Predicting Word Choice using Distributional Information in a Cooperative Language Game

Abhilasha A. Kumar, Mark Steyvers, David A. Balota

Abstract

Simple language games are considered central to understanding contextual language use, although the role of distributional information in predicting word choice and cooperative interactions in language games remains relatively understudied. We introduce a novel two-player language game, *Connector*, based on the Codenames boardgame, and evaluate the performance of three embedding models in the game. Our results indicate that embedding models effectively capture search processes and also predict shifts in word choice based on distributional information, although there are limits to the explicit predictive power of pure embedding models. These results highlight the different sources of information (distributional, perceptual, categorical, etc.) that speakers and listeners recruit to generate novel word associations in a cooperative setting and indicate that distributional models trained on purely linguistic corpora may be limited in the extent to which they can model this behavior.

1 Introduction

Language games represent a useful tool to investigate contextual language use and have been used to assess the performance of artificial language systems and neural agents (Ferrucci et al., 2010; Hill et al., 2018). Complex language games such Codenames (Chvátil, 2016) represent a unique opportunity to study cooperative interactions and associative reasoning, two relatively understudied aspects of language. We introduce a novel two-player cooperative language game, *Connector*, based on the boardgame Codenames, to study how players use contextual and distributional information during the game. To enable players to freely search their semantic space, we do not place any hard constraints on word choice, allowing us to track natural search and retrieval processes in the game. Additionally, we investigate indirect referential communication and cooperative interaction through multiple attempts at retrieving the correct answers in the game. Finally, we evaluate the performance of three embedding models (word2vec, fastText, and GloVe) in accounting for player behavior in the *Connector* game.

2 Related Work

Shen et al., (2018) recently adapted Codenames to evaluate the role of associational information in predicting responses in a simple reference game. Shen et al. found that bigram collocations (derived via Google Ngram probabilities) were more predictive of word choice, compared to word embeddings (derived from word2vec). However, Shen et al.'s task limited word choice through a predetermined set of clues, and also did not examine direct player interactions, both of which were primary goals of the current work. In another related study, Xu and Kemp (2010) examined observational data from the television game show *Password*. They modeled word choice using forward and backward associations derived from free association norms (Nelson et al., 2004) and found that player responses were cooperative and calibrated. However, Xu and Kemp did not examine the role of distributional or lexical information in predicting responses in the game. Therefore, the present work adds to previous studies by directly examining real-time player interactions in an unconstrained game task and evaluating the role of distributional information in modeling word choice.

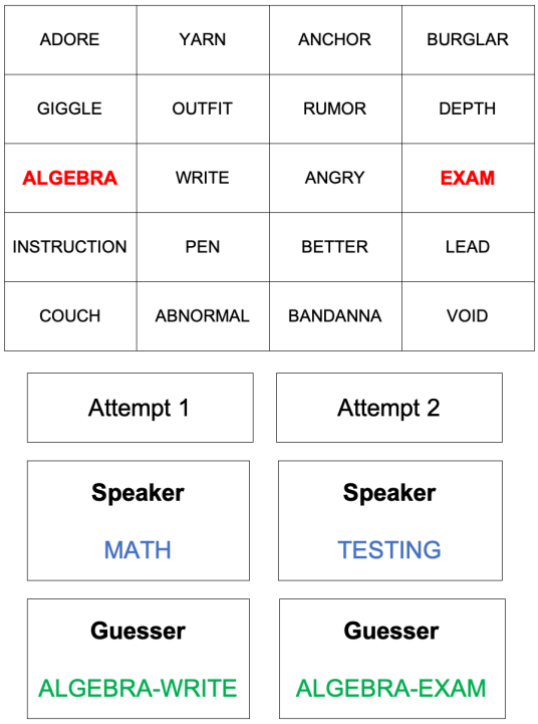

Figure 1. The *Connector* game. Speakers and guessers view a 20-item board. Speakers are provided a word-pair from the board (highlighted in red only for the speaker) and come up with a one-word clue. Guessers receive the clue and guess the word-pair the clue refers to on the board. Speakers can provide two additional clues if the first attempt is unsuccessful (only two attempts are shown here).

3 The Connector Game

The *Connector* game is a modified version of the boardgame Codenames (Chvátil, 2016). As shown in Figure 1, two players viewed a 20-item board and were randomly assigned the role of the speaker or guesser. Speakers and guessers were introduced to each other in-person before the game, after which they played the game in adjoining rooms. On each trial, the speaker was presented with a word-pair from the board (e.g., *exam-algebra*) and asked to freely generate a one-word clue corresponding to the word-pair. This clue (e.g., *math*) was then communicated to the guesser, who selected the two words it most likely referred to on the 20-item board (e.g., *algebra-write*). If the first attempt was unsuccessful, speakers could provide two additional clues to the guesser. Three word-pairs of varying difficulty were selected for clue generation and guessing on each board, and players viewed 10 boards throughout the game, resulting in 30 trials per dyad.

We conducted two experiments for which we recruited 156 young adults in dyads from Washington University in St. Louis, who were compensated via course credit for their participation. Data from both experiments was largely consistent, and we therefore report results combined across both experiments. We make all data and analysis code freely available[1]. Speakers and guessers successfully played the game, with an overall success rate of 85% across three attempts. Table 1 provides some examples of clues generated by the speaker in the game.

4 Models of Word Choice

We examined the predictive power of three embedding models, GloVe (Pennington et al., 2014) skip-gram trained word2vec (Mikolov et al., 2013) and fastText (Bojanowski et al., 2017) in predicting speaker and guesser word choice in the *Connector* game. All models were pretrained on Wikipedia corpora and produced 300-dimensional vector representations available from the *magnitude* python package (Patel et al., 2018) and Wikipedia2Vec (Yamada et al., 2018). To calculate word similarity, we computed cosines between vector representations.

5 Modeling the Search Space

To limit the search space of the models to a tractable yet large set of words in the game, we constrained the search space of all models to the topmost 10,000 primary free association responses in the Small World of Words (SWOW; De Deyne et al., 2019) database. This database contains free association responses to over 12,000 cue words collected from over 80,000 participants. We used

Word Pair	Top 3 first clues (frequency)
exam – algebra	math (22), test (3), school (2)
giggle – abnormal	snort (4), funny (3), laugh (3)
breeze – bubble	blow (7), air (5), float (3)
toes – dracula	blood (5), vampire (4), count (3)
war – quiet	peace (5), fight (3), ceasefire (2)

Table 1: Examples of clues provided by the speaker

[1] Uploaded as .zip file, will share via github in final version

the SWOW database as the potential search space for all models because the speaker and listener tasks are similar to associative retrieval from semantic memory. Therefore, responses in SWOW database are likely to include most responses produced in the game. Indeed, across both experiments, 80% of all clues provided by the speakers and 95% of the words on the boards fell within the 10,000-word space. Several of the remaining clues were over one word (e.g., *raisinbran*, *outcropping*, etc.), infrequent (e.g., *photosynthesis*, *pantomime*, etc.), or proper names (e.g., *quercus*, *riverdale*, etc.) and model predictions for these clues were treated as missing. For all word-pairs, word vector representations were averaged to get a composite word-pair vector (i.e., $v_{exam\text{-}algebra} = (v_{exam} + v_{algebra})/2$). We examined the closest words to this composite vector within the 10,000 words, ranked by cosine distance within each model[2].

6 Scoring

6.1 Speaker Task Scoring

For the speaker task of generating a clue for a given word-pair on the board, three scores were computed: *top answer*, *top distance*, and *rank correlation*. The *top answer* corresponded to whether the top prediction by the model matched with the most frequent response among speakers, scored as a 0 or 1. The *top distance* corresponded to the standardized distance of the top model prediction to the most frequent response, to evaluate how "close" the models were to speaker responses. The *rank correlation* corresponded to the Kendall's tau correlation between the ranks of speaker responses, sorted by frequency, and the ranks of these responses in each model, sorted by cosine distances, to evaluate general symmetry in ranking of responses beyond the top prediction.

Word Pair	Top Second Clue (Frequency)	
Word 1 – Word 2	**Word 1 incorrect**	**Word 2 incorrect**
exam – algebra	test (2)	number (1)
giggle – abnormal	laugh (2)	weird (7)
breeze – bubble	wind (3)	bath (2)
toes – dracula	feet (5)	vampire (2)
war – quiet	battle (1)	silence (5)

Table 2: Examples of second clues provided by the speaker based on which word was not guessed in the first attempt for selected word-pairs in the *Connector* game

[2] We also computed a non-linear measure, i.e., $m = min(d(v_{exam}, c)_2 + d(w_{algebra}, c)_2)$, where $c \in$ SWOW, which resulted in similar predictions

6.2 Guesser Task Scoring

For the guesser task of selecting two words on the board, given a clue, we obtained the topmost two guesses predicted by each model based on largest cosine of each word on the board from the clue provided by the speaker. These predictions were then compared to the guesser's actual responses and scored as 0 (for no matches with the guesser's responses), 1 (for a single word match), or 2 (for both words matching). The proportion of correct predictions for each model was averaged to obtain a *guess score* per model.

6.3 Second Attempt Scoring

For each failed first attempt, the speaker provided at most two additional clues to the guesser. As shown in Table 2, the second clue was indeed influenced by initial guesser responses, such that when the guesser correctly identified one word, the second clue was closer to the incorrect word. To evaluate which model best captured this pattern, given a particular failed guess (e.g., *algebra-write*), we computed new composite word vectors for each model, with an increased weight for the unguessed word (e.g., *exam*) in the composite (i.e., $2/3\ v_{exam} + 1/3\ v_a$). Predictions were then computed as before, to yield *top answer*, *top distance*, and *rank correlations* for each model. We excluded word-pairs that did not have a clear ranking in responses in computing rank correlations.

7 Results

Table 3 displays the performance of the models for the speaker task in the first attempt. As shown,

Model	Speaker			Guesser
	Top Answer (SEM)	Top Distance (SEM)	Rank Correlation (SEM)	Guess Score (SEM)
fastText	**.08 (.04)**	**.93 (.04)**	**.17 (.07)**	**.59 (.04)**
word2vec	.08 (.04)	.86 (.49)	.14 (.06)	.54 (.03)
GloVe	.10 (.04)	.82 (.07)	.14 (.07)	.56 (.03)

Table 3: First attempt performance in the *Connector* game

fastText's predictions were closest to the most frequent response given by speakers, and also correlated most with the speaker ranked responses, followed by word2vec and GloVe. fastText also outperformed word2vec and GloVe models in predicting guesser responses. In the second attempt, as shown in Table 4, fastText again best predicted the top second clue, although the results for the rank correlations and top distances were slightly noisier due to fewer data points.

Model	Top Answer (SEM)	Top Distance (SEM)	Rank Correlation (SEM)
fastText	**.18 (.04)**	**0.86 (.06)**	**0.28 (.05)**
word2vec	.09 (.29)	0.87 (.05)	0.30 (.05)
GloVe	.13 (.33)	0.87 (.05)	0.30 (.05)

Table 4: Second attempt performance in the *Connector* game

8 Discussion

We introduced a novel language game, *Connector*, and showed that distributional information derived from fastText embeddings were most predictive (and closest) to speaker and guesser responses. fastText enriches its learning through character-level information in the form of n-grams (Bojanowski et al., 2017), which likely improves the model's ability to infer word relationships. Additionally, it is possible that fastText is likely more sensitive to other lexical variables, such as word frequency (Scarborough et al., 1977). To investigate this possibility, we obtained spoken and textual word frequency estimates from the English Lexicon Project (Balota et al., 2007) for each first clue. These estimates were correlated with model ranks and compared to the correlation between speaker clue ranks and frequency. Speaker clue ranks were moderately correlated with frequency (r=-.12), and fastText's rank frequency correlations aligned closest to this estimate (r = -.10), followed by word2vec (r = -.07), and GloVe (r = -.24). Importantly, GloVe overestimated word frequency and word2vec underestimated frequency in the rankings. Therefore, it appears that fastText's learning mechanism was better at calibrating word frequency against semantic information, and therefore better captured how speakers use word frequency to guide their search, although these results do require more detailed investigation.

Another finding from this work was that model accuracy in predicting speaker responses was relatively low (~10-20%), compared to guesser responses (~55%) suggesting that speakers use strategies and relationships that are difficult for embedding models to capture. One example is the word-pair *sun-bowl*, where over 50% of the speakers chose the clue *round*, successfully inferring a perceptual relationship between the words. However, none of the embedding models predicted this clue, and predictions leaned more towards the word *sun* (e.g., *warm*, *peach*, *moon*, etc.). A second example is the word-pair *happy-sad*, where over 80% of the speakers chose the clue *emotion*, successfully inferring a superordinate category relationship between the words. The embedding models predicted clues related to only one of the words, but not both (e.g., *glad*, *joyful*, *miserable*, *sorry*, etc.). These examples highlight how individuals use different semantic relations (perceptual, superordinate, etc.) to generate associations between seemingly unrelated words and embedding models may be disadvantaged at inferring such associations due to training on non-hierarchical linguistic corpora.

A third important result from this work is that speakers were able to adjust their subsequent clues based on the initial performance of the guesser, and word embedding models were able to successfully capture this pattern. Indeed, language games such as *Connector*, Codenames, and Taboo may be particularly suited to examining questions related to referential communication and the cooperative nature of language. The present study introduces a tractable yet unconstrained task through which the social component of language can be studied.

9 Conclusion

We introduce a cooperative language game *Connector*, based on the game Codenames to examine real-time speaker and guesser semantic search processes. We find that vector embeddings can successfully model behavior in this task, although there appears to be a ceiling to their ability to explicitly predict speaker responses, indicating that there may be limitations to modeling language learning using purely linguistic corpora. Additionally, we show how player interactions can be modeled using vector word representations, and how language games may be particularly suited to investigate the social and cooperative nature of language.

References

Balota, D., Yap, M., Cortese, M., Hutchison, K., Kessler, B., Loftis, B., . . . Treiman, R. (2007). The English Lexicon Project. *Behavior Research Methods, 39*,, 445-459.

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics, 5*, 135-146.

Chvátil, V. (2016). *Codenames.* Retrieved from https://boardgamegeek.com/boardgame/178900/codenames

De Deyne, S., Navarro, D. J., Perfors, A., Brysbaert, M., & Storms, G. (2019). The "Small World of Words" English word association norms for over 12,000 cue words. *Behavior Research Methods, 51*(3), 987-1006.

Ferrucci, D., Brown, E., Chu-Carroll, J., Fan, J., Gondek, D., Kalyanpur, A. A., . . . John Prager, e. a. (2010). Building Watson: An overview of the DeepQA project. *AI Magazine, 31*(3), 59-79.

Hill, F., Hermann, K. M., Blunsom, P., & Clark, S. (2018). Understanding grounded language learning agents.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). *Efficient Estimation of Word Representations in Vector Space.* Retrieved from https://arxiv.org/pdf/1301.3781.pdf%C3%AC%E2%80%94%20%C3%AC%E2%80%9E%C5%93

Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (2004). The University of South Florida free association, rhyme, and word fragment norms. *Behavior Research Methods, Instruments, & Computers, 36*(3), 402-407.

Patel, A., Sands, A., Callison-Burch, C., & Apidianaki, M. (2018). *Magnitude: A fast, efficient universal vector embedding utility package.* Retrieved from https://arxiv.org/pdf/1810.11190.pdf

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. *2014 conference on empirical methods in natural language processing (EMNLP)*, (pp. 1532-1543).

Scarborough, D. L., Cortese, C., & Scarborough, H. S. (1977). Frequency and repetition effects in lexical memory. *Journal of Experimental Psychology: Human Perception and Performance, 3*(1), 1–17. https://doi.org/10.1037/0096-1523.3.1.1

Shen, J. H., Hofer, M., Felbo, B., & Levy, R. (2018). Comparing Models of Associative Meaning: An Empirical Investigation of Reference in Simple Language Games. *Conference on Natural Language Learning.*

Xu, Y., & Kemp, C. (2010). Inference and communication in the game of Password. *Advances in neural information processing systems* , (pp. 2514-2522).

Yamada, I., Asai, A., Shindo, H., Takeda, H., & Takefuji, Y. (2018). *Wikipedia2vec: An optimized implementation for learning embeddings from wikipedia.* Retrieved from https://arxiv.org/pdf/1812.06280.pdf