NATURAL LANGUAGE AT A CROSSROADS: FORMAL AND PROBABILISTIC APPROACHES IN PHILOSOPHY AND COMPUTER SCIENCE

PAULO PIROZELLI[1]
[1]https://orcid.org/0000-0002-4714-287X
University of São Paulo
Institute of Advanced Studies
São Paulo, S.P.
Brazil
paulo.pirozelli.silva@usp.br

IGOR CÂMARA[2]
[2]https://orcid.org/0000-0002-1831-1750
University of São Paulo
Institute of Mathematics and Statistics
São Paulo, S.P.
Brazil
igorcsc@ime.usp.br

Article info
CDD: 401
Received: 17.03.2021; Revised: 23.08.2021; Accepted: 21.10.2021
https://doi.org/10.1590/0100-6045.2022.V45N2.PI

Keywords
Philosophy of Language
Meaning
Use
Natural Language Processing

Abstract: Philosophy of language and computer science, despite being very distinct fields, share a great interest in natural language. However, while philosophy has traditionally opted for a formalist approach, computer science has been increasingly favoring probabilistic models. After presenting these two approaches in more detail, we discuss some of their main virtues and limitations. On the one hand, formalist models have trouble in acquiring

semantic information from corpora and learning from large amounts of data. Probabilistic approaches, on the other hand, have difficulty in operating with compositionality, in dealing with contrast sets and hierarchical relations, and in distinguishing normative and descriptive views of meaning. We argue that a more fruitful dialogue between philosophers and computer scientists may help to produce a better approach to natural language and stimulate the integration of logical and probabilistic methods.

1. Introduction

At first glance, philosophy and computer science appear to lie on opposite sides as regards to knowledge. Philosophy has always been the paradigm of a humanistic discipline, involving abstract discussions on highly theoretical problems, whereas computer science is part applied science, and part formal science, such as mathematics and logic.

Despite their very distinct sets of problems and methods, philosophy and computer science both share a strong interest in *natural language*. The two disciplines have been studying this subject for a long time and dealing with many of the same topics, such as the nature of meaning, the compositionality of sentences, and inferentiality. Unfortunately, philosophers and computer scientists are usually not as informed of each other's advances as they could, and perhaps, should be. This paper intends to discuss some of the theoretical affinities between these two fields and the common problems and challenges they currently face. We take this as an opportunity to engage philosophers and computer scientists in a dialogue that can benefit all participants. This is an attempt to stimulate a more interdisciplinary research in AI, similarly to Miller (2017) and Grimsley et al. (2020).

2. Formal Approaches in Philosophy of Language

The interest of philosophers in natural language has often been connected to a concern for the logical structure of language. The first systematic inquiry in that direction was conducted by Aristotle and concerned the nature of inference. He was particularly interested in a specific class of arguments called syllogism, formed by a pair of premises (major and minor) and a conclusion. The problem, for him, was to understand why conclusions were always true in some of them, provided that the premises were also true. Those were the *valid* syllogisms — arguments in which the conclusions followed from premises.

As Aristotle found out, validity had nothing to do with the semantic content of the nouns and adjectives in the sentences, but only with the *form* of the arguments, given by the combination of words like "all" and "not". Thus, in order to attest the validity of a syllogism, one could simply replace "(certain) of the expressions in their premises and conclusions with schematic letters thereby abstracting away from what the arguments are about" (Lepore & Ludwig 2001, p. 3). The translation of a syllogism to this semi-formal language could then display the argument's underlying structure. For example, a syllogism as:

> All men are mortal
> Socrates is a man
>
> ———————————
>
> Socrates is mortal ∴

could be showed to be an instance of the general argument

All A are B
c is A

---

c is B ∴

which represents a class of valid arguments.

In the late 19th century, a new step towards a full formalization of language was given with the invention of symbolic logic. Using some notions from mathematics, Frege (1879) developed a powerful framework that enabled the reconstruction of subject-predicate relations as function-argument structures. His intent was to put aside surface grammar and look for what he thought was the underlying logical form of natural language. Thus, instead of analyzing a sentence such as "Socrates is mortal" as being of the form "*s* is *M*", in which *s* represents the subject Socrates and *M* the predicate "being mortal", Frege preferred to treat it as being of the form *G(s)*, in which *s* is the object Socrates and *G*, a concept that maps mortal things to the value True, and non-mortal ones to the value False.

The use of a function-argument structure, together with mathematical quantifiers, proved to be a powerful resource to philosophical analysis. It allowed rewriting natural language sentences in a way that could immediately unveil the inferential links beneath arguments. Thus, with the use of formal logic, the syllogism just mentioned could be expressed (in a contemporary notation) as $\forall x(((M(x) \rightarrow H(x)) \wedge M(z)) \rightarrow H(z))$. The validity of the argument is, then, conspicuously displayed.

But it was perhaps with Russell that logical formalization showed all its power in the application to natural language. In his classic article "On denoting" (1905), Russell dealt with the problem of definite descriptions—expressions of the form "the *x* which is *y*". Definite descriptions have the same syntactic function as proper nouns — they are nominal

phrases —, and at first glance, seem to be objects of the same kind. A more careful analysis, however, reveals that this interpretation may lead to serious ontological problems. In a sentence that mentions a proper noun *A*, we take the existence of *A* as granted. The sentence "John is tall", for instance, does not appear to have any sense if there is no John. Definite descriptions, on the other hand, are not subject to these requirements. We understand the meaning of the sentence "the current king of France is bald", even though there is no current king of France.

The question Russell investigated was how a sentence that referred to a definite description could be meaningful, and how sentences that denied the existence of a definite description were possible. Do we need to assume that they refer to some sort of ethereal existence—a "subsistence", as Meinong (1904) defended? The problem seemed of far-reaching consequences, since a great part of our linguistic constructions have the form of descriptions like that.

Russell's solution to this problem was to deny that definite descriptions worked like names at all. In his view, the logical syntax underlying natural language did not necessarily (or even often) coincide with regular syntax. In order to show the true behavior of sentences with definite descriptions, therefore, he offered a logical reconstruction of these expressions. "The current king of France is bald", his famous example, could be rewritten, for instance, as "currently there is at least one x that is king of France, there is no other y that is king of France, and this x is bald". More formally, it could be expressed as $\exists x((K(x) \wedge \forall y(K(y) \rightarrow x=y)) \wedge B(x))$, in which K(x) stands for "*x* is the current king of France", and B(x) for "*x* is bald".

If that was the true structure of the sentence, then there was no problem in denying the existence of an object, since when affirming that "there is no current king of France", we would be simply stating that "there is no *x* who is the current

king of France and is bald". No "king of France", nor any other sort of object, would have to be assumed.[1]

The power exhibited by the application of symbolic logic to the comprehension of natural language, as well as the achievements it promised, seemed irresistible to many philosophers. Russell, for instance, thought that "some kind of knowledge of logical forms, though with most people it is not explicit, is involved in all understanding of discourse. It is the business of philosophical logic to extract this knowledge from its concrete integuments, and to render it explicit and pure" (Russell 1914, p. 35). His most famous pupil, Ludwig Wittgenstein, similarly claimed that the source of all philosophical problems is that "the logic of our language is misunderstood" (Wittgenstein 1921, p. 3).

Russell's and Wittgenstein's most radical views, reducing philosophical activity to a logical analysis of language, were the result of a euphoria with the surprising novelty of symbolic logic. Most philosophers, of course, never agreed that all philosophical problems could be reduced to language analysis. Nonetheless, this logical approach had enduring effects on the way philosophers came to see natural language, particularly in the idea that formalization, through the use of symbolic logic, contributed to the understanding of how human language worked.

Later, philosophers extended this formalist approach to language in various ways. Kaplan (1989), for example, gave a classic account on context-dependent expressions, those that make use of elements such as indexicals (e.g., "me" and

[1] An alternative, pragmatic solution to definite descriptions was given by Strawson (1950), for whom referring is not the property of an expression, but an act performed by a speaker on a given occasion. As consequence, a sentence whose reference fails (as when we say nowadays that "the current king of France is bald") is one which simply lacks a truth-value.

"now") and demonstratives ("this" and "that"). Another important line of investigation was on modalities, with the application of possible world semantics to natural language (Kratzer 1977). Possible world semantics proved itself a valuable tool for tackling linguistic and philosophical problems, such as the inner workings of counterfactual conditionals, analyzed by Lewis (1976) in terms of similarity between possible worlds.

3. The Rise of Distributional Semantics

Similar attempts to formalize natural language were followed in the field of linguistics and, more recently, in Natural Language Processing (NLP). The connection to philosophy was not accidental, since formal approaches to natural language semantics were pioneered by Montague, a well-known philosopher and logician. This formalist approach in linguistics sought to model natural language semantics by means of formal languages and logical reasoning. The idea was that only a system with well-defined rules could satisfactorily explain the infinite set of sentences that human language is potentially able to generate. Formalist approaches to language and knowledge gave birth to several research areas in computer science, many of which are still active nowadays, such as expert systems and computational ontologies. Even some non-formalist methods employ formalist tools in their process, such as wordnets and knowledge graphs.

Formal Semantics made important progress over time, with interesting attempts at processing natural language through formal methods. The use of categorial grammars in order to represent sentence meaning, for instance, facilitated the use of inferential methods, such as adding new propositions to a knowledge base or determining the truth

of implicit propositions (Moot & Retore, 2012). Another successful technique was referring expression generation, a subtask that dates back to the 1970s, which allowed the production of descriptions that enabled the identification of an entity in a given context (Krahmer & van Deemter 2012).

However, Formal Semantics also demonstrated some strong limitations. Particularly, Formal Semantics had difficulty in dealing "with descriptive content, despite the large amount of work done on lexical semantics and formal ontology" (Boleda & Herbelot 2016, p. 620). As Boleda & Herbelot argue, because Formal Semantics must limit the phenomena it deals with, it must ignore large portions of natural language. Moreover, it is not yet clear how a purely Formal Semantics-oriented approach would be able to tackle problems closely related to the vast sea of lexical meaning in a more automatic way, such as the one championed by machine learning methods.

The theoretical issues in Formal Semantics, allied with the progress in the manipulation of data brought by technology, inspired a more data-driven approach in NLP, broadly called Distributional Semantics. Historically, the first attempts to represent meaning as probabilistic distributions were proposed around the same time as the Montagovian and Chomskian formalist theories. The idea of representing words as vectors, and the relation of such representations to the notion of semantic similarity, can already be found in Spärck-Jones' thesis (1964).[2] However, it was only with the recent growth in computational power and the exponential growth in available data that Distributional Semantics became the most prominent approach in NLP.

Distributional Semantics encompasses many techniques for modeling natural language. What they all have in

[2] We thank an anonymous reviewer for bringing this work to our attention.

common is that they are based on the hypothesis that the meaning of a word can be inferred from its context. More formally, Distributional Semantics assumes that "semantic representations of lexical items can be built by recording their distribution in linguistic contexts" (Lenci, 2018, p. 160). In practice, "context" is usually defined as the window of *n* words that occur in the vicinity of the target.

The idea of finding meaning representations through contexts does not fit well with Frege's and Russell's logical approach, but it had an important and influential predecessor in philosophy, too — Wittgenstein (1953). Like any field, philosophy of language deals with its own unique set of problems. This includes questions such as what makes certain marks or sounds meaningful; how human beings can produce meaningful sentences; how people are able to understand phrases they have never heard before; and many others. Its most important subject, though, regards the nature of meaning—a question that can almost be confused with philosophy of language itself. Philosophers have given many answers to that question over time. The meaning of a term has been understood as being an object to which the word refers (Wittgenstein 1921); an abstract content-bearing object (Frege 1892); an idea on the subject's mind (Locke 1690); a hidden description (Russell 1905); something explainable through facts regarding the speakers' intentions in uttering something (Grice 1989); among other views.

Wittgenstein also gave an original and powerful answer to that problem. According to him, the meaning of an expression can be associated with "its use in language" (Wittgenstein 1953, §43). Meaning and usage, in this sense, are related: knowing the meaning of a word *is* knowing how to properly use it in language — that is to say, being able to correctly use the word in sentences; giving adequate explanations for its usage when requested; and being able to correct mistakes and provide standards for its usage.

Although Wittgenstein would have denied that he was advancing any general thesis (Kuusela 2006), it is not hard to see that his notion of meaning is closely related to that of Distributional Semantics. Some of his writings reveal ideas strictly related to Distributional Semantics' main hypotheses. If the meaning of an expression is its usage, then its meaning could be simply reduced to a description of its possible usages. Naturally attached to this notion of meaning is also a plausible definition of synonymy: words that are used in the same way have the same meaning.

Distributional Semantics models are based on the implementation of this contextual notion of meaning. They can roughly be divided into two main categories—counting and prediction models (Baroni et al. 2014). Each of them develops Distributional Semantics' assumption in a slightly different manner, but with similar results (Lenci 2018). Counting models are built by directly extracting statistics of co-occurrence of words. A very simple implementation is to build a co-occurrence matrix A, where each time that $word_1$ appears with $word_2$, the value of a_{ij} increases by one. The outcome, then, is a collection of vectors, one for each word in the vocabulary. Those word representing vectors are also known as "word embeddings". Word-word matrices, however, are not the only possible implementation of counting models. Alternative schemes include, for example, word-document matrices, where rows represent words, and columns represent the documents in which the words are used.

Predictive models are more sophisticated. They produce representations as a byproduct of particular NLP tasks, such as word prediction — whose objective is to find out the correct word given a certain context, or the other way around, *i.e.*, to find the right context given a single word. The algorithms that comprise *Word2Vec* provide a typical example of the aforementioned learning scheme (Mikolov et

al., 2013a). In its two versions, CBOW and Skip-gram, a network with a single hidden layer is initialized with random weights. The parameters of the model are then adjusted through training in order to optimize this predictive task. After this process, the hidden layer serves as a vector representation of the word. The dimension of the hidden layer is a hyperparameter of the model that usually goes between the hundreds and a few thousands.

While the vectors generated by prediction models are *dense* and usually have reasonable dimensionalities, those stemming from count-based models are *sparse* and have a very high number of dimensions. In sparse matrices, each entry a_{ij} represents the frequency of $word_i$ and $word_j$ occurring in the same window. Because they represent each co-occurrence explicitly, they are called *explicit vectors*. Their size is not optimal because most of the words never occur together, or co-occur just a few times; hence, several techniques were developed to reduce the size of such vectors so that the algorithms that operate on them may perform better.[3]

Compact vectors are said to be *implicit*, as the information they carry about a word is not readily interpretable. One does not know what each dimension "represents" in them. On the one hand, this operation produces highly informative compact vectors, which makes them more computationally tractable. On the other hand, though, the model loses some of its interpretability (Lipton, 2016); we do not know for certain what the vectorial space generated by the vocabulary is, and what each dimension means.

[3] Some of these methods of dimensionality reduction, such as Singular Value Decomposition (SVD), are also employed outside natural language processing.

4. The Virtues of Distributional Semantics

Data-driven approaches to natural languages have gained a lot of popularity in recent times. More importantly, their success has changed the landscape of NLP: from automatic translation to dialog generation, a wide range of applications thrived under Distributional Semantics' assumptions.

The main advantage of Distributional Semantics over Formal Semantics is its ability "to acquire semantic representations directly from natural language data" (Boleda & Herbelot 2016, p. 623). Such representations are employed in several tasks with impressive results, such as in machine translation, natural language generation, and summarization. One fact that attests to the quality of Distributional Semantics-based models is its capacity of preserving relations of "semantic similarity"—an umbrella term that encompasses several phenomena. Words that have similar meanings according to human judgment are kept close in vector space. Moreover, Distributional Semantics gives rise to intuitive geometric implementations of this semantic similarity: the similarity of a pair of words, w_1 and w_2, is given by the cosine distance or the Euclidean distance of the vectors, v_1 and v_2, that represent them.[4]

This technique allows for a surprising richness in the information extracted. In addition, word embedding methods can be constructed in such a way as to be sensitive to changes in context. For example: in one context, a cat is closer to a lion than to a dog, as both are felines; in a different

[4] Cosine distance between vectors is, by no means, the only similarity measure available; nor is it the best in every case, although it is a popular and effective one. The choice depends on several factors, including how the vectors were built. Depending on the technique employed, other similarity measures can be more effective.

context, a cat is closer to a dog than to a lion, because both dogs and cats are domestic animals. Semantic relations such as these can be captured by Distributional Semantics, in a way that formal means are not able to do. This can be done by modifying the original embeddings to produce context-sensitive representations of words (Thater et al. 2011; Erk & Padó 2008), or more recently, as exemplified by the BERT framework, through the use of an attention mechanism that generates representations of context-based words (Devlin et al. 2018).

Distributional Semantics models have several other advantages, too. For instance, they can grasp subtle differences expressed in the use of co-referential terms (e.g., "cop" and "police officer"). Furthermore, they can deal with cases of polysemy, as in the difference between the expressions "tall postdoc"—which refers to a person by her actual job—and "long postdoc"—which refers to the duration of an academic research position (Boleda & Herbelot 2016).

Finally, Distributional Semantics models allow the automatic extraction of semantic relationships that are surprisingly close to human ones, as that "queen" is to "woman" what "king" is to "men", and "Paris" is to "France" what "Rome" is to "Italy". These semantic relationships can be found by simply subtracting word embeddings ("Paris" - "France", as a country-capital relation) and then adding the result to the vector of a third word, in order to find out the unknown member of the analogy (from "Rome" to "Italy") (Mikolov et al. 2013a; Mikolov et al. 2013b).[5]

[5] Given that semantic spaces are continuous, we will rarely land on a point that represents an actual word of the vocabulary. The solution proposed by Mikolov et al. (2013a) and Mikolov et al. (2013b) is taking the word with the largest cosine similarity in relation to the calculated point. The success of this procedure,

5. The Shortcomings of Distributional Semantics

Despite its impressive achievements, Distributional Semantics currently presents serious limitations. We will briefly discuss some of the most important ones in this section.

5.1 Normative versus Descriptive View of Meaning

Distributional Semantics assumes that the meaning of a word is a probabilistic function of its linguistic context. If this is the only criterion for meaning attribution, then what we call the incorrect usage of a word (or a sentence) is simply (relatively) a deviant usage; i.e., a situation in which the word does not usually occur and where other words would be expected with higher probability. Distributional Semantics is incapable, in principle, of distinguishing normatively sanctioned formal language and colloquial uses.

---

however, may be often due to features of the neighboring structure, and not due to the discovery of a true similarity relation among words. If a similarity relation results in a small vector, then adding it to a word embedding may not move it enough from the starting point. Thus, by choosing the nearest word we may be simply picking its closest neighbor, irrespective of the similarity relation. For instance, after adding a base-to-gerund similarity to "scream" (which is small), and then looking for the closest word to it, we get the correct answer "screaming". But because "screaming" is much closer to "scream" than any word in the vocabulary, any other small similarity relation would also result in picking "screaming" as the word with the closest cosine similarity to "scream". Linzen (2016) proposes the use of different measures as baselines to calculate the amount of similarity that is being actually captured by the analogy task.

Together with semantic and syntactic information, these models also incorporate biases and implicit prejudices present in actual discourses. This is a consequence of the fact that Distributional Semantics approaches absorb *all* the contextual information present in the data, without any normative guidance. Bolukbasi et al. (2016), for example, trained word embeddings on a Google News corpus and found several sexist gender stereotypes — such as "man" is to "computer programmer" what "woman" is to "homemaker", and "woman" is to "bookkeeper" what "man" is to "warrior". Even more dangerously, algorithms trained on biased corpora may not only reproduce those biases, but may also amplify them (Zhao et al. 2017).

As can be seen, the consequences of this kind of generalization go well beyond the theoretical debate on what constitutes linguistic meaning. As AI algorithms' usage spreads to tasks that go from university admissions to credit analysis, the potential of such systems to reinforce existing biases and coat them under an aura of unavoidable rationality poses a serious problem for society. The fact that many of these algorithms are huge "black-boxes", involving complex calculations and non-interpretable features, makes the issue even more problematic (Rudin, 2019).

Although computer scientists are becoming increasingly aware of possible implicit biases, those problems are seen mostly as matters of data collection and curation, rather than a shortcoming in the assumptions of modern NLP. But we can easily consider forms of biases so prevalent that they can be present in almost any dataset. In this case, from a radical Distributional Semantics point-of-view, meaning representations seem to be unavoidably subject to harmful forms of biases.

5.2 Antonymy, Hyperonymy and Hyponymy

Distributional Semantics has an enormous amount of difficulty in dealing with a series of linguistic phenomena that human beings treat very naturally in daily life, as antonymy, hyperonymy, and hyponymy.

In linguistic theory, antonymy is defined as a pair of words that expresses opposite meanings, often accommodated in the extremes of a spectrum (e.g., "good" and "bad" or "long" and "short"). Hyperonymy and hyponymy are two faces of a hierarchical semantic class-genus relation, such as the one present in "dog" and "domestic animal" — "dog" is a hyponym of "domestic animal" and "domestic animal" is the hypernym of dog.

These phenomena challenge Distributional Semantics models because pairs of words in one of these relations occur in virtually the same contexts. This threatens the usual interpretation of semantic similarity as measured by its corresponding vectors, as they are calculated from very similar contexts (e.g., "this coffee is *hot*" and "this coffee is *cold*). There are also more specific difficulties engendered by the inability of Distributional Semantics in dealing with these linguistic phenomena. Whereas replacing a term for a close word tends to produce a small change in meaning, as in "The student is clever" and "The student is intelligent", replacing a term for its antonymous produces a radical change in the meaning of the sentence, despite these terms being close in the semantic space, as in "The student is dumb". Also, as discussed in more detail in section 5.3, hyperonymy and hyponymy entail specific forms of inferences that are not *prima facie* captured by word embeddings. Important attempts have been conducted to deal with these problems but none of them is entirely satisfactory yet (Lenci, 2018).

5.3 Contrast sets

Understanding the meaning of a word often involves understanding the set of possible alternatives to it. Suppose someone answers negatively when asked if she has a cat. It would sound odd if, following that answer, the person added: "but I have a chair". Implicitly, the context of the utterance indicates and delimits what "fillers" are acceptable. In this case, the person could explain that she has a dog, a bird, or no animal at all, but a chair does not sound as an appropriate answer. This is especially true for words that are part of hierarchical structures or ontologies.

Contrast sets are also involved in the construction of inferences. To give a simple example, in a standard context: if it is true that "*A* is blue", then one can conclude that "*A* is not green". Those kinds of inferences are regularly performed by people because they make use of shared norms of language usage; in this specific case, the rule that governs the use of color ascription, which determines that we cannot attribute more than one color to the same object at the same time in the same spot.[6] Failing to operate appropriately with contrast sets results in syntactically admissible but semantic senseless inferences, as Chomsky's famous "colorless green ideas sleep furiously". Although they are ultimately derived and connected to our world knowledge (ideas do not have colors, do not go to bed, and do not get angry), those norms are an essential element of our understanding and manipulation of language.

[6] One could say that there are plenty of objects with more than one color, as in the expression "a black and white shirt". This, however, is not rigorously what is being said. A black and white shirt is something that has black parts and white parts, but each part has exactly one color.

Inferences backed by implicit information on properties can be formalized in different ways. Suppose that an object *x* is blue and that objects can have only a single color; then, one would be justified in concluding that *x* does not have any color that is not blue. Resorting to second-order logic with equality, one could formalize that by: $\forall x((Blue(x) \wedge Color(Blue) \wedge \forall C(Color(C) \hat{} C \neq Blue)) \rightarrow \neg C(x))$, *i.e.*, every object that is blue, a color, is not C-colored, where C is a color different from blue. These constructions, however, are only rarely explicitly ascertained by speakers and, therefore, cannot easily be directly captured by probabilistic distributions based on word occurrences. More importantly, these are necessary rules of grammar, not empirical regularities.

Recently, some researchers have tried to tackle this problem from the perspective of Distributional Semantics. Kruszewski et al. (2016) modeled contrast sets via the plausibility of alternative sentences. Returning to the cat example, a sentence as "no, I have a dog" would be considered more plausible than "no, I have a chair". The model relies on simple cosine similarity between word embeddings, and even a simple experiment, which employed only unsupervised learning, achieved consistent results. Although their results show that it is by no means impossible to investigate this kind of phenomenon within Distributional Semantics, there are still some questions. The solution works in a constrained environment, tackling a very specific problem; it is not obvious that this kind of technique is scalable to a more general one that encompasses the problem of common sense. Another issue is that there is no explanation for the plausibility: one sentence is more plausible than the other simply because its cosine similarity is higher. A solution that includes some formalized knowledge could pinpoint *reasons* that explain why one sentence is preferable to another. A possibility would be that

"no, I have a dog" is more plausible that "no, I have a chair" because cats and dogs are very common house pets, a topic that fits neatly with the question "do you have a cat?".

Problems of this nature are a fertile ground for hybrid approaches, combining distributional methods with formal ones, and where philosophy, linguistics, and areas of expert knowledge could have valuable input. Those models could employ the powerful methods of distributional semantics as a first step in the construction of robust models of knowledge that are in some sense explicit and that encode aspects of common sense reasoning. Such a strategy could lift techniques that solve well-delimited problems in order to reach a more general framework that enables robust reasoning, combining traditional logical-based methods with common-sense.

5.4 Inference

NLP methods are often related to predictive goals, such as predicting the next word in a sequence or guessing what term would better fill a sentence. For these sorts of tasks, NLP methods have showed incredible results. However, their capacity to deal with more complex phenomena, such as inference, is still quite limited.

An inference $A \vDash B$, in which A and B are pieces of information expressed in some natural language, is valid when the information in A allows one to derive the information in B. This can be defined precisely through some logical calculus, or in a more relaxed way, by taking those inferences that competent speakers with common sense knowledge would judge valid.

Distributional Semantics' difficulty in dealing with inferences is symptomatic of its limited capacity of representing meaning as use, which does not encompass

things as "use in inferences". The problem is that the notion of inference is strictly connected to that of meaning: inferences are not something that we can do when we know words but are *part of the meaning* of words themselves. Inferences, in other words, are not something that is attached later to words *a posteriori*; it is something that is constitutive of words, and so should be present in their representations. Brandom (1995), who proposed an approach known as inferentialism, defends that knowing the meaning of a word is, among other things, being able to draw correct inferences from it.

There are several challenges to model inference employing Distributional Semantics. One of them is how to represent complex information and the compositionality problem, which will be addressed later in this section. Another challenge is how to represent background knowledge.[7] For a person to conclude that "Paris is in Europe" from "Paris is in France", for example, she has to know that France is in Europe and that the property of being physically contained in another place is transitive.

The meaning of words is generally defined as an independent unity, having at most statistical correlation with other words and specific morpho-syntactic features. Inferences are then seen as an external relation that must be attached through some correlation. If we follow the inferentialist view, however, inferential properties should be

[7] Although this background knowledge is not necessarily a *linguistic knowledge*, it is closely related with linguistic competence, e.g., in the Natural Language Inference (NLI) problem. Competent speakers of a language usually agree on what counts as a valid textual entailment. Bos & Markert (2005) measured this agreement and found the astounding agreement rate of 95.25% between humans. Therefore, even if linguistic and world knowledge are theoretically separate, in actual human subjects they are closely related with the faculty of language.

searched simultaneously with semantic and syntactic correlations. This, in turn, indicates that mere vectorial representations may not be enough to represent the meaning of words — we also need to access the meaning that is located at the "glue" that connects the expressions in the language.

As said before, the context in which a word appears may vary, and it is possible for its meaning to change accordingly. There are countless examples of this general phenomenon. Some examples are quantifiers ranging over different domains depending on the context — e.g., "*every* student will take a test" said in a classroom or by a government official — and judgements of qualities — e.g., a tall man is not akin to a tall elephant and may even depend on the particular region of the world where this is uttered. In that case, it is also fundamental that inferential relations be equally adaptable. This suggests that we should look for grammatical rules that are *occasion-specific—i.e.*, varying according to particular circumstances of use (Dobler 2013).

5.5 Implicature

Inferences are generally thought of as entailments — a relation in which the truth of the antecedent leads to the truth of the consequent. In natural language, however, some implications are not entailed from a strictly logical point of view. This is what Grice (1975) called "conversational implicature": something suggested or implied by an utterance, but which is not literally expressed. A good example is someone who in a conversation says "It is getting late". A reasonable speaker understands that as a polite way of ending the conversation (for some of the kinds of implicature, see Davis 2019).

This linguistic phenomenon involves more than hidden or reconstructed logical relations of inference. It is also grounded in world knowledge, such as people's behavior, social norms, and facts about nature. While some of the phenomena related to this more subtle communication can be reduced to frequencies and probabilities — *i.e.*, some politeness rules are just arbitrary conventions that may be emulated by pattern-matching —, there are more complex instances that require robust world knowledge. Left by itself, probabilistic models of meaning have little chance of achieving success in those cases. A possible research path is to enrich NLP models with some kind of world knowledge, sided with formal models of how particular kinds of implicatures work. Here, philosophy and linguistics could be of great help, as both areas have been consistently dealing with problems of this nature for decades.

5.6 Lack of generalization

One of the most remarkable features of natural language is its generalization power. Distributional Semantics, however, is shorthanded in this respect. This is a fundamental problem with the whole set of assumptions backing this approach and not only a matter of fine-tuning algorithms. If meaning stems entirely from context, words that were not seen before are in principle devoid of meaning. This is particularly problematic for proper nouns, which may appear only a few times in very large corpora. Augenstein et al. (2017) demonstrate the difficulties that some machine learning algorithms have in generalizing beyond previously seen features.

In NLP, generalization is often considered as a matter of availability of corpora. The problem, it is thought, is simply that of producing large enough corpora that are

representative of language. Generalization, however, affects more than just lexical semantics. In fact, one of the main obstacles Distributional Semantics faces is giving adequate treatment for one of the most distinctive features of human language: compositionality.

Roughly speaking, compositionality is the thesis that the meaning of complex sentences is built from the meanings of their components down to a fundamental level. One of the main arguments supporting compositionality rests upon the fact that human beings with finite cognitive power and limited access to data can, in principle, generate and understand an infinite number of sentences. When reading a book, a good portion of the sentences is seen for the first time and will never be read again. This, however, does not prevent understanding, even if we have not seen those sentences before. Humans can make sense of unprecedent constructions that they have never faced before and imagine the scenarios described by them.

According to Frege, "the possibility of our understanding sentences which we have never heard before rests evidently on this, that we can construct the sense of a sentence out of parts that correspond to words" (1914, p. 79). Indeed, this is the kind of task in which formal approaches thrive. With a well-defined set of rules for deriving meaning from the combination of smaller parts, programs can easily generate new grammatically correct sentences. Despite the advances mentioned above, in modeling some semantic phenomena (e.g., as capturing the difference between "tall postdoc" and "long postdoc") and impressive performances in natural language inference tasks (Liu et al. 2019; Radford et al. 2018), it is still not obvious how one can make complex logical operations with propositions based on vector representations of word meanings without any explicit knowledge representation.

6. Conclusion

Traditionally, philosophy aims to explain the nature of meaning and how people can understand and produce discourses with sense. On the other hand, NLP is generally more inclined to constructing models for specific tasks, which can be evaluated according to objective performance metrics. Although none of these trends should be taken at face value, they reflect some of the more fundamental aspirations of their respective fields. Through this paper, we discussed some of the advantages and limitations of the main approaches to natural language in both fields. Now, we want to argue that philosophy and computer science could both grow with a more intense dialogue.

The fact that philosophy possesses a more theoretical character while computer science is more prone to applications could foster the exchanges even more. Philosophers could pay attention to the models being developed by computer scientists and reflect on which causes generate results. As Manning & Schütze write, "while practical utility is something different from the validity of a theory, the usefulness of statistical models of language tends to confirm that there is something right about the basic approach" (1999, p. 4). Philosophers could also focus more on stipulating and testing their language models, stating more clearly the empirical consequences of their theories.

Computer scientists, on the other hand, even though they are ultimately concerned with improving mathematical models, could take inspiration in philosophical theories of language, enriching their computational models by integrating them with more formalist approaches, as first and second-order logic. They could also gain a lot from philosophers' more theoretical discussions on the nature of meaning and theoretical problems associated with standard approaches. Brandom's inferentialist approach, which treats

propositions as the basic unity of meaning, may be especially promising as a source for new avenues in NLP. It represents an alternative to the traditional representations of meaning that treat words as isolated objects.

We would like to give some more concrete examples from our previous discussions of how this interaction could benefit both philosophers and computer scientists. Philosophy could gain a lot by incorporating some of Distributional Semantics' typical probabilistic approaches. The theory of meaning as use implies that if two words have the same usage, they have the same meaning. Put differently, it assumes that "the semantic properties of a lexical item are fully reflected in appropriate aspects of the relations it contracts with actual and potential contexts" (Cruse 1986, p. 1). But is the inverse always true? Does synonymity imply equality of usage? According to some philosophers, the answer is negative: sameness of meaning may actually co-exist with a difference in usage (Glock 1996, "use"). Despite being used in different contexts, words such as "cop" and "policeman" seem to have the same meaning. This situation poses theoretical and methodological problems for both philosophers and computer scientists: how large must a change in usage be to result in a difference of meaning? Moreover, what is meaning, if not simply sameness of usage?

Distributional Semantics' operative notion of "synonymy" could help to solve these difficulties. Philosophy could get rid of some of its enduring problems regarding the nature of meaning and its relation to usage by treating synonymy not as a categorical relation but as a matter of degree, a definition that can be measured empirically. This probabilistic approach is also more compatible with the fact that semantic similarity may vary from one context to another. At the same time, philosophers could help computer science to think of ways of

computationally determining more sophisticated criteria for similarity of linguistic context.

Another case that suggests that usage and meaning of a word may not always coincide is that some words have a usage, but no apparent meaning, as in "abracadabra". According to linguists, some words such as expletives have no propositional meaning. Nevertheless, they do have *some* meaning — an expressive meaning, which conveys emotions, expectations, and attitudes (Cruse, 1986, ch. 12). A probabilistic approach allows overcoming the gap between meaning and usage by reducing these words to a function of their contexts; overcoming limitations of a static definition of meaning as usage. It is not obvious how to extract non-propositional meaning from large corpora, but the answer may lie on the search for inferential relations, as discussed above.

Current computational models, in turn, have been facing difficulties in expanding information extracted from a particular corpus to another, showing some of the important limitations of Distributional Semantics models (although there were recently some advances in this area with the transfer learning paradigm, which aims at solving problems with knowledge gathered from one domain to another). Two paths, strongly tied to philosophy, can offer interesting insights and open new lines of investigation. First, the Wittgensteinian idea that language is a rule-governed activity suggests that we should focus on standards of logical and linguistic inferences rather than correlations of words. Second, compositionality indicates that we should look for approaches that mix probabilistic models and formal languages, as defended by Boleda & Herbelot (2016). Also known as "neuro-symbolic" approach, the combination of probabilistic and formal models is a promising field in AI (Garcez & Lamb, 2020). In both cases, philosophy offers a good antidote for simplistic approaches to natural language,

as well as highly technical solutions to a series of specific problems, as exemplified by the use of indexicals and counterfactual propositions; simultaneously, computer scientists can offer the experience and methods they have acquired through the persistent inquiry of natural language.

Acknowledgment

We would like to thank Marcos Lopes and Fabio G. Cozman for helpful discussion and critical commentary, as well as to two anonymous reviewers for their suggestions. We would also like to thank the Center for Artificial Intelligence (C4AI-USP) and the support from the São Paulo Research Foundation (FAPESP grant #2019/07665-4) and from the IBM Corporation; and the Conselho Nacional de Desenvolvimento Científico e Tecnológico – CNPq (grant #168270/2018-8).

References

Augenstein, I.; Derczynski, L.; Bontcheva, K. "Generalisation in Named Entity Recognition: A Quantitative Analysis". *Computer Speech & Language*, v. 44, pp. 61-83, 2017.

Baroni, M.; Dinu, G.; Kruszewski, G. "Don't Count, Predict! A Systematic Comparison of Context-Counting vs. Context-Predicting Semantic Vectors". In: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 238-247, 2014.

Boleda, G.; Herbelot, A. "Formal Distributional Semantics: Introduction to the Special Issue". *Computational Linguistics*, v. 42, n. 4, pp. 619-635, 2016.

Bolukbasi, T. et al. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, v. 29, p. 4349-4357, 2016.

Bos, Johan; Markert, Katja. Recognising textual entailment with logical inference. In: *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pp. 628-635, 2005.

Brandom, R. *Making It Explicit: Reasoning, Representing, and Discursive Commitment*. Harvard University Press, Cambridge, MA, 1995.

Cruse, D. A. *Lexical Semantics*. Cambridge University Press, 1986.

Davis, W. "Implicature". In: E. N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy (Fall 2019 Edition)*, . Available at: https://plato.stanford.edu/archives/fall2019/entries/implicature/.

Devlin, J. et al. "Bert: Pre-Training of Deep Bidirectional Transformers for Language Understanding". arXiv preprint arXiv:1810.04805, 2018.

Dobler, T. "What Is Wrong with Hacker's Wittgenstein? On Grammar, Context and Sense-Determination". *Philosophical Investigations*, v. 36, n. 3, pp. 231-250, 2013.

Erk, K.; Padó, S. "A Structured Vector Space Model for Word Meaning in Context". In: *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pp. 897-906, 2008.

Frege, G. "Letter to Jourdain". In: *Philosophical and Mathematical Correspondence*. University of Chicago Press, pp. 78-80, 1980. [1914]

_____. "Begriffsschrift, a Formula Language, Modeled upon that of Arithmetic, for Pure Thought". In J. van Heijenoort (ed.), *From Frege to Gödel: A Source Book in Mathematical Logic, 1879-1931*. Harvard University Press, 2002. [1879]

_____. "On Sense and Reference". In: P. Geach & M. Black (eds.), *Translations from the Philosophical Writings of Gottlob Frege*. Basil Blackwell: Oxford, 1980, p. 56–78. [1892]

Garcez, Artur D'avila; Lamb, Luís C. "Neurosymbolic AI: the 3rd Wave." *arXiv preprint arXiv*: 2012.05876, 2020.

Glock, H-J. *A Wittgenstein Dictionary*. Wiley-Blackwell, Hoboken, New Jersey, 1996.

Grice, H. P. "Logic and Conversation". *Syntax and Semantics: Vol. 3: Speech Acts*. Academic Press, Cambridge, MA, 1975.

_____. *Studies in the Way of Words*. Cambridge, MA, Harvard University Press, 1989.

Grimsley, et al. (2020). "Why attention is not explanation: Surgical intervention and causal reasoning about neural models." Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 1780–1790.

Hacker, P. M. S.; Baker, P. *Wittgenstein: Rules, Grammar and Necessity*. Wiley-Blackwell, Hoboken, New Jersey, 2009.

Kaplan, D. "Demonstratives". *Themes from Kaplan*. Oxford University Press, pp. 481-563, 1989.

Krahmer, E.; Van Deemter, K. "Computational Generation of Referring Expressions: A Survey". *Computational Linguistics*, v. 38, n. 1, pp. 173-218, 2012.

Kratzer, A. "What 'Must' and 'Can' Must and Can Mean". *Linguistics and philosophy*, v. 1, n. 3, pp. 337-355, 1977.

Kruszewski, G. et al. "There is no Logical Negation Here, But There Are Alternatives: Modeling Conversational Negation With Distributional Semantics". *Computational Linguistics*, v. 42, n. 4, pp. 637-660, 2016.

Kuusela, O. "Do the Concepts of Grammar and Use in Wittgenstein Articulate a Theory of Language or Meaning?". *Philosophical Investigations*, v. 29, n. 4, pp. 309-341, 2006.

Lenci, A. "Distributional Models of Word Meaning". *Annual Review of Linguistics*, v. 4, pp. 151-171, 2018.

Lepore, E.; Ludwig, K. A. "What is Logical Form?". In: G. Preyer (ed.), *Logical Form and Language*. Oxford University Press, 2001.

Lewis, D. "General Semantics". *Montague Grammar*. Academic Press, pp. 1-50, 1976.

Linzen, T. "Issues in Evaluating Semantic Spaces Using Word Analogies". *arXiv preprint* arXiv:1606.07736, 2016.

Lipton, Z. C. (2016). "The Mythos of Model Interpretability". CoRR, abs/1606.03490.

Liu, Y. et al. "Roberta: A Robustly Optimized Bert Pretraining Approach". *arXiv preprint* arXiv:1907.11692, 2019.

Locke, J. *Essay Concerning Human Understanding*. Chicago, Encyclopaedia Britannica, 1955. [1690]

Manning, C.; Schutze, H. *Foundations of Statistical Natural Language Processing*. MIT press, 1999.

Meinong, A. "The Theory of Objects". In R. M. Chisholm (ed.), *Realism and the Background of Phenomenology*. Free Press, pp. 76-117, 1981. [1904]

Miller, T. (2017). "Explanation in Artificial Intelligence: Insights from the Social Sciences". *CoRR*, abs/1706.07269.

Mikolov, T. et al. "Distributed Representations of Words and Phrases and Their Compositionality". *arXiv preprint arXiv*: 1310.4546, 2013. [2013a]

_____. et al. "Efficient Estimation of Word Representations in Vector Space". *arXiv preprint arXiv*:1301.3781, 2013. [2013b]

Moot, R.; Retore, C. *The Logic of Categorial Grammars: A Deductive Account of Natural Language Syntax and Semantics*. Springer Verlag, 2012.

Radford, A. et al. "Improving Language Understanding by Generative Pre-Training". 2018. Available at: https://www.cs.ubc.ca/~amuham01/LING530/papers/radford2018improving.pdf.

Rudin, C. "Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead". *Nature Machine Intelligence*, v. 1, n. 5, pp. 206-215, 2019.

Russell, B. "Our Knowledge of the External World". Routledge: Abingdon, 2009. [1914]

_____. "On Denoting". *Mind*, v. 14, n. 56, pp. 479-493, 1905.

Spärck-Jones, K. *Synonymy and Semantic Classification*. Edinburgh University Press, 1986. [1964]

Strawson, P. F. "On Referring". *Mind*, v. 59, n. 235, pp. 320-344, 1950.

Thater, S.; Fürstenau, H.; Pinkal, M. "Word Meaning in Context: A Simple and Effective Vector Model. *Proceedings of 5th International Joint Conference on Natural Language Processing*, pp. 1134-43, 2011.

Wittgenstein, W. *Philosophical Investigations*. Wiley-Blackwell: Hoboken, New Jersey, 2009. [1953]

_____. *Tractatus Logico-Philosophicus*. Routledge, Abingdon, 2001. [1921]

Zhao, J. et al. "Men Also Like Shopping: Reducing Gender Bias Amplification Using Corpus-Level Constraints". *arXiv preprint arXiv*:1707.09457, 2017.