Wittgenstein's Philosophy of Language
The Philosophical Origins of Modern NLP Thinking

Robert Bain * Andrew Festa † Ga Young Lee ‡ Andy Zhang §

Abstract

Ludwig Wittgenstein's theory of language games introduced a philosophically unorthodox understanding to the meaning and the use of language. Modern natural language processing (NLP) approaches focus on context-derived models of meaning, avoidance of syntactically defined rules, and rely on large bodies of data to statistically approximate our real-world context. This paper traces the origin, development, and intersection of Wittgenstein intellectual legacy and its relevancy to NLP, assesses the ways his thoughts have influenced it, and examines how his work could be better applied in future directions of NLP.

Keywords Ludwig Wittgenstein, Philosophical Investigations, Philosophy, Language games, Language model, Context, Embedding, Natural Language Processing, Cognitive Science

*e-mail: bainro@oregonstate.edu
†e-mail: festaa@oregonstate.edu
‡e-mail: leegay@oregonstate.edu
§e-mail: amornsrb@oregonstate.edu
[1]All authors contributed equally.

1 Opening Remarks

Artificial intelligence is inherently an interdisciplinary pursuit. It pulls inspiration from many different perspectives and domains to build out models for representing knowledge and reasoning. Computer vision draws from theoretical neuroscience as the fundamental underpinnings for convolutional neural networks [26], and reinforcement learning builds on the early work of Ivan Pavlov surrounding classical conditioning [42]. Natural language processing is well posed to borrow from linguistics and language philosophy in order to enhance the capabilities of models. Many early machine translation models in the '70s and '80s looked to linguistic frameworks, such as Chomsky's Universal Grammar, as inspiration for building translation models [14, 28].

Ludwig Wittgenstein was one of the major natural language philosophers of the 20^{th} century and is considered one of the great thinkers of his time [7]. When looking to language philosophy for a muse, Wittgenstein stands poised to fulfill this role. However, his works were not of a sole focus or even a sole narrative. Rather, his works are often characterized as Early or Late Wittgenstein. Early Wittgenstein asserted meaning lay only in those things that could be succinctly stated [64]. This philosophical view is captured in the *Tractatus Logico-Philosophicus* (often referred to simply as *the Tractatus*), published in 1921. Later Wittgenstein took a very different view. In his second work, Wittgenstein argued that language was a game played to describe our thoughts and experiences to others in a similar domain [63]. The work surrounding this view, *Philosophical Investigations* (PI), was published posthumously in 1953.

This paper explores the intellectual influence of later Wittgenstein, traces their lineage from Frege to reinforcement learning to transformers, and finds their spiritual descendants in the field of modern natural language processing.

2 Later Wittgenstein

The ideas of later Wittgenstein are largely focused in his text, *Philosophical Investigations*. In his book, Wittgenstein presents several high-level ideas that have influenced modern science and philosophy which will be uncovered in the following subsections.

1. Context matters
2. Language games
3. Words as tools
4. Language as a form of life
5. Family resemblance

2.1 Context Matters

This idea came to Wittgenstein from Frege's *Context Principle*. Wittgenstein directly mentions Frege in 3 places in *Philosophical Investigations*. The first in §22-23[2] is inconsequential to our thesis. Then, in §49, Wittgenstein expresses that he agrees with Frege's context principle; words do not represent meaning by themselves. Instead, both Wittgenstein and Frege (and later discussed, John Firth) are semantic holists, believing that the other terms surrounding a word also affect its semantics. Paraphrasing John Searle on this idea from his talk with Bryan Magee [29], the basic atomic unit of meaning is not the word but is instead to be found in the context of a whole sentence.

The next time Frege appears in PI is in the context of naming. Naming an object or an event is just part of language. Suppose that we just named things and gave them straightforward mappings

[2] § denotes a citation to a specific section in the first part of PI

between terms and their definitions. Then you might ask, "Who cares?" Wittgenstein would agree; we would be missing most of what constitutes language with such a myopic definition (§49). Naming only makes sense when you take the many other forms of language into account. (§1, §5, §16, §27, §30-31, §38, §206)

The final mention of Frege by Wittgenstein is to make a clear distinction between their philosophies (§71). Frege rejects the utility of fuzzy bordered concepts where Wittgenstein leverages analogies heavily when discussing family resemblance. For further intuition, Wittgenstein points out that we deal with uncertainty all the time in language, and that a lot of common sense dictated by cultural context seems to sneak in and make that uncertainty tractable to us (§76-79).

2.2 Language Games

The second idea of Wittgenstein's we highlight is that of *language games* (*sprachspiel* in his original German manuscripts). He bolsters intuition of his meaning by giving several examples of language games: parents pointing out things and naming them with words (with the hope that their child repeats it), giving and obeying orders, describing properties of objects, drawings conditioned on descriptions of scenes, reporting the news, speculating on said news, crafting jokes, telling them, etc. (§8, §16, §23, §27, §77, §206) Language games are governed by social rules. They take form in living things, and are an inextricably social activity.

Wittgenstein likens this concept of games to an ever-growing and evolving city. For instance, the language of math and chemistry could be seen as distinct suburbs. Each of them has old buildings, new additions, imprecise boundaries, and its own slang (§18, §23). Language games resemble these constantly changing dynamics. Wittgenstein thus rejects the idea that we are all striving towards an ideal language (e.g., logical systems) that is static and universal (§81).

Thus far, we have only given a cursory explanation of what Wittgenstein meant by language games. Later, we will circle back to further clarify this definition after introducing some necessary tools. Wittgenstein attempted to find a satisfactory definition for over 20 years only to fail without a clear-cut explanation. In this paper, we will investigate the fundamentals of Wittgenstein's philosophy around language games and distill them more succinctly in his absence.

2.3 Words as Tools

Next, we move on to introducing *words as tools* (e.g. hammers, screwdrivers, saws, ...) namely called the *use theory of meaning*. This idea is in contrast to his earlier work in the *Tractus*, where he thought that words reflect the structure of reality [64]. Instead, he argues that the literal use of words in the real world, as a social form of life, is the true meaning of words. Using words between people in a reasoned way gives them meaning. Words are actions in and of themselves (§11, §12, §14, §15, §17, §19, §27). John Wisdom credited Wittgenstein as having said at a sciences club, "Don't ask for the meaning, ask for the use." [62].

His utilitarian turn in semantics has been partially attributed to the reading of William James's *Varieties of Religious Experience* (1902) [31]. Nowadays these ideas tend to exemplify pragmatism, a philosophical tradition defining terms of meaning with respect to their practical application.

2.4 Language as a Form of Life

An important concept of language games is captured in what we will refer to as *language as a form of life.* This term, form of life, is brought up only three times in PI, but connects a lot of Wittgenstein's ideas about language together. He views the act of language as inextricably social and as a part of nature.

The first mention of form of life in PI shows that Wittgenstein thought other philosophers were often missing the idea that language serves our condition (§23). We should not separate the investigation of our words and their meanings from natural human existence and everyday usage. Context, usage, and grammar matter in understanding language. Instead, if you remove them, you will move to a bizarre metaphysical space with little or no traction towards the real problems (§38, §107).

2.5 Family Resemblance

The final concept to unpack is that of *family resemblance*. It is a tool of analogies for describing the fuzzy borders around categories. Wittgenstein observes that language games are often unaligned and subject to change. When comparing games, certain elements drop out, and new ones arise arbitrarily. However, there is seemingly no underpinning attribute across them all (see Fig. 1). Words and games are too unrelated and multifaceted to support a logical or clear-cut definition (§65-67, §72-74). Instead, games' and words' meanings have likeness, which Wittgenstein refers to as family resemblance (§10, §12, §65-67, §72-74, §130, §185). In other words, there are degrees of belonging to a category where elements in each set share some common attributes.

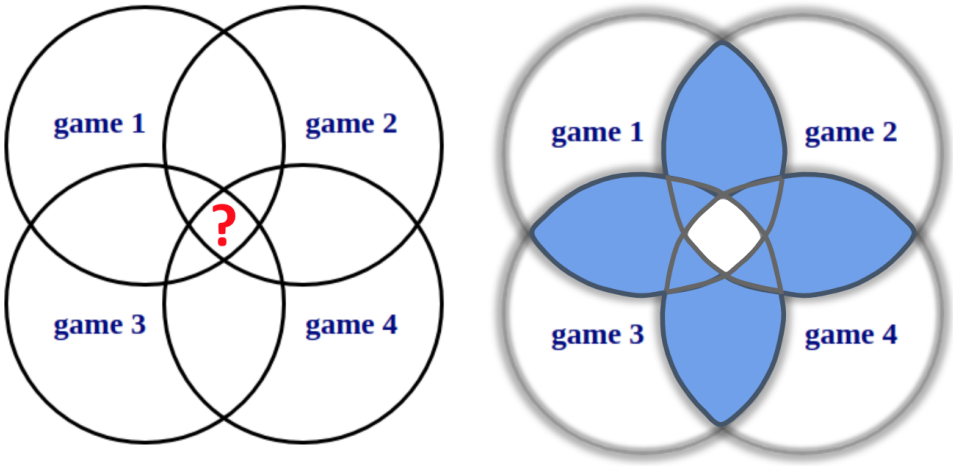


Figure 1: *Left*: Overlap of games, demonstrating that a logical definition of *language games* is difficult since there are few shared attributes (perhaps none) between all games. *Right*: The later Wittgenstein thought that fuzzy borders, where family resemblance resides, was enough for traction in language philosophy.

Words and their meanings are not set in stone. Instead, they are growing, multifaceted, and fuzzy, requiring some inherent and unconscious similarity measure (§10, §18, §23, §69-71, §76-77, §106). There are many different aspects that can be shared, a crisscrossed fabric of relations that Ludwig likens to family resemblance. Some offspring might have its dad's nose, or its mom's eyes, but members of a family often don't have single unifying features. Meanings and uses of words are similar. They can have little pockets of overlap, forming subgroups, but there might not be a single essence that they all share.

The set of all games is one big family (i.e., category). Numbers are given as an example of a smaller sub-concept (§10, §68). The closest thing to a common feature shared by all games in the whole family might be how they are socially governed, contextually bounded, and subject to change over time (§258-261, §268-269, §271).

Wittgenstein hopes to prove that not only does word context matter, but so does the social context in which they appear. At several points in PI, he points out a philosophic tradition of plucking words out of context and then over-analyzing them. He calls this practice "language going on holiday" and asserts that this creates superficial problems for philosophers, and gets them lost in a metaphysical space with no pragmatic use. We cannot stand outside of language to try and make sense of it (§38, §107).

3 Wittgenstein's Kinder

This section details modern ideas in AI that share a family resemblance with those of the late Wittgenstein.

3.1 Context in Natural Language Generation

Firth's and later Wittgenstein's definition of meaning is essentially the same [21]. Firth's most cited quote in NLP is, "You shall know a word by the company it keeps." Firth invokes Wittgenstein's language games in the sentences preceding and following the quote [58]. This notion of context can be traced from Frege and Malinowski to Wittgenstein and Firth [48]. Though Wittgenstein may not have been the originator of this view of context, Wittgenstein was thorough in articulating it and developing it into a theory of language around the notion of context mattering.

OpenAI's GPT-3 transformer model serves as a modern exemplar of Wittgenstein's idea that context matters [11]. Their GPT-3 model has recently made large waves in the natural language processing (NLP) community, building off of a tradition of unsupervised pre-training like word2vecs [37], GLoVe [43], and other word-space models [50]. These methods rely on word meaning correlating with statistics of word usage derived from large text corpora. Unfortunately, the length of context is usually a fixed size, which is at odds with our intuitions of human language understanding.

GPT-3 ended up being more of a hardware success than a software one. GPT-3 is basically a scaled up version of GPT-2. It does very well at grammar, yet its ability to reason is pointillistic [35], as is typical of large deep neural networks (DNNs) used in NLP [6, 40].

3.2 Evolutionary Algorithms

Given that Wittgenstein's idea that language is a form of life, there must be a way to analyze language with the tools of evolution. Looking at culture, which encompasses language, through the lens of evolution from the 1960s to the 1980s came to be known as dual-inheritance theory [13, 65].

The conditions for natural selection outlined in Charles Darwin's *On the Origin of Species* are as follows [15]:

1. Individual variation
2. Variable success at reproduction
3. Heritability

This list applies to more than just genes or biological species. These conditions are intended to detail sufficient and necessary requirements for evolution by natural selection in any medium. One can begin to speculate if they apply to cultural ideas as well. This question prompted the development of the primary idea behind memetics, the study of memes. At an introductory level,

memes are ideas that need human brains to live [9,17,18,19]. In this process, memes get copied into a variable number of other human minds (i.e., they have variable fitness), and an idea is never the same for very long. They mutate like genes do.

Richard Dawkins was the first to coin the term, "memes" [17]. In the preface to Susan Blackmore's *The Meme Machine*, Dawkins recalls the following anecdote. He was joking around and mimicking a style of thinking employed by one of his former students [9]. The student would put her head down for quite a while, and then surface again after a minute or so and clearly articulate the points she had just thought of. The listeners recognized this sequence of actions as one of Wittgenstein's memes. Dawkins had unknowingly been a 4th generation Wittgenstein meme replicator, even though he was doing his interpretation in jest, the meme persisted and replicated. The resulting meme was not an exact copy, but definitely had a strong family resemblance!

Daniel Dennett is another champion of memetics, a cognitive scientist, and one of the most influential philosophers alive today. In his 2017 book, *From Bacteria to Bach and Back* [19], he suggests that we domesticated words like we domesticated dogs, pigeons, and mustard plants. Words probably began as inconsequential memes and could have gone unnoticed for quite some time. Dennett mentions how babies are similarly the beneficiaries of words long before they realize what they are.

> Eventually [infants] reach a state where the words in their manifest image become their own words, affordances that belong to their kit, like a club or a spear...

This is reminiscent of Wittgenstein's pragmatic beliefs about meaning, language, and words as tools. Darwin was also thinking about the origin of language in such terms in 1871 [16].

> I cannot doubt that language owes its origin to the imitation and modification, aided by signs and gestures, of various natural sounds, the voices of other animals, and man's own instinctive cries.

Darwin believed that language developed as tools to address the challenges experienced by early hominid forms of life on their long journey to becoming human.

3.3 Reinforcement Learning Applied to Memes

To expand the notion of evolution from genes to cultures, we must ask how our preferences shaped the evolution of our ideas. Our preferences are dictated by our goals, and a maturing field of tools for understanding human goals is the reward system in neuroscience and psychology. Neuroscientist Peter Dayan has been largely responsible for introducing the language games of reinforcement learning (RL) to neuroscience. In a 1997 paper Schultz, Montague, and Dayan make the case that the majority (55 - 80%) of dopamine neurons in the ventral tegmental area (VTA) of monkey brains encode an error signal from the RL models of Richard Sutton and Andrew Barto [53]. Their models were developed about a decade prior and named Temporal Difference (TD) models [47]. Their work on TD models began at least as early as 1982 [5], where they tried making a mathematical framework to explain classical conditioning (e.g. the data collected by Pavlov's experiments, dating back to the 1920's [42]).

Sutton and Barto's TD model has become quite significant in RL and AI communities nowadays, as it served as the direct predecessor of SARSA [49], Q-Learning [60], and the critic component of Actor-Critic methods. An actor-critic method A3C was used to train DeepMind's AlphaGo to be the best player as of yet at the board game Go [55].

$$V(t) = \mathbb{E}\left[\gamma^0 r(t) + \gamma^1 r(t+1) + \gamma^2 r(t+2) + \ldots\right]$$

Equation 1. Value function from reinforcement learning.

We can use the value function from RL (see Eq. 1) to discuss the utility of our memes. The value function takes the current timestep into the environment as input and returns how much immediate reward and discounted future reward an optimal agent would get on average from the given starting configuration. The discounting factor $\gamma \in [0, 1]$ controls how short sighted the agent is. As an example, death is an apparent threat to all our values since it removes the possibility of future rewards. Part of the success of the Abrahamic religions' memes surely comes from the reassurance provided by replacing eternal nothingness with an ever-loving parent. The great majority of people are seemingly well intended. Their goals are moral, yet their memes are poor heuristics with respect to their underlying values. The price we pay for comfort is often ignorance.

3.4 Sapiens Memed Best

Yuval Harari's book, *Sapiens* [23], offers further insight into the usefulness that fictitious memes might have played in our evolution. Language is often thought to have evolved to communicate accurate and pragmatic information among individuals, including gossip. Models of gossip suggest that it enables group sizes not much larger than 150 people [56], yet many humans today go well beyond that when associating themselves with any of the hundreds of millions of people in their nation that they will never meet. There had to be something more and new that allowed further cooperation among conspecifics. Harari postulates that it was imagined beliefs that made all of the difference between *Homo sapien's* first failed effort at combating neanderthals 100,000 years ago and their success only 30,000 years later (a blink of an eye in evolutionary timescales).

The success is speculated by Harari to have come from a cognitive revolution that began up to 70,000 years ago. The invention of tools, fire, and an ever-expanding neocortex enabled *Homo sapiens* to revise its behaviour rapidly via (in this paper's terms) meme selection. Cultural evolution's rate of change far outpaces that of genetic evolution. With it, *Homo sapiens* quickly became the most effective cooperators on the planet. This enabled success on sapien's second venture north into Neanderthal territory.

> In a one-on-one brawl, a Neanderthal would probably have beaten a Sapiens. But in a conflict of hundreds, Neanderthals wouldn't stand a chance. Neanderthals could share information about the whereabouts of lions, but they probably could not tell stories about tribal spirits. Without an ability to compose fiction, Neanderthals were unable to cooperate effectively in large numbers, nor could they adapt their social behaviour to rapidly changing challenges.
>
> —Yuval Harari in *Sapiens*

3.5 Language, Analogies, and Neuroscience

The first chapter of Douglas Hofstadter and Emmanuel Sander's book *Surfaces and Essences* is all about theories of concepts and categories in language, of which they attributes modern conceptions to later Wittgenstein [34]. They also briefly cover the long history of analogy usage in philosophy. Plato and Aristotle liked the idea of using analogies, but did warn about it being a slippery slope

(amusing considering Wittgenstein thought much of philosophy was tractionless without them). Immanuel Kant and Friedrich Nietzsche were both adamant supporters of analogies. Empiricists and positivists traditionally admonish them (e.g., Thomas Hobbes and John Locke). Hobbes unironically used metaphors when denouncing their usefulness. A more recent exploration of analogy usage is Hofstadter's and his previous Ph.D. student Melanie Mitchell's *Copycat* [24], a model of analogy making and human cognition. Hofstadter is perhaps more famous for introducing many people to the field of AI through his 1979 book *Gödel, Escher, Bach* and for asserting that analogy is the core of cognition [25].

One of humanity's most celebrated thinkers, physicist Roger Penrose, also has theories about cognition. He believes that it involves quantum mechanical (QM) effects in mitotic cell division. Much of QM is not known, assuredly even to Penrose, and he knew even less about the biology of mitosis he referenced, and **even less** about cognition. Penrose appealed to ignorance, while at the same time reaching for his golden memes (i.e golden hammer: everything looks like a nail if you have a hammer) in an attempt to fill in gaps of understanding. As more and more fields develop their suburbs of language games to describe cognition, we should feel progress from all sides. The heavily interdisciplinary field of neuroscience seems to be the arena where this is happening today with language. The uncertainty surrounding measuring memes, language meaning, and understanding the human ability to incessantly make analogies underscores the importance of future inquiries into their respective neural substrates.

3.6 New Memes Still Required

Perhaps Wittgenstein made a similar move to Penrose's when employing analogies. They are not understood at a satisfactory level and can afford fallacies, yet were offered as a tool for philosophical problems. Is that not passing the problem off to someone else such as Mitchell, Hofstadter, or us? Perhaps we also used analogies in a philosophical context too judiciously by tying these ideas together under the guise of family resemblance.

Modern approaches to NLP put the onus on the data and hardware used to train the model. The datasets are required to be fully encompassing, and the hardware is simply scaled up. Yet, they are only so capable. There is seemingly a gap between its ability to model language and the desired behavior we would expect of such a model. That is, there comes a point where simply throwing more resources at a problem is insufficient to yield significant results. For example, GPT-3 was just a scaled up version of GPT-2, but required a demoralizing amount of compute and energy considering the improvement in the results [11]. Deep RL is similar in that it is often wasteful and over-hyped. Since Deep RL agents adopt all of the problems of DNNs, they have narrow intelligence and cannot generalize outside the training distribution, or effectively leverage previous tasks' training when learning new ones [20, 51, 45, 1]. DNNs are already data wasteful, adding RL just compounds that wastefulness [46]. E.g. AlphaStar took 600,000 years of in-game self-play to train, and it still was not the best player at StarCraft [59].

We should aim to create machine intelligence more efficiently than biological intelligence was found via evolution, testing trillions and trillions of policies in parallel over astronomical timescales. In the age of climate change, as entropy becomes more and more scarce, the value of compute will only continue to climb.

While Wittgenstein definitely left an impressive mark on many fields through his strange inversions of reasoning, his spiritual decedents in modern AI still leave much room for improvement.

4 Wittgenstein's Presence in NLP

4.1 A Wittgensteinian View of Language

Wittgenstein's theories reveal useful insights about how to approach language, not just philosophy. Wittgenstein's theory of language encompasses a way of thinking about language, as opposed to a set of concrete techniques, so it is more valuable to think of what a Wittgensteinian approach to thinking about language looks like. Some principles that could fit into this are as follows:

- Language derives its meaning from its usage and context
- Language uses are irreducibly complex
- The flexibility of language usage makes it unamenable to rules

These Wittgenstein characteristics are present in varying ways in the modern field, but were not the dominant approach until the latter part of the 20 century. A brief history of the field is useful in understanding this change.

4.2 Historical Background

It is important to contextualize Wittgenstein's way of thinking about language, and to contrast it with some major trends in NLP during the early 20th century.

Intellectually, linguistics were powerfully influenced by structuralism, the belief that a component could be broken into constituent components, or that there was a fundamental structure to language independent of experience [38]. These views are counter to Wittgenstein's later views which emphasized the importance of language learning from experience. On the engineering side, it's probable that scientists were encouraged too much by the success of cryptography and code-breaking, and likewise sought a rule-based approach to tackle translation. In a 1947 letter, American scientist Warren Weaver wrote [61].

> When I look at an article in Russian, I say: 'This is really written in English, but it has been coded in some strange symbols. I will now proceed to decode.'

Early approaches sought to expand the relatively rule-friendly field of cryptography to natural languages [61]. The 1954 Georgetown-IBM experiment demonstrated Russian-English translation on an IBM 701, based on only six rules [27]. This model was based on several fundamental logical rules, and its output was limited to a carefully selected corpus. Its creators claimed that machine translation would be solved within a decade, but the following half-century was defined by gradual progress based on the proliferation of various, reductive-rules. For a while, new models continued demonstrating modest improvements, but on the whole they still struggled with homonyms, polysemy, ambiguity, and idioms. Machines continued to appear a long ways away from human understanding.

4.3 The Paradigm Shift: Rules to Context

The shift towards context began in earnest during the 1980s, with the introduction of Statistical Machine Learning (SMT) which applied probabilistic methods over large bodies of text in the absence of a large set of handwritten syntactic rules [39]. IBM was one of the early adopters of SMT, and used it to great success in machine translation [10]. This was followed by Neural Machine Learning (NMT), introduced to the world in 2011, based on an recurrent neural network (RNN) framework [52].

Google Neural Machine Translation (GNMT), introduced in 2016, was Google's first switch from SMT to NMT. It considers entire sentences as input rather than decomposing them into their constituent parts [30]. As of 2020, the original GNMT model has been replaced by a updated version with a transformer encoder and decoder, as well as incorporating web crawl and noise modeling. These changes further improve translation tasks of both high-resource and low-resource languages [12].

In addition to running into practical limitations with rule-based machine learning, the shift to context is in part owed to the titanic improvements in computational power and Big Data in the 21st century. Machine learning is computationally demanding and requires large amounts of data, prerequisites that simply did not exist in the mid-20th century. In some sense, the digital realm advanced to the point that it can better simulate the type of real-life context that Wittgenstein described.

There are inherent limits to a rules-based approach. The wisdom of the last half-century has demonstrated that contrary to intuition and scholarship, a context based-approach (or even a model without syntactic rules) outperforms a traditional rules-based approach.

None of this discredits the alternative theories, but it does demonstrate that when it comes to NLP, Wittgenstein's view, typified by his works and the works of his intellectual descendants, has been more pivotal.

4.4 The Current State of the Field

Modern NLP has borrowed Wittgenstein's view that a word's meaning lies in its use rather than seeking a universally true rule. The field has moved away from reductionist, syntactical formulations of language, towards context-driven NLP tools. This is partly a response to the limited success of aforementioned approaches.

Context is everywhere in NLP. word2vec applies distributional semantics to infer word meanings based on embeddings [37]. attentions assigns emphasis to word vectors in a corpus-based on context [4]. These are all examples of the momentum in that context. Models no longer treat words in isolation but instead consider them within the context of entire bodies of text.

Another area in which NLP has moved towards Wittgenstein is its embracing of Big Data. Hardware improvements and the Internet have resolved the lack of compute and digital text that has that stymied the field for years. Now, models can be tested on billions of documents and process just as many parameters. Vendors such as Google Translate and Voice Assistants, rely on a large-base of user-analytics to further improve ("contextualize," if we are speaking in Wittgensteinian terms) and improve upon their pre-existing models.

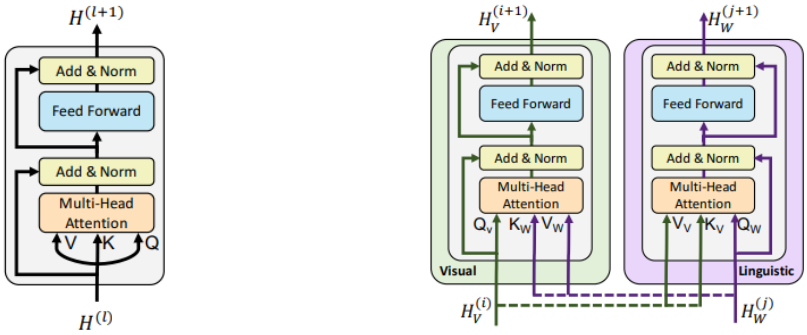
Wittgenstein's ideas have their direct descendants in modern context-based techniques such as word2vec and Attention, but in addition to these concrete influences, there exists a distinctively Wittgenstein-ian way of *thinking about language* that is pervasive in the field of modern NLP, which has reasserted itself.

5 Context-driven NLP Techniques

5.1 Moving Away from Definition to Inference

There are some areas where modern approaches fall short of Wittgenstein's ideals. For example, Wittgenstein's view of context is situational, whereas modern NLP techniques primarily rely on textual context. NLP models infer meaning and relationships mainly based on what is knowable in the text, a more limited definition to the variety of different types of context through which language

learning and use takes place. This is not critical to just improving the state of art NLP models, but also to address social issues such as bias [54].



(a) **(b)**

Figure 2: a) A regular transformer layer. b) ViLBERT's co-attention layer. Note that the stack on the right is processing visual information, but the attention is guided by processing from the textual context (and vice versa). Figure adopted from [33]

A recent trend in the literature is emerging of using multiple modalities, and thus conditioning on more than just textual context. A prime example is ViLBERT [33], a variant of BERT that pretrained on visual captioning and is then fine-tuned for a variety of downstream tasks. It takes as input both text and image context. Separate transformer layers process them, but they are combined cleverly through a co-attention transformer layer. Here, the two different modal transformer backbones swap information in the form of key-value pairs (but not query vectors), allowing image-conditioned language attention in the visual backbone and language-conditioned image attention in the text backbone. Pretraining on 3.1 million image and caption pairs, and then fine-tuning on specific tasks allowed them to achieve state-of-the-art performance on visual question answering, visual commonsense reasoning, referring expressions, and caption-based image retrieval datasets. Thus, more context beyond textual is proving useful in modern NLP.

5.2 Temporal Logic as a Basis of Language Game

The latest NLP techniques consider varying degrees of textual context, but Wittgenstein's view of context is broader, encompassing cultural norms and physical situations. While written and spoken language reveal important aspects of human intelligence and communications, even humans sometimes struggle to understand when dissociated from a particular physical context.

Temporal probabilistic distribution of words matters because each word is represented in a single vector that considers the frequency of appearance regardless of the neighboring words. Therefore, the sequence of events is essential in deciphering the meaning of context. Depending on the surrounding text, a word's meaning can change significantly. Polysemy is a characteristic of words that can have multiple meanings and functions in a sentence when placed with other words [41]. Example 1 demonstrates this edge case of language.

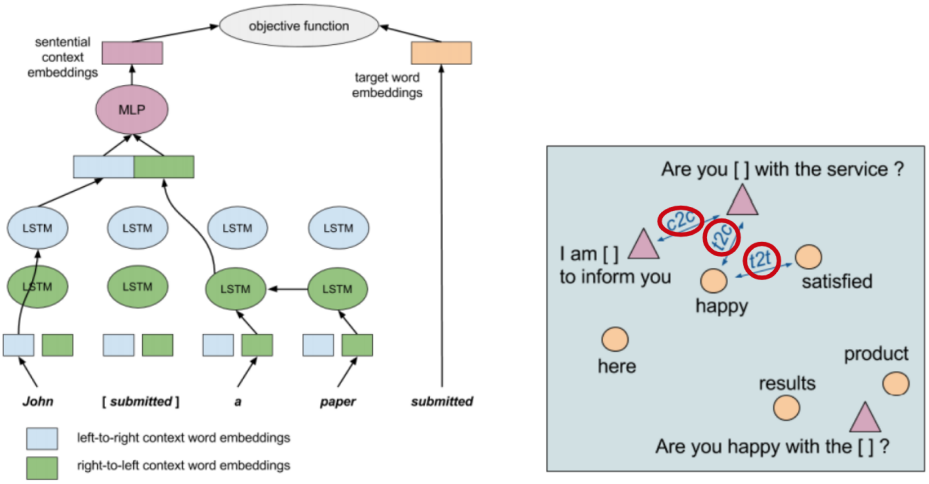


Figure 3: *Left:* (context2vec architecture) bidirectional LSTM and cloze prediction objective, *Right:* (context2vec embedding rules) *c2c*: relationship between context and context. *t2c*: relationship between text and context. *t2t*: relationship between text and text. Figures from [36].

$$\text{f(play|The kids play a game in the park.)} \neq \text{f(play|The Broadway play premiered yesterday.)}$$

Equation 2. Distributed vector representation fails to consider different meanings of play

In Eq. 2, *play* followed by *a game* is a verb while the same word preceding *premiered* is a noun describing theatrical art. Such examples of polysemy illustrate the importance of considering temporal logic in defining language games. As a way to incorporate temporal logic, event extractions methods show a promising result in connecting relevant word entities to identify hidden or implied information that is not explicit in the text [22].

5.3 knowledge graph (KG) to Represent Facts About the World

KG for Encoding Common Sense: However, event extractions is not widely applicable as the amount of information to feed the encoder can exponentially grow as there is an infinite number of events that can occur sequentially. Therefore, there is a clear need to distill the casualty and represent the relations efficiently.

Knowledge graph (KG) can offer a solution. KG is useful for representing complex information that entails lots of relations in a simplified and elegant form [2]. KG's are regular models that incorporate contextual information into a model. This approach

augments inference ability by teaching the relationships between entities rather than teaching individual entities [57]. In Figure 4, World-Knowledge Base is encoded in the feature input to enrich the descriptiveness of implicit and explicit information to improve prediction accuracy.

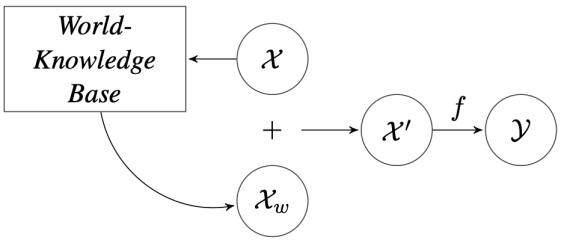


Figure 4: The relevant world knowledge for the task χ_w augments feature existing knowledge χ to improve the final prediction, $\mathcal{Y}$. Figure taken from [3].

KG Applied to CNN: One of the most promising breakthroughs in the use of KG in language models is KG applied convolution-based entity and relation vector relations [3]. In this model, the entity relationships about general knowledge about the world are represented in a graph format to be used for training CNN with pooling [32]. With the K-Nearest Neighbor algorithm, CNN is able to identify clusters of closely related entities [8], which effectively removes the redundancies in processing the whole text corpus. In Figure 5, the mechanism of adding KG into CNN is illustrated. This clustering method is effective for reducing the attention space by learning the representation of similar entities or relation vectors and focusing on them using reduced search space.

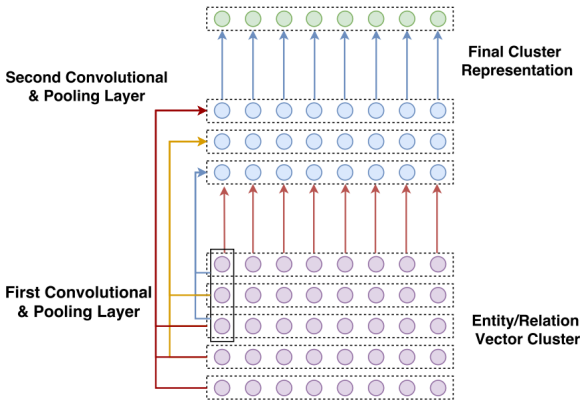


Figure 5: Convolution model cluster application. This figure is originally from [3].

KG Applied to Pre-trained LM: Pre-trained language representation models, such as BERT capture a general language representation from large-scale corpora, but it is

significantly lacking in the ability to infer domain-specific knowledge. When handling a domain-specific text, experts leverage their prior knowledge to make inferences. A novel method to inject KG into a pre-trained model is recommended to enable machines to accomplish this. However, too much knowledge incorporation may divert the sentence from its correct meaning, which is called knowledge noise (KN) issue. This phenomenon is different from a conventional idea of noise as dirty data such as erroneous or missing values in databases [44]. K-BERT suggests a framework to use a visual matrix to limit and control the impact of knowledge injected and set a soft position to overcome KN. Another benefit of K-BERT is that it can augment specific domain knowledge into the model without pre-training, which reduces the time required for running the model. This enhanced capability to load parameters is from the pre-trained BERT.

6 Closing Remarks

Throughout Wittgenstein's life, he interacted with numerous philosophers, linguists, and scientists across disciplines, and challenged the traditional notions of language and conceptual systems. He coined novel concepts such as language games, family resemblance, and words as tools during his lifetime. He offered a new viewpoint of language as its own entity, evolving organically like a living organism or a burgeoning city. Based on the principles of Wittgenstein's philosophy, the latest advances in the NLP field were examined and traced back to Wittgenstein's philosophy of language.

Though Wittgenstein's influence is not always overt, NLP has come a long ways by drawing from Wittgenstein's legacy. However, there is still a ways to go before NLP captures the full complexity of language. Existing approaches, such as word2vec and neurosymbolic knowledge graphs, are not on par with human-level language comprehension and generation ability. To overcome this limitation, the Wittgenstein paradigm of context-focused language can lay the foundation for understanding, modeling, and building a natural language system that accurately depicts the intricacies of language. Furthermore, for future research directions, language models can bridge the gap between word representations and reality by augmenting inference capability by applying various techniques such as event extraction and knowledge graphs.

References

[1] Alessandro Achille, Michael Lam, Rahul Tewari, Avinash Ravichandran, Subhransu Maji, Charless Fowlkes, Stefano Soatto, and Pietro Perona. Task2vec: Task embedding for meta-learning, 2019. arXiv:1902.03545.

[2] K. M. Annervaz, Somnath Basu Roy Chowdhury, and Ambedkar Dukkipati. Learning beyond datasets: Knowledge graph augmented neural networks for natural language processing. *CoRR*, abs/1802.05930, 2018. URL: http://arxiv.org/abs/1802.05930, arXiv:1802.05930.

[3] K. M. Annervaz, Somnath Basu Roy Chowdhury, and Ambedkar Dukkipati. Learning beyond datasets: Knowledge graph augmented neural networks for natural language processing. *CoRR*, abs/1802.05930, 2018. URL: http://arxiv.org/abs/1802.05930, arXiv:1802.05930.

[4] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate, 2014. cite arxiv:1409.0473Comment: Accepted at ICLR 2015 as oral presentation. URL: http://arxiv.org/abs/1409.0473.

[5] Andrew G. Barto and Richard S. Sutton. Simulation of anticipatory responses in classical conditioning by a neuron-like adaptive element. *Behavioural Brain Research*, 4(3):221–235, 1982. URL: https://www.sciencedirect.com/science/article/pii/0166432882900018, doi:https://doi.org/10.1016/0166-4328(82)90001-8.

[6] Emily M. Bender and Alexander Koller. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198, Online, July 2020. Association for Computational Linguistics. URL: https://aclanthology.org/2020.acl-main.463, doi:10.18653/v1/2020.acl-main.463.

[7] Anat Biletzki and Anat Matar. Ludwig Wittgenstein. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Spring 2020 edition, 2020.

[8] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[9] Susan Blackmore. The meme machine, 1999.

[10] Peter F. Brown, John Cocke, Stephen A. Della Pietra, Vincent J. Della Pietra, Fredrick Jelinek, John D. Lafferty, Robert L. Mercer, and Paul S. Roossin. A Statistical Approach to Machine Translation. *Computational Linguistics*, 16(2):79–85, 1990. URL: https://aclanthology.org/J90-2002.

[11] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020. arXiv:2005.14165.

[12] Isaac Caswell and Bowen Liang. Recent Advances in Google Translate, June 2020. URL: http://ai.googleblog.com/2020/06/recent-advances-in-google-translate.html.

[13] Luigi Luca Cavalli-Sforza and Marcus W. Feldman. *Cultural Transmission and Evolution (MPB-16), Volume 16: A Quantitative Approach. (MPB-16).* Princeton University Press, 2020. URL: https://doi.org/10.1515/9780691209357, doi:doi:10.1515/9780691209357.

[14] Noam Chomsky. *Aspects of the Theory of Syntax.* The MIT Press, Cambridge, 1965. URL: https://mitpress.mit.edu/books/aspects-theory-syntax.

[15] Charles Darwin. *On the Origin of Species by Means of Natural Selection.* Murray, London, 1859. or the Preservation of Favored Races in the Struggle for Life.

[16] Charles Darwin. *The Descent of Man.* Princeton University Press, 1981.

[17] Richard Dawkins. *The selfish gene.* Oxford University Press, New York, 1976.

[18] Richard Dawkins. Richard dawkins — memes — oxford union, 2014. URL: https://www.youtube.com/watch?v=4BVpEoQ4T2M&t=45s.

[19] Daniel Dennett. *From Bacteria to Bach and Back: The Evolution of Minds.* 12017.

[20] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks, 2017. arXiv:1703.03400.

[21] Béatrice Godart-Wendling. Firth and Wittgenstein via Malinowski : a shared influence leading to differences. In *Henry Sweet Society Colloquium*, Sheffield, United Kingdom, September 2010. URL: https://hal.archives-ouvertes.fr/hal-01314169.

[22] Rujun Han, Qiang Ning, and Nanyun Peng. Joint event and temporal relation extraction with shared representations and structured prediction. *CoRR*, abs/1909.05360, 2019. URL: http://arxiv.org/abs/1909.05360, arXiv:1909.05360.

[23] Yuval Noah Harari. *Sapiens.* Harper; Illustrated edition, 2015.

[24] D. R. Hofstadter and M Mitchell. The copycat project: A model of mental fluidity and analogy-making. In *Fluid Concepts and Creative Analogies.*, chapter 5, pages 205–267. Basic Books, 1995.

[25] Douglas Hofstadter. Epilogue: Analogy as the core of cognition. In Dedre Gentner, Keith J. Holyoak, and Boicho N. Kokinov, editors, *The Analogical Mind: Perspectives from Cognitive Science*, pages 499–538. MIT Press, 2001.

[26] David H. Hubel and Torsten N. Wiesel. Receptive fields of single neurons in the cat's striate cortex. *Journal of Physiology*, 148:574–591, 1959.

[27] W. John Hutchins. The Georgetown-IBM Experiment Demonstrated in January 1954. In Robert E. Frederking and Kathryn B. Taylor, editors, *Machine Translation: From Real Users to Research*, Lecture Notes in Computer Science, pages 102–114, Berlin, Heidelberg, 2004. Springer. doi:10.1007/978-3-540-30194-3_12.

[28] William J. Hutchins. Machine translation: A brief history. 1995.

[29] Bryan Magee John Searle. Ludwig wittgenstein. URL: https://www.youtube.com/watch?v=xCZ3Qnf6DsI&t=353s.

[30] Quoc V. Le and Mike Schuster. A Neural Network for Machine Translation, at Production Scale. URL: http://ai.googleblog.com/2016/09/a-neural-network-for-machine.html.

[31] Catherine Legg and Christopher Hookway. Pragmatism. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Summer 2021 edition, 2021.

[32] Weijie Liu, Peng Zhou, Zhe Zhao, Zhiruo Wang, Qi Ju, Haotang Deng, and Ping Wang. K-bert: Enabling language representation with knowledge graph. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(03):2901–2908, Apr. 2020. URL: https://ojs.aaai.org/index.php/AAAI/article/view/5681, doi:10.1609/aaai.v34i03.5681.

[33] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks, 2019. arXiv:1908.02265.

[34] Chris A Mack. *Surfaces and Essences*. Basic Books, 2013.

[35] Gary Marcus. The next decade in ai: Four steps towards robust artificial intelligence, 2020. arXiv:2002.06177.

[36] Oren Melamud, Jacob Goldberger, and Ido Dagan. context2vec: Learning generic context embedding with bidirectional LSTM. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 51–61, Berlin,

Germany, August 2016. Association for Computational Linguistics. URL: https://aclanthology.org/K16-1006, doi:10.18653/v1/K16-1006.

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

[38] Daniele Moyal-Sharrock. Universal Grammar: Wittgenstein Versus Chomsky. In Michael A. Peters and Jeff Stickney, editors, *A Companion to Wittgenstein on Education: Pedagogical Investigations*, pages 573–599. Springer Singapore, Singapore, 2017. doi:10.1007/978-981-10-3136-6_38.

[39] Prakash M Nadkarni, Lucila Ohno-Machado, and Wendy W Chapman. Natural language processing: an introduction. *Journal of the American Medical Informatics Association*, 18(5):544–551, September 2011. doi:10.1136/amiajnl-2011-000464.

[40] Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. Adversarial nli: A new benchmark for natural language understanding, 2020. arXiv:1910.14599.

[41] Qiang Ning, Hao Wu, Haoruo Peng, and Dan Roth. Improving temporal relation extraction with a globally acquired statistical resource. *CoRR*, abs/1804.06020, 2018. URL: http://arxiv.org/abs/1804.06020, arXiv:1804.06020.

[42] I. P Pavlov. *Conditioned reflexes*. Oxford University Press, 1927.

[43] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. volume 14, pages 1532–1543, 01 2014. doi:10.3115/v1/D14-1162.

[44] Jose Picado, John Davis, Arash Termehchy, and Ga Young Lee. Learning over dirty data without cleaning. *CoRR*, abs/2004.02308, 2020. URL: https://arxiv.org/abs/2004.02308, arXiv:2004.02308.

[45] Kate Rakelly, Aurick Zhou, Deirdre Quillen, Chelsea Finn, and Sergey Levine. Efficient off-policy meta-reinforcement learning via probabilistic context variables, 2019. arXiv:1903.08254.

[46] Kate Rakelly, Aurick Zhou, Deirdre Quillen, Chelsea Finn, and Sergey Levine. Efficient off-policy meta-reinforcement learning via probabilistic context variables. *CoRR*, abs/1903.08254, 2019. URL: http://arxiv.org/abs/1903.08254, arXiv:1903.08254.

[47] Andrew G. Barto Richard S. Sutton. A temporal-different model of classical conditioning. URL: http://incompleteideas.net/papers/sutton-barto-TD-87.pdf.

[48] R. H. Robins. John rupert first (1890-1960). In *Portraits of Linguists*, volume 2. Indiana University Press.

[49] G. Rummery and Mahesan Niranjan. On-line q-learning using connectionist systems. *Technical Report CUED/F-INFENG/TR 166*, 11 1994.

[50] Magnus Sahlgren. The word-space model : Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces. 2006.

[51] Adam Santoro, Sergey Bartunov, Matthew M. Botvinick, Daan Wierstra, and Timothy P. Lillicrap. Meta-learning with memory-augmented neural networks. In *ICML*, 2016.

[52] Teven Le Scao. A brief history of machine translation paradigms, May 2020. URL: `https://medium.com/huggingface/a-brief-history-of-machine-translation-paradigms-d5c09d8a5b7e`.

[53] Wolfram Schultz, Peter Dayan, and P. Read Montague. A neural substrate of prediction and reward. *Science*, 275(5306):1593–1599, 1997. `doi:10.1126/science.275.5306.1593`.

[54] Deven Shah, H. Andrew Schwartz, and Dirk Hovy. Predictive biases in natural language processing models: A conceptual framework and overview. *CoRR*, abs/1912.11078, 2019. URL: `http://arxiv.org/abs/1912.11078`, `arXiv:1912.11078`.

[55] David Silver, Aja Huang, Christopher Maddison, Arthur Guez, Laurent Sifre, George Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, Sander Dieleman, Dominik Grewe, John Nham, Nal Kalchbrenner, Ilya Sutskever, Timothy Lillicrap, Madeleine Leach, Koray Kavukcuoglu, Thore Graepel, and Demis Hassabis. Mastering the game of go with deep neural networks and tree search. *Nature*, 529:484–489, 01 2016. `doi:10.1038/nature16961`.

[56] Stephan Slingerland, Marco Mulder, Elsbeth E. van der Vaart, and Laurina Verbrugge. A multi-agent systems approach to gossip and the evolution of language. 2009.

[57] Robyn Speer, Joshua Chin, and Catherine Havasi. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, AAAI'17, page 4444–4451. AAAI Press, 2017.

[58] Robert P. Stockwell. Studies in linguistic analysis. J. R. Firth. *International Journal of American Linguistics*, 25(4):254–259, 1959. `doi:10.1086/464540`.

[59] Oriol Vinyals, Igor Babuschkin, Wojciech M Czarnecki, Michaël Mathieu, Andrew Dudzik, Junyoung Chung, David H Choi, Richard Powell, Timo Ewalds, Petko Georgiev, Junhyuk Oh, Dan Horgan, Manuel Kroiss, Ivo Danihelka, Aja Huang,

Laurent Sifre, Trevor Cai, John P Agapiou, Max Jaderberg, Alexander S Vezhnevets, Rémi Leblond, Tobias Pohlen, Valentin Dalibard, David Budden, Yury Sulsky, James Molloy, Tom L Paine, Caglar Gulcehre, Ziyu Wang, Tobias Pfaff, Yuhuai Wu, Roman Ring, Dani Yogatama, Dario Wünsch, Katrina McKinney, Oliver Smith, Tom Schaul, Timothy Lillicrap, Koray Kavukcuoglu, Demis Hassabis, Chris Apps, and David Silver. Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature*, 575(7782):350–354, 2019. doi:10.1038/s41586-019-1724-z.

[60] Christopher Watkins. Learning from delayed rewards. 01 1989.

[61] Warren Weaver. Translation. In Sergei Nirenburg, H. L. Somers, and Yorick Wilks, editors, *Readings in Machine Translation*, pages 15–23. MIT Press, Cambridge, Mass, 2003.

[62] J. Wisdom. Ludwig Wittgenstein, 1934-1937. *Mind. Vol. 61*, pages 258–260, 1952.

[63] Ludwig Wittgenstein. *Philosophical Investigations*. Basil Blackwell, Oxford, 1953.

[64] Ludwig Wittgenstein. *Tractatus logico-philosophicus. Logisch-philosophische Abhandlung*. Suhrkamp, [Frankfurt am Main, 1963. URL: http://www.worldcat.org/search?qt=worldcat_org_all&q=9783518100127.

[65] Adrienne L. Zihlman, Robert Boyd, and Peter J. Richerson. Culture and the evolutionary process. *BioScience*, 1985.

Glossary

attention A part of a neural architecture that enables to dynamically highlight relevant features of the input data, which, in NLP, is typically a sequence of textual elements. It can be applied directly to the raw input or to its higher level representation.

CNN Convolutional Neural Networks (CNN) is artificial neural network with the use of pooling layers, typically applied after the convolutional layers, mostly used in Computer Vision.

distributional semantics The quantification and categorization of similarity between linguistic contents based on their distributional properties in a large sample.

embeddings Embeddings, or word embeddings, is a vector representation of a word that maps its relationship to other words in a corpus.

Empiricist A person who supports the theory that all knowledge is based on experience derived from the senses.

event extraction A method to identify structured events, including event triggers and their corresponding arguments, from unstructured text using labeled training data in Natural Language Comprehension and Generation

knowledge graph A representation of a network of real-world entities (i.e. objects, events, situations, or concepts) that illustrates the relationship between them. This information is usually stored in a graph database and visualized as a graph structure.

knowledge noise A major obstacle in certain knowledge integration techniques where too much supplementary information leads the model to deviate from the correct semantics

PI *Philosophical Investigations.* The primary work of later Wittgenstein, published posthumously in 1953.

sprachspiel The original name used by Wittgenstein in the PI to describe language games. It is notable that spiel is not a direct translation of play. Instead, it encompasses ideas of play as well.

word2vec An algorithm that takes a text corpus as input and produces the word vectors as output. It first constructs a vocabulary from the training text data and then learns vector representation of words. The resulting word vector file can be used as features in many natural language processing and machine learning applications.