Defining Words with Words: Beyond the Distributional Hypothesis

Iuliana-Elena Parasca[*] **Andreas Lukas Rauter**[*] **Jack Roper**[*] **Aleksandar Rusinov**[*]
Guillaume Bouchard **Sebastian Riedel** **Pontus Stenetorp**
{iuliana.parasca,andreas.rauter,jack.roper,aleksandar.rusinov}.13@ucl.ac.uk
{g.bouchard,s.riedel,p.stenetorp}@cs.ucl.ac.uk
Department of Computer Science, University College London

Abstract

The way humans define words is a powerful way of representing them. In this work, we propose to measure word similarity by comparing the overlap in their definition. This highlights linguistic phenomena that are complementary to the information extracted from standard context-based representation learning techniques. To acquire a large amount of word definitions in a cost-efficient manner, we designed a simple interactive word game, *Word Sheriff*. As a byproduct of game play, it generates short word sequences that can be used to uniquely identify words. These sequences can not only be used to evaluate the quality of word representations, but it could ultimately give an alternative way of learning them, as it overcomes some of the limitations of the distributional hypothesis. Moreover, inspecting player behaviour reveals interesting aspects about human strategies and knowledge acquisition beyond those of simple word association games, due to the conversational nature of the game. Lastly, we outline a vision of a *communicative evaluation* setting, where systems are evaluated based on how well a given representation allows a system to communicate with human and computer players.

1 Introduction

The distributional hypothesis (Harris, 1954) is at the core of many modern Natural Language Processing (NLP) techniques. It is based on the following assumption:

Words are similar if they have similar contexts.

While powerful, the assumption of context is not always convenient, nor satisfactory. For example, antonyms (black vs. white) and hypernyms (laptop vs. computer) tend to appear in the same context, but they cannot naively replace each other. Similarly, implicit or prior knowledge is difficult to capture by only referring to word contexts. One rarely writes that a banana is yellow, while this is one of the main adjectives one would use when defining it.

In this paper, we describe a novel and complementary framework to capture information that is difficult to obtain by exploiting the distributional hypothesis. It is based on a *relaxed* variant of a dictionary-based hypothesis that assumes that words are the same if they have the same definition. We soften our dictionary-based definition by introducing the notion of "similar definition":

Words are similar if they have similar definitions.

The issue with using word definitions is that it depends on the ability for people to define words. In principle, coming up with proper coherent definitions is costly, as it requires multiple linguistic experts. However, if what we aim to capture is *the ability to identify a word*, we can come up with a more cost-effective data acquisition technique. Our key contribution is the use of crowdsourcing and gamification to show that creating a simple interactive game can generate a huge amount of "short definitions" at very low cost, with the potential to lead to an exciting new data source to evaluate or improve existing word representations. What we mean by "short definition" is a short sequence of words that enables a human to uniquely identify a word.

We will now describe such a game, *Word Sheriff*, which is based on the interaction between a narrator and several guessers, the narrator being a human who implicitly creates definitions. Be-

[*] Contributed equally to this work.

Proceedings of the 1st Workshop on Evaluating Vector Space Representations for NLP, pages 122–126,
Berlin, Germany, August 12, 2016. ©2016 Association for Computational Linguistics

fore going into the details of the game, we should point out that there are many variants or alternative "definition games" that could be designed in a similar spirit, the main idea being that "word definitions matter" because there is some unwritten knowledge that is hard to capture by a static analysis of already existing textual data.

2 Word Sheriff

Our game is loosely based on the Pyramid game show franchise (Stewart, 1973), and for each round, one player (narrator) is presented with a target word or phrase known only to herself. The player must then give the partners (guessers) a series of clues in order to lead them to guess the correct word. After receiving each clue, guessers are allowed to make one guess. The game terminates when one of the guessers find the target word. To incentivise the narrator to use a minimal number of salient clues, the total number of allowed clues is decided beforehand by the narrator, where a lower number of clues lead to a higher reward. An initial web-based prototype of the game was created by four undergraduate students as a part of a project-based course over eight weeks.

Illustrations of successful and unsuccessful game sessions are shown in Tables 1 and 2. In the first session, the narrator decided on limiting herself to 2 clues as she thought that banana is easily uniquely identifiable by yellow and fruit. In fact, this was somewhat risky, as lemon would have been an equally correct answer. While in the second session, a larger number of clues were selected by the narrator, yet the guessers did not arrive at the target word weather. Interestingly, the narrator used a syntactic clue noun that was supposed to guide the guessers to the right type of word. This shows the two-way communicative aspect of the game, as this word was probably chosen because both guessers were proposing adjectives in the second round. Another interesting aspect of the game appears in the first round, where Guesser 1 proposed a word with an opposite meaning (sun when rain is given as the first clue), and Guesser 2 tried to complete a common n-gram (rain jacket).

3 Initial Limited Release

By analysing the logs generated by the game played by human players, we can make interesting linguistic insights and observe player behavioural patterns. Ultimately, in order to be successful in the game, any player, human or computer, must be able to account for the linguistic phenomena that we observe.

Round	Narrator's clue	Guesser 1	Guesser 2
1a	fruit		
1b		orange	apple
2a	yellow		
2b		lemon	**banana**

Table 1: Successful game in 2 rounds for banana

Round	Narrator's clue	Guesser 1	Guesser 2
1a	rain		
1b		sun	jacket
2a	sunny		
2b		cloudy	windy
3a	noun		
3b		cloud	umbrella

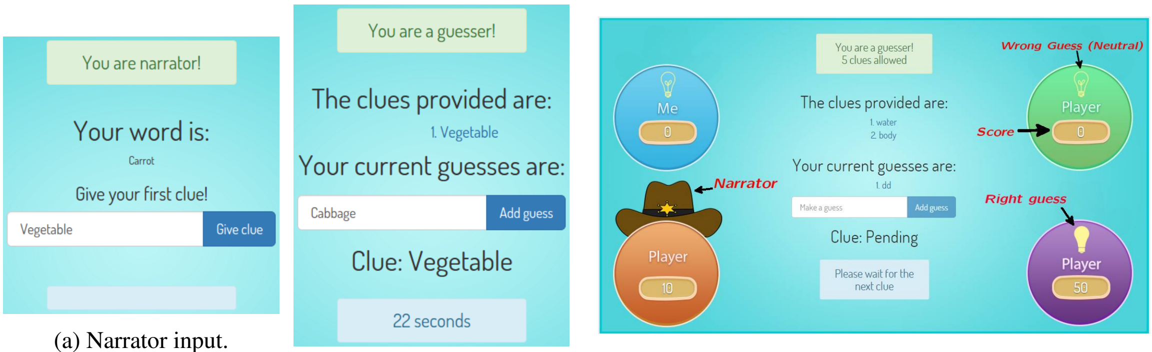
Table 2: Unsuccessful try (3 rds., weather)

To seed our game, we annotated 241 words with clues to be used as gold data for bots that could be introduced if not enough players were online to start a game. We then performed a limited release over a handful of days within our computer science department, where members could play the game freely. All in all, 246 games were played by roughly 100 individual players, where 85% stated that they would consider playing the game again when answering a voluntary anonymous survey.

4 Data Analysis

To better understand what linguistic phenomena that can be observed when playing our game, we qualitatively analysed the annotations collected from the players during our initial limited release. For brevity, we only report phenomena that are difficult to account for using the distributional hypothesis, namely:

- **Hypernymy**: One of the most common strategies is to use two clues involving one hypernym and one distinguishable feature, such as animal + horn for rhinoceros or country + oslo for norway. Perhaps surprisingly, we did not observe any hyponym relations, but this might be due to the limited amount of data analysed.



You are narrator!

Your word is:

Carrot

Give your first clue!

Vegetable

(a) Narrator input.

You are a guesser!

The clues provided are:

1. Vegetable

Your current guesses are:

Cabbage

Clue: Vegetable

22 seconds

(b) Guesser input.

You are a guesser!
5 clues allowed

The clues provided are:

1. water
2. body

Your current guesses are:

1. dd

Make a guess

(c) User Interface overview.

Figure 1: Screenshots of our web-based prototype.

	1st	2nd	3rd
hyena	animal	laugh	dog
wasabi	japanese	spice	
sausage	meat	pig	
anaesthesiologist	doctor	sleep	

Table 3: Compositional strategies.

- **Antonymy**: When observing the guesses given after the first clue, it was interesting to see that players sometimes strategically use antonyms, such a `win` ↦ `lose`. We speculate that experienced players will tend to use antonymy more often than beginners, as it has the potential to uniquely identify a word using single clue, but this intuition would have to be statistically validated on a larger dataset.

- **Prior Knowledge**: Many clue words are related to the target words based on prior knowledge about the world, such as the physical proximity, functional properties or other types of common sense knowledge. One interesting example appears when the target word is `mouth`: guessers tend to use the *Container/Containee* relation and propose `teeth` or `tongue` as clues. Another interesting example is `guacamole`, for which some clues are `avocado` and `burrito`, which are related to the subject or the object of the relation *IsIngredientOf*. Another clue is `condiment`, which relate to the *Typical Usage* of the target word.

The previous observations were mainly focusing on individual words, but another interesting aspect is the compositional nature of the clue words. In Table 3 we report several examples of compositional strategies used by the narrators. This strategy is primarily enabled by the conversational nature of our game, which unlike traditional word association games allow for more than a single response.

5 Related work

For NLP, games have been proposed for numerous tasks such as anaphora resolution (Hladká et al., 2009) and word sense disambiguation (Jurgens and Navigli, 2014). From the literature, Verbosity (von Ahn et al., 2006) is the game most closely related to ours. However, unlike Verbosity our game does not impose ontological restrictions on the input given by the narrator since the end result of the annotations produced by our game does not seek conform with an ontology. Our game also has an adversarial component (guesser-guesser), which we argue is essential for player enjoyment (Prensky, 2007).

Despite a plethora of proposed games, the ones that remain available online have a waning or non-existing player base, why? Our hypothesis is that this is due to the games constraining player creativity to conform with annotation guidelines, leading to less enjoyment, or because of attempts to mimic existing games and adding seemingly unrelated annotation elements to it, to which the player naturally asks the question "Why should I play a variant with a convoluted annotation element, as opposed to a variant without it?".

Thus, we took inspiration from Boyd-Graber et al. (2012) that gathered annotations using an online quiz bowl game and found that the annotators needed no financial incentives and even implemented their own version of the game once the authors had taken their version offline.[1] Our starting-point was thus, can we build upon an existing game that is enjoyable in its own right and with only minor modifications make it sustainable and yield annotations that are useful for evaluating NLP methods?

There are clear parallels between our game and word association games that date back several hundred years and has been of interest to the field of psycholinguistics. One can thus see our goal to be a natural extension of word associations work such as Nelson et al. (2004). In regards to using dictionary definitions, there is the work of Hill et al. (2016), that used dictionary definitions to learn word representations.

6 Future Directions and Challenges

Given the promising results of our prototype implementation and data acquired from our initial limited release, we believe that there are several interesting directions to take our work:

- In our initial release we did not account for the demographic background of our players. An interesting experiment would be to collect such data and inspect it to see if players with different backgrounds would use different explanations.

- Since the data we collected indicate that our model can avoid several pitfalls of the distributional hypothesis, it would seem that retrofitting existing word representations could in fact lead to better word representations for both intrinsic and extrinsic tasks.

- Ultimately, what we find to be the most exiting application would be to use our data and game to perform what we term *communicative evaluation*. Most evaluation of NLP systems is performed in a setting where a system is presented with an input and is asked to *infer* some aspect of the input such as its sentiment, topic, or linguistic structure. However, a key aspect of language is that its purpose is communication, something which our game captures in that players are not only asked to infer the intent of the narrator but also to *communicate* the meaning of the target word when they themselves act as the narrator. Given a representation, a system should be able to learn to perform both the guesser and narrator role, evaluating how well the representation aids the communicative task. This is similar to existing work in computer to computer communication, where two systems learn to communicate about the world, but our setting is different in that as long as a portion of the data is given by human players the way of communicating that is learnt is grounded in human communication.

However, we do believe that there are several hurdles to overcome if we are to succeed in our efforts and we highlight two issues in particular:

- Firstly, our game being a multi-player game, we are reliant on a large player base in order to be sustainable. Not only is it necessary for a significant number of players to be online at any given point in time, it can also be argued that the quality of our annotations are reliant on the players coming from diverse backgrounds, so as not to bias our data.

- Secondly, running a large-scale online game requires large-scale infrastructure. Such infrastructure would also need to me maintained over a large period of time, potentially longer than what a research grant may offer.

Our strategy to overcome these issues is to seek partnership with a commercial actor that can give us access to a wider audience and provide infrastructure. Such a commercial actor would be compensated by more immediate access to the data generated by the players of the game and by the value the game itself can provide for its users, for example as an educational application for language learners.

7 Conclusions

In this work, we showed how to generate an interesting dataset that captures linguistic phenomena such as antonymy, hypernymy and common sense knowledge that are difficult to capture by standard approaches based on the distributional hypothesis. Not only is this data complementary to existing word-similarity datasets, but they can come at nearly no cost as their are obtained as a by-product of a game that is actually very fun to play.

Apart from direct applications of such datasets to psycholinguistics, there are several applications for which the data generated by "definition games", but it could be useful in applications

[1] Personal communication.

where prior knowledge plays an important role, such as question answering involving reasoning about the physical world. It is also likely that it will help to improve machine translation by using the word with the right definition, when there is no one-to-one correspondence between words in the two different languages.

Lastly, we outlined future directions that we seek to take our research in and described several challenges and how we seek to overcome them.

Acknowledgments

This work was supported by a Marie Curie Career Integration Award and an Allen Distinguished Investigator Award. We would like to thank the over one hundred players that took part in the initial limited release of our game, Jordan Boyd-Graber for answering questions regarding their online quiz bowl game, and the anonymous reviewers for their helpful feedback.

References

[Boyd-Graber et al.2012] Jordan Boyd-Graber, Brianna Satinoff, He He, and Hal Daume III. 2012. Besting the quiz master: Crowdsourcing incremental classification games. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1290–1301, Jeju Island, Korea, July. Association for Computational Linguistics.

[Harris1954] Zellig S Harris. 1954. Distributional structure. *Word*, 10(2-3):146–162.

[Hill et al.2016] Felix Hill, Kyunghyun Cho, and Anna Korhonen. 2016. Learning distributed representations of sentences from unlabelled data. *arXiv preprint arXiv:1602.03483*.

[Hladká et al.2009] Barbora Hladká, Jiří Mírovskỳ, and Pavel Schlesinger. 2009. Play the language: Play coreference. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 209–212. Association for Computational Linguistics.

[Jurgens and Navigli2014] D. Jurgens and R. Navigli. 2014. It's all fun and games until someone annotates: Video games with a purpose for linguistic annotation. *Transactions of the Association for Computational Linguistics*, 2:449–464.

[Nelson et al.2004] Douglas L Nelson, Cathy L McEvoy, and Thomas A Schreiber. 2004. The university of south florida free association, rhyme, and word fragment norms. *Behavior Research Methods, Instruments, & Computers*, 36(3):402–407.

[Prensky2007] M. Prensky. 2007. *Fun, Play and Games: What Makes Games Engaging*. Paragon House, St Paul, MN, USA.

[Stewart1973] Bob Stewart. 1973. The $10,000 Pyramid.

[von Ahn et al.2006] Luis von Ahn, Mihir Kedia, and Manuel Blum. 2006. Verbosity: a game for collecting common-sense facts. In *Proceedings of the SIGCHI conference on Human Factors in computing systems*, pages 75–78. ACM.