Novel Aficionados and Doppelgängers: a referential task for semantic representations of individual entities

Andrea Bruera
Queen Mary University of London
School of Electronic Engineering and Computer Science
a.bruera@qmul.ac.uk

Aurélie Herbelot
University of Trento
Center for Mind and Brain Sciences

Abstract
In human semantic cognition, proper names (names which refer to individual entities) are harder to learn and retrieve than common nouns. This seems to be the case for machine learning algorithms too, but the linguistic and distributional reasons for this behaviour have not been investigated in depth so far. To tackle this issue, we show that the semantic distinction between proper names and common nouns is reflected in their linguistic distributions by employing an original task for distributional semantics, the Doppelgänger test, an extensive set of models, and a new dataset, the Novel Aficionados dataset. The results indicate that the distributional representations of different individual entities are less clearly distinguishable from each other than those of common nouns, an outcome which intriguingly mirrors human cognition.

1 Introduction
Learning and retrieving semantic representations for proper names is a task which, unlike other cognitive processes that are much more challenging for computers than for humans (e.g. Lake et al., 2015), seems difficult for both human beings (Semenza, 2009, Brédart, 2017) and machine learning algorithms (Herbelot, 2015, Gupta et al., 2015a, Aina et al., 2019, Almasian et al., 2019, Balasubramanian et al., 2020). Cognitive studies on the subject abound: it has been consistently found that proper names are both more difficult to acquire and retrieve from memory than common nouns and that, as a result of neurodegenerative diseases or vascular lesions, one category can be cognitively impaired independently of the other (Cohen, 1990, Martins and Farrajota, 2007). On the contrary, the linguistic properties which make proper names more difficult than common nouns for computers are a relatively unexplored field in computational linguistics and NLP. ¹

In contrast to common nouns, proper name representations are difficult to evaluate in computational settings (Chen et al., 2019). They cannot be assumed to be ‘known’ by human annotators, so the standard evaluations (e.g. similarity, analogy) cannot be applied without extensive and costly annotation (Newman-Griffis et al., 2018). Further, it is unclear whether such evaluations are appropriate: the meaning of a proper name is exclusively the unique individual entity it refers to, whereas common nouns refer to classes of individuals (Kripke, 1972). So proper names are by nature extensional and should perhaps receive extensional treatment in the course of their evaluation. The main hypothesis of our work is that this difference in semantic properties between proper names and common nouns, found in human cognition, can be retrieved by distributional representations of meaning, when tested over an appropriate referential task.

To show that this is the case, we propose an original referential task, the Doppelgänger test, associated with a new dataset, the Novel Aficionados dataset, made of 59 novels. The Doppelgänger test evaluates whether each entity representation learned in one subcorpus (one half of a novel) can be correctly matched to its co-referring entity representation from another subcorpus (the second half of the same novel), choosing among all the other entity representations (see figure 1). The task is challenging in that the model must distinguish between very similar entities (people and entities engaged in shared activities in a common

¹Since names of places, objects or events have been reported in cognitive studies to dissociate from proper names of conspecifics (Lyons et al., 2002, Crutch and Warrington, 2004), in order to avoid confounds, these other sorts of names won’t be considered. Therefore, in the following, the expression ‘individual entity’, ‘individual’ and ‘proper name’ will be used to indicate human individuals and their names.
universe) using scarce data.

Using the Doppelgänger test, we compare the distributional representations of the proper names referring to the novels’ characters and those of similarly frequent common nouns mentioned in the novel. For robustness, we use several models (ELMO Peters et al., 2018, BERT: Devlin et al., 2018, Word2Vec Mikolov et al., 2013, and Nonsense2Vec Herbelot and Baroni, 2017). Our approach to the task is unsupervised, and in this respect it can be considered a special case of a language model probing task, focused on referential semantic information (Rogers et al., 2020, Sorodoc et al., 2020). By employing the same, controlled semantic representation learning procedure for proper names and common nouns within a novel, we show that distinct patterns of results for the two linguistic categories emerge.

As further analyses, we look at three levels of possible discrepancies between the two categories in our setup. First, we look at low-level, distributional differences in part-of-speech neighbourhood, which confirm that they have different distributional signatures. Then, we turn to mid-level differences in terms of narrative features of the novels, with a correlational analysis. This highlights the disruptive effect of competing semantic representations, which disproportionately affect reference resolution for proper names, drawing a parallel with effects found in human semantic cognition (Abrams and Davis, 2017). Finally, we analyze the higher-level structural differences between the obtained vector spaces, by way of a Representational Similarity Analysis study (Kriegeskorte et al., 2008) which indicates that common nouns give rise to structurally more coherent spaces than proper names.

Finally, in order to show how the Doppelgänger test can be adapted to texts from a different domain, we present the Quality test, a challenging variation on the Doppelgänger test which requires linking entities across different corpora (the original novels and Wikipedia).

Overall, these results suggest that proper names, when modelled by way of distributional semantics algorithms such as language models and word embeddings, require specific computational strategies in order to capture their referential properties.

2 Related work

2.1 Reference in NLP

The topic of proper names and that of reference have been going hand in hand in language studies for a long time, at least ever since Mill (1884), Frege (1892) and later Strawson (1950) and Kripke (1972). With respect to our approach, the most closely related tasks in computational linguistics are entity linking (EL) and anaphora (or co-reference) resolution.

Entity Linking, also called Named Entity Disambiguation, is a NLP task where the correct
reference of mentions of proper names in a text has to be found in a knowledge database (Balog, 2018, Onoe and Durrett, 2020). Entity Linking models have no interest in modelling in any way cognitive processes (indicated for instance by the fact that they often use strong supervision), whereas our model is kept unsupervised, in order to obtain evidence which can be theoretically interpretable.

Anaphora resolution is the name of the process by which a competent speaker naturally gets to understand that in the sentence “Saul and Tina went to the market: he bought a pin and she bought a fake gun” the word ‘he’ refers to the same individual ‘Saul’ refers to, and ‘she’ refers to the referent of ‘Tina’. Various algorithms and tasks have been proposed in order to model this linguistic phenomenon in computational linguistics (Poesio et al., 2016). However, as anaphora resolution can be modelled by employing the same strategies for both common nouns and proper names, the two linguistic categories have, to our knowledge, not been investigated separately (Clark and Manning, 2016).

A perspective more akin to the present one is that of Herbelot (2015) where, given the poor quality of distributional semantics representations of characters as extracted from two novels, the author presented ad-hoc techniques in order to improve those semantic representations. It is important to underline, however, that the focus of the present work is different from the one of Herbelot (2015): here the goal is not at all that of finding ways to extract better representations for proper names in distributional semantics. Rather, the aim is that of studying, from a distributional and cognitively-oriented perspective, why proper names are more difficult than common nouns for computational semantic processing in the first place.

In this sense, as it focuses on theoretical investigation, our work is more similar to (Gupta et al., 2015b) and (Gupta et al., 2018) which respectively try to extract attributes and categories for proper names from distributional models.

2.2 Characters in novels

Work on individual entities - the kinds of entities proper names refer to - in computational linguistics and NLP has often made use of novels. However, such approaches have concentrated mainly on conceptually different tasks: learning character types (Bamman et al., 2013, Flekova and Gurevych, 2015), inferring characters’ features (Louis and Sutton, 2018), relations (Iyyer et al., 2016, Elson et al., 2010), networks (La-batut and Bost, 2019) or on broader natural language understanding tasks (Fremann et al., 2018), such as inferring plot structure (Elser, 2012).

Here, instead, the focus is on the investigation of the semantic and referential distinction between proper names and common nouns, whose distinct categorical status is a solid cross-linguistic phenomenon (Van Langendonck and Van de Velde, 2016).

3 Data

In order to carry out our experiments, we collected a new dataset, the Novel Aficionados dataset. The core material of the dataset is made of 59 novels, collected from the Project Gutenberg website, an online repository of free ebooks. They were selected from the list of the 100 most downloaded ebooks of the month at the time of data collection, by excluding non-fiction ebooks. All novels are not protected by copyright anymore.

Narrative literature is particularly suited to our approach, because of the importance of characters in narration. Fiction plots are built around characters, which are (often non-existing) human individuals, and around their thoughts and actions. In this sense, novels are written precisely in order to allow the creation of semantic representations of the individual entities by way of text only (Bamman et al., 2019).

The dataset consists of an augmented and annotated version of the novels. First, all character mentions, which often take various forms despite referring to the same entity (e.g. ‘Mr. Darcy’, ‘Darcy Fitzwilliam’ and ‘Darcy’), are substituted by a unique label (in our example, ‘mr_darcy’) and marked by two ‘$’ symbols, before and after the mention ('$mr_darcy$'). For the analyses, only characters occurring more than 10 times are retained. Secondly, the most frequent common nouns (considering their lemmas) for each novel are selected in order to be used for the Doppelgänger test. Their number was matched to the...
amount of characters previously annotated. The rationale for choosing the most frequent common nouns as the counterpart to the characters in the dataset is that they arguably capture the novel’s main themes and topic. To distinguish them from the characters’ names, the selected common nouns are surrounded by two ‘#’ symbols (e.g. ‘#hound#’).

For this process of data augmentation and annotation, we used BookNLP (Bamman et al., 2014), a full NLP pipeline optimized for novels which importantly includes both Named Entity Recognition and co-reference resolution modules. Finally, the dataset is enriched with the matched Wikipedia pages for each one of the 59 novels, processed using the same annotation style. This data is included as it constitutes a non-narrative, encyclopedic source of information about the characters, themes and topics present in the novels. An example of use of this portion of the dataset is presented in section 7.4.

Each file in the dataset was split into sentences by using Spacy\(^{4}\), so that each line of the resulting files contains a single sentence. Also, all punctuation was removed and the letters were turned into lower case.

\(^{4}\)https://spacy.io/

4 Task

The Doppelgänger test aims at probing referential information contained in distributional semantic representations. It starts from the intuition that we should be able to match two different representations of the same referent, even if they are obtained from distinct data. In order to reduce confounds, in the Doppelgänger test we take a single document where multiple entities appear - in our case, a novel, where entities are referred to by either proper names and common nouns. Then, the document is first split into two sub-corpora (Part A and Part B), both containing mentions of all the entities (see figure 1). Subsequently, for each part, a semantic representation for each entity is obtained by way of a distributional semantics model. Finally, by taking the two separate sets of representations, containing the same entities but coming from different parts of the document, the Doppelgänger test probes to what extent it is possible to match the co-referring vectors with one another. It is a purely referential, extensional task, as it evaluates word vectors on the basis of their ability to model the extension (the reference) of a word. Because of this, it is naturally suited to comparing the capabilities of distributional semantic models with respect to two categories, proper names and common nouns, whose referential properties are different: proper names refer to unique entities, whereas common nouns refer to classes of individuals. This is the question that we focus on in
this work - however, the Doppelgänger test can be used as a generic probing task for distributional semantic models, and computational models of semantics at large.

In this work we decided to keep a strictly unsupervised approach to the Doppelgänger test. This is in line with recent work in language models probing (Broscheit, 2019, Petroni et al., 2019, Talmor et al., 2020), whose goal is to investigate as directly as possible the behaviour of the models’ representations on the task at hand.

More precisely, in the current setup, given a novel \(N\), we split it in two halves, \(N_A\) and \(N_B\). We experiment with both splitting a novel in two parts at the original midpoint, and with first randomizing the list of sentences, then splitting the randomized sentences in two halves, averaging the results for 100 iterations. No difference in results emerged, so we chose the former, simpler approach. Entities present in only one of the two parts were not retained for further analyses. We found this had very little impact on the final amount of entities used.

The analyses for proper names and common nouns are carried out separately, making sure to employ the same number of entities for each category. The two categories are compared only at the end. From each part, we obtain a matched set of word vectors \(E_{part} = \{\vec{e}_1^...\vec{e}_n^\}\), either referring to characters or to common nouns’ referents. In order to probe the performance at the Doppelgänger test, we use a simple unsupervised ranking approach. For each vector in \(E_A\) (and then conversely in \(E_B\)), the query \(\vec{e}_A^A\), we compute the pairwise cosine similarities with all vectors in \(E_B\), then we rank the vectors in \(E_B\) according to their similarity to the query \(\vec{e}_A^A\). The position in the ranking of the coreferring vector \(\vec{e}_B^A\) constitutes the model’s performance with respect to the current entity. The median of the per-entity scores is the per-novel score, and the median of the 59 per-novel scores constitutes the final score for the model at hand.

Scores for all the models and semantic categories (common nouns or proper names) are compared in figure 2.

5 Models

We employed a broad range of distributional semantic models, so as to avoid biases inherent in specific implementations: they all rely on the Distributional Hypothesis (Firth, 1957), which states that words found in similar contexts have similar meanings, but they all differ in their realization (Pilehvar and Camacho-Collados, 2020). We used three kinds of models: count-based, prediction-based (following the terminology of Baroni et al. (2014)), and contextualized language models. In all models, and for each novel, both sets of vectors \(E_A\) and \(E_B\) are initialized as two sets of vectors filled with zeroes, and they are then updated by using the novel’s data.

The count model is based on simple word co-
occurrence counts, transformed to PPMI measures, a correction which has been shown to drastically improve performances (Goldberg and Levy, 2014). Co-occurrences were counted by considering a sliding window of 5 words to the right and to the left of each target word.

The prediction-based models are Word2Vec (W2V), a very successful language model consisting of a feed-forward neural network (Mikolov et al., 2013), and Nonce2Vec (N2V), a modified version of Word2Vec, specialized for small datasets such as novels (Herbelot and Baroni, 2017). First, we pre-trained a Word2Vec model on the English version of Wikipedia in its Python Gensim implementation (Rehurek and Sojka, 2010), using the skip-gram training method and default parameters.

For Word2Vec, each entity mention was modeled as the average of the pre-trained model’s vectors for the words surrounding it, again within a window of 5 words on each side. The final set of vectors $E_{part}$ was obtained by representing each entity by the average of its mentions’ vectors.

In the case of Nonce2Vec, the same pre-trained Word2Vec model was used. However, Nonce2Vec allows to adapt the skip-gram training regime to the reduced amount of data offered by a novel, thus creating new entity representations by exploiting the pre-trained weights.

As contextualized models we used BERT (both BERT-BASE and BERT-LARGE (Devlin et al., 2018), in their Python huggingface implementation (Wolf et al., 2019)) and ELMO (Peters et al., 2018). They have been shown to have comparable performances, but they differ in several respects. In order to predict a target word, ELMO first models separately left and right context by two separate LSTMs, eventually merging their representations. BERT, instead, employs the Transformer architecture (Vaswani et al., 2017), considering at the same time all the words around the target one. In our setup, for both models, given a sentence containing an entity mention, we first mask the mention, thus making it the unknown target word to be predicted. Then we provide the full sentence, with the masked entity mention, to the model for a forward pass. Finally the vector corresponding to the masked entity mention is extracted from the last hidden layer of the contextualized language model. As in Word2Vec, the final representation for an entity is obtained by averaging all the vectors for its mentions.

6 Results

Results are shown in figure 2. All models perform better when matching representations for common nouns, than for proper names. An inspection of the distributions of the scores for each model (see appendix A) confirms that whereas common nouns most often obtain a score of 1, indicating that the referential task was carried successfully, the distribution for proper names is much less skewed towards 1 and has a much longer tail.
This pattern of results is strikingly consistent across models, indicating that the semantic, referential distinction between proper names and common nouns emerges in the acquisition of semantic representations even when using exclusively textual, distributional linguistic information.

7 Further analyses

In order to understand what drives such a consistent pattern of results, we carried out three separate investigations, focusing on three levels: low-level distributional features, by way of a part-of-speech neighbourhood analysis; novel-level variables such as length in words, number of characters involved and differences in characters’ mentions; vector space-level analyses, by way of Representational Similarity Analysis.

7.1 Part-of-speech neighbourhood

To quantify the differences in the distributional properties of proper names and common nouns, we looked at the part-of-speech occurrences around the characters’ names and the chosen common nouns in the Novel Aficionados dataset. We used a sliding window of 2 words on both sides of each mention, and we kept track of the co-occurrences in a matrix. The matrix had two rows, one for proper names and one for common nouns, and six columns corresponding to six parts-of-speech categories: adjectives (ADJ), adverbs (ADV), determiners (DET), nouns (NOUN), pronouns (PRON) and verb (VERB). The part-of-speech tagging was carried out with the Spacy toolkit. Aside from the obvious difference in the frequency of determiners (in English proper names can’t have a determiner before them), this analysis shows that proper names are more frequently found in the vicinity of nouns and verbs than common nouns, confirming that there are low-level differences in surrounding word distributions amongst the two categories.

7.2 Correlational analysis

Novels may be characterized by structural features which make it more difficult to match co-referring word vectors for characters: some novels may be very short, thus not providing enough data; some may have a larger amount of characters, making it more difficult to correctly discriminate among different characters’ representations; finally, in some novels some characters may receive much more attention than others, a case of uneven data split which may affect results. In order to understand the importance of these variables in the Doppelgänger results, we looked at the correlation between the models’ scores and the three variables (novel length, number of characters, standard deviation of mentions across characters). Results are shown in figure 4. Both novel length and number of characters correlate strongly with results,
with the latter dominating in all models. This entails that, as the number of characters increases, the representations for the characters get progressively confused with one another, and that distributional models have a hard time with correctly establishing reference for numerous entities. This result dovetails with both cognitive (Abrams and Davis, 2017) and computational findings (Ilievski et al., 2018).

7.3 Representational similarity analysis

Finally, for each novel, we compared the properties of the vector spaces corresponding to the two portions of the novels. We wanted to find out whether the resulting vector spaces across the two parts of the novels, $E_A$ and $E_B$, were significantly different in their structural properties between proper names and common nouns. An ideal framework to carry out such analyses is that of Representational Similarity Analysis (Kriegeskorte et al., 2008), originally proposed in cognitive neuroscience.

In this approach two different vector spaces (having the same, matched amount of vectors) are not compared directly, but rather by way of the vectors of their within-space pairwise similarities. These two pairwise similarity vectors encode the representational structure of each space; and if two such vectors correlate, then they are taken to be similar in their representational structure. For each novel, and separately for each word category, we look at the correlation of the matched pairwise similarities for the two vector spaces $E_A$ and $E_B$. Results are shown in figure 5.

Proper names exhibit lower representational similarities across vector spaces in all models. It seems reasonable to speculate that this structural difference must play an important role in the Doppelgänger scores, where structurally less similar pairs of vector spaces (those for proper names) perform worse.

7.4 Going beyond the novels: the Quality test

It is important to understand whether our results are specific to novels, or can generalize to other domains and kinds of text. As a first step, we include a different implementation of the Doppelganger test, that we call the Quality test. In this test, for each set of entities, instead of using two sub-documents, we use two different kinds of documents: a novel and a Wikipedia page on the novel, which is included in the Novel Aficionados dataset for each novel.

The Wikipedia description of a novel includes information about the same characters and entities as the novel itself, but it presents it with both a different purpose (short presentation of fundamental features) and a different style (non-narrative). Therefore, the task should be more difficult than the original Doppelgänger test, because of the added difficulty due to the difference between the two documents used for creating the sets of entity representations $E_A$ and $E_B$.

As it can be seen from figure 6, results have a par-
tially different pattern with respect to the Doppelgänger test. Contextualized models perform similarly to the original test, confirming their ability to encode semantic information solidly, even in challenging conditions. In this case too, contextualized models show worse performance for proper names. This confirms that this semantic category poses peculiar challenges to distributional semantic models. The models based on Word2Vec models perform very poorly, and on a par for proper names and common nouns, indicating that they are not very robust to this experimental manipulation. Finally, the performance for the count-based model show the reversed pattern: a puzzling result which calls for an application of the Doppelgänger test to different types of texts.

8 Conclusion

Using as a starting point the distinction between proper names and common nouns, fairly well studied in the neuro-cognitive and formal semantics literature, but almost ignored in computational linguistics and NLP, we proposed a new evaluation for computational representations of entities, the Doppelgänger test. This task probes in particular for referential, extensional semantic information encoded in those representations, which is of paramount importance specifically for proper names. It does so by first splitting a document into two sub-documents, then obtaining two matched sets of semantic representations for the entities contained in the document, and finally evaluating to what extent it is possible to match the pairs of co-referring vectors.

We compared the performances of an extensive set of distributional semantic models by using an original dataset, the Novel Aficionados dataset, tailored to comparing the models’ performances on proper names and common nouns. By means of the Doppelgänger test the semantic distinction between the two categories emerged in strikingly different patterns of results. What’s more, the models’ performances mirrored human cognition, with common nouns being consistently easier to match according to their reference than proper names. By way of further analyses, we showed that the distinction between the two categories is present both at the level of textual distributional properties, in the form of part-of-speech co-occurrence differences, and at the level of vector space structure, which is more similar across matched sets of vectors for common nouns than proper names. Also, models were shown to be gradually degrading their performance as more individual entities were considered.

Finally, by using the Doppelgänger test on different data, we demonstrated how it can become, beyond the current setup, a valuable evaluation framework for probing for referential information in semantic representations of individual entities.

References

Lise Abrams and Danielle K Davis. 2017. Competitors or teammates: how proper names influence each other. Current Directions in Psychological Science, 26(1):87–93.

Laura Aina, Carina Silberer, Ionut Sorodoc, Matthijs Westera, and Gemma Boleda. 2019. What do entity-centric models learn? Insights from Entity Linking in multi-party dialogue. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 3772–3783.

Satya Almasian, Andreas Spitz, and Michael Gertz. 2019. Word embeddings for entity-annotated texts. In European Conference on Information Retrieval, pages 307–322. Springer.

Sriram Balasubramanian, Naman Jain, Gaurav Jindal, Abhijeet Awasthi, and Sunita Sarawagi. 2020. What’s in a name? Are bert named entity representations just as good for any other name? In Proceedings of the 5th Workshop on Representation Learning for NLP, pages 205–214.

Krisztian Balog. 2018. Entity Linking, pages 147–188. Springer International Publishing, Cham.

David Bamman, Brendan O’Connor, and Noah A Smith. 2013. Learning latent personas of film characters. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 352–361.

David Bamman, Sejal Popat, and Sheng Shen. 2019. An annotated dataset of literary entities. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 2138–2144.

David Bamman, Ted Underwood, and Noah A Smith. 2014. A bayesian mixed effects model of literary character. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), volume 1, pages 370–379.

Marco Baroni, Georgiana Dinu, and Germán Kruszewski. 2014. Don’t count, predict! A
systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 238–247.

Serge Brédart. 2017. The cognitive psychology and neuroscience of naming people. *Neuroscience & Biobehavioral Reviews*, 83:145–154.

Samuel Broscheit. 2019. Investigating entity knowledge in bert with simple neural end-to-end entity linking. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 677–685.

Mingda Chen, Zewei Chu, Yang Chen, Karl Stratos, and Kevin Gimpel. 2019. Enteval: A holistic evaluation benchmark for entity representations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 421–433.

Kevin Clark and Christopher D Manning. 2016. Improving coreference resolution by learning entity-level distributed representations. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 643–653.

Gillian Cohen. 1990. Why is it difficult to put names to faces? *British Journal of Psychology*, 81(3):287–297.

Sebastian J Crutch and Elizabeth K Warrington. 2004. The semantic organisation of proper nouns: the case of people and brand names. *Neuropsychologia*, 42(5):584–596.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Micha Elsner. 2012. Character-based kernels for novelistic plot structure. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 634–644. Association for Computational Linguistics.

David K Elson, Nicholas Dames, and Kathleen R McKeown. 2010. Extracting social networks from literary fiction. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, pages 138–147. Association for Computational Linguistics.

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

Lucie Flekova and Iryna Gurevych. 2015. Personality profiling of fictional characters using sense-level links between lexical resources. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1805–1816.

Gottlob Frege. 1892. *Ueber Sinn und Bedeutung*. *Zeitschrift für Philosophie und philosophische Kritik*, 100:25–50.

Lea Freremann, Shay B Cohen, and Mirella Lapata. 2018. Whodunnit? Crime drama as a case for natural language understanding. *Transactions of the Association of Computational Linguistics*, 6:1–15.

Yoav Goldberg and Omer Levy. 2014. word2vec explained: deriving Mikolov et al.’s negative-sampling word-embedding method. *arXiv preprint arXiv:1402.3722*.

Abhijeet Gupta, Gemma Boleda, Marco Baroni, and Sebastian Padó. 2015a. Distributional vectors encode referential attributes. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 12–21.

Abhijeet Gupta, Gemma Boleda, Marco Baroni, and Sebastian Padó. 2015b. Distributional vectors encode referential attributes. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 12–21.

Abhijeet Gupta, Gemma Boleda, and Sebastian Padó. 2018. Instantiation. *arXiv preprint arXiv:1808.01662*.

Aurélie Herbelot. 2015. Mr Darcy and Mr Toad, gentlemen: distributional names and their kinds. In *Proceedings of the 11th International Conference on Computational Semantics*, pages 151–161.

Aurélie Herbelot and Marco Baroni. 2017. High-risk learning: acquiring new word vectors from tiny data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 304–309.

Filip Ilievski, Piek Vossen, and Stefan Schlobach. 2018. Systematic study of long tail phenomena in entity linking. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 664–674.

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

Nikolaus Kriegeskorte, Marieke Mur, and Peter A Bandettini. 2008. Representational similarity analysis—connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, 2:4.

Saul A Kripke. 1972. Naming and necessity. In *Semantics of natural language*, pages 253–355. Springer.
Vincent Labatut and Xavier Bost. 2019. Extraction and analysis of fictional character networks: A survey. *ACM Computing Surveys (CSUR)*, 52(5):1–40.

Brenden M Lake, Ruslan Salakhutdinov, and Joshua B Tenenbaum. 2015. Human-level concept learning through probabilistic program induction. *Science*, 350(6266):1332–1338.

Annie Louis and Charles Sutton. 2018. Deep Dungeons and Dragons: Learning character-action interactions from role-playing game transcripts. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 708–713.

Frances Lyons, J Richard Hanley, and Janice Kay. 2002. Anomia for common names and geographical names with preserved retrieval of names of people: A semantic memory disorder. *Cortex*, 38(1):23–35.

Isabel Pavão Martins and Luisa Farrajota. 2007. Proper and common names: A double dissociation. *Neuropsychologia*, 45(8):1744–1756.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Radim Rehurek and Petr Sojka. 2010. Software framework for topic modelling with large corpora. In *In Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. Citeseer.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in bertology: What we know about how bert works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

Mohammad Taher Pilehvar and Jose Camacho-Collados. 2020. Embeddings in natural language processing: Theory and advances in vector representations of meaning. *Synthesis Lectures on Human Language Technologies*, 13(4):1–175.

Massimo Poesio, Roland Stuckardt, and Yannick Versley. 2016. *Anaphora Resolution*. Springer.

John Stuart Mill. 1884. *A system of logic, ratiocinative and inductive: Being a connected view of the principles of evidence and the methods of scientific investigation*, volume 1. Longmans, Green, and Company.

Denis Newman-Griffis, Albert M Lai, and Eric Fosler-Lussier. 2018. Jointly embedding entities and text with distant supervision. In *Proceedings of The Third Workshop on Representation Learning for NLP*, pages 195–206.

Yasumasa Onoe and Greg Durrett. 2020. Fine-grained entity typing for domain independent entity linking. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473.
