A Zero-Shot Transformer Model For an Attribute-Guided Challenge

Nicos Isaak[a]
Computational Cognition Lab, Cyprus

Keywords: Language Models, Deep Learrning, Story Understanding, Natural Language Processing, Word Guessing, Taboo Challenge, Zero-Shot Setting.

Abstract: The Taboo Challenge competition, a task based on the well-known Taboo game, has been proposed to stimulate research in the AI field. The challenge requires building systems able to comprehend the implied inferences between the exchanged messages of guesser and describer agents. A describer sends pre-determined hints to guessers indirectly describing cities, and guessers are required to return the matching cities implied by the hints. Climbing up the scoring ledger requires resolving the highest number of cities with the smallest number of hints in a specified time frame. Here, we present TabooLM, a language-model approach that tackles the challenge based on a zero-shot setting. We start by presenting and comparing the results of our approach with three studies from the literature. The results show that TabooLM achieves SOTA results on the Taboo challenge, suggesting that it can guess the implied cities faster and more accurately than existing approaches.

1 INTRODUCTION

Since the late fifties (McCarthy et al., 2006), numerous studies have investigated ways to develop systems that automate or enhance basic human abilities. Based on the McCarthy et al. (2006) proposal, where the term Artificial Intelligence (AI) was coined, the research community listed several topics to be tacked, such as neural nets, abstraction, NIP, etc. In this regard, various challenges have been proposed, such as the Turing test or the Winograd schema challenge, for developing systems humans can relate to and interact with. Given the above, researchers have increasingly become interested in identifying ways to endow machines with the necessary knowledge (Michael, 2013; Isaak and Michael, 2016) that would allow machines to perform as humans do.

One of those challenges, the Taboo Challenge Competition (TCC) (Rovatsos et al., 2018), is concerned with the ability to resolve hints to cities, which in certain cases is argued to require the use of commonsense knowledge. The challenge refers to games between guessers and describers, where a describer sends hints that need to be resolved to city names by the guessers. Like with the traditional Taboo game, the tricky part of the challenge is that it requires the guessers to speculate the domain of the describers so that players from the same region have better chances of getting higher scores. Given the above, the challenge requires the development of diversity-aware guessers to tackle games previously played with interactive decision-making strategies by human players. According to Rovatsos et al. (2018), although humans can easily solve Taboo-like games, the performance of automated approaches is still significantly lacking.

It is believed that the challenge will lead to the development of diversity-aware agents, something partially overlooked by other AI challenges. On another, the challenge allows the development of agents based on various solutions, such as classic AI, modern AI, or blended hybrid solutions that combine the best of two worlds.

Motivated by the difficulty of having AI agents tackle Taboo Games, we introduce TabooLM, a language model approach for the Taboo Challenge competition. To date, this is the first published work to report results on the feasibility of a zero-shot setting approach. According to our results, TabooLM outperforms the other systems in prediction accuracy, suggesting that it can guess the implied cities faster and more accurately than previously used approaches.

Our work shows in detail how out-of-the-box language models can be utilized in a zero-shot setting to tackle an inference kind of challenge by determining whether or not group words refer to the same entities, for which classic techniques necessitated the utiliza-

[a] https://orcid.org/0000-0003-2353-2192

Isaak, N.
A Zero-Shot Transformer Model For an Attribute-Guided Challenge.
DOI: 10.5220/0011782000003393
In *Proceedings of the 15th International Conference on Agents and Artificial Intelligence (ICAART 2023) - Volume 3*, pages 821-828
ISBN: 978-989-758-623-1; ISSN: 2184-433X

tion of various tools like dependency parsers, the removing of stop-words, or the use of similarity scores.

In the next section, we present the challenge itself by explaining how a describer interacts with guessers. Then, a comprehensive review of the related work is presented, along with the results of the first and only Taboo Challenge. After that, we present what Language Models are all about. In the follow-up section, we present our system architecture by analyzing its parts in detail, showing at the same time how they interact between them to tackle taboo games. Finally, we present our experimental evaluation and conclusions in the last two sections.

2 THE TABOO CHALLENGE

The challenge refers to Taboo-like games where players exchange request-response messages describing a concept without using taboo words, that is, words making the concept too easy to guess (Rovatsos et al., 2018). Instead of having humans, the challenge refers to a new type of game[1] played between machines. Specifically, the challenge consists of request-response messages between describer and guesser agents trying to guess city names. To illustrate, a game starts with a guesser sending an online request to a describer (see Figure 1). Next, a describer returns a hint referring to a city and waits for the guesser's response, that is, the implied city name. All subsequent responses to incorrect guesses contain the answer "no" and a new hint. On the other hand, the response to a correct guess is a simple yes. In case no more hints are available, the answer "no more hints" is returned.

The evaluation consists of running the guesser agents a predefined number of times, each time playing a different game. A guesser's score is derived by the number of guesses it submits (see game rules in Figure 1). If a guesser fails to answer, it obtains a score of number_of_hints+5. A guesser's total score is the sum of the individual game scores it played. The winner is the guesser with the total (lowest) score.

According to Rovatsos et al. (2018), to reduce time complexity:

- The domain of concepts is limited to include only popular cities —unknown to the guessers.
- The hints describing cities are limited to simple noun phrases —plus adjectives and/or adverbs.
- Each game had to be answered in the limited time frame of twenty minutes.

[1] https://www.essence-network.com

3 RELATED WORK

The first and only Taboo Challenge took part as a side event of IJCAI 2017 (Rovatsos et al., 2018). According to the organizers, a describer (server script) was programmed to interact with several guesser agents via an API. The hints and Taboo words were previously collected from games played by eighty-two crowd workers (based in the UK and USA). Starting from an initial set of 300 cities, they ended up with 226 cities, for which more than one worker generated eight to twelve taboo words —cities with no more than three taboo words were eliminated. Next, via a Web and a mobile application (see Figure 1), based on 226 cities, thirty native English speakers generated 283 games. Based on the collected hints, around 25% of the games were solved right after one hint, and around 50% used two to four hints. Finally, the challenge evaluation procedure was fully automated, and each guesser agent was tested on a subset of 109 games (see Table 1). Prior to the competition, participants were given access to several games in order to train their guesser agents.

Isaak and Michael (2017b) tackled the challenge through a commonsense reasoning system originally designed for the Winograd Schema Challenge. Given a predefined list of cities, for any hint, the system searches the English Wikipedia to build semantic scenes, that is, relations between nouns, verbs, and adjectives found in English sentences. Then, it feeds the scenes to a Learner and a Reasoner, and through chaining, it outputs logical inference rules (e.g., city name, LearnerWeight). Finally, the engine responds to a describer with the city with the bigger LearnerWeight. For instance, the scene, mahal([variable:0]):-mausoleum([variable:0, exists:0]), tells us that *mahal* is a mausoleum. Based on the results, the system won only six games out of 109 (5.50%) with 197 guesses and a total score of 816.

Dankers et al. (2017), via word embeddings, they associate hints with cities. The idea behind word embeddings is that words that appear in similar contexts tend to have related meanings. To build their embeddings' semantic space, they utilize and filter data from online sources such as Wikipedia, Wikivoyage, and NomadList. Basically, based on multiple game strategies and cosine similarity, they calculate and store the association between the vectors of hints and cities. In the end, they return the city with the highest score. In case a hint is not presented in the semantic space, they employ another strategy via searching pre-trained word2Vec vectors, previously trained on Google News. Based on the results, the system won thirteen games out of 109 (11.9%) with 293 guesses

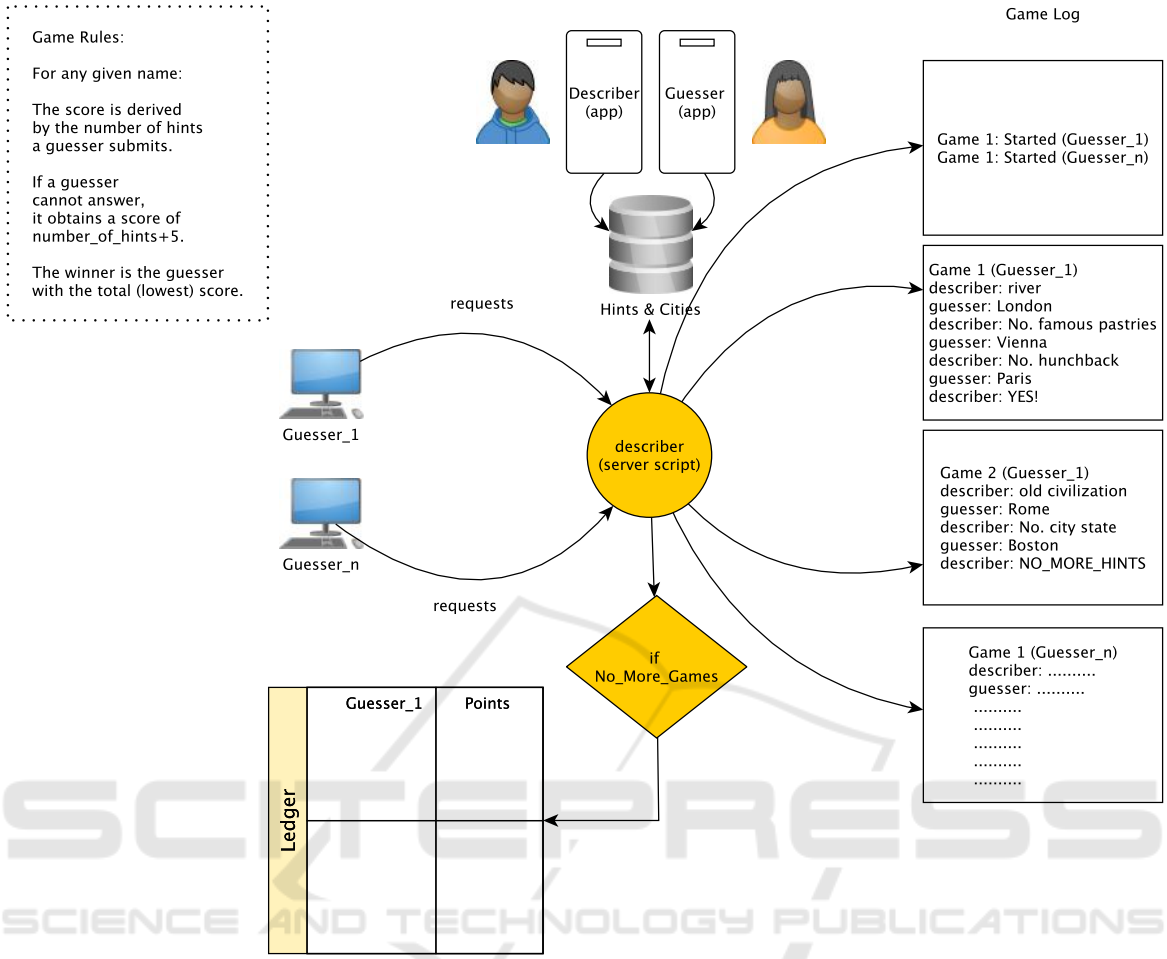

Figure 1: Data Collection and Evaluation of the Taboo Challenge Competition.

and a total score of 773.

Similar to Dankers et al. (2017)'s approach, Koksal (2017) tackles the challenge via similarities between hints and cities based on the pre-trained Skip-gram Word2Vec architecture. Moreover, instead of a simplified city-level approach, they employ a different country-level approach. In short, given a list of countries, for any hint, they calculate the country-hint similarity to return a list of the top three countries. Similarly, for each city, they estimate the city-hint similarity. To calculate the similarities between cities and hints, they use a heuristic approach that increases the likelihood of an answer by multiplying the similarity factor by each city population. According to the authors, this approach returns a country's capital or most famous city. According to the results, this system was ranked 1st by winning eighteen games out of 109 (16.5%) with 290 guesses and a total score of 745.

Our work differs from previous works mainly in three key aspects. Firstly, it tackles the challenge by utilizing an out-of-the-box language model approach. Secondly, it handles the challenge as a zero-shot classification problem without training or fine-tuning. Thirdly, it does not make us of standard NLP techniques, like removing stop-words, utilizing dependency parsing, or using any kind of similarity score to enhance its decision-making mechanism. The sections below explain each of these tasks along with our system's architecture.

Table 1: Results of the First and Only Taboo Challenge.

Team	Games Won	Guesses	Score
Koksal (2017)	18 (16.5%)	290	745
Dankers et al. (2017)	13 (11.9%)	293	773
Isaak and Michael (2017b)	06 (5.5%)	197	816

4 LANGUAGE MODELS

Language models (LMs) refer to neural approaches that use sizeable pre-trained models, mostly fine-tuned on downstream tasks, to maximize their performance. These models make it possible for machines to answer questions, write poems or music, and play games, sometimes even better than humans.

Language models such as BERT (Devlin et al., 2018), RoBERTa (Liu et al., 2019), GPT-2 (Radford et al., 2019), GPT-3 (Brown et al., 2020), DeBERTa (He et al., 2020), PaLM (Chowdhery et al., 2022), ChatGPT[2], or embeddings and architectures such as ELMO (Peters et al., 2018) and ULMFiT (Howard and Ruder, 2018) are increasingly important to the AI research community as they revolutionized the NLP field.

Basically, these models learn the probability of word occurrence in text sequences. From this standpoint, we can have large models with billions of parameters that can be either autoregressive, like GPT-3, or able to predict sequences of sentences or a missing token from a sentence, like BERT. According to the literature, the parameters of these models, an enhanced form of word embeddings, appear to store a form of knowledge that can help tackle various NLP tasks, such as question answering, pronoun resolution, text summarization, token classification, text similarity, and zero-shot classification tasks.

For instance, in a recent work, language models have been utilized to output knowledge to construct knowledge graphs (Wang et al., 2020). For another, in the Winograd Schema Challenge (WSC), a challenging task for pronoun resolution, we can use the embedded knowledge of language models to resolve pronouns.

All in all, language models seem to catch the relationships between words in sentences and phrases that can be used in various tasks. In the case of challenging NLP problems, the parameters of these models can be further fine-tuned to downstream tasks without having to train them from scratch. In this regard, models with millions or mostly billions of parameters can tackle challenges that good-old fashion AI (GOFAI) has struggled with for many years. However, we must keep in mind that LM's computational innards are so complex that nobody understands how they really work, meaning that they are not transparent solutions —their achievements do not seem to relate to a deeper understanding of the natural language text they are dealing with.

[2]http://chat.openai.com

4.1 Zero-Shot Classification

Zero-shot classification refers to techniques for applying language models to downstream tasks so that no further training or fine-tuning is needed. Although traditional zero-shot methods require providing some kind of descriptor to help the model predict the required task, the aim is to classify unseen classes (Xian et al., 2017; Romera-Paredes and Torr, 2015; Wei et al., 2021). In a simplified way, zero-shot learning refers to tasks when we have annotated data for only a limited subset of our classes. Given that language-model learning refers to a specified vector space (called semantic space), zero-shot refers to the ability of these models to match unseen classes to the seen classes' vector space, which acts as a bridging source mechanism.

According to Radford et al. (2019), when trained on large datasets, language models start learning various relations between text sequences without requiring explicit supervision. As a result, these models seem to have the necessary relations needed to tackle various tasks in zero-shot settings. Though not always necessary, in some cases, some instruction tuning is needed to minutely tailor the dataset to a specific zero-shot setting Wei et al. (2021) —e.g., large LMs like GPT-3 showed to perform better in few-shot than in zero-shot learning.

Nevertheless, in this paper, we approach the Taboo challenge as a natural language inference (NLI) problem (MacCartney, 2009), which can be tackled via a zero-shot setting approach. Below, we will show how language models trained on datasets for determining whether a premise and a hypothesis are connected via entailment or contradiction can be utilized in a zero-shot setting to tackle the Taboo challenge (see Figure 2).

5 SYSTEM ARCHITECTURE

We start by briefly discussing the main elements of our approach by presenting how it works and how it handles its semantics to guess cities given a predefined set of hints (see Figure 3).

Based on the constrained definition of the challenge, our system requests Taboo games from a describer, meaning lists of hints, one at a time, to return cities implied by the given hints. However, given that i) the Essence's describer service is no longer available, and ii) we only have access to the final 109 Taboo games from our previous system (Isaak and Michael, 2017b), TabooLM was designed to tackle games from a local-built service. In this regard, in its

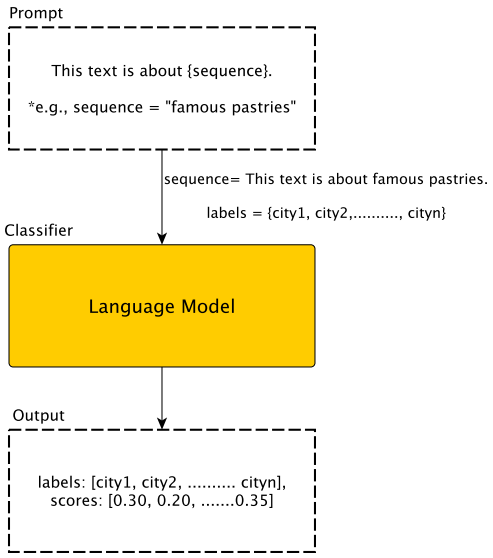

Figure 2: Zero-Shot Classification via Language Models. The model accepts a sequence —in some cases modified by a hypothesis_template such as "This text is about *sequence*", followed by a list of candidate labels. The result is a list of probabilities for entailment and contradiction for each label based on the hypothesis_template.

current version, TabooLM has access only to a testing set of 109 Taboo games. In this sense, from now on, when we refer to the local describer or guesser, we refer to components of our system's architecture — although they appear to be completely different, they are, ultimately, parts of the same system.

Moreover, with respect to the first Taboo challenge, the evaluation process provided convincing evidence of a strong association between each Taboo player's biases and beliefs, according to which pairs of hint-cities were collected. In this regard, to build TabooLM, we focused on utilizing large pre-trained language models[3] in a zero-shot setting. Knowing that no training set was available, our system allows access to a broad collection of language models and handles the describer's hints in various ways.

5.1 Loader

At first, our guesser starts its interaction with the local describer by requesting a list of Taboo games, meaning pairs of games, hints, and correct answers (see *part-1* in Figure 3). This interaction occurs each time no more hints are available for the current game.

[3]https://huggingface.co

5.2 Initializer

This component, which relates to initializing some key variables, takes part in every new game round. For instance, a new city list is initialized with zero weights in every new game. This is like a feature vector that, for each city-hint pair, holds the returned values of our zero-shot classification setting. Additionally, given that each city's country might improve our system's results (Koksal, 2017), the same process is repeated for each country (city)-hint pair (see part-2 in Figure 3).

5.3 Zero-Shot Evaluator

Part-3 of Figure 5 illustrates how our zero-shot setting works. To instruct the model to classify cities, we start by modifying a premise which in our case is a specified sequence of text in the form of "This text is about hint". The aim is to classify whether the premise entails the hypothesis, meaning a city name. For instance, given a list of hints ('tea', 'whiskey', 'kilt', 'crocodile') and a list of cities (e.g., 'Dundee', 'Athens'), the system instructs the model to predict how much the premise "This text is about tea" relates to each city. In short, for each city, the system automatically runs the classifier with the same premise to output a numerical value in the range of 0-1. The classifier runs either with the flag true_labels "true" or "false", where in the case of the former more than one city could be true.

Additionally, in each game, the system either runs the classifier for each hint independently or with an aggregation mechanism in the form of hint+=hint. To illustrate, in the second run —for the second hint "whiskey', the system either runs the premise "This text is about whiskey" or combined with the first hint as "This text is about tea, whiskey", against each city to output a probability list (e.g., ['Dundee', 0.05485], ['Athens', 0.0029]).

Afterward, via an API, the system loads each city's country and repeats the same procedure (e.g., ['UK', 0.30], ['Greece', 0.10], ...). In the final step, it matches cities and countries, adds their probability scores, and produces a final city list (e.g., ['Dundee', 0.35485], ['Athens', 0.1029], ...). The variability in probability values generally stems from the relation between the hypothesis and the premise values. In every game, for every new hint, the system has the option to either initialize or accumulate the weights.

In order to represent the cities according to their significance (see part-4 in Figure 3), the system sorts the city list in descending order with the highest probability values at the start. Once the sorting is made, it

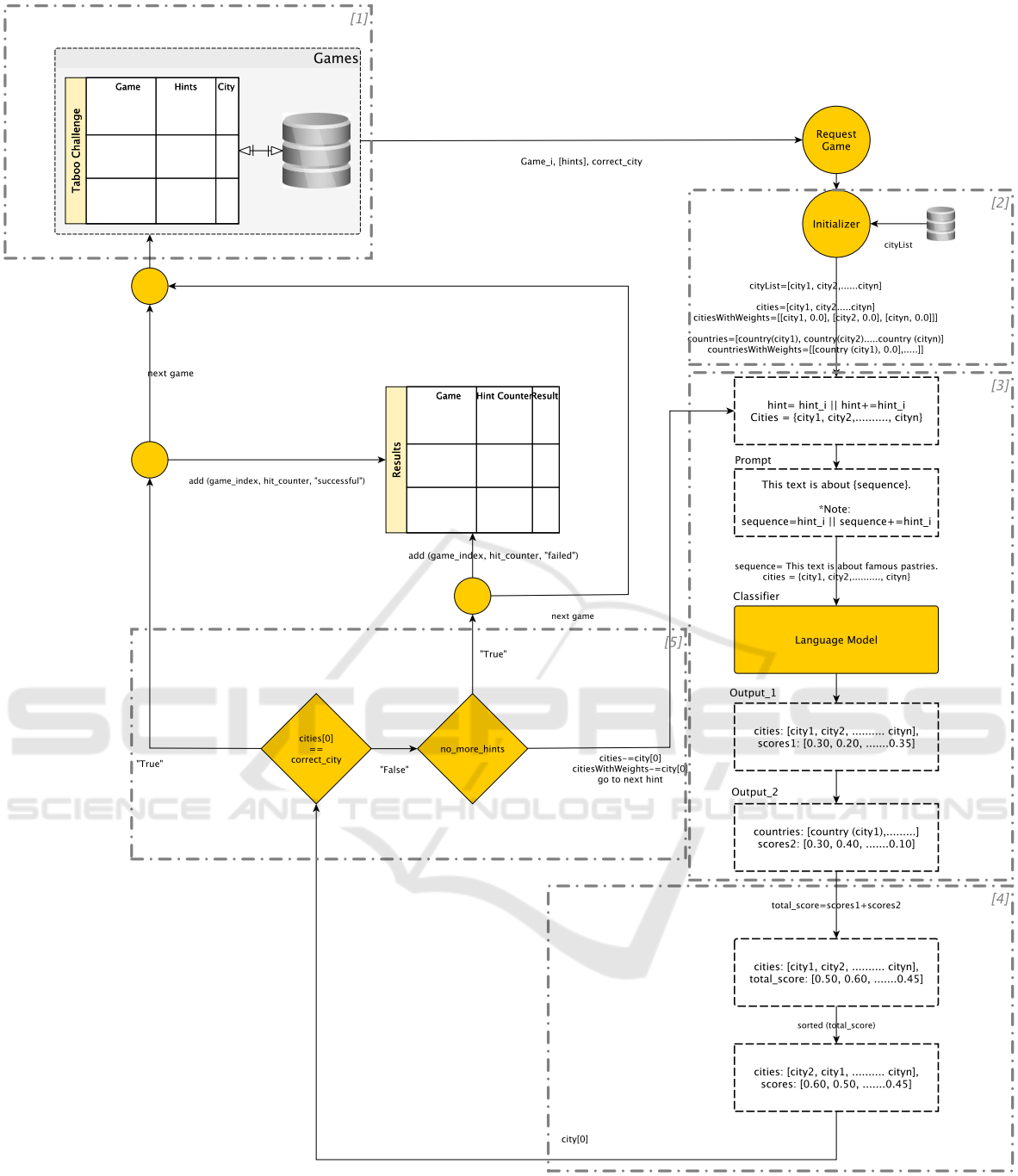

Figure 3: Full Pipeline of TabooLM: A Zero-Shot Language Model for a Word Guessing Game. The reasoning process starts with each hint, modified by the hypothesis_template: This text is about *hint*. The various parts of the architecture are marked with dashed rectangles and are discussed in Section 5.

sends the first city as the answer to the given hint.

Once the city is sent, the local describer replies with a successful or an unsuccessful message (see part-5 of Figure 3). In case we were able to guess the city correctly, the system adds the results to a local scoring file in the form of *game, hint, Flag*, where game and hint refer to the current game and hint indexes, and Flag to whether we were successful or unsuccessful. On the other hand, if we cannot predict the correct city, the system proceeds to the following actions: Firstly, in case no more hints are available, it adds the results to the local scoring file and moves to

the next game. Secondly, if more hints are available, it proceeds to the next hint and removes the city from the current city list along with its probability value — a wrongfully guessed city cannot be the answer to the current game.

As an example of this, consider the case of sending the city of Dundee from the final list (e.g., ['Dundee', 0.35485], ['Athens', 0.1029]). In case of an unsuccessful guess, i) the city of Dundee is removed, and ii) the system proceeds to the next hint with the left cities (e.g., ['Athens', 0.1029], ...]).

6 EXPERIMENTAL EVALUATION

This section presents the results obtained by applying the methodology described in this paper. We describe the design along with the results of the experiments we undertook to evaluate the system performance on guessing cities from a predefined list of 109 games.

To evaluate the zero-shot performance setting, we started by investigating whether our system could produce similar or better results than previously used systems Rovatsos et al. (2018). For the purposes of this experiment, the *weights' initialize* Flag was enabled, meaning that with every new hint, city and country weights were initialized with zero values. Furthermore, within every game, each hint was added to the previous one in the form of hint+=next_hint.

Our experiments ran under the DeBERTa-V3 model (He et al., 2021; Laurer et al., 2022), fine-tuned on the multiNLI, adversarial-NLI (ANLI), fever-NLI, lingNLI, and wanli datasets. DeBERTa V3 is an enhanced version of DeBERTa, which combines BERT and RoBERTa models in an efficient novel way. In short, each model, from BERT, RoBERTa, DeBERTa, and DeBERTa V2 to DeBERTa V3, can be considered an enhanced version of the previous one. Finally, all datasets combined result in 885.242 premise-hypothesis pairs.

6.1 Results and Discussion

The general picture emerging from the analysis is that TabooLM correctly tackled 53 games out of 109, achieving a success rate of 49% (see Table 2). A cursory glance at Table 2 reveals that our approach significantly outperformed all previously used systems. This is in line with recent results where language models significantly outperformed other methods in various NLP tasks Brown et al. (2020).

A more detailed analysis of our results shows that TabooLM not only tackled more games but it also achieved the best score of 417 points —recall that less is more. Specifically, it outperformed Koksal (2017)'s system by 330 points, Dankers et al. (2017)'s by 358 points, and finally, our previous approach by 301 points (Isaak and Michael, 2017b). An interesting finding was that it achieved the best score with a minimum number of 267 guesses, which is very important considering the challenge difficulties (see Table 2). Moreover, compared to our previous work, which due to its reasoning engine, was unable to answer all the games as it timed out frequently (Rovatsos et al., 2018), TabooLM tackled 109 games in 25 minutes — experiments ran under an NVIDIA Tesla K80 GPU. Specifically, compared to an average time of 20 minutes for each game, TabooLM was found to be 77% faster than our previously used approach.

These findings are less surprising if we consider not only the advances of deep learning approaches in the NLP field but also recent work showing ways to utilize LMs to output semantic relations to help tackle NLP tasks (Wang et al., 2020). It seems that LMs could be utilized in zero-shot settings to achieve state-of-the-art results (Radford et al., 2019).

Further experiments we undertook revealed that the capacity of the language model in our zero-shot setting relates to its training size. For instance, we further analyzed the relationship between models trained on different datasets and their success in the Taboo challenge competition. The data provide convincing evidence of a link between accuracy and the variety of the training datasets. In short, as the training data increases, we can get better semantics for better word guessing. This is in line with other work in which leveraging larger language models or training with larger datasets improve system performance (Radford et al., 2019; Isaak and Michael, 2017a). A cursory look at Table 3 reveals that as the number of training datasets increases, the number of unanswered games decreases. It seems that the increase in accuracy was due to the increase in the number of training datasets, meaning that a variety of training datasets limits the situations where a hint-city relation is unlike anything an LM has met in the training phase (see Table 3).

Table 2: Results of TabooLM Compared to Systems Participated in the First Taboo Challenge).

Team	Games Won	Guesses	Score
TabooLM	53 (49%)	267	417
Koksal (2017)	18 (16.5%)	290	745
Dankers et al. (2017)	13 (11.9%)	293	773
Isaak and Michael (2017b)	06 (5.5%)	197	816

Table 3: Results of TabooLM Based on Various Models.

Model	Games Won
facebook/bart-large-mnli	27
bart-large-mnli-yahoo-answers	38
DeBERTa_v3_large_mnli-_fever_anli_ling-_wanli	53

7 CONCLUSION

We have shown TabooLM, a system that takes queries in a city-hint format and utilizes language models in a zero-shot setting to return ranked lists of cities implied by the given hints. Given a list of hints, it iterates from top to bottom and matches those hints with popular cities worldwide. Although it was built explicitly for the Taboo challenge competition, the system can be used with any task involving a word-guessing problem.

Compared to previous work, the results provide convincing evidence that our system can achieve state-of-the-art results. In this regard, the results suggest that solutions utilizing language models in a zero-shot setting can be used to tackle challenging NLP tasks. However, given that the computational innards of these kinds of models are complex, further gains could be achieved via transparent solutions that employ additional semantic analysis of city-hint pairs.

Future studies could blend both modern and classic AI in order to build transparent hybrid solutions. Among possible directions, systems that construct the building of knowledge graphs from language models could offer a better solution.

REFERENCES

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H. W., Sutton, C., Gehrmann, S., et al. (2022). Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.

Dankers, V., Bilgin, A., and Fernández, R. (2017). Modelling word associations with word embeddings for a guesser agent in the taboo city challenge competition.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

He, P., Gao, J., and Chen, W. (2021). Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing.

He, P., Liu, X., Gao, J., and Chen, W. (2020). Deberta: Decoding-enhanced bert with disentangled attention.

Howard, J. and Ruder, S. (2018). Universal language model fine-tuning for text classification.

Isaak, N. and Michael, L. (2016). Tackling the Winograd Schema Challenge Through Machine Logical Inferences. In Pearce, D. and Pinto, H. S., editors, *STAIRS*, volume 284 of *Frontiers in Artificial Intelligence and Applications*, pages 75–86. IOS Press.

Isaak, N. and Michael, L. (2017a). How the Availability of Training Material Affects Performance in the Winograd Schema Challenge. In *Proceedings of the (IJCAI 2017) 3rd Workshop on Cognitive Knowledge Acquisition and Applications (Cognitum 2017)*.

Isaak, N. and Michael, L. (2017b). Tackling the taboo challenge with machine logical inferences. page 46.

Koksal, A. (2017). Skip-gram model for simulation of taboo game.

Laurer, M., van Atteveldt, W., Casas, A., and Welbers, K. (2022). Less annotating, more classifying–addressing the data scarcity issue of supervised machine learning with deep transfer learning and bert-nli.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

MacCartney, B. (2009). *Natural language inference*. Stanford University.

McCarthy, J., Minsky, M. L., Rochester, N., and Shannon, C. E. (2006). A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence, August 31, 1955. *AI magazine*, 27(4):12–12.

Michael, L. (2013). Machines with Websense. In *Proc. of 11th International Symposium on Logical Formalizations of Commonsense Reasoning (Commonsense 13)*.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018). Deep contextualized word representations.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. (2019). Language models are unsupervised multitask learners.

Romera-Paredes, B. and Torr, P. (2015). An embarrassingly simple approach to zero-shot learning. In Bach, F. and Blei, D., editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 2152–2161, Lille, France. PMLR.

Rovatsos, M., Gromann, D., and Bella, G. (2018). The Taboo Challenge Competition. *AI Magazine*, 39(1):84–87.

Wang, C., Liu, X., and Song, D. (2020). Language models are open knowledge graphs.

Wei, J., Bosma, M., Zhao, V. Y., Guu, K., Yu, A. W., Lester, B., Du, N., Dai, A. M., and Le, Q. V. (2021). Finetuned language models are zero-shot learners.

Xian, Y., Schiele, B., and Akata, Z. (2017). Zero-shot learning-the good, the bad and the ugly. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4582–4591.