insider subtask and the outsider subtask, respectively.

4.2 Baselines

Word Embedding Similarity Our task is naturally similar to the task of word similarity (Jin and Wu, 2012). We select FastText (Grave et al., 2018), SGNS (Li et al., 2018) (trained with mixed large corpus), DSG (Song et al., 2018) and VCWE (Sun et al., 2019a) as word embedding baselines. For each word embedding baseline, we first check if the cant is in the vocabulary; if it is not, we try to use a word tokenizer[4] to break it into words. If there is still any out-of-vocabulary token, we then break it into characters. For the *insider* subtask, we take the average of the word vectors to represent the cant and select the hidden word with the smallest cosine distance in the embedding space. For the *outsider* subtask, we take the average of the history cant for each hidden word as the representation. Then we predict the label by selecting the smallest distance

[4] We use Jieba, a popular Chinese tokenizer: `https://github.com/fxsjy/jieba`

Model	*Insider*		*Outsider*	
	dev	test	dev	test
Human Performance	87.5	88.9	43.1	43.1
Random Guessing	25.0	25.0	25.0	25.0
FastText (300D) (2018)	52.6	53.3	29.8	30.3
SGNS (300D,large) (2018)	52.3	52.3	30.6	30.8
DSG (200D) (2018)	56.3	56.2	31.4	31.4
VCWE (50D) (2019a)	46.0	46.2	28.0	28.0
BERT-base (2019)	73.5	74.1	33.7	33.7
RoBERTa-base (2019)	73.5	74.1	34.0	34.1
ALBERT-base (2020)	72.6	73.0	33.6	33.7
ERNIE-base (2019b)	73.4	73.9	34.0	34.1
RoBERTa-large (2019)	74.8	75.4	34.2	34.3
ALBERT-xxlarge (2020)	75.4	76.1	34.6	34.6

Table 2: Accuracy scores of human performance and baselines for the two subtasks of `DogWhistle`, *insider* and *outsider*. For word embedding baselines, the number of dimensions is marked, e.g., (300D).

between the representation of the cant and the history cant. Note that for word embedding baselines, the cant context is omitted and the evaluation is under a zero-shot setting (without any training).

Pretrained Language Models We use BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), ALBERT (Lan et al., 2020), and Baidu ERNIE (Sun et al., 2019b) as baselines.[5] The implementation is based on Hugging Face's Transformers (Wolf et al., 2020). Specifically, for the *insider* subtask, we construct the input sequence for each choice by concatenating its context, cant, and candidate hidden words with a special token `[SEP]`. We then concatenate the input sequences for all candidate hidden words with `[SEP]` and feed it into a BERT-like model. Finally, we use the hidden representation of the first token `[CLS]` to output the final prediction with a linear layer. For the *outsider* subtask, we replace the hidden words with the cant history. We fine-tune the models on the training set and report the results on the development and test sets. We use Adam (Kingma and Ba, 2015) with a learning rate searched over {2e-5, 3e-5, 5e-5} and a batch size of 64 to fine-tune the models for 3 epochs. We warm-up the learning rate for the first 10%

4.3 Quantitative Analysis

We show the experimental results in Table 2. For word embedding similarity baselines, DSG (Song et al., 2018), which is trained with mixed characters, words and n-grams on a diverse large corpus, drastically outperforms other word embeddings. For pretrained language models, large-size models, with more computational capacity, remarkably outperform base-size models on the *insider* subtask. Both RoBERTa-base and ERNIE-base outperform BERT-base while ALBERT-base, which employs parameter sharing, slightly underperforms BERT on both tasks. Notably, the best-performing model still trails human performance by a large margin of 12.8 and 8.5 on the *insider* and *outsider* subtasks, respectively. It indicates that `DogWhistle` is a very challenging dataset, providing a new battleground for next-generation pretrained language models.

4.4 Qualitative Analysis

We list some representative samples that BERT fails to predict but that are correctly predicted by human players in Table 3. For example #1, "Dancing Pallbearers"[6] is a recent meme that went viral after the release of the models. Thus, it is likely that the pretrained models have little knowledge about the subject. For example #2, "007" refers to James Bond films[7], in which the protagonist often cracks passwords in a mission. This kind of reasoning requires a high understanding of world knowledge instead of overfitting shallow lexical features, which has been pointed out as a major drawback in natural language inference (Poliak et al., 2018; Zhang et al., 2019). For example #3, "孩子都可以打酱油了" (the child can buy sauce) is a Chinese slang that means a child has grown up. To successfully predict this example, the model must have extensive knowledge of the language.

4.5 Intermediate-Task Transfer

Intermediate-Task Transfer Learning (Pruksachatkun et al., 2020) exploits an *intermediate* task to improve the performance of a model on the target task. As we analyzed before, `DogWhistle` contains rich world knowledge and requires high-level reasoning. Therefore, we can strengthen the ability of a model by leveraging our dataset

[5] The pretrained weights for BERT are from the official BERT repository: https://github.com/google-research/bert. Pretrained weights for other models are provided by CLUE: https://github.com/CLUEbenchmark/CLUEPretrainedModels.

[6] https://en.wikipedia.org/wiki/Dancing_Pallbearers

[7] https://en.wikipedia.org/wiki/James_Bond

	Hidden words	Cant context	Cant to decode	BERT	Human
#1	合作, 死神, 密码, 机械 cooperation, Grim Reaper, password, machinery	黑人抬棺, 007, 握手 Dancing Pallbearers, 007, handshaking	黑人抬棺 Dancing Pallbearers	密码 ✗ password	死神 ✓ Grim Reaper
#2	合作, 死神, 密码, 机械 cooperation, Grim Reaper, password, machinery	黑人抬棺, 007, 握手 Dancing Pallbearers, 007, handshaking	007	死神 ✗ Grim Reaper	密码 ✓ password
#3	破产, 日历, 轴, 熊孩子 bankruptcy, calendar, kids	酱油, 零, 字 sauce, zero, digits	酱油 sauce	日历 ✗ calendar	熊孩子 ✓ kids

Table 3: Some cases that BERT fails to predict but that human players predict correctly for the *insider* subtask.

Model	AFQMC		LCQMC	
	orig.	trans.	orig.	trans.
BERT-base (2019)	74.2	**74.5** (+0.3)	89.4	**89.7** (+0.3)
RoBERTa-base (2019)	73.8	**74.4** (+0.6)	89.2	**89.7** (+0.5)
RoBERTa-large (2019)	74.3	**74.8** (+0.5)	89.8	**90.0** (+0.2)

Table 4: Accuracy scores *(dev set)* of the original performance and intermediate-task transfer performance.

as an intermediate task. Specifically, we transfer `DogWhistle` for a semantic similarity task. We first fine-tune the models on the *insider* subtask, then re-finetune the models on two real-world semantic matching datasets, Ant Financial Question Matching Corpus (AFQMC) (Xu et al., 2020d) and Large-scale Chinese Question Matching Corpus (LCQMC) (Liu et al., 2018). As shown in Table 4, on both datasets, `DogWhistle` helps models significantly obtain better performance ($p < 0.05$).

5 Conclusion and Future Work

In this paper, we propose `DogWhistle`, a new Chinese dataset for cant creation, understanding and decryption. We evaluate word embeddings and pretrained language models on the dataset. The gap between human performance and model results indicates that our dataset is challenging and promising for evaluating new pretrained language models. For future work, we plan to leverage this dataset to train agents to compete against each other, to better understand verbal intelligence and teach agents to reason, guess and deceive in the form of natural language to make new progress at higher levels of World Scope (Bisk et al., 2020).

Ethical Considerations

During data collection, the game has a guideline that asks the players not to use any offensive content when playing the game. However, like all user-generated language resources, there would inevitably be bias and stereotyping in the dataset. We consider this as a double-edged sword, which provides opportunities for computational social science research of bias in human language, but also requires responsible use of these data. We would also like to warn that there would inevitably be potentially toxic or offensive contents in the dataset. Likewise, this dataset could be abused to generate dog-whistle phrases and political propaganda; Being aware of the risks, we have set terms to restrict the use to be for research purposes only.

Acknowledgments

We would like to sincerely thank Ren Wuming, the game developer of Decrypto Online, for his full support for this research. We appreciate all anonymous reviewers, especially the meta-reviewer, for their insightful comments. Canwen wants to thank all members of the Board Game Club of Microsoft Research Asia for the inspiration. Tao Ge is the corresponding author.

References

Bethany Albertson. 2006. Dog whistle politics, coded communication, and religious appeals. *American Political Science Association and International Society of Political Psychology. Princeton, NJ: Princeton University*.

Bethany L Albertson. 2015. Dog-whistle politics: Multivocal communication and religious appeals. *Political Behavior*, 37(1):3–26.

Prashanth Bhat and Ofra Klein. 2020. Covert hate speech: white nationalists and dog whistle communication on twitter. In *Twitter, the Public Sphere, and the Chaos of Online Deliberation*, pages 151–172. Springer.

Yonatan Bisk, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Chai, Mirella Lapata, Angeliki Lazaridou, Jonathan May, Aleksandr Nisnevich, Nicolas Pinto, and Joseph P. Turian. 2020. Experience grounds language. In *EMNLP*.

Zhenpeng Chen, Sheng Shen, Ziniu Hu, Xuan Lu, Qiaozhu Mei, and Xuanzhe Liu. 2019. Emoji-powered representation learning for cross-lingual sentiment classification. In *WWW*.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*.

Kristen K. Dergousoff and Regan L. Mandryk. 2015. Mobile gamification for crowdsourcing data collection: Leveraging the freemium model. In *CHI*, pages 1065–1074. ACM.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*.

Daniel J Dieterich. 1974. Public doublespeak: Teaching about language in the marketplace. *College English*, 36(4):477–481.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning word vectors for 157 languages. In *LREC*.

Robert Henderson and Elin McCready. 2019a. Dogwhistles and the at-issue/non-at-issue distinction. In *Secondary content*, pages 222–245. Brill.

Robert Henderson and Elin McCready. 2019b. Dogwhistles, trust and ideology. In *Proceedings of the 22nd Amsterdam Colloquium*.

Robert Henderson and Elin McCready. 2020. Towards functional, agent-based models of dogwhistle communication. In *PaM*.

Robert Henderson and Eric McCready. 2017. How dogwhistles work. In *JSAI-isAI Workshops*, volume 10838 of *Lecture Notes in Computer Science*, pages 231–240. Springer.

Longtao Huang, Ting Ma, Junyu Lin, Jizhong Han, and Songlin Hu. 2019. A multimodal text matching model for obfuscated language identification in adversarial communication? In *WWW*.

Longtao Huang, Lin Zhao, Shangwen Lv, Fangzhou Lu, Yue Zhai, and Songlin Hu. 2017. KIEM: A knowledge graph based method to identify entity morphs. In *CIKM*.

Peng Jin and Yunfang Wu. 2012. Semeval-2012 task 4: Evaluating chinese word similarity. In *SemEval@NAACL-HLT*.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *ICLR*.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard H. Hovy. 2017. RACE: large-scale reading comprehension dataset from examinations. In *EMNLP*.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. ALBERT: A lite BERT for self-supervised learning of language representations. In *ICLR*.

Shen Li, Zhe Zhao, Renfen Hu, Wensi Li, Tao Liu, and Xiaoyong Du. 2018. Analogical reasoning on chinese morphological and semantic relations. In *ACL*.

Xin Liu, Qingcai Chen, Chong Deng, Huajun Zeng, Jing Chen, Dongfang Li, and Buzhou Tang. 2018. LCQMC: A large-scale chinese question matching corpus. In *COLING*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

Ian Haney López. 2015. *Dog whistle politics: How coded racial appeals have reinvented racism and wrecked the middle class*. Oxford University Press.

Tom McArthur, Jacqueline Lam-McArthur, and Lise Fontaine. 2018. *Oxford companion to the English language*. Oxford University Press.

Mario Pei. 1973. Double-speak in america.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018. Hypothesis only baselines in natural language inference. In **SEM@NAACL-HLT*.

Ajeng Ramadhani Prasetyo. 2019. *Euphemism Used by Trevor Noah in Stand-up Comedy Show Trevor Noah: Son of Patricia (2018)*. Ph.D. thesis, Universitas Airlangga.

Yada Pruksachatkun, Jason Phang, Haokun Liu, Phu Mon Htut, Xiaoyi Zhang, Richard Yuanzhe Pang, Clara Vania, Katharina Kann, and Samuel R. Bowman. 2020. Intermediate-task transfer learning with pretrained language models: When and why does it work? In *ACL*.

Jerry L Pulley. 1994. Doublespeak and euphemisms in education. *The Clearing House*, 67(5):271–273.

Alan H Sommerstein. 1999. The anatomy of euphemism in aristophanic comedy. *Studi sull'Eufemismo, Bari: Levante*, pages 183–217.

Yan Song, Shuming Shi, Jing Li, and Haisong Zhang. 2018. Directional skip-gram: Explicitly distinguishing left and right context for word embeddings. In *NAACL-HLT*. Association for Computational Linguistics.

Lauren Squires. 2010. Enregistering internet language. *Language in Society*, 39(4):457–492.

Chi Sun, Xipeng Qiu, and Xuanjing Huang. 2019a. VCWE: visual character-enhanced word embeddings. In *NAACL-HLT*.

Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Xuyi Chen, Han Zhang, Xin Tian, Danxiang Zhu, Hao Tian, and Hua Wu. 2019b. Ernie: Enhanced representation through knowledge integration. *arXiv preprint arXiv:1904.09223*.

Sharon Henderson Taylor. 1974. Terms for low intelligence. *American Speech*, 49(3/4):197–207.

Niels van Berkel, Jorge Gonçalves, Simo Hosio, and Vassilis Kostakos. 2017. Gamification of mobile experience sampling improves data quality and quantity. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 1(3):107:1–107:21.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *EMNLP (Demos)*.

Canwen Xu, Tao Ge, Chenliang Li, and Furu Wei. 2020a. Unihanlm: Coarse-to-fine chinese-japanese language model pretraining with the unihan database. In *AACL-IJCNLP*.

Canwen Xu, Jiaxin Pei, Hongtao Wu, Yiyu Liu, and Chenliang Li. 2020b. MATINF: A jointly labeled large-scale dataset for classification, question answering and summarization. In *ACL*.

Canwen Xu, Wangchunshu Zhou, Tao Ge, Furu Wei, and Ming Zhou. 2020c. Bert-of-theseus: Compressing BERT by progressive module replacing. In *EMNLP*.

Liang Xu, Hai Hu, Xuanwei Zhang, Lu Li, Chenjie Cao, Yudong Li, Yechen Xu, Kai Sun, Dian Yu, Cong Yu, Yin Tian, Qianqian Dong, Weitang Liu, Bo Shi, Yiming Cui, Junyi Li, Jun Zeng, Rongzhao Wang, Weijian Xie, Yanting Li, Yina Patterson, Zuoyu Tian, Yiwen Zhang, He Zhou, Shaoweihua Liu, Zhe Zhao, Qipeng Zhao, Cong Yue, Xinrui Zhang, Zhengliang Yang, Kyle Richardson, and Zhenzhong Lan. 2020d. Clue: A chinese language understanding evaluation benchmark. In *COLING*.

Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. 2018. SWAG: A large-scale adversarial dataset for grounded commonsense inference. In *EMNLP*.

Boliang Zhang, Hongzhao Huang, Xiaoman Pan, Heng Ji, Kevin Knight, Zhen Wen, Yizhou Sun, Jiawei Han, and Bülent Yener. 2014. Be appropriate and funny: Automatic entity morph encoding. In *ACL*.

Boliang Zhang, Hongzhao Huang, Xiaoman Pan, Sujian Li, Chin-Yew Lin, Heng Ji, Kevin Knight, Zhen Wen, Yizhou Sun, Jiawei Han, and Bülent Yener. 2015. Context-aware entity morph decoding. In *ACL*.

Guanhua Zhang, Bing Bai, Jian Liang, Kun Bai, Shiyu Chang, Mo Yu, Conghui Zhu, and Tiejun Zhao. 2019. Selection bias explorations and debias methods for natural language sentence matching datasets. In *ACL*.

Wangchunshu Zhou, Canwen Xu, Tao Ge, Julian J. McAuley, Ke Xu, and Furu Wei. 2020. BERT loses patience: Fast and robust inference with early exit. In *NeurIPS*.