UNIVERSITÀ DEGLI STUDI DI TRENTO
CIMeC - Center for Mind/Brain Sciences

Master's Degree in Cognitive Science

2018-2019

Referentiality in distributional representations of named events

Supervisor: Dr. Aurélie Herbelot

Student: Gosse Minnema

Co-supervisor: Dr. Yannick Parmentier

Referentiality in distributional representations of named events

Gosse Minnema | MSc Thesis
Supervised by Aurélie Herbelot and Yannick Parmentier

Erasmus Mundus Joint Master's Degree in
Language & Communication Technologies

University of Lorraine (2017-2018)
University of Trento (2018-2019)

Co-funded by the
Erasmus+ Programme
of the European Union

UNIVERSITÀ
DI TRENTO

Abstract

Events are an important part of natural language meaning, but pose a challenge to distributional semantics, a popular approach in computational linguistics that models the meanings of words and sentences based on their co-occurrence contexts in large corpora. This thesis investigates named events (e.g., 'Hurricane Sandy', 'Battle of Waterloo') and proposes distributional representations derived from encyclopedic definitions of these events, as well as from the textual contexts of the event names themselves. We investigate to what extent these distributional representations encode referential information about the events that they represent (e.g., when the event happened, where it happened, the event's participants, etc.). To do this, we train classification models that take as input distributional representations of events and predict referential attributes of these events. In line with earlier work about predicting referential information from distributional representations, our results show that many event attributes can be successfully predicted. Finally, we perform a qualitative analysis of the event description representation space to find out what semantic properties of events it encodes.

Acknowledgements

This thesis is the conclusion of two very special years in my life, during which I studied in the Erasmus Mundus Joint Master's Degree in Language and Communication Technologies. This program did not just let me discover computational linguistics, but also two new places in Europe, and helped me grow a lot, both personally and academically.

Here, I would like to thank one person in particular: my main supervisor, Dr. Aurélie Herbelot. Thank you for accompanying me on this strange journey through computational semantics, for your endless stream of ideas, and of course for your friendliness and your (almost) English sense of humour. I would also like to thank the other members of the 'Meaning@CLIC' working group, which was a place where I felt very much at home intellectually, and also a great platform for sharing ideas and problems. Finally, I am grateful to the city of Rovereto for being the most beautiful place where I have lived so far, and to the friends I made there.

This thesis was made possible by an Erasmus Mundus scholarship from the European Union.

Contents

List of Figures

List of Tables

1. Introduction

1.1. Motivation

At the core of this thesis is an old and difficult question: how does the way in which we talk about the world reflect what that world is like? Even though most people would probably intuitively agree that when we talk, we usually convey some type of information about the outside world, exactly how this happens (and even whether it happens at all) is something that philosophers and language scientists have been arguing about for centuries. This thesis focuses on a particular aspect of the outside world, namely events ('things that happen in the world') and investigates how distributional and neural approaches to computational semantics can be used to model how we talk about events.

1.1.1. Vector space representations and referentiality

An important problem in computational linguistics and related fields is how to represent the meanings of words, sentences, and texts in a way that is useful for computer algorithms. While it is possible to do this using symbolic representations (large-scale symbolic knowledge bases include WordNet [Miller 1995] for word meaning, and the Parallel Meaning Bank [Abzianidze et al. 2017] for sentence meaning), these often have the disadvantage of being very labour-intensive to produce and sometimes lacking in nuance. As an alternative, continuous representations have been proposed that represent meanings as vectors in a high-dimensional space and can be automatically derived from corpus data using statistical methods. Such vectors can be computed in many different ways, for example directly from word-context co-occurrence statistics (e.g. Latent Semantic Analysis [LSA, Landauer and Dumais 1997], Distributional Memory [Baroni and Lenci 2010], Global Vectors for Word Representation [GloVe, Pennington et al. 2014]), as a byproduct of neural network language models (e.g. NNLM, Bengio et al. 2003), or using neural networks specifically design to learn word representations (e.g. word2vec, Mikolov et al. 2013). However, all of these vector-space approaches have in common that they implicitly or explicitly rely on co-occurrence patterns, placing vectors of words (or larger linguistic expressions) that occur in similar contexts close to each other in the vector space.

In the theoretical literature, this principle is known as the 'distributional hypothesis', which was first formulated in Harris (1954) and holds that linguistic expressions with similar meanings occur in similar contexts (Turney and Pantel 2010). This hypothesis can be interpreted in many different ways; 'strong' versions claim that context and meaning are identical and that mental meaning representations are also contextual in nature, while 'weak' versions only say that contextual information is useful for inferring semantic information (Lenci 2008). There is a lot of evidence for at least the weak version: vector-space representations are very successful

at capturing conceptual information, such as the fact that *cat* is more similar to *dog* than to *piano* or that the difference between *king* and *man* is similar to the difference between *queen* and *woman* (Mikolov et al. 2013). Hence, while distributional representations are solely based on corpus data, without any explicit information about the 'outside world', they implicitly encode a certain amount of world knowledge.

While vector representations are very successful at capturing the lexical meaning of individual words, it cannot be easily extended to capture the semantics of larger linguistic expressions (such as noun phrases or sentences) in a compositional way. By contrast, the formal semantics tradition (which started with work like Montague 1973), relying on tools from mathematical logic, elegantly models the overall semantic structure of phrases and sentences, but is less suitable for computationally modeling lexical meaning as it relies on complex, manually defined definitions. Under the umbrella term of formal distributional semantics (Boleda and Herbelot 2017), recent work has started trying to unite these two parts of computational semantics and model lexical and compositional aspects of linguistic meaning at the same time. So far, work in formal distributional semantics has successfully addressed several phenomena in the noun phrase domain such as quantification (Baroni et al. 2012; Herbelot and Vecchi 2015), noun-adjective pairs (Baroni et al. 2012), and (adjectival) negation (Hermann et al. 2013).

While a unified theory of semantics would be very attractive, a key problem is what exactly combined logical and distributional representations would represent. In formal semantics, expressions are always evaluated against a model of the world (e.g., *cat* denotes the set of cats in the world), while in distributional semantics, meanings are only defined relative to other meanings; it is not clear how to combine these two kinds of meaning in a consistent way.

Although efforts have been made to rigorously define what distributional representations refer to (e.g., Erk 2013; Erk 2016), these focused on the referential properties of kinds (e.g. the different properties of *alligators* and *crocodiles*). Modeling descriptions individual entities (e.g. *my cat*, *the alligator I saw yesterday*, or *Harry*) in a distributional way is much trickier because, by nature, corpus-derived vector representations are an 'average' of from many different texts about many different situations, which means that any reference to individual entities gets lost. An interesting 'trick' to get around this limitation is to model expressions that, within a given corpus, always refer to the same entity, such as proper names of famous people or of geographical entities (Herbelot and Vecchi 2015; Gupta et al. 2015). These studies computed vector representations for names such as 'Mr. Darcy' or 'Italy', and investigated the referential properties of these vectors (see section 2.3 for extended discussion). This works because, if a name occurs frequently enough in a corpus, the contexts in which it is used might tell us certain things about what the referent of that name is like. For example, it is likely that the name 'Italy' is occurs much more frequently in the context of 'European Union' than 'Argentina' does, which could help us infer that Italy is a member of the EU while Argentina is not. In thesis, we will use a similar strategy, but for events rather than for entities.

1.1.2. Events in natural language

When we use language, we sometimes describe the world in a static way (*My alligator's underside is cream-colored*) but we often also talk about what happens or changes in the world (*Your alligator killed my chickens*). An interesting observation from the formal semantics literature

is that in some respects, descriptions of events behave similarly to individuals:

(1) a. Your alligator killed my chickens
    b. Your alligators killed my chickens [slowly]
    c. Your alligators killed my chickens [slowly] [around 4am] [in a Parisian suburb]

(2) a. The death of my chickens [was slow]
    b. The death of my chickens [was slow] and [happened around 4am]

In these examples, (1b) refers to the same event as (1a), but additionally assigns the predicate 'slowly' to this event. In (1c), we see that any number of such predicates can be added and will be interpreted conjunctively (the events happened slowly and it happened around 4am, ...). When the event is expressed by a noun phrase rather than by a full sentence as in (2), these properties (predication and conjunction) are even clearer. The idea that events are a kind of 'things' that can be the subject of predicates was first expressed in Davidson (1967) and led formal semanticists to represent sentences like those in (1) with logical forms like (3):

(3) $\exists e[\text{eat}(e) \wedge \ldots \wedge \text{happened\_slowly}(e) \wedge \text{happened\_at\_4am}(e) \wedge \ldots]$[1]

This similarity between events and entities means that individual events in distributional semantics is difficult for the same reason that modeling individual entities is difficult: most information about individual events would get lost in representations derived from large corpora. However, just like for entities, there are events that have a unique and well-known name, and whose distributional properties might contain clues about their referential properties. For example, the expression 'The battle of Waterloo' could be seen as referring to the same event as a sentence like (4):

(4) British and Prussian forces fought Napoleon's army on 18 June 1815 near Brussels.[2]

1.2. Problem statement and approach

In this thesis, we investigate how named events can be represented in corpus-based vector space models of meaning and what referential information these representations can encode. In particular, we are interested in the following three subproblems:

1. Which possible ways of constructing an 'event space' are there?
2. How well can referential information be predicted from event vectors?
3. How is referential information encoded in the vector space?

[1] Patient and agent arguments omitted for simplicity.
[2] Cf. https://en.wikipedia.org/wiki/Battle_of_Waterloo

We aim to address these problems by first (chapter 3) constructing a dataset of named events, along with attributes and textual descriptions, derived from Wikipedia. Then, we start addressing problem #1 by proposing two methods for constructing vectors for named events. In chapter 4, we introduce a method that does not model event names directly, but approximates doing this by modeling encyclopedia definitions of the events. This has the advantage of being possible even for events with infrequent or ambiguous names, and can be done using existing (pre-trained) vector spaces, which makes the vector construction process simpler and less computationally costly. We experiment with different ways of representing event descriptions, both using more traditional distributional methods and with contextual text embeddings derived from BERT (Devlin et al. 2018), a deep learning-based language understanding model. However, we also explore the possibilities for directly creating distributional representations for event names (chapter 5).

Then, we move on to testing what referential properties our event spaces encode (#2). In chapter 6, we perform an experiment inspired by Gupta et al. (2015)'s work: we train classification models to predict referential properties of individual events, including traditional semantic roles such as event participant information (where possible), TIME and LOCATION, as well as event type-specific attributes (e.g., wind speed for hurricanes). Finally, in chapter 7, we address #3 by performing an qualitative analysis of our event spaces and trying to interpret their dimensions.

2. Theoretical background

2.1. Language and the world

Intuitively, there has to be some kind of relationship between language and the real world, if only because language is used not just in our own minds but also in the real world, and because we often use language to refer to things in the outside world. However, there exist completely different opinions about what that relationship is like and how important it is. On one end of the spectrum is the view defended by Noam Chomsky, which holds that "natural language has no semantics in the sense of relations between symbols and mind-independent entities. Rather, it has syntax (symbol manipulation) and pragmatics (modes of use of language)" (Chomsky 2013, p. 44). Under this view, meaning is something that exists only within the mind and relates to the outside world only in a very indirect way. Another extreme are truth-theoretic approaches to semantics, that are based around models of the outside world (or possible worlds) against which linguistic expressions are evaluated.

Somewhere in the middle between these extremes is an approach called Natural Language Ontology (NLO) (Moltmann 2018). NLO is not a theory of semantics but a framework for investigating what ontology (i.e., theory of 'what there is' in the world) is implicit in how speakers use natural languages. An example of this would be sentences like in (5), which are taken as evidence that objects (whether physical or abstract) and events belong to separate ontological categories.

(5) (adapted from Moltmann 2018, p. 2)
   a. The building described in the guide exists.
   b. The smallest prime number exists.
   c. ?? The inauguration of the president exists.
   d. The inauguration of the president happens.

NLO does not agree with Chomsky's rejection of referential semantics, although it is not necessarily incompatible with the view that meaning is purely mental: the proposed ontologies are only meant to reflect how reality 'appears' to speakers, not what it 'really is'. On the other hand, NLO is also different from model-theoretic semantics, which studies how language refers to the world but is not necessarily interested in the world itself, and tries to make only minimal assumptions about what the world is like.

The approach developed in this thesis is part of a line of research within distributional semantics (e.g. Herbelot and Vecchi 2015; Gupta et al. 2015; Kuzmenko and Herbelot 2019; see section 2.3) that investigates how corpus-based representations reflect the real world. In some sense, our work can be thought of as a version of NLO, as we share the goal of investigating how language reflects what the world is like (or appears to be like). However, there are two

crucial differences. First, we use different methods and sources of data: we use corpus-based computational models, unlike NLO, which uses the methods of traditional formal semantics. In other words, NLO models linguistic competence, whereas we model language use ('performance') and its relation to world knowledge. Secondly, we investigate different kinds of phenomena. Whereas NLO is typically concerned with what categories of entities (e.g., objects, properties, events) the ontology implicit in language should contain, we focus mostly on properties of specific entities. For example, our experiments in chapter 6 will tell us something about what usage data about the event name 'Battle of Waterloo' encodes about that event. On the other hand, in chapter 7 we will briefly look at what distributional data can tell us about the distinction between objects and events; this is more in line with 'standard' NLO and could be seen as a large-scale version of example (5) above.

2.2. What are events?

Events are (or rather, happen) everywhere in the world around us, but it is hard to define what exactly they are, and how what they are relates to how we talk about them. The aim of this thesis is to predict referential information from distributional representations of named events. In this section, we will first discuss what named events are and how event names are different from other linguistic expressions of events. Then, we describe the view on the ontological status of events that we will adopt (i.e., what is the referential information that we will be trying to predict), how we connect event names to event argument structure. Finally, we combine all of these insights into a unified view on named events and relate them to our experiments.

2.2.1. Linguistic expressions of events

For the moment, let's define events informally as 'things that happen' during a certain span of time and that can bear a relationship to entities in the world, such as participants or places. In this sense, the bracketed expressions in both sentences in (6) clearly describe an event, possibly the same one:

(6) a. [$_S$ Angela and Mark dined in a fancy restaurant yesterday evening.]
b. After [$_{NP}$ yesterday evening's dinner in a fancy restaurant], Angela kissed Mark.

However, it is not so clear *how* these expressions describe the event, and exactly *where* in the expressions the event is expressed. Arguably, information about the type of event that is described (i.e., a dining event) is carried by the verb 'dined' in (6a) and by the noun 'dinner' in (6b), while the other constituents ('Angela and Mark', 'in a fancy restaurant', 'yesterday evening') further specify the event's participants and its location in space and time. Both event expressions contain these elements; however, the denotation of the expressions as a whole, at least in standard model-theoretic semantics, is different: sentences denote truth values (type $< t >$) and noun phrases/determiner phrases (depending on your terminology) denote either entities (type $< e >$) or functions from properties to truth values (type $<< e, t >, t >$). To further complicate the situation, there also exist other ways to describe events that are, syntactically, in between noun phrases and sentences:

(7) a. [$_S$ That Angela and Mark dined in a fancy restaurant yesterday evening] is good news.

b. After [$_{NP}$ (their) dining in a fancy restaurant yesterday evening], Angela kissed Mark.

c. After [$_{NP}$ (their) eating sushi in a fancy restaurant yesterday evening], Angela kissed Mark.

The event in (7a) is expressed by the same sentence as in (6a), but embedded in a 'that-clause'. While the sentence itself has not changed, 'that' makes it more 'noun-y' because it allows it to occur in the same syntactic context where a noun phrase could also occur (cf. 'The officials claimed falsehoods', 'No news is good news'). Meanwhile, (7b-7c) are similar in meaning to (6b) but is more 'verb-y' because 'dining' and 'eating' are transparently derived from a verb while 'dinner' is not;[1] moreover, 'eating' has a direct object, which normally only verbs can have.

Thus, there are at least three types of event expressions that, syntactically, behave to some extent as nominals ('real' nouns, that-clauses, and gerunds). However, semantically, they behave quite differently:

(8) a. [$_{NP}$ Yesterday evening's dinner] (was/took place) in a fancy restaurant

b. ?? [$_S$ That Angela and Mark dined yesterday evening] (was/took place) in a fancy restaurant.

c. ?? [$_{NP}$ Dining in a fancy restaurant yesterday evening] (was/took place) in a fancy restaurant.

(9) a. ?? The officials claimed [$_{NP}$ yesterday evening's dinner in a fancy restaurant]

b. The officials claimed [$_S$ that Angela and Mark dined in a fancy restaurant yesterday evening.]

c. ? The officials claimed [$_{NP}$ Angela and Mark's dining in a fancy restaurant yesterday evening]

d. Angela and Mark claimed [$_{NP}$ having dined in a fancy restaurant yesterday evening]

e. [$_S$ Angela and Mark dined in a fancy restaurant yesterday evening]. At least, the officials claimed so.

Apparently, event-denoting noun phrases containing a 'real noun' (8a) can be the subject of predicates that further specify characteristics of the event, but the other types of 'nominal' expressions (8b-8c) cannot. By contrast, to 'claim a dinner' sounds strange. This is not the case for that-clauses and gerunds (9b-9d), although gerund sentences where the (implicit) subject of the gerund agrees with the main clause subject (9d) sound more natural than when there is an explicit subject (9c). Denial can also be achieved using a pronoun that refers back to a full (unembedded) sentence, as in (9e). Contrasts as in (8-9) are widely taken as evidence that

[1] Following the terminology first introduced by the American linguist Zeno Vendler, 'dinner' would be either a 'perfect' or 'derived' nominal, while 'dining' would be an 'imperfect' nominal; for discussion see Bennett (2002) and Casati and Varzi (2015).

although both sentences and noun phrases can *express* events, only 'real' noun phrases can directly *denote* events, whereas sentences, that-clauses, and gerund noun phrases denote facts (Bennett 2002; Casati and Varzi 2015).

To our knowledge, named events have not been previously addressed in the formal semantics literature.[2] However, if we see event names as simply proper names that denote events rather than other entities, fitting them into the typology of event expressions becomes trivial. As an example, let's assume there is an event name uniquely identifying the event described in (6):

(10) [$_{NP}$ Yesterday evening's dinner in a fancy restaurant] made a lasting impression on Mark. He would later refer to it as [$_{NP}$'The Dinner'].

In formal semantics, proper names are generally defined as special cases of noun phrases referring to entities. Hence, we expect the event name we introduced in (10) to show exactly the same behaviour as the NP in (6):

(11) a. After [$_{NP}$ The Dinner], Angela kissed Mark.
b. [$_{NP}$ The Dinner] (was/took place) in a fancy restaurant.
c. ?? The officials claimed [$_{NP}$ The Dinner].

This is indeed the case; in what follows, we will assume event names have the same semantic properties as any other event-describing NP.

2.2.2. Events and reference

In this subsection, we discuss two different, but, in our opinion, complementary views on what events refer to. First, we look at the (Neo-)Davidsonian approach to event semantics, and then at the more metaphysics-oriented view that events are property instances.

Semantics: events as entities

In the introduction, we already briefly discussed Davidson (1967)'s observation that events behave in some ways as if they were a kind of 'things'. Consider the following examples and sketches of their logical forms:

(12) a. There is a hungry cat in the house.
$\exists x.[\text{cat}(x) \wedge \text{hungry}(x) \wedge \text{in\_the\_house}(x))]$
b. There is a cat in the house.
$\exists x.[\text{cat}(x) \wedge \text{in\_the\_house}(x))]$

(13) a. A brutal fight took place yesterday.
$\exists x.[\text{fight}(x) \wedge \text{brutal}(x) \wedge \text{occurred\_yesterday}(x)]$

[2] However, the 18th-century philosopher Leibniz already noted there existence: "In certain cases, though, there has been a need to remember an individual accident, and it has been given a name. [...] Religion provides us with some, for instance, the birth of Jesus Christ, the memory of which we celebrate every year; the Greeks called this event 'Theogeny' [...]." (quoted in Bennett 2002, p. 3).

b. A fight took place yesterday.
$\exists x.[\text{fight}(x) \wedge \text{occurred\_yesterday}(x)]$

c. The Battle of Waterloo took place yesterday.
occurred_yesterday(battle_of_waterloo)

(14) a. The brutal Battle of Waterloo took place yesterday.
$a = \text{battle\_of\_waterloo} \wedge \text{brutal}(a) \wedge \text{occurred\_yesterday}(a)$

b. The Battle of Waterloo took place yesterday.
$a = \text{battle\_of\_waterloo} \wedge \text{occurred\_yesterday}(a)$

(15) a. Napoleon fought brutally yesterday.

i. $\text{fought\_brutally\_yesterday}(\text{napoleon})$

ii. $\exists e.[\text{fight}(e) \wedge \text{brutal}(e) \wedge \text{occurred\_yesterday}(e) \wedge \textsc{agent}(e, \text{napoleon})]$

b. Napoleon fought yesterday.

i. $\text{fought\_yesterday}(\text{napoleon})$

ii. $\exists e.[\text{fight}(e) \wedge \text{occurred\_yesterday}(e) \wedge \textsc{agent}(e, \text{napoleon})]$

In each of these examples, sentence (a) entails sentence (b). In (12), where the subject is an NP, this entailment is easily captured in the logical form: intersective adjectives like 'hungry' are modeled with conjunctions, so that $\text{cat}(x) \wedge \text{hungry}(x)$ automatically entails both $\text{cat}(x)$ and $\text{hungry}(x)$. The same strategy can be used for event-describing NPs and event names (13-14). For events described in sentences (15), we have the same entailment (the set of individuals who 'fight brutally' is a subset of those who 'fight'), but using standard predicate logic it is not possible to model this directly. In a pre-Davidsonian analysis of (formulas (i)), the verb phrases have to be modeled in a monolithic way that does not take into account the relationship between the two sentences. To solve this, we would need to model the adverbials 'brutally' and 'yesterday' as separate predicates (i.e., $\text{brutal}(x), \text{occurred\_yesterday}(x)$), but the problem is that it is not clear what these should be predicated of (i.e., what the x stands for). Davidson (1967) proposed event-describing sentences implicitly assert the existence of an event entity, which is a 'silent' argument of adverbials in these sentences. Parsons (1980) went one step further and proposed that not only modifiers, but also the arguments (agent, patient, ...) of verbs are predicates over events. This analysis (called 'Neo-Davidsonian semantics') led to logical forms like the ones in (ii), which allow the verb phrases to be 'split up' into separate predicates, so that we can capture entailments such as 'fight brutally' $\vDash$ 'fight'.

Ontology: events as properties

A related view from philosophical ontology holds that events are *tropes* (Bennett 2002). Tropes are individual instances of properties; for example, if 'redness' is a (universal) property, the redness of a particular red object (e.g., one of the roses in my backyard) is a trope. Similarly, an event type such as 'dining' can be seen as a property of 'spatio-temporal zones' (e.g. sets of points in space and time), and particular dining events are tropes of this property. For example, the event described in (6) corresponds to the trope possessed by the points in space and time occupied by Angela and Mark on yesterday evening.

Under this view, an important question is what happens to complex properties. For example, suppose a particular event is described by the sentences in (16):

(16) a. Angela and Mark dined happily.
b. Angela and Mark dined.

To how many tropes do these sentences correspond? Some proponents of the events-as-tropes theory, such as Kim (1966), hold that for complex properties, each of the components of these properties has its own trope. Hence, 'dine' and 'dine happily' in (16) correspond to different tropes, even if these tropes occupy the same spatio-temporal zone. Under this theory, events are very similar to facts,[3] because every fact will describe slightly different properties, which means that for every fact there is a different event (trope). By contrast, according to Bennett (2002), a 'rich event' corresponds to a single, composed trope. For example, there could be a single event, described by the sentences in (16), which is an instance of the combination of the properties 'dining', 'being done happily', and potentially many other properties. Thus, event descriptions are heavily underspecified: they refer to rich events that are the combination of many different properties, but only describe a small subset of these.

This second view sounds quite familiar: under Davidsonian event semantics, a formula like (17) asserts the existence of at least one event that is an instance of a particular set of properties (in this case, the property of being a dining event, the property of being performed in a happy way, the property of having Angela and Mark as the agents, etc.).

(17) $\exists e.[\text{dine}(e) \wedge \text{performed\_happily}(e) \wedge \textsc{agent}(e, \text{angela\_and\_mark}) \ldots]$

Instead of just determining the truth value of such a formula, we could also think of it as the set of events that satisfy it: $E = \{e \in W \mid \text{dine}(e) \wedge \ldots\}$ (where W is the set of all event entities). Each of the events in E is likely to have many other properties as well. Thus, both in Davidsonian event semantics and in Bennett's theory, events can be seen as instances of the conjunction of a set of properties. The main difference between the two approaches is that in Davidsonian semantics, events themselves are purely semantic entities without any internal structure, while in Bennett's theory, tropes are possessed by spatial-temporal zones (in the physical world).

2.2.3. Events and argument structure

In natural language, a crucial feature of event descriptions is that they relate an event to entities involved in that event (the *participants*) as well as to properties of the event. If the event is expressed by a verb, the participants are usually syntactic arguments of the verb. The relationships between a verb expressing an event and its arguments are often referred to as *semantic* or *thematic roles*. For example, in example (16a), Angela and Mark execute the action of dining; thus, the noun phrase 'Angela and Mark' is said to have the AGENT role relative to the verb.

[3] Informally speaking, facts are whatever it is that makes a true proposition true. According to Bennett (2002), the main characteristic that distinguishes facts from events is that events do not have a location in space and time (the fact that Angela and Mark dined yesterday is still true today and also on the other side of the world, whereas the event happened yesterday at a specific place).

For events expressed by nouns, the syntactic relationship between the event noun and the participants is less clear; it is sometimes expressed by adjectives (e.g. 'a brutal fight' (13a), where 'brutal' expresses the manner in which the event happens), or by genitives (e.g. 'Napoleon's death')

There exist different typologies for labeling semantic roles; some of these try to cover all possible event relations with a small set of 'universal' roles. However, designing such a set is very difficult; in practice, it is more useful to use a more fine-grained set of roles for specific kind of events. A theoretical framework that uses this second approach is *frame semantics* (Fillmore 1977). According to frame semantics, we conceptualize the world using *frames*, which are abstract representations of kinds of events and situations and the participants associated with these. In linguistic expressions, certain words are said to *evoke* (bring to mind) the frame related to the situation that the expression refers to. In computational linguistics, frame semantics is best known for FrameNet (Baker et al. 1998), a lexical database that consists of more than 1200 *frames* and more than 13,000 word senses that are connected to these frames. An example[4] of a frame is *Revenge*, which represents a situation involving two participants, A and B, where A has harmed B and B takes revenge. The frame can be evoked by various lexical units (or *targets*), e.g., 'revenge' (noun), 'retaliate' (verb), or 'vengeful' (adjective). The semantic roles of the frame (called *frame elements*) include *avenger*, *offender*, *injury*, *injured_party*, and *punishment*. Not all frame elements have to be expressed in a given sentence; for example, in (18), we only find *punishment*, *avenger*, and *injury*; the targets is 'avenge':

(18) (FrameNet code: `429-s20-rcoll-death`)
[PUNISHMENT With this], [AVENGER El Cid] at once *avenged* [INJURY the death of his son] and showed that any attempt to reconquer Valencia was fruitless while he still lived.[5]

The theory of frame semantics, as well as FrameNet, is useful to the vision on named events that we are developing here, for two reasons. First of all, while much of the literature on event argument structure is about verb phrases, frames specify a set of semantic roles that can be evoked by frame elements of any syntactic type (i.e. also by event names). Thus, we can use the concept of frames to connect our work about named events to theories about events in general. Second, we will use the idea of event type-specific semantic roles when defining our event dataset and in part make use of existing FrameNet frames.

2.2.4. Putting it all together

In this section, we have discussed different elements from linguistic and philosophical theories that are useful for understanding the semantics of named events. In this final subsection, we try to put these elements together to form a consistent whole. This is by no means meant as a 'grand theory' of events, but rather as a framework that helps us interpret our findings about named events in a broader perspective. This framework is schematized in Figure 2.1.

[4] Taken from the slides at `https://framenet.icsi.berkeley.edu/fndrupal/CJFFNintroPPT`

[5] From `https://framenet2.icsi.berkeley.edu/fnReports/data/lu/lu6056.xml?mode=annotation&banner=`

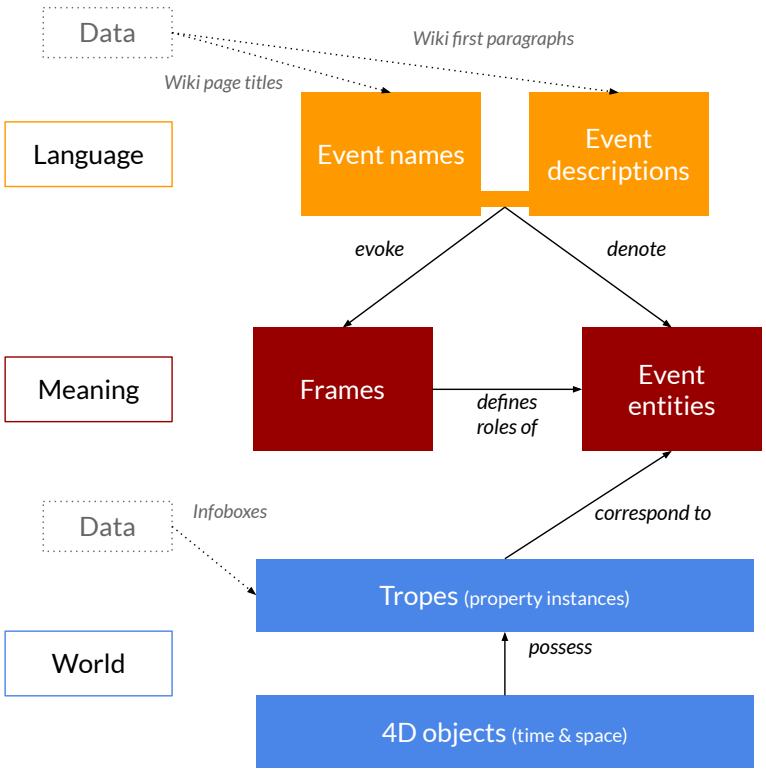


Figure 2.1.: Schema of our theoretical framework

Our framework consists of three components: LANGUAGE, MEANING, and WORLD. WORLD, at the bottom of the scheme, stands for the physical world, as perceived by humans,[6] which contains objects that are defined in terms of space (three dimensions) and time (i.e., four dimensions in total). These objects possess various tropes (instances of properties). We assume that tropes include events (following Bennett 2002) and also properties associated with these events. As an example of a named event, let's consider the Battle of Waterloo. This battle would correspond to an object that can be defined as a set of points in 4D-space, that together include all of the participants in the battle (e.g., the soldiers, the weapons, the battlefield, destroyed buildings, etc.) during the time spans that these were involved in the conflict. Then, the tropes of this object would include the 'battle' event itself, as well as properties associated to the event such as that the French army was one of the parties in the conflict, that n many people were killed or injured, etc. We interpret the raw, factual data that we extract from Wikipedia infoboxes (see section 3.2) as referring to tropes.

Next, MEANING is an intermediate level between WORLD and LANGUAGE, and can be thought of as a mental, more abstract representation of the objects in World. We adopt a Neo-Davidsonian

[6] We assume that, while it might be impossible to have objective knowledge of the 'real world', to a certain extent, people perceive the world in a similar way. We define WORLD to stand for this shared perception of the physical world.

view of the representations of events. Each of the event tropes in WORLD corresponds to an event entity in MEANING, and the other tropes correspond to predicates over this event. There can be, in theory, indefinitely many tropes associated to every event, but not all of these are important for the mental representation of the event. In order to select the tropes that form the basis for the semantic roles associated with the event, for each of the event types in our dataset we define a frame specifying a set of frame elements. Following Bos and Nissim (2008), we equate frame elements with Neo-Davidsonian semantic role predicates and frame targets (i.e., event names) with event entities. Moreover, we make the tropes themselves more abstract. For example, we define a frame *Battle* that defines semantic roles for battle events; one of these roles, *strength*, describes the intensity of the battle (in terms of how many soldiers participated in it). In WORLD, a trope corresponding to *strength* could be an instance of the property of there being 10,000 soldiers participating in the event. We do not assume conceptual representations of battles to contain this level of detail,[7] and thus, in Meaning, this trope is mapped to a predicate *has_large_strength*. In our prediction experiments (Chapter 6), we will try to learn mappings from distributional representations to these predicates (which we call 'attributes' in the context of the experiments).

Finally, LANGUAGE is the level of the scheme defining linguistic expressions of named events. In our experiments, we work with two types of expressions of these events: the event names themselves, and encyclopedic definitions of the names. For example, for the Battle of Waterloo, our dataset includes the name 'Battle of Waterloo' and the definition "The Battle of Waterloo was fought on Sunday, 18 June 1815, [...] The battle marked the end of the Napoleonic Wars". These linguistic expressions are linked to MEANING in two ways: they can invoke a particular frame, and they relate to a particular event entity.[8]

2.3. Distributional representations of individual entities

Unlike formal, truth-theoretic models of semantics, distributional models do not make direct reference to the extra-linguistic world. However, distributional representations can still, indirectly, encode world knowledge. Gupta et al. (2015) investigated the referential properties of Word2Vec (Mikolov et al. 2013) representations of country and city names. For each country or city, a set of numerical and categorical attributes (e.g. number of inhabitants, continent) was retrieved from FreeBase[9] and a logistic regression model was learned mapping the distributional space to a vector space where every dimension represents a referential attribute. In this space, categorical attributes are represented with a dimension for every class (e.g. `member-of::world_bank`, `member-of::UN`, etc.), whose value (in $\{0, 1\}$) indicates member-

[7] Although this might differ from person to person, for example, historians would be expected to have a much more detailed mental representation of battles than laypeople.

[8] We assume here that event descriptions have the same meaning as the event names that they define, even though this is not necessarily true: event descriptions are (sequences of) sentences and, as such, under Neo-Davidsonian event semantics they correspond to a proposition following the pattern 'there exists an event e such that ...', which in turn corresponds to a set of entities that make this formula true. To simplify our framework, we make the assumption that the definitions are always sufficiently detailed to make sure that this set only includes one entity, i.e. the one that the corresponding event name refers to.

[9] `https://developers.google.com/freebase/`, now deprecated.

ship of that class. Numerical and categorical attributes were evaluated separately, using rank scores and accuracy, respectively. The results were generally very positive, although there was much variation between the predictability of different attributes. Amongst the attributes with the best prediction scores were geo-location attributes (longitude and latitude, continent) and economy-related statistics (e.g. GDP, CO_2 emissions). Especially the finding about geo-location is interesting for our purposes, since 'place' is also a relevant feature for many events.

Several other studies (Johns and Jones 2012; Făgărășan et al. 2015; Herbelot and Vecchi 2015) have related distributional representations to information about the real world by predicting features from the McRae norms (McRae et al. 2005), a dataset describing properties of animals and inanimate objects, as judged by human participants (e.g. 'airplanes have wings', 'alligators are scary'). Note that this is world knowledge of a different kind than the country and city attributes in Gupta et al. (2015)'s study: the McRae norms are more subjective, and, more importantly, describe information on the level of concepts (or kinds) rather than of individual entities such as cities and countries. Herbelot and Vecchi (2015) is the McRae-related study whose approach is most comparable to ours: the authors propose a model for mapping distributional representations of concepts to a 'truth-theoretic' space whose dimensions represent quantifiers (e.g. if the vector for 'cat' has the value 0.95 for the attribute `has_four_legs`, this corresponds to the natural language sentence 'most cats have four legs'). The mapping is evaluated by calculating the Spearman correlation between the predicted and true quantifier values.

3. Event dataset

3.1. Event types

The event dataset includes three types of events: hurricanes, concert tours, and battles. These categories were selected based on two criteria: (1) the availability of a large number of Wikipedia articles for individual events and (2) the availability of a consistent and easily parseable set of referential attributes in the 'infoboxes'[1] in each article (see Figure 3.5 for an example). In this section, for each of the three event types, we discuss what the events are like in the real world, how they can be conceptualized in terms of frames and semantic roles, and introduce a set of finer-grained predicates based on these semantic roles. For example, if a particular frame includes the element Time, the associated predicates could be *recent_year*, *non-recent_year*, etc. In Chapter 6, these predicates will be used as feature-attribute pairs in our classification experiments (e.g. the predicate *non-recent_year* would correspond to the attribute-value pair ⟨*year, non-recent*⟩). The full list of attributes is given in Table 3.1.

3.1.1. Hurricanes

Hurricanes are extreme weather events with a number of special characteristics that distinguish them from other storms: for example, they involve very strong winds, have a special shape (the winds move in a circle around the 'eye' of the hurricane), and only occur in specific areas of the planet. Hurricanes are not stationary but form in oceans and then move towards land, gaining energy while moving until they hit the shore. Figure 3.1 shows the pathways of all hurricanes in the recent era; most hurricanes occur on the Atlantic coast of North America, on the the Pacific coast of East-Asia (where they are known as 'typhoons'), and in the South Pacific (where they are known as 'cyclones'). Hurricanes are classified on the Saffir-Simpson scale[2] which has five categories based on wind speeds. Additionally, there are similar, but less extreme weather events: tropical depressions and tropical storms. For the sake of simplicity, we will refer to all of the events as 'hurricanes'.

Hurricanes are interesting for our study because they are given names. This is done by official bodies (in modern times, this is done by the World Meteorological Organization[3]) in order to make it easier for the media and the public to refer to a particular hurricane. Lists are chosen from a pre-defined lists in alphabetical order. There is a separate list for every year; lists are re-used every six years. Thus, hurricane names are not strictly unique; however, the names of very destructive hurricanes (e.g., 'Katrina', 'Sandy') are retired from the lists.

Wikipedia infoboxes for hurricane events include the entries:

[1] On Wikipedia, infoboxes are the tables found in the upper right-hand corner of an article containing structured information, often using a fixed format for a particular kind of article.

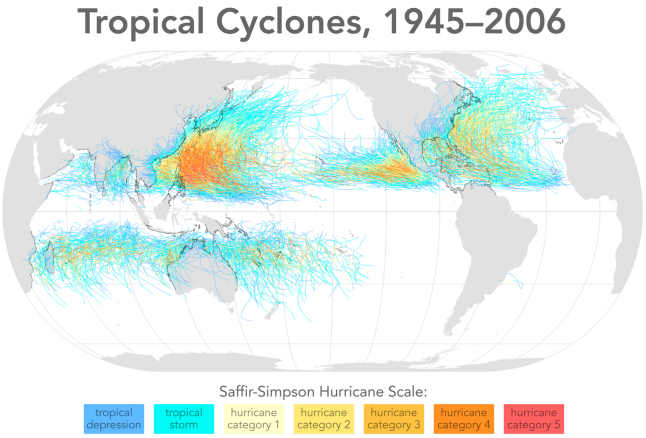
[2] https://en.wikipedia.org/wiki/Saffir-Simpson_scale

[3] https://public.wmo.int/en/About-us/FAQs/faqs-tropical-cyclones/tropical-cyclone-naming

Event/Attribute	Abbreviation	Type of underlying data	Num. events	Example
Hurricanes			*1241*	*Hurricane Isodore*
Area (binary)*	ArNS, ArWE	Categorical (hemispheres)	1185 / 1144	north / west
Area (four-way)*	Ar	Categorical (quadrants)	1137	north-west
Category (binary)	Ca	Categorical (weak/strong)	1186	strong
Category (SSHWS scale)	Ca	Categorical (7 categories)	1186	3 ('major')
Damage	Da	Numerical (US$)	1053	1.28 bln
Duration	Du	Numerical (days)	1216	13
Fatalities	Fa	Numerical (#)	1200	22
Pressure	Pr	Numerical (hPa)	1191	934
Wind speed	Wi	Numerical (km/h)	1208	125
Year	Ye	Numerical	1216	2002
Concert tours			*1978*	*McCartney World Tour*
Duration	Du	Numerical (days)	1884	306
Legs (cities)	Le	Numerical (#)	1650	9
Year	Ye	Numerical	1884	1989
Battles			6138	*Battle of Waterloo*
Area (binary)	ArNS, ArWE	Categorical (hemispheres)	3264	north / east
Area (four-way)	Ar	Categorical (quadrants)	3264	north-east
Belligerents	Be	Numerical (#)	5981	8
Involved US	InUS	Categorical (true/false)	5981	false
Involved France	InF	Categorical (true/false)	5981	true
Involved Spain	InS	Categorical (true/false)	5981	false
Strength (ratio)**	StR	Numerical (%)	1294	n/a
Strength (total)**	StT	Numerical (#)	1294	n/a
Year	Ye	Numerical	6111	1815

Notes: * hemispheres: east/west of the Greenwhich meridian or north/south of the equator. Quadrants: north-east, south-east, ... ** Strength = number of soldiers; 'total' is the strength summed over both sides of the conflict; 'strength' is the strength of the weakest side as a percentage of the strength of the strongest side.

Table 3.1.: Events and attributes



From https://en.wikipedia.org/wiki/Tropical_cyclone#/media/File:Tropical_cyclones_1945_2006_wikicolor.png, used under a BY-SA 3.0 licence (https://creativecommons.org/licenses/by-sa/3.0/)

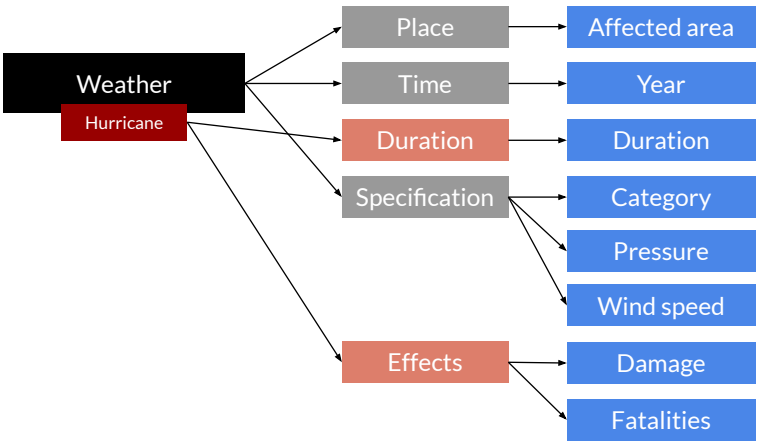
Figure 3.1.: Map of historical cyclones

- At the top of the infobox: name, hurricane category
- Dates: 'formed', 'dissipated', 'extratropical after' (optional)
- Meteorological properties: 'highest winds' (organized by duration, e.g., "10-minute sustained: 185 km/h; 1-minute sustained: 230 km/h"), 'lowest pressure'
- Consequences: 'fatalities', 'damage' (in dollars), 'areas affected'

Following the theoretical framework sketched in the previous chapter (section 2.2.4), we interpret the infobox attributes as defining the event in World. The raw values given in the infoboxes (as well as the hurricane event itself) as tropes of the four-dimensional physical object O within which the event took place. For example, if a hurricane has the value '925 hPa' for the attribute 'lowest pressure', this is interpreted as an instance of the property 'O having a minimal air pressure of 925 hPa'. Note that these properties are very fine-grained: if two hurricanes have slightly different air pressure values, this will automatically lead to them having instances of different properties as well.

Next, we connect these properties to semantic roles in the Meaning of the event. Our roles are similar to the Weather frame[4] in FrameNet. This frame is defined as "Ambient conditions of temperature, precipitation, windiness, and sunniness pertain at a certain Place and Time. Further Specification of the conditions that pertain may also be indicated" (where Place, Time, and Specifications are frame elements/semantic roles). An example of a sentence annotated with this frame is (19):

[4] https://framenet2.icsi.berkeley.edu/fnReports/data/frameIndex.xml?frame=Weather&banner=



Legend: Black/gray boxes correspond to FrameNet frames and elements; red/light red boxes correspond to our extension of the Weather frame. Blue boxes correspond to attributes.

Figure 3.2.: Conceptual scheme for hurricanes

(19) (FrameNet code: `429-s20-lgov-after`)
Some of the 500 people forced to leave their homes after [TIME Thursday's] [SPECIFICATION freak] *storm* [PLACE in the Llandudno and Conwy areas] believe it could be several months before they are able to return to their homes.

We propose a new Hurricane frame, which is an extended version of the Weather frame, but with an extra frame element Effects corresponding to negative consequences (such as killed or injured people and material damage) of the hurricane event. Furthermore, we propose predicates that form a bridge between the fine-grained tropes discussed above and the coarse-grained semantic roles. For example, the air pressure tropes correspond to the predicates *has_high_air_pressure*, *has_medium_air_pressure*, etc. This leads to the conceptual structure schematized in Figure 3.2. The following is an example of how this conceptual structure can be applied to the first sentence of the Wikipedia definition of hurricane Sandy[5] (we indicate both frame elements and predicates):

(20) *Hurricane Sandy* (unofficially referred to as Superstorm Sandy) was the [EFFECTS/*high_fatalities* deadliest] and [EFFECTS/*high_damage* most destructive] hurricane of the [TIME/*recent_year* 2012] Atlantic hurricane season.

In our prediction experiments, our hypothesis will be that both distributional representations of event descriptions (as in (20)) and of event names can implicitly capture this conceptual structure.

[5] `https://en.wikipedia.org/wiki/Hurricane_Sandy`

3.1.2. Concert tours

A concert tour is "a series of concerts by an artist or group of artists in different cities, countries or locations."[6] Concert tours can take a long time (months or years) and consist of several *legs*: subdivisions of the tour that are based on the locations, dates, and/or content of the concerts in the tour. Wikipedia infoboxes for concert tours have the following entries:

- At the top of the infobox: tour name, artist name, associated album
- Dates: 'start date', 'end date'
- Route/length: 'number of legs', 'number of shows' (often split by country/continent)
- Results: 'box office' (in dollars), 'number of visitors'

Unfortunately, some of these entries are only present for a very limited number of articles (such as 'box office' and 'number of visitors') or too difficult to parse (such as 'number of shows'); others, such as the associated album or the artist, are interesting from a semantic point of view but would be too difficult to predict and are therefore excluded from our dataset.

Again, we assume these entries to correspond to tropes in the WORLD part of our theoretical framework. Moving to MEANING, the most relevant frame in FrameNet for concert tours is Travel, which defines a "Traveler [who] goes on a journey, an activity, generally planned in advance, in which the Traveler moves from a Source location to a Goal along a Path or within an Area. [...] The Duration or Distance of the journey, both generally long, may also be described as may be the Mode_of_transportation. [...]"[7] The frame includes many frame elements, most of which are not very salient for concert tours (e.g. 'co-participant', 'baggage', 'mode of transportation') and for which we do not have any data. Thus, the Concert_Tour frame that we propose here uses a subset of the semantic roles in the Travel frame. The conceptual structure of Concert_Tour and its associated predicates is given in Figure 3.3; example (21) is a sentence annotated with the original Travel frame, evoked by the lexical unit *tour* and example (22) is the first sentence of the Wikipedia definition of the Paul McCartney World Tour[8] annotated using Concert_Tour and its associated predicates.

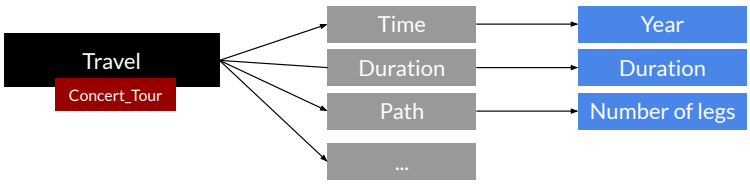
(21) (FrameNet code: `250-s20-ppthrough`)
[TIME In autumn 1903] [TRAVELER she] made a [MEANS walking] *tour* [PATH through France] [CO-PARTICIPANT with Dorelia McNeill, later Augustus John 's companion.

(22) The *Paul McCartney World Tour* was a [PATH/*high_number_of_legs* worldwide] concert tour by Paul McCartney [TIME/*medium_duration* during [TIME/*medium_recent_year* 1989] and 1990].

3.1.3. Battles

A battle "is a combat in warfare between two or more armed forces [...] [and are] generally are well defined in duration, area, and force commitment."[9] Battles are usually part of a war,

[6] `https://en.wikipedia.org/wiki/Concert_tour`
[7] `https://framenet2.icsi.berkeley.edu/fnReports/data/frameIndex.xml?frame=Travel`
[8] `https://en.wikipedia.org/wiki/The_Paul_McCartney_World_Tour`
[9] `https://en.wikipedia.org/wiki/Battle`



Legend: as in Figure 3.2.

Figure 3.3.: Conceptual scheme for concert tours

which is much longer and less-well defined type of event. Wikipedia battle infoboxes contain the following entries:

- Above the infobox: battle name, campaign/war that the battle is part of;
- General information: 'date', 'location' (sometimes including geo-coordinates), 'result'
- Parties: 'Belligerents', 'Commanders and leaders' (split by side)
- Size: 'Strength' (number of soldiers, machinery, etc.; split by side)
- Consequences: 'Casualties and losses' (split by side)

We connect this factual information (i.e., tropes in WORLD) to a frame in MEANING. The most relevant frame in FrameNet is Hostile_encounter, which describes "a hostile encounter between opposing forces (Side_1 and Side_2, collectively conceptualizable as Sides) over a disputed Issue and/or in order to reach a specific Purpose."[10] An example of this frame from FrameNet is (23):

(23) (FrameNet code: `250-ppagainst`)
He was later reputed to have played a role in the [PLACE sea] *battle* [SIDE_2 against the French] [PLACE off Sandwich] [TIME in 1217].

The frame has many different frame elements, only a few of which are directly applicable to the information that is available to us. For example, the frame element Explanation would be very relevant to include in our scheme, but Wikipedia infoboxes for battles do not include this information. Thus, our Battle frame uses a subset of the Hostile_Encounter frame. Figure 3.4 shows our Battle frame and its associated predicates/attributes, and example (24) applies this frame to the first sentence of the definition of the Battle of Waterloo:[11]

(24) The *Battle of Waterloo* was fought [TIME/*medium_recent_year* on Sunday, 18 June 1815] [PLACE/*north_east_hemisphere* near Waterloo in Belgium, part of the United Kingdom of the Netherlands at the time].

[10] `https://framenet2.icsi.berkeley.edu/fnReports/data/frameIndex.xml?frame=Hostile_encounter&banner=`
[11] `https://en.wikipedia.org/wiki/Battle_of_Waterloo`

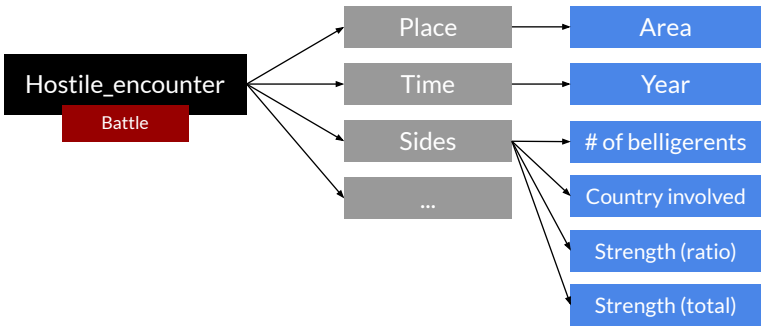


Figure 3.4.: Conceptual scheme for battles

3.2. Wikipedia scraping

We compiled the dataset by retrieving articles for events of our three types from Wikipedia. First, we found relevant articles using Wikipedia's category hierarchy. Then, we retrieved each of these articles and extracted its first paragraph and its infobox. In some cases, we also retrieved extra information about an article, either from that article itself or by following links to other pages. Finally, we parsed the infoboxes and processed them in order to obtain as much structured, machine-readable information as possible.

3.2.1. Finding relevant pages

Relevant articles were found using category pages: special pages that provide lists of articles and/or other category pages about related topics. The articles for each of our event types have a different category structure; see Figure 3.6 for an overview. We adapted our strategy for finding articles to each of these categories.

Hurricanes

Hurricanes were found by doing starting from the category page 'Tropical cyclones by region',[12] which forms the root of a somewhat complicated hierarchy of category pages and hurricane articles. Directly below the root are categories for hurricanes by continent. The structure and the depth of these subtrees varies from continent to continent; for example, for Europe, most hurricanes are listed directly at the continent level, while for North America the hierarchy is deeper. In many cases, category pages and hurricane pages are listed on the same level; for example, the category 'Hurricanes in the Caribbean' contains both the sub-category 'Hurricanes in the Caribbean by country' and the hurricane page 'Hurricane Matthew'. We found the set of hurricane articles by doing a full traversal of the tree, at every level saving the URLs of hurricane articles and following links to sub-categories until there were none left.

To get a list of unique hurricane pages, two types of filtering were needed: removing duplicate URLs, and removing pages that were listed in the hierarchy but do not describe an

[12] https://en.wikipedia.org/wiki/Category:Tropical_cyclones_by_region

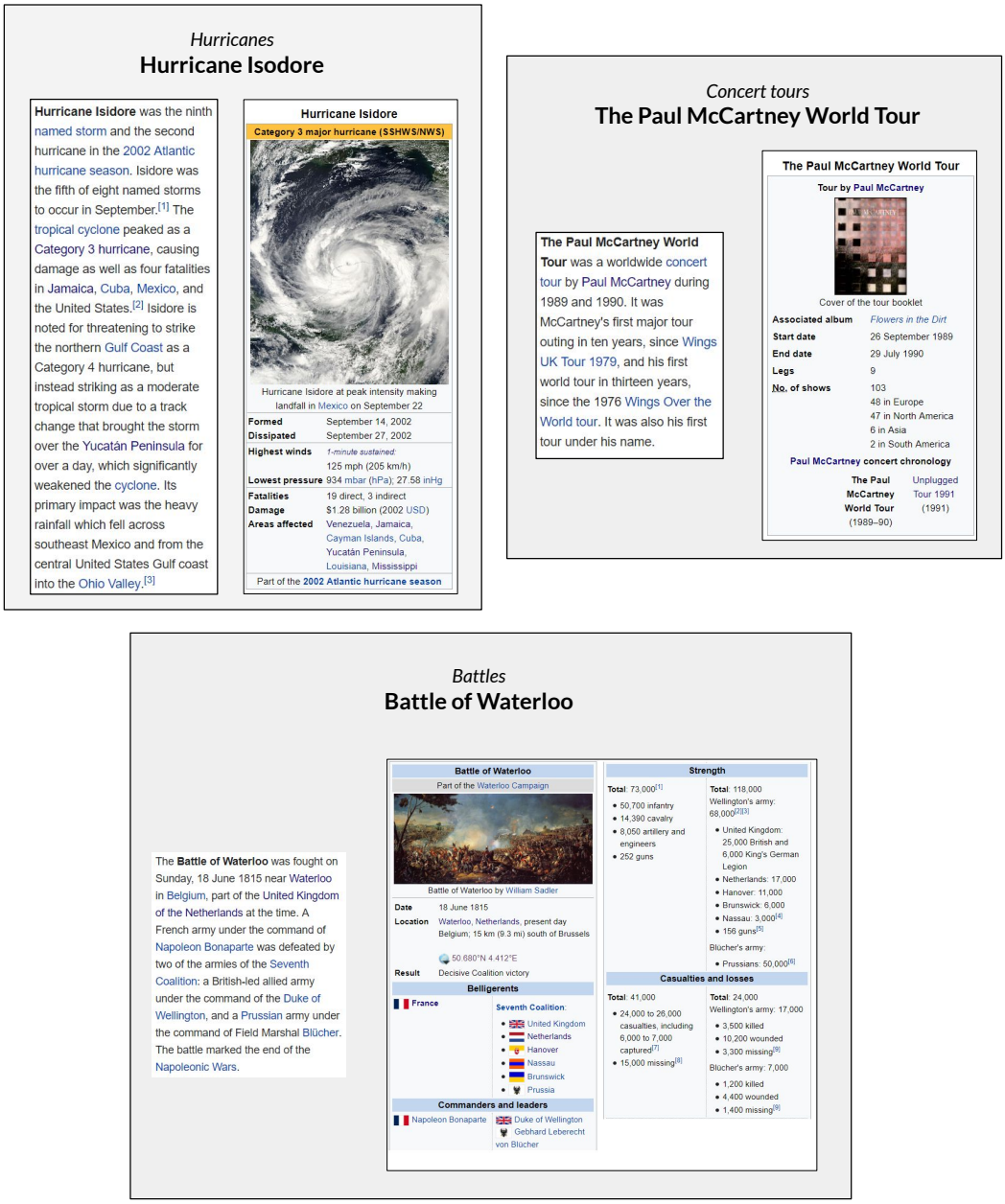


Figure 3.5.: Example event descriptions and infoboxes (https://en.wikipedia.org/wiki/Hurricane_Isidore, https://en.wikipedia.org/wiki/The_Paul_McCartney_World_Tour, and https://en.wikipedia.org/wiki/Battle_of_Waterloo)

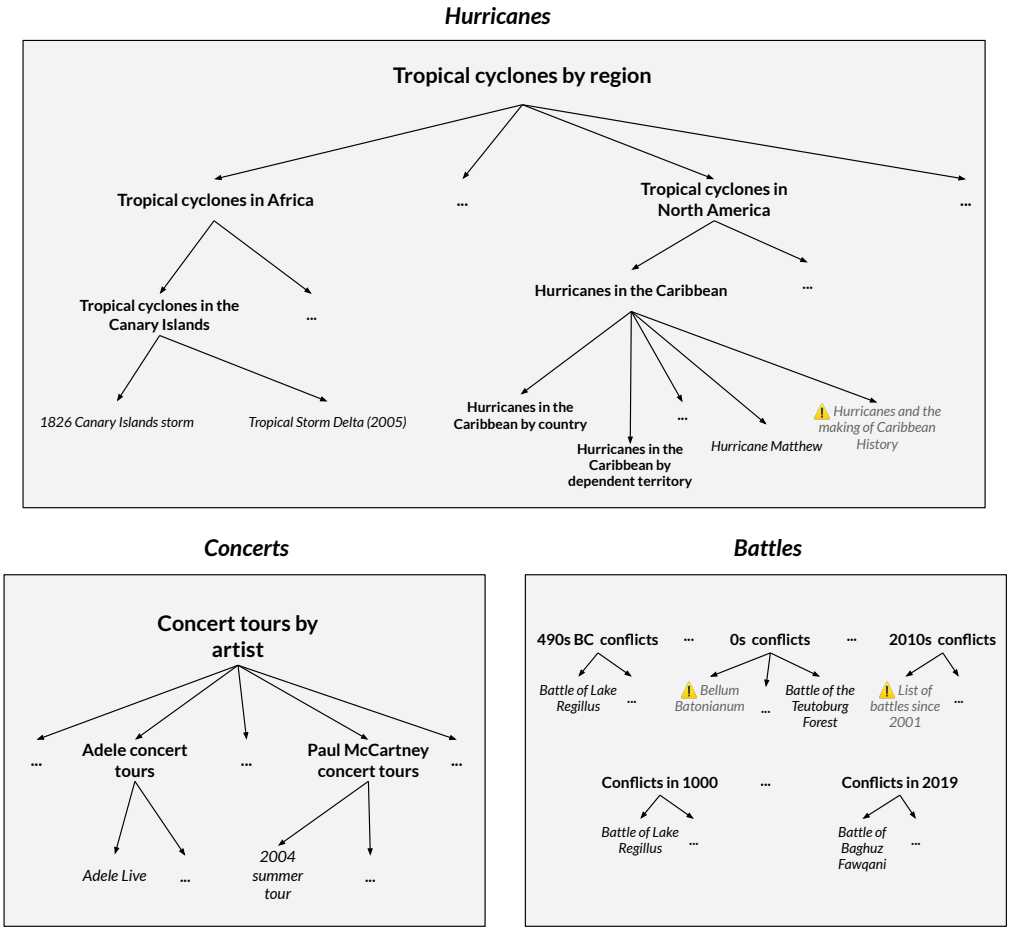


Figure 3.6.: Tree structure of event categories and pages. Category pages printed in **bold**, event pages printed in *italics*, and pages to be removed marked with a warning sign.

individual hurricane. An example in figure 3.6 is 'Hurricanes and the making of Caribbean History', which is about hurricanes in general and hence not interesting for our purposes. Other examples include pages listing hurricanes that do not have their own page and pages summarizing hurricane seasons; these pages are potentially interesting but do not follow the same format as pages for individual hurricanes and were discarded for that reason. Filtering these out happened at two different stages: first, by checking the URLs of article pages against a set of heuristics (e.g. whether the URL starts with 'Timeline:' or whether it includes the word 'season'), and later, when retrieving pages to extract information, checking whether or not they have a hurricane infobox and excluding them when they do not.

Concert tours

Concert tours were found using the category page 'Concert tours by artist'[13] and finding all the subcategories for tours of specific artists listed on that page. Then, for each artist category page, we retrieved all of the tour pages listed on that page. Some artist category pages had further subcategories, but we ignored these as they were in many cases duplicates (e.g., the category 'Paul McCartney concert tours' contains the subcategory 'The Beatles concert tours', which is also listed directly under the 'Concert tours by artist' category).

Battles

Battles were found through two types of category pages: category pages for years (e.g. 'Conflicts in 2019') and category pages for decades (e.g. '1550s conflicts'). The former exist for many years, especially relatively recent ones (from the middle ages); decade pages go back further in time (until the 490s BCE). We found all year and decade pages by generating the URLs of all possible pages (from the year 0 for year pages and from the 500s for decade pages) and then checking which URLs exist. Then, for all of the category pages, we found the list of conflict pages that each page linked to. The next step was removing duplicates and pages for conflicts that are not battles (but lists of battles, or more long-term conflicts such as wars). For deciding which conflicts were actual battles, we first checked whether the URL contained 'battle' or not, and then, after retrieving the page, we checked whether there was a battle infobox or not.

3.2.2. Extracting information

For each of the event pages that we found, we were interested in at least two parts of the page: the first paragraph, and the infobox. For battles, we additionally looked for the geo-coordinate component located in the upper-right of the page (which is a simple hyperlink listing the GPS coordinates of a location where the event took place, e.g. *Coordinates: 48.8566°N 2.3518°E*; the component links to a GeoHack[14] page that gives more information about the coordinates as well as links to online mapping services). For extracting these components of the page, we first retrieved the HTML code of the page and then parsed it using the BeautifulSoup library

[13] https://en.wikipedia.org/w/index.php?title=Category:Concert_tours_by_artist
[14] https://tools.wmflabs.org/geohack/geohack.php

for Python.[15] We found the first paragraph by looking for all `<p>` tags and selecting the first one with more than 10 words (we added this condition because on some pages, there are short pieces of text marked with `<p>` tags before the first real paragraph). Infoboxes were found by looking for a `<table>` with class `infobox vevent`. Finally, where applicable, geotags were find by looking for a `span` with the ID `coordinates`.

Once the infoboxes were found, we extracted as many referential attributes as possible from them, within the limits of machine-readability. An example of a hurricane infobox together with the structured information that we parsed from it is given in Figure 3.7. Our main challenge was that not all infoboxes have the same entries, and entries do not always have a uniform format. For example, dates are sometimes given in the American format ('September 14, 2002') and sometimes in the British format ('14 September 2002'). We addressed this by first using regular expressions to detect which of these formats (if any) a date was formatted in, and then passing the date, together with the appropriate format string, to Python's `strptime()` function.[16] More difficult were numerical attributes consisting of several, varying sub-components such as casualties (hurricanes and battles) or army strength (battles). For these attributes, sometimes a total is given, but in other cases, only individual components are listed; for example, casualties could be listed as 'x casualties, including y captured' (for battles), 'x deaths, y missing', (for battles and hurricanes) x direct, y indirect' (only for hurricanes). These expressions are too complex to extract in a structured way; hence, we had to make (rather arbitrary) decisions about how to extract a machine-readable value for them. For hurricane fatalities, the strategy we chose was to sum all of the listed numbers. Doing this implies that our numbers do not have a consistent interpretation across the dataset; however, we think this is not a problem for our purposes: our interest is in the approximate 'magnitude' of a hurricane event (rather than the precise number of victims) as perceived by observers (and presumably reflected in language use).

Another challenge were attributes that whose values are always structured in the same way, but have a complex structure; for hurricanes, examples of this are 'highest winds' and 'affected areas'. At the parsing stage, we aimed to preserve as much information as possible (for example, from the 'affected areas' field of hurricanes we extracted both a list of the names of the areas and a list of the URLs of the corresponding Wikipedia pages), while later, when defining our classification problems, we simplified the information to make it more feasible to predict. For example, for 'affected areas', we mapped the extracted lists of areas to earth quadrants (i.e., 'north-east', 'south-east', 'south-west', 'north-west') by retrieving their Wikipedia pages and parsing their geo-coordinates.

[15] https://pypi.org/project/beautifulsoup4/

[16] https://docs.python.org/3.6/library/datetime.html#strftime-strptime-behavior


```
"/wiki/Hurricane_Isidore": {
  "title": "Hurricane Isidore",
  "category": "Category 3 major hurricane (SSHWS/NWS)",
  "category_num": 3,
  "formed_date": "2002-09-14T00:00:00",
  "dissipated_date": "2002-09-27T00:00:00",
  "highest_winds": [
      {
          "duration": 1,
          "speed": 205
      }
  ],
  "lowest_pressure": 934,
  "fatalities": 22,
  "damage": 1280000000.0,
  "affected_areas": [
      "Venezuela",
      "Jamaica",
      "Cayman Islands",
      "Cuba",
      "Yucat\u00e1n Peninsula",
      "Louisiana",
      "Mississippi"
  ],
  "affected_areas_urls": [
      "/wiki/Venezuela",
      "/wiki/Jamaica",
      "/wiki/Cayman_Islands",
      "/wiki/Cuba",
      "/wiki/Yucat%C3%A1n_Peninsula",
      "/wiki/Louisiana",
      "/wiki/Mississippi"
  ]
},
```


Figure 3.7.: Example of extracted structured information from an infobox (colored dots indicate corresponding attributes)

4. Representations 1: Event description modeling

4.1. Background

The first approach to representing named events that we experimented with is to construct distributional representations of encyclopedic descriptions of these events. This approach is based on the hypothesis that a description of a particular event can be used as a proxy for the contexts in which the event occurs. For example, if, in natural text, the name of a particular hurricane co-occurs frequently with words like 'destructive', 'devastating', ..., we could use this to make inferences about the (perceived) severity of the hurricane. These words are also likely to occur in an encyclopedic description of that same hurricane. Our assumption is that, because encyclopedic descriptions are designed to capture the most salient information about a given event, the lexical content of the description will, to some extent, be representative of the contexts that an event will frequently occur in. Following this assumption, we can construct a composed representation of an event description and use it as if it was a distributional representation of the event name.

Our approach is not new but draws on several ideas from the literature about distributional representations for infrequent words. First, the hypothesis that the distributional representation of a word can be approximated by a composed representation of a text fragment in which it occurs comes from Lazaridou et al. (2017). This study investigated how well distributional semantic models can approximate the human ability to learn new concepts from only a few examples. The authors did this by constructing a corpus of short text passages describing non-existing 'chimera' words (whose meaning is a combination of the meanings of two existing, related words), and representing these words by summing distributional vectors of the words in the passages in which the chimera words occur.[1] These vectors were then tested in a similarity judgement task where chimera words are compared to existing words; impressively, the distributional model approximated human performance on this task. Moreover, the idea of using encyclopedic definitions as a proxy for a maximally informative context was introduced in Herbelot and Baroni (2017), who construct a 'definitional nonce dataset' consisting of one-sentence Wikipedia definitions. This dataset was designed for testing the performance of distributional representations derived from small data; this was done by calculating the distances between representations derived from the Wikipedia definitions and a standard representation of the same word derived from a large corpus. One of the models that was tested was Lazaridou et al. (2017)'s summing approach, which again turned out to work well.

We had several reasons for first trying definition-based representations, before moving on to directly computing distributional vectors for event names (see the next chapter). First of all,

[1] In Lazaridou et al. (2017)'s original study, multi-modal word vectors were used that incorporate both distributional and visual information; however, Herbelot and Baroni (2017) showed that the summing approach can also be successfully applied to purely distributional vectors.

event names are relatively rare; for example, as we will see in the next chapter, only a small subset of the battle events in our database occur more than 50 times in the Wikipedia corpus. Another practical reason is computational cost: training high-quality representations requires processing large corpora (e.g. Google News), which is costly and which we wanted to avoid in our initial experiments.

A more fundamental reason is that, while this thesis is about named events, we are are not exclusively interested in event names, but rather in comparing different linguistic forms that can express events. In everyday language, most of the events that we talk about ('Mark kissed Angela') do not have a name but are denoted by sentences or even discourses. Named event descriptions can be seen as a special case of event-describing sentences, so modeling these distributionally is interesting in and of itself.

4.2. Approach and methods

We try several techniques for distributionally representing the content of event definitions. Composing distributional representations of individual words is a complicated problem (Baroni 2013). While in theory, approaches uniting formal and distributional approaches would be most attractive as they combine the strengths of both approaches (formal model of semantics are designed to capture compositionality, whereas distributional approaches are good at capturing lexical meaning) (Baroni et al. 2014b), such unified models are still underdeveloped. By contrast, simplistic approaches to compositional distributional semantics, such as summing the vectors for individual words to obtain a representation for a larger unit such as a sentence are known to work surprisingly well (Mitchell and Lapata 2008). However, an obvious downside of such models is that they are unable to capture any syntactic or contextual information (e.g. the representation of 'Mark kissed Angela' will be identical to that of 'Angela kissed Mark').[2]

Recent research in deep learning could offer an alternative way to capture contextual and syntactic information without explicitly using any formal machinery: contextualized deep neural embedding models such as ELMo (Peters et al. 2018) or BERT (Devlin et al. 2018). These models, unlike traditional word embeddings, do not assign a fixed representation to each word in the vocabulary, but compute word representations that take into account the context in which the word occurs, implicitly capturing information about word senses and syntactic structure.[3] We think that experimenting with these representations is interesting for our task for two reasons. On one hand, such models have been shown to produce good semantic representations for various semantic tasks, so it is not implausible that they could work well for our task as well. On the other hand, little is known yet about what semantic information these models capture and how (but see Gulordava et al. 2018; Baroni 2019). Since our task is exactly aimed at 'probing' the semantic content of distributional representations, it might help us better un-

[2] This is true for summing standard distributional vectors; however, there have been efforts to incorporate information about the syntactic contexts of words in distributional representations (Padó and Lapata 2007; Baroni and Zamparelli 2010).

[3] Note that these models did not 'invent' contextualized word embeddings but merely popularized them; the idea of incorporating contextual information in distributional representations goes back to much earlier work such as Erk and Padó (2008), Erk and Padó (2010), and Thater et al. (2010).

derstand how contextual models represent meaning and for what kind of semantic tasks they work well.

We performed experiments with summed representations and several types of representations extracted from BERT. In the remainder of this section, we explain both of these approaches in more detail.

4.2.1. Summing

Summing word vectors, and related methods such as averaging them, have been investigated as ways to obtain compositional representations since the very early days of distributional semantics (Landauer and Dumais 1997; Landauer et al. 1997). If the dimensions of a distributional space are thought of as semantic features, summing word vectors corresponds to finding the union of the semantic features, yielding a representation of all of the lexical content in a sentence or text passage. While this representation does not take into account word order, it has been hypothesized that in many cases, 'scrambled' texts could be reconstructed even without this information (Landauer et al. 1997). For our purposes, additive models are interesting for testing how much semantic information can be reconstructed from such 'poor' representations that only take into account lexical content, and nothing else.

Our approach to creating summed vectors for event descriptions (from now on referred to as **GloVe-Summed**) takes as a starting point a pre-trained GloVe model (Pennington et al. 2014).[4] Vectors from this model are retrieved for the set of content words in each event description, and then computing a simple (unweighted) sum over these. Finding content words is done through a simple pipeline, implemented using NLTK (Bird et al. 2009), consisting of the following steps: (1) word tokenization; (2) lemmatization using a WordNet lemmatizer; (3) lowercasing; (4) removal of stopwords, digits, and duplicate words. This last step is important as it makes sure that only purely lexical information survives into the vectors. Particularly, we hypothesize that removing numbers will make it harder for predictive models to 'cheat' by making use of year numbers for predicting temporal attributes. By impoverishing the summed vectors in this way, we hope to get an idea of the lower bound on the amount of non-lexical information needed for encoding referential properties of events.

4.2.2. BERT

BERT (Devlin et al. 2018) is a recent deep learning model that takes as input a sentence (or an arbitrary piece of text of up to 512 word tokens) and produces a representation for that text. The defining characteristic of BERT is that it is built around a *transformer encoder*. Transformers (Vaswani et al. 2017) are a new type of neural network model for learning patterns in sequential data, and are intended as a replacement for existing sequence models such as Recurrent Neural Networks (RNNs). A key mechanism of transformers is *self-attention*: in each of the layers in the model, the representation of every input token is computed as a weighted sum of the learned representations of all of the other tokens in the input sequence. A highly simplified example could be the following: in a sentence like 'Mary went to the bank to withdraw money', the model might learn that, for the word 'bank', the words that are most relevant

[4] Version `42B.300d`, obtained from `https://nlp.stanford.edu/projects/glove/`.

to its contribution to the sentence are 'bank' itself, but also 'money' and 'withdraw', and assign relatively high weights to these words while assigning lower weights to the other words. These weights are then used to compute the representation of 'bank' that will be passed on to the next layer of the model; this representation will not just contain information from the previous representation of 'bank' itself, but also from the representations of 'money' and 'withdraw'. Hence, unlike in traditional word embedding models like GloVe, word representations in the output of BERT are *contextualized* and can implicitly encode information about word senses and syntactic dependencies.

BERT is designed to be pre-trained on two tasks: masked language modeling (i.e., given a sentence in which one or more words are 'hidden', predict which words these are) and next-sentence prediction (i.e., given two sentences, how likely is it that the first sentence is the continuation of the next one?). There are two methods for using BERT: *fine-tuning* and *feature extraction*. When using fine-tuning, a model designed for a specific task (which could be anything from POS-tagging to sentiment analysis) is added on top of BERT, taking BERT's output as input. When training for this specific task, all the combined model's parameters (both those of BERT and those of the 'extra' model) are optimized. By contrast, using feature extraction, BERT's outputs are used as-is: they can be used as the input for another model, but during training, only that model's parameters will be updated, leaving BERT itself 'frozen'. Choosing which of these methods to use depends on the specific situation at hand (see Peters et al. 2019 for an overview of how to use both methods); here, we will only use the feature extraction approach since we would like to compare BERT's representations with traditional distributional methods.

Given its newness and complicated architecture, much is still unclear about how exactly BERT works and for which tasks it works well. While (fine-tuned) BERT models achieved state-of-the-art performance on semantic tasks such as the SQuAD question answering task (Rajpurkar et al. 2016) and the GLUE benchmarks (Wang et al. 2018), not much is known yet about the semantic properties of word and sentence embeddings directly extracted from the pre-trained model. Neither is it clear what the best way is for extracting sentence embeddings from BERT. Each of BERT's hidden layers contains an embedding for each of the tokens in the input sequence, as well as for the special `[CLS]` token, which is used as a representation of the complete sentence when fine-tuning BERT for classification tasks. Liu et al. (2019) tested BERT word embeddings from different layers on a set of 'probing tasks' and found that word embeddings extracted from hidden layers in the middle of the model encode the most general linguistic and semantic information. Meanwhile, informal experiments reported on a discussion forum in Google's official GitHub repository for BERT [5] compare different methods for extracting sentence representations, with and without fine-tuning. These experiments suggest that, without fine-tuning, embeddings for the `[CLS]` token do not perform well in a sentence-similarity task, but that sentence representations obtained by averaging the embeddings for each of the individual tokens in the sentence lead to better results. While such unpublished results should be interpreted with great caution, they did help us formulate hypotheses when experimenting with different BERT-derived representations.

We tried to make our BERT-derived vectors as information-rich as possible, by feeding our

[5] https://github.com/google-research/bert/issues/276 (consulted on 2019-05-07)

paragraphs to the model 'as-is', without removing any information. Based on our review of the limited available literature on BERT, we propose several possible ways of extracting paragraph embeddings. [6] Our first approach, **BERT-Pooled-Paragraph**, uses the representation of the `[CLS]` token from a 'pooling layer' (with pre-trained parameters) that takes the final hidden state of the model corresponding to this token as input and transforms it to a representation used for next-sentence classification. We also tested a variation of this approach, **BERT-Pooled-Sentence**, that first segments the event descriptions into sentences, then feed these to the model separately, and then sums the pooled embeddings for each sentence in order to produce a paragraph representation.

Our second approach does not make use of the pooling layer but instead uses hidden layer states corresponding to individual word tokens and composes these (by averaging the values for each dimension) to obtain a paragraph representation. We experiment with several hidden layers: the fifth (**BERT-Mean-5**), the ninth (**BERT-Mean-9**), and the (final) twelfth layer (**BERT-Mean-12**). Following the previous literature, we expect layers 5 and 9 to work best as these are around the middle of the model, but we also include layer 12 for comparison with the pooled representations, which are also derived from the final layer.

[6] In all cases, we use the pre-trained version of the 12-layer $\text{BERT}_{\text{BASE}}$ model made available by Google (https://github.com/google-research/bert) and its PyTorch reimplementation available at https://github.com/huggingface/pytorch-pretrained-BERT.

5. Representations II: Event name modeling

This chapter introduces the second type of event embeddings that we use in our experiments: representations of the co-occurrence contexts of event names. In section 5.1 we briefly motivate our general approach, and in section 5.2 we describe the methods that we used for obtaining our representations.

5.1. Motivation

To model the co-occurrence contexts of event names, we need to construct a distributional model in which event names are treated as single word tokens. There are two types of possible approaches for doing this: (i) finding and tokenizing a corpus in which our event names occur, and then training a distributional model on this corpus; or (ii) somehow making use of existing, pre-trained models. In either case, our models should satisfy the following constraints as much as possible:

1. **Quantity**: We would like to compute vectors for as many of the events in our database as possible (to maximize the number of training samples in our experiments in Chapter 6).

2. **Quality**: The vectors in our model should be sufficiently detailed, i.e., based on a large number of occurrences of the event name.

3. **Interpretability**: We would like the vectors to have interpretable dimensions (so that they are usable in our experiments in chapter 7).

4. **Computational cost**: We prefer simpler models that are less expensive to train (or off-the-shelve models that have already been trained).

There is a clear trade-off between some of these constraints: for example, to increase the quality of the vector space, we could choose to include only event names with an occurrence count above a certain threshold, but this would obviously decrease the quantity of the vectors. There is also a trade-off between quality and interpretability: predictive models such as Word2Vec (Mikolov et al. 2013) have been shown to outperform count-based models (Baroni et al. 2014a), but are also harder to interpret. Finally, sophisticated models trained on very large corpora might be able to produce better results, but also require more computational resources and thus have a higher cost in terms of time, financial resources and environmental impact (Strubell et al. 2019). Since there is no method that satisfies all of the constraints, we experimented with several approaches, each prioritizing certain constraints over others.

	# with $\geq$ 50 occurrences	# with $\geq$ 10 occurrences	# in dataset
Hurricanes	50	231	1241
Concert tours	77	468	1978
Battles	419	1901	6138

Table 5.1.: Occurrences of event names in the Wikipedia corpus

5.2. Methods

We compare three different approaches: simple count-based models trained on the Wikipedia corpus (section 5.2.1), pre-trained Word2Vec vectors for entities in Freebase (section 5.2.2), and pre-trained vectors from the 'Wikipedia2Vec' model (section 5.2.3).

5.2.1. Simple count-based vectors

Models based on word-word co-occurrence matrices are the oldest and 'purest' kind of distributional model. While their performance on standard distributional tasks is not as good as that of more recent, prediction-based models, their advantages are that they are relatively cheap to train and are conceptually simple.

Our count-model was obtained using a four-step process. The first step was choosing and obtaining a corpus that contains our event names. Since our event dataset is derived from Wikipedia, using Wikipedia itself as a corpus seemed a logical choice. We used a dump of the full text of Wikipedia[1], which has a length of over 2 billion words. An advantage of the Wikipedia corpus is that we can be certain that our event names occur in it, at least in the articles where they are defined but also in articles that refer to these events (e.g., we would expect the 'Battle of Waterloo' to be mentioned in the article about 'Napoleon Bonaparte'). However, an important question is how frequently the event names appear. As shown in Table 5.1, most event names appear less than ten times. In order to make sure our representations are of sufficient quality, we only consider events that have a frequency of at least 50. This amounts to around 5% of the dataset across event types.

The second step was pre-processing and tokenizing the corpus in such a way that every event name corresponds to a unique token. To do this, we first created a mapping assigning every event to a unique ID token (e.g., `Battle_of_Waterloo` ⇒ `__EVENT_06585__`). Next, we identified all instances of event names in the corpus and replaced them by their respective ID tokens. Note that this search was based on the event names as given in the URL and title of Wikipedia articles, which are unique. Events with the same name are distinguished using year numbers (e.g. 'Battle of Mogadishu (1993)' and 'Battle of Mogadishu (2006)' are two different events); in such cases, we look only for the full title (including the year number) of the event. This has as a disadvantage that all instances of 'Battle of Mogadishu' without a year number have to be discarded; unfortunately, this is inevitable given that we do not have a way of decid-

[1]Dump date: November 20, 2018; text extracted and cleaned using Wiki Extractor (https://github.com/attardi/wikiextractor).

ing to which unique battles these instances correspond. After finding all event name instances and replacing them by token IDs, we tokenized the corpus using the default `word_tokenize()` function from NLTK (Bird et al. 2009).

Next, we constructed a raw word-word co-occurrence matrix from our tokenized corpus. This involved several sub-steps: first, we found unigram counts for each word $w \in V$ in the corpus and filtered out non-frequent words ($n < 50$) to get a vocabulary of frequent words V_{freq}. We also assigned a unique ID to each word in V_{freq}. Then, we started building a co-occurrence matrix C of size $V_{\text{freq}} \times V_{\text{freq}}$ in which each entry (i, j) is the number of times that context word j is found within a window of s words of target word i. The window size s is a hyperparameter; we experimented with $s \in \{2, 5\}$.

As a last, optional step, we weighted the raw counts using Positive Pointwise Mutual Information (PPMI). PPMI expresses how 'surprising' it is if two words x and y occur together, given how frequently each of them occur on their own:

$$PMI(x, y) = \log \frac{P(x, y)}{P(x)P(y)}$$

$$PPMI(x, y) = \max(0, PPMI(x, y))$$

where $P(x, y)$ is the probability of two words occurring together and $P(x)$ is the unigram probability of a word. Doing this has the advantage of reducing the influence of very frequent words. For example, if the word 'hurricane' co-occurs very frequently with the context words 'the' and 'he', this tells us nothing special about hurricans, because almost every word occurs frequently with these context words. On the other hand, the words 'hurricane' and 'damage' co-occurring frequently is more informative because neither of these words have a very high unigram frequency.

5.2.2. Word2Vec Freebase vectors

An interesting, ready-to-use resource for distributional representations of named entities, also including named events, are the pre-trained 1000-dimensional Freebase vectors released by Google as part of Word2Vec[2] (Mikolov et al. 2013). Freebase (Bollacker et al. 2008) is a (now deprecated) database containing structured knowledge about various kinds of named entities; the Word2Vec model contains distributional representations of the names of 1.4 million of these. These vectors have two main advantages: they are pre-trained and thus do not have any computational cost to us, and they are the same vectors that were used in Gupta et al. (2015), so using them makes it easier to compare our study with theirs. Two a-priori disadvantages, however, are transparency (Word2Vec dimensions do not have any inherent meaning), and the impossibility of comparing named entity vectors and other words (the vectors are released as a separate space that exclusively contains named entity vectors). Finally, while vectors are available for a fair number of events in our dataset, we are not sure of the quality of these as we do not have access to the underlying corpus and cannot check how frequent each of the entity names was in that corpus.

[2] See `https://code.google.com/archive/p/word2vec/`

	# in Freebase	# in Wikipedia2Vec	# in dataset
Hurricanes	158	821	1241
Concert tours	154	999	1978
Battles	593	4493	6138

Table 5.2.: Occurrence of named events in Freebase and Wikipedia2Vec

We constructed our Freebase space by by converting the Wikipedia URLs of all of our named events to the Freebase identifier format and then retrieving the Freebase vectors whose identifier actually existed. The identifier conversion was straightforward and involved only lowercasing the Wikipedia URLs and replacing the prefix `/wiki/` by `/en/` (e.g. `/wiki/1991_Bangladesh_Cyclone` ⇒ `/en/1991_bangladesh_cyclone`). Table 5.2 lists the numbers of events that were found both in Wikipedia and in Freebase. While this is true for only a small subset of the events in Wikipedia, we have slightly more Freebase vectors than count vectors.

5.2.3. Wikipedia2Vec vectors

As an alternative to the Freebase vectors, we also experiment with vectors from Wikpedia2Vec (Yamada et al. 2016; Yamada et al. 2018), an extension of Word2Vec's SkipGram algorithm that takes into account not just word co-occurrence statistics, but also the link graph structure of Wikipedia. This means that entities whose Wikipedia pages have similar incoming links and whose textual contexts are similar will be close to each other in semantic space. Entities and other words are projected into the same semantic space. This approach has two main advantages over the Freebase vectors: the fact that they are based on Wikipedia means that they should (in theory) contain representations for all of the events in our dataset, and the fact that entities and words are projected into the same space means that they are comparable. Furthermore, Yamada et al. (2016) show that a system based on their Wikipedia2Vec embeddings achieves state-of-the-art performance on a named entity disambiguation (NED) task, suggesting that these vectors also capture referential information about entities. However, a main drawback of these vectors is that they are not just based on corpus data but also on structured knowledge, which might give them an unfair advantage vis-à-vis other entity representations.

Event names in Wikipedia2Vec are identical to their corresponding Wikipedia counterparts, except that the prefix `/wiki/` is replaced by `/ENTITY/` (e.g. `/wiki/1991_Bangladesh_Cyclone` ⇒ `/ENTITY/1991_Bangladesh_Cyclone`). Table 5.2 shows the number of our named events for wich we could retrieve vectors; while this was possible in a majority of cases, for a significant number of events in Wikipedia we could not find a corresponding entry in Wikipedia2Vec. We have not been able to find out why this is the case.

6. Testing referentiality 1: Attribute prediction

In this chapter, we describe our experiments with predicting referential properties from our two types of distributional representations of named events. In section 6.1, we describe how we approach the prediction task and which models we use, and in section 6.2 we discuss the results of our experiments.

6.1. Methods

6.1.1. Tasks

Our experimental setup is based on the one used in Gupta et al. (2015), but departs from it in several ways. Gupta et al. frame the attribute prediction problem as a zero-shot learning task: their goal is to learn a mapping from distributional space to a set of referential attributes, that are predicted all at once. To do this, they use a logistic regression model that takes a distributional vector as input and outputs predictions for each referential attribute. For binary attributes, the predicted value is the probability that an entity has that attribute; for numerical attributes the normalized value of the attribute is predicted directly.

The first difference between Gupta et al.'s work and ours is that we train a separate model for each attribute. Under Gupta et al.'s approach, both the inputs and outputs are multi-dimensional vectors. A potential theoretical advantage of this approach is that one could frame the task as a cross-space mapping problem, in which distributional and referential information have a similar status (i.e., both distributional and referential entities 'live' in a similar kind of space). Although it is not clear whether or not Gupta et al. actually view their specific task in this way, cross-space mappings are a popular approach elsewhere in formal distributional semantics, for example in (Herbelot and Vecchi 2015), where a mapping is learned from distributional vectors to a 'truth-theoretic space' whose dimensions are quantifiers (represented as numerical values corresponding to 'all', 'most', 'some', and 'none'). A more practical reason to choose for a cross-space mapping rather than predicting attributes separately would be to make use of correlations between different attributes.

However, we have still chosen to predict attributes separately because doing so makes our approach more flexible and allows us to make better use of the available data. In our event database, there is much variation between events as to how many attributes they have. For learning a cross-space mapping, we would have to restrict our dataset to only those events that have all of the attributes that we would like to predict, which would imply discarding a large part of this dataset. By contrast, when learning to predict attributes separately, we can train and test on all of the data available for each attribute. Another advantage of learning separate models is that it allowed us to work incrementally: we could add new attributes to our system 'on the fly' without having to re-train the models for the attributes we already had.

Moreover, we expect that we can achieve good results even without being able to take advantage of correlations between attributes, given that Gupta et al. (2015) obtained good results with a regression model that learns parameters for every output dimension separately and does not capture relationships between the output dimensions. This means that, in practice, their approach is equivalent to learning separate models but with the same hyperparameters. We approximate this property by tuning on only one of the attributes and using the same hyperparameters for all of the attributes.

A second difference is that we use classification rather than regression. This has several advantages: first of all, it allows us to evaluate predictions for numerical and categorical attributes in the same way, using standard measures for classification tasks such as accuracy and F1 scores (Gupta et al. 2015 use accuracy for binary attributes and a rank-based score for numerical attributes). Moreover, some of our attributes, such as hurricane category and affected areas are modeled more naturally using classification than regression; while it is possible to approximate these using binary attributes (e.g., *is_category_1*, *is_category_2*, ...), interpreting would be more difficult. Finally, from preliminary regression experiments on our dataset, we expect that trying to predict exact values for numerical attributes is too difficult a task, even when evaluated in a rank-based way as in Gupta et al. (2015). Based on this expectation, we decided to set a somewhat lower bar for our model and classify numerical attributes into broad categories rather than predict precise values.

We considered several strategies for defining classes for numerical attributes. An earlier idea was to find the range of the values found in the dataset (e.g. 2000 BCE to 2000 CE for year attributes) and divide this range into a given number of equal-sized 'windows' (e.g. 2000 BCE to 1000 BCE, 1000 BCE to 0, etc.). However, this approach is very sensitive to outliers, and yields very unbalanced class distributions (e.g. almost all events are in a single class, and the other classes only contain a few events each). Instead, we opted for an approach that relies on the distribution of event values in the dataset and sets thresholds based on percentiles (e.g., the 50th percentile, or the median, is the value below which 50% of the data points lie). The percentile thresholds for each of the numerical attributes are given in Table 6.1.

In order to test how the level of detail of the classes affects prediction accuracy, we define two classification problems for every attribute (except for categorical attributes that are binary by definition): a binary problem and a multi-class problem. For numerical attributes, the binary problem is defined as classifying events as either below the median, or above it, while for the multi-class problem, the boundaries are set at the 25th, 50th, and 75th percentiles. By definition, classes defined in this way will have the same size (implying that a majority-class or similar baseline algorithm would be expected to achieve 25% or 50% accuracy, respectively), although class sizes can vary a bit if there are many events with values exactly at one of the thresholds. For categorical attributes, we tried to find 'natural' boundaries: for area attributes we divide the world into hemispheres (east and west, or north and south) for binary classification, or into quadrants (north-east, south-east, south-west, north-west) for multi-class classification; for hurricane categories we used the full Saffir-Simpson scale (plus two categories for tropical storms and tropical depressions) for multi-class classification and the distinction between hurricanes and sub-hurricane storms for binary classification. The other three categorical attributes (*involves_France*, *involves_US*, *involves_Spain*), are binary by definition.

	Percentiles		
	25%	50%	75%
Hurricanes			
Da (damage, $)	2,960,000.00	45,000,000.00	350,000,000.00
Du (duration, days)	5.00	8.00	12.00
Fa (fatalities, #)	1.00	7.50	40.00
Pr (air pressure, mbar)	935.00	960.00	986.00
Wi (wind speed, km/h)	117.50	175.00	230.00
Ye (year)	1985.75	2002.00	2010.00
Concert tours			
Du (duration, days)	79.00	205.00	368.00
Le (legs, #)	1.00	3.00	5.00
Ye (year)	1996.00	2007.00	2013.00
Battles			
Be (belligerents, #)	2.00	3.00	4.00
StR (strength ratio)	0.30	0.52	0.75
StT (strength total, #)	5505.00	19921.00	50103.50
Ye (year)	1520.00	1796.00	1901.00

Table 6.1.: Percentile-based class thresholds for numerical attributes. The 50th percentile (=median) is used for binary classification, and the 25th, 50th, and 75th percentiles are used for multi-class classification.

6.1.2. Models

Earlier work predicting referential properties from distributional representations (e.g. Gupta et al. 2015; Herbelot and Vecchi 2015) obtained good results with simple regression models (logistic regression, partial least squares regression), suggesting that simple (linear) classification models might also be sufficient for our task. We try two different models: **SVM**, a Support Vector Machine classifier, and **MLP**, a feedforward neural network model with a single hidden layer. SVM classifiers compute a hyperplane that separates instances of two classes; we used the `SVC` implementation from the Scikit-learn Python package Pedregosa et al. 2011, with a linear kernel[1] and using the one-against-one approach for multi-class classification. MLP learns weight matrices such that it optimizes a cross-entropy loss function using gradient descent. We implemented our model using PyTorch (Paszke et al. 2017). We use an Adam optimizer Kingma and Ba (2014), and implemented 'early stopping': we train the model for 200 epochs, or until the loss on the validation set does not improve over a window of 10 epochs. We also experiment with applying 'dropout' (Srivastava et al. 2014) on both the input layer and the hidden layer in order to reduce overfitting. We tested both models against a 'stratified' baseline algorithm that makes random predictions consistent with the class distribution of the training set (e.g. if 40% of the training samples are in class X, 40% of the test samples will be assigned to class X).

[1] We also experimented with an RBF kernel, but this did not produce good results.

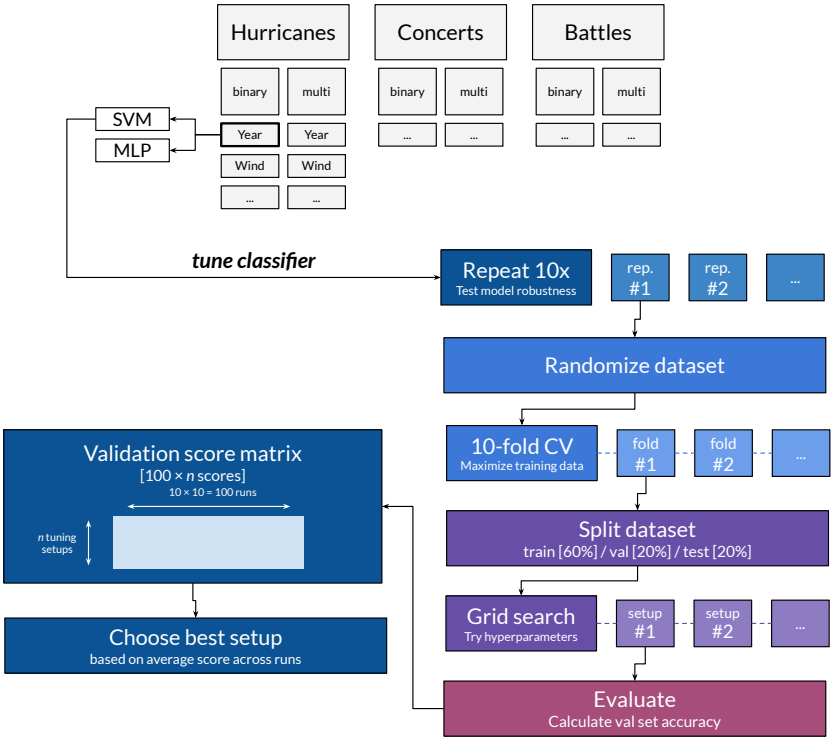


Figure 6.1.: Tuning pipeline (only applied to hurricane/year attribute)

For both models, we implemented a grid search to find the best hyperparameter settings. For SVM, we only tuned the C parameter, which influences the width of the margin of the hyperplane that is learned. For MLP, we tuned the initial learning rate, L2 penalty (regularization), dropout rate (on the input layer and on any hidden layers), batch size, and hidden layer size for MLP. Given the large number of experimental setups (37 attribute classification tasks $\times$ 2 models $\times$ 6 embedding types = 444 experiments for the experiments with description embeddings alone), doing a full hyperparameter grid-search for each setup was not feasible. For this reason, and to test the robustness of our models, we only tuned on the models for one attribute and used the best settings for the other attributes.

6.1.3. Tuning and training pipelines

The pipeline for tuning our models is summarized in Figure 6.1. Tuning is done on each of the setups for the 'year' attribute of the hurricane dataset (there is one setup for every combination of model type, embedding type, and binary/multi-class; e.g., we tune SVM/Summed/Binary, MLP/BERT-Mean-5/Multi-class, etc. separately). While our choice for the 'year' attribute is largely arbitrary, we preferred it over other attributes because it is present in all three event

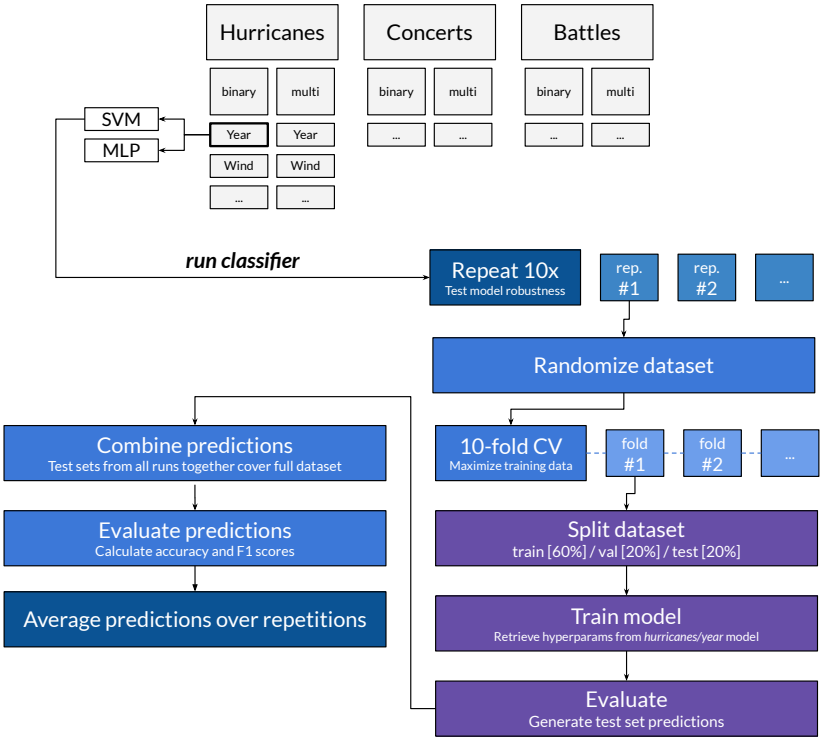


Figure 6.2.: Training and evaluation pipeline (applied to all attributes)

datasets, has a large number of training samples, and because we see it as a 'core' attribute of events in general. Our pipeline is somewhat complicated as it involves applying cross-validation and repeating the training procedure several times. We use 10-fold cross-validation because our dataset is relatively small, and we would like to maximize the amount of data available for use in both training and testing. On the other hand, we added repetitions to the dataset because initial experiments with our MLP model suggested that, while generally performing well, its performance varied a lot (in part depending on hyperparameter settings). Repeating the training procedure makes it possible to quantify our models' stability and to get a more robust estimate of the model's performance.

For doing cross-validation, we created ten different (randomized) partitions of the dataset ('folds'), each of which consisted of 60% training items, 20% validation items, and 20% test items. For each fold, we ran each of the hyperparameter setups in our search grid, and calculated accuracy scores on the validation set. We then repeated the cross-validation procedure ten times. Doing this yielded an accuracy score matrix with 100 validation scores (10 folds $\times$ 10 repetitions) for every hyperparameter setup; we selected the setup with the highest average validation score as the 'winner'.

Next, having found optimal hyperparameter settings for the 'year' attribute, we can train

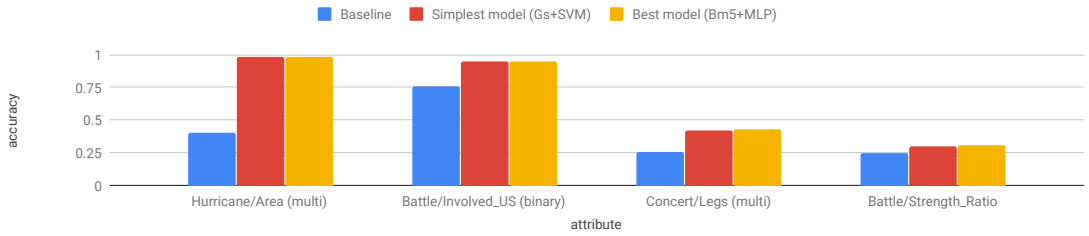


Figure 6.3.: Accuracy scores for strong (easy-to-predict) and weak (hard-to-predict) attributes

all of our models using these settings. Our training pipeline is schematized in Figure 6.2, and is performed for each of our experiments. Again, we use 10-fold cross-validation (with the same split as before); the validation sets are used only for calculating validation loss for the MLP early stopping algorithm. Evaluation was done by combining the test sets from each of the folds and calculating accuracy and F1 scores. As before, we repeated the cross-validation process 10 times and average the scores over predictions to produce final results.

It should be noted that, while cross-validation is used both for tuning and training, it is applied in both cases on the full dataset. Since tuning is performed for one attribute only, in the majority of cases there is not overlap in tuning and testing data. However, for the 'year' attribute in the hurricane dataset, test results are obtained on data that is already seen during tuning, which means that the results for this attribute should be interpreted with some caution.

6.2. Results

In this section, we report and discuss the results of our attribute prediction experiments. We first look at the results of our description representation experiments (section 6.2.1, before moving on to event name modeling (section 6.2.2).

6.2.1. Description representations

We first look at how well each attribute is predicted, and how accuracy varies across models and embedding types. Next, we provide a brief qualitative analysis of the mistakes that our models make and evaluate the influence of hyperparameter choices and of random factors.

Prediction accuracy

We evaluate every combination of attribute, model and embedding type with accuracy scores and macro-averaged F1 scores.[2] Tables A.1 and A.2 in the appendix give our full set of results. For the sake of simplicity, in the discussion in this paragraph we focus mainly on the results for our 'simplest model' (summed GloVe embeddings [Gs], with an SVM for predicting attributes) and our 'best model' (the best-performing BERT embeddings [Bm5], with an MLP classifier).

[2]We calculate macro-averaged F1 scores by first computing F1 scores for individual classes and then averaging these.

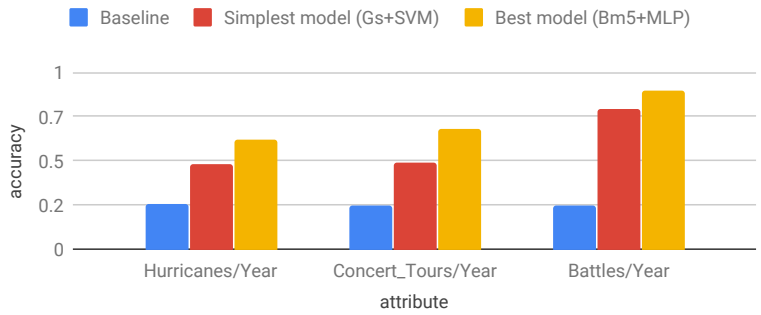
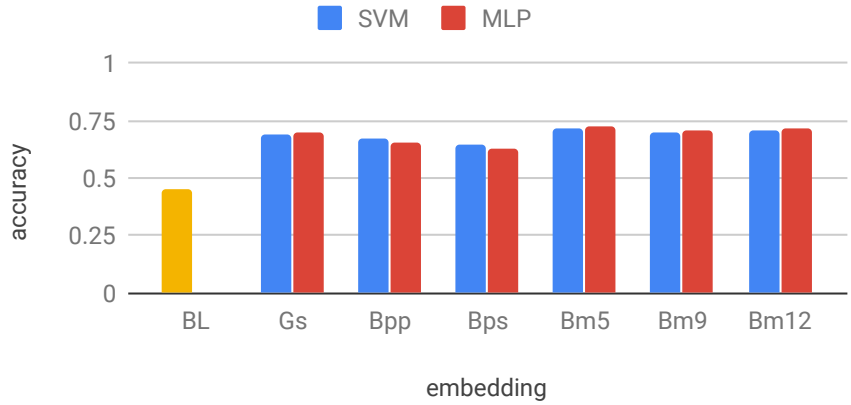


Figure 6.4.: Accuracy scores for year attributes (multi-class), split by event type

We achieved above-baseline performance for all attributes. However, as shown in Figure 6.3, there are large differences between how well every attribute is predicted. Near-perfect accuracy is achieved for area attributes (ArNS, ArWE, Ar; for hurricanes and battles), and for the 'year' attribute and the 'country involved' attributes (InF, InS, InUS) for battles. While, due to the unbalanced distribution of these attributes, baseline performance on these attributes is also high, our models still outperform the baseline by a wide margin. On the other hand, numerical attributes have an equal class distribution; the best accuracy scores for most of them lie between 0.69-0.91 (binary) or 0.43-0.71 (multi-class). A negative outlier is 'Strength_Ratio' (i.e. the difference is army sizes between the two sides in the conflict), which is predicted with only slightly above-baseline accuracy by most models.

Comparing the predictability of the different event types is difficult because every event type has different attributes. The only attribute that all event types have in common is 'Year' (see Figure 6.4). For both hurricanes and concert tours, the best models achieve a score of up to around 0.85, but for battles we get up to 0.96. A possible explanation for this might be the higher number of training samples. Also, unlike for most other attributes, for the 'year' attribute there is a large difference in performance between the 'simple' and the 'best' models. This difference is likely due to the fact that for creating GloVe-Summed vectors, we exclude numerical tokens (and hence, year numbers) to make the representation as 'poor' as possible. For BERT representations we did not do this; hence, we hypothesize that the BERT model was able to use its representation of year numbers to improve its guess of when an event happened. Interestingly, the difference between GloVe-Summed and BERT is smaller for battles than for the other event types.

Moreover, hurricanes and concert tours both have a 'Duration' attribute; prediction on hurricanes works slightly better (up to 0.74 vs. up to 0.69). Also, hurricanes and battles both have area attributes; interestingly, performance is better (up to 0.98 vs. up to 0.93 for the multi-class problem) for hurricanes even though there is much less training data available. Note that what 'area' means has a slightly different interpretation for both event types: for battles, we usually have GPS coordinates for the specific location where the event took place, while for hurricanes we rely on the list of countries and regions that were affected, found the coordinates given on the Wikipedia page for that country or region, and inferred the hemisphere from those coordinates. However, given the coarseness of our predictions (on the hemisphere/quadrant level),



Legend: BL=baseline, Gs=GloVe-Summed, Bpp=Bert-Product-Paragraph, Bps=Bert-Product-Sentence, Bm5/Bm9/Bm12=Bert-Mean-5/9/12

Figure 6.5.: Comparison of event description embeddings; averaged scores across attributes

it is unclear whether this could have an influence on prediction accuracy.

There is some variation between which model performs best for which problem, but overall, the SVM and MLP model work about equally well. On the GloVe-Summed embeddings, SVM and MLP have accuracy scores of 0.689 and 0.694, respectively, averaged across all attributes. This difference is statistically significant (a two-tailed paired T-test gives $T = -3.35$ with $p = 0.002 < 0.05$) but negligible. More interesting is the variation in performance between different representations. As shown in Figure 6.5, the BERT-Mean representations work best, but GloVe-Summed comes very close. For the SVM model, GloVe-Summed has an average score of 0.69, whereas Bert-Mean-5 has an average score of 0.71. Again, this difference is significant but negligible ($T = -2.8$, $p = 0.008 < 0.05$). Comparing the different BERT embeddings, we find that combining representations for individual word tokens (BERT-Mean) works much better than using pooled representations (from the CLS token) for the entire paragraph (Bert-Pooled-Paragraph) or combining pooled representations for individual sentences (Bert-Pooled-Sentence). Furthermore, zooming in on BERT-Mean, we do not find large differences between the three layers.

Confusion analysis

We performed a confusion analysis to see what kinds of mistakes our models make. Figure 6.6 shows confusion diagrams for the subset of multi-class problems that we think is most relevant to discuss. First, we consider the 'Area' attribute, because it exists for both hurricanes and battles and because it has high accuracy scores for both event types. Hurricane area predictions are (almost) perfect for the two largest classes, North-East (Europe and most of Asia) and North-West (mostly North-America). The smaller South-East category is also predicted well, but South-West seems to have too little training data to make meaningful predictions (however, for most hurricanes the model still correctly predicts that they occurred in the southern hemi-

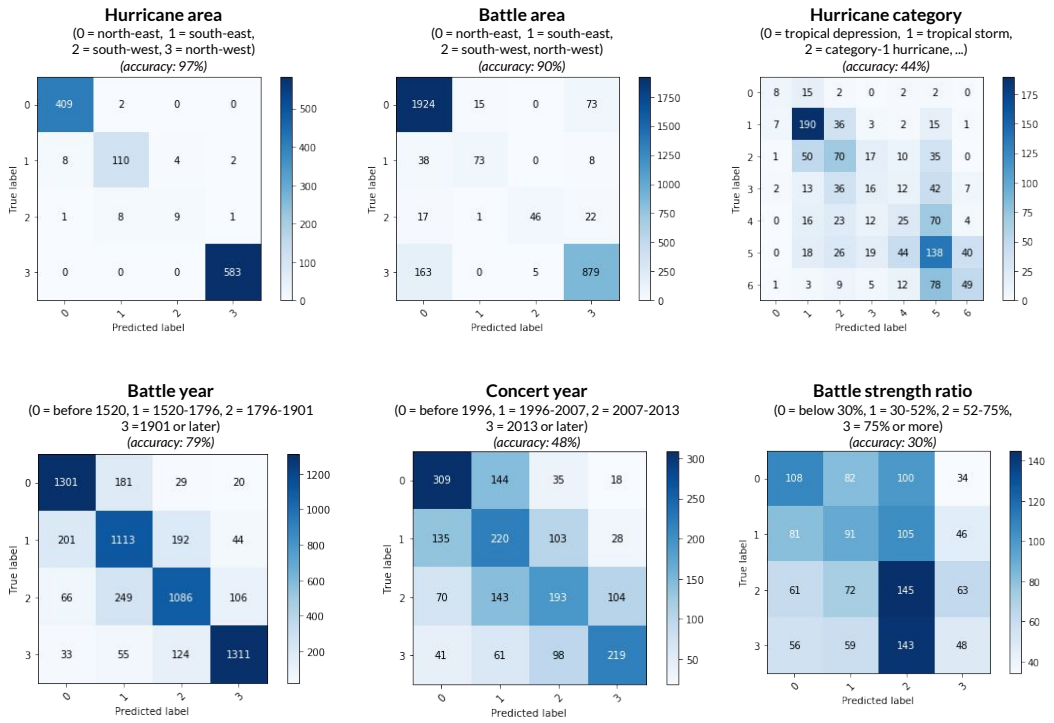


Figure 6.6.: Confusion matrices for a subset of multi-class event description classification problems (model: SVM, embedding: GloVe-Summed)

sphere). Battle area has a slightly lower prediction area; here, we find some confusion even in the two largest classes (again, these are North-West and North-East), for example, about 15% of battles in the North-West get classified as having occurred in the North-East.

Next, we look at hurricane category, which is interesting because it is ordinal, and because it has more categories than other attributes, and also at the numerical attributes 'year' (for battles and hurricanes) and 'strength ratio' (for battles). 'Year' predictions are excellent for battles, but less so for concerts; 'strength ratio' is not well predicted at all (only slightly above the baseline accuracy of 25%). In all cases, we hoped to find that most mistakes are between the correct class and adjacent classes. We indeed see this pattern for the 'Year' attributes, and in part for hurricane category. However, for some less-well predicted hurricane category classes, mistakes are more evenly distributed, and this is also the case for battle strength ratio. Another interesting phenomenon is that hurricane categories 1 ('tropical storm') and 5 ('category-4 hurricane') 'attract' a large share of the predictions; this could in part be explained by the relatively high frequencies of these classes.

Model	Average presence in 25% best setups						
BS	***50***	***500***	***full***		***LR***	***0.001***	***0.005***	***0.01***
MLP	0.03	0.39	0.58			0.46	0.33	0.21
DO	***0***	***0.1***	***0.5***		***HS***	***50***	***150***	
MLP	0.40	0.39	0.21			0.48	0.52	
C	***0.001***	***0.01***	***0.1***	***1***	***10***	***100***		
SVM	0.14	0.21	0.50	0.07	0.07	0.09		

Table 6.2.: Hyperparameter results (BS=batch size, LR=initial learning rate, DO=dropout rate, HS=hidden layer size)

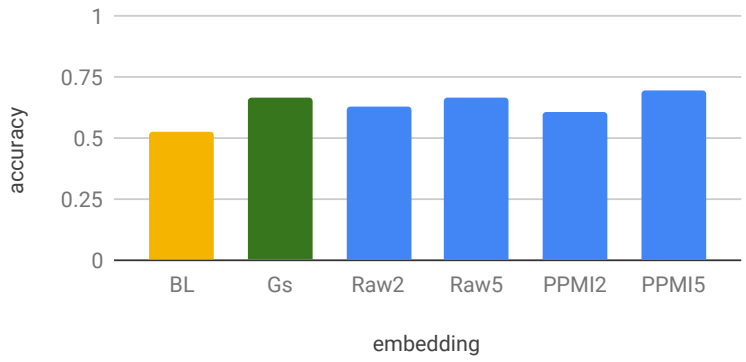
Hyperparameters and stability

Hyperparameter settings were analyzed by going back to the score matrix (which compares the accuracy of our setups across cross-validation folds and repetitions; see Figure 6.1) and finding the 25% best-performing settings for each of our tuning setups (i.e., combinations of attribute, binary/multiple, model and embedding options). Then, we calculated, for every hyperparameter value, in what proportion of best setups (on average) it is present. As shown in Table 6.2, there is quite a lot of variation between tuning setups, but there are some clear patterns: such as that full batch gradient descent works best and that dropout does not help much. We also evaluated the stability of the models by testing how much variation in accuracy there is between repetitions; SVM is the most stable model of the two: its accuracy varies with a standard deviation between $\sigma = 0.001$ and $\sigma = 0.02$ (depending on the experiment), while for MLP we find $0.001 < \sigma < 0.03$.

6.2.2. Event name representations

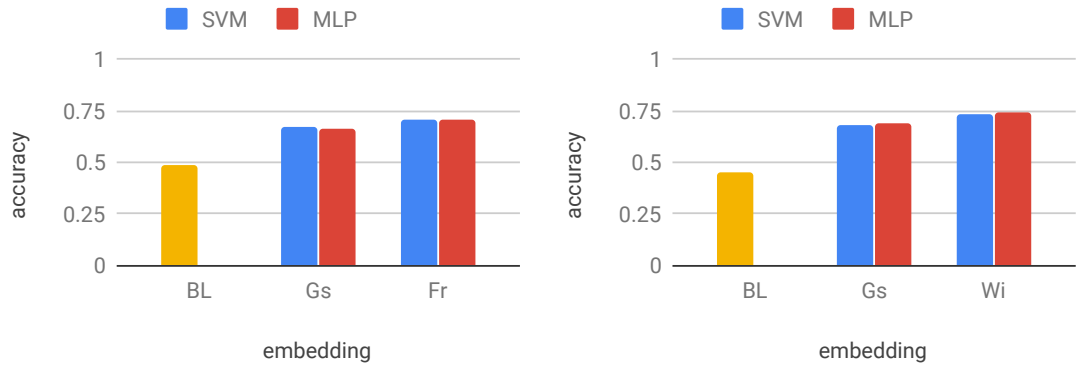
In this subsection, we discuss the performance of our three types of event name representations: count-based vectors, Freebase vectors and Wikipedia2Vec vectors. The full results of our experiments are given in the appendix (Tables A.3 through A.5). Because these vectors are available for different subsets of our event dataset, we cannot compare their results directly against each other. However, we will compare each of these methods against results for description embeddings: for each of the three event name representation types, we also trained and tested our models on description embeddings (GloVe-Summed) for the subset of events for which the name representations were available. The overall patterns of performance of event name representations are very similar to those of description embeddings; here, we only discuss those of our findings that are specific to event names.

Figure 6.7 shows that the best count-based representations are PPMI-weighted vectors with window size 5, which (with an average score of 0.70) outperforms both the other count-based embeddings and the GloVe-Summed description embeddings (which get an average score of 0.66; a two-tailed paired T-test shows that this difference is significant with $T = 3.7$ and $p = 0.0008$). Interestingly, PPMI-weighting increases performance for window size 5, but



Legend: BL=baseline, Gs=GloVe-Summed, Raw2/Raw5=Raw (unweighted) counts, window size 2/5, PPMI2/PPMI5=PPMI-weighted counts, window size 2/5

Figure 6.7.: Comparison of count-based name representations; averaged scores across attributes



Legend: BL=baseline, Gs=GloVe-Summed, Fr=Freebase, Wi=Wikipedia2Vec

Figure 6.8.: Freebase and Wikipedia2Vec vs. description vectors; averaged scores across attributes

not for window size 2. Note that these results are only from SVM models, which work better for sparse, high-dimensional representations than MLPs. Meanwhile, in Figure 6.8, we see that both Freebase vectors and Wikipedia2Vec vectors also perform slighly better than GloVe-Summed vectors.

While, overall, all types of event embeddings perform well above baseline, the margin with which they do can vary: description embeddings and Wikipedia2Vec, on average, perform between 20-28 percentage points above baseline, while this is only between 8-17% for count embeddings. Freebase vectors are in the mid range and add around 22% to baseline performance. Moreover, event name vectors, unlike description vectors, do not beat the baseline for all attributes. This is especially true for count vectors; while the best embedding type, PPMI5, beats baseline accuracy on all attributes, other representations perform at or below baseline level for 1-3 out of 33 attributes; interestingly, these are not the same attributes for every representation type. When we also look at F1 scores, the picture becomes more pessimistic: PPMI5 representations have an F1 score below the baseline for almost half of attributes (15/33). Strangely, raw count representations, which have lower accuracy scores than PPMI representations, do better: Raw2 scores below baseline in only 6 cases. Meanwhile, Freebase vectors perform better, and get a below-baseline F1 score for only 3/37 attributes; Wikipedia2Vec never performs below baseline.

We suspect that the lag in performance of count vectors is due more to the limited availability of training data than to an inherent weakness of the representations, for several reasons. First of all, models trained on GloVe-Summed representations for the same set of events as the count vectors have the same performance issues (with a sub-baseline F1 score in 10/33 cases). Second, the fact that different count-based models have problems with different attributes suggests that these problems are more or less random and not due to an inherent inability for count models to capture certain kinds of referential information. Finally, Freebase vectors capture the same kind of distributional information as count vectors but achieve better performance (with more training data), which again points in the same direction.

7. Testing referentiality II: Analyzing the event space

In chapter 6, we found that we can successfully predict many different referential attributes from the events in our datasets. However, what we do not know yet is how our models do this: what kind of distributional cues are helpful for inferring referential information? And how is the resulting event space structured?

7.1. Motivation and approach

A challenge in distributional semantics is that word representations are not transparent: the position of a word in semantic space relative to other words tells us something about its meaning, but it is usually unclear exactly how each dimension contributes to the overall meaning of the word. This is especially the case for vectors learned using techniques like Word2Vec and GloVe, where the dimensions of the space do not have any inherent (interpretable) meaning. In count-based models (without dimensionality reduction), individual dimensions have a clear interpretation (i.e., co-occurrence counts with particular context words), it is not clear how meaning emerges from the combination of these dimensions. For our purposes, we are not just interested in how distributional representations encode the meanings of words in general, but more particularly in how named event representations are represented and how this might reflect there referential properties.

Here, we propose a method of analyzing event spaces based on Principal Components Analysis (PCA), which is a technique for transforming a space (S) whose dimensions are possibly correlated into another space (S') whose dimensions (called 'principal components') are not correlated. The dimensions of S' are ordered by how much of the variance in the original data they explain. In many cases, the first few dimensions can account for a large portion of the variance. Because of this, PCA is popular as a method for dimensionality reduction: by keeping only the first few dimensions of S', most of the information in S can be represented but in a form that is much more compact and easier to process computationally.

Here, we are not interested in dimensionality reduction but in combining information from different dimensions into a single dimensions. For example, suppose that there are a number of events in one of our event spaces that have a certain referential property in common, such as that they all took place in the same region). If this referential property is reflected in the distributional representation, we might expect that these events have similar values in one or several dimensions. However, such patterns are difficult to find since these dimensions also encode many other properties besides spatial location. By 'un-correlating' the dimensions in the event space, we hope to get single dimensions that encode only spatial location (or other referentially relevant features).

Our procedure is as follows: we take the space of all GloVe-Summed representations G_s for a certain event type (e.g. hurricanes) and then fit a PCA algorithm on this space, so that we get

a transformed space G'_s.[1] Next, we take the space G_w of GloVe vectors for all of the individual words that occur at least once in the event descriptions from which G_s was derived, and then project these vectors into G'_s. Then, for every dimension in G'_s, we sort the words in G_w by their value on that dimension. Our hypothesis is that, if a certain word has a high absolute value on a given dimension, that dimension is important for distinguishing that word from other words. By inspecting which words from the hurricane descriptions are 'activated' by particular dimensions, we hope to find out something about what that dimension encodes. For example, suppose that in the G'_s space for hurricanes, there is a dimension that encodes wind speed, we might expect wind-related words to have a high absolute value on this dimension. This idea is (in part) inspired by earlier work on embedding space interpretation such as Shin et al. (2018).

7.2. Results

In Table 7.1, for each of the first five PCA dimensions, we give the words with the ten lowest activations (the 'blue zone' of the table, indices 0 - 9) and the words with the ten highest activations (the 'red zone', indices -10 to -1). For some dimensions we indeed find very clear patterns: in those dimensions, the words with the lowest activations and/or with the highest activations belong to a certain semantic category. Interestingly, the patterns that we find on the 'positive side' and the 'negative side' of the dimension are often not related to each other. Table 7.2 gives our interpretation of these patterns: for each dimension, we give the most common semantic category of the words that are activated by that dimension, along with how many words in the top-10 belong to that category. Of course, grouping words into categories is always somewhat arbitrary, and our interpretation is fundamentally subjective; however, we believe that some of the categories that we found are so clear that they should be uncontroversial.

For hurricanes, we find at least two geography-related dimensions: all of the (top-10) positively activated words in PCA dimension 4 are related to Asia ('asia', 'malaysia', 'korean', 'japan'), whereas most of the postively activated words in PCA dimension 1 are related to U.S. geography ('louisiana', 'texas', 'boston', but also 'county'). Furthermore, there is a 'change' dimension (dimension 3, negative activations) that contains verbs, adverbs and adjectives describing how hurricanes can change ('deepened', 'emerged', 'increasingly', 'rapidly'). Finally, there is a dimension with damage-related words (dimension 3, negative: 'damage', 'erosion', 'catastrophic') and a dimension with finance-related words and symbols (dimension 2, positive: 'crore', '$', 'php', 'rs' are currency symbols, and '%' and 'amount' are also finance-related).

Concerts also have geography-related dimensions: in dimension 1 all negatively activated words are countries and continents, and in dimension 2 most positively activated words are related to the north-western Europe ('iceland', 'scotland', 'scandinavia', 'celtic'). There are also two dimensions (dimension 1, positive activations; dimension 4, negative activations) that activate only music-related terms (e.g. 'bassist', 'sound' in dimension 1; 'singer-songwriter',

[1] We use the implementation from Scikit-Learn (Pedregosa et al. 2011) (documented at`https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html`), which in turn uses the LAPACK implementation which performs Singular Value Decomposition (SVD) on the full dataset (rather than using a randomized approximation)

	Dim	Var	Activated words (-)	Activated words (+)
Hurricanes	0	0.70	*francelia, bogale, gaivota, kuring, sendang, koryn, kulap, nomoi, o7f, gorio*	*first, water, [,], new, would, around,'s, [.], one, time*
	1	0.03	*temperature, inhg, convective, km/h, humidity, gusting, precipitation, southwesterly, extratropical, barometric*	*louisiana, texas, boston, county, orleans, florida, st., killed, $, million*
	2	0.03	*atlantic, northward, southward, shores, seaboard, westward, eastward, landfall, coast, caribbean*	*error, use, crore, %, $, total, php, amount, rs, per*
	3	0.02	*deepened, emerged, intensified, worsened, deepen,increasingly, widespread, evolved, plagued, rapidly*	*hpa, gusting, mb,),(, ft, mbar, km, mph, km/h*
	4	0.02	*damage, moisture, erosion, damages, catastrophic, excessive, residual, caused, minimal, adverse*	*asia, malaysia, korean, japan, shanghai, philippines, kong, korea, taiwan, hong*
Concert tours	0	0.76	*2017-2018, blá, cómplices, 279.2, f.u.c.k, blankensee, ahoi, 237.8, samppa, albums*	*last, n't, [.], [,], year, time, first, new, one, s*
	1	0.03	*netherlands, emirates, spain, oceania, switzerland, asia, france, portugal, colombia, uruguay*	*bassist, sounds, chord, song, band, sound, drummer, guitarist, songs, guitar*
	2	0.01	*consumption, increasing, continent, worldwide, technological, china, asia, significantly, profitable, achieving*	*drummer, february, march, guitarist, october, april, august, july, june, bassist*
	3	0.01	*], [, $,), (, x, #, +, blige, citation*	*iceland, scotland, rock, europe, scandinavia, celtic, england, toured, thrash, band*
	4	0.01	*singer-songwriter, music, singer/songwriter, composer, artists, songwriter, orchestra, album, ensemble, orchestral*	*accident, ended, avoid, died, months, injured, injury, prevent, passed, knee*
Battles	0	0.68	*zizhi, tongjian, на, râmnicu, tín, xasan, jiā, zhī, michel-ange, maserfield*	*new, people, [.], two, time, [,], would, first, one,'s*
	1	0.02	*de, william, la, st, thomas, charles, saint, henry, del, st.*	*airstrikes, sunnis, insurgency, al-qaeda, qaeda, islamist, iraqi, insurgents, hamas, taliban*
	2	0.02	*marine, harbor, creek, patrol, aircraft, maine, pacific, aviation, carolina, aerial*	*ptolemy, heraclius, claudius, habsburg, emperors, charlemagne, visigoths, constantinople, antiochus, emperor*
	3	0.01	*],), [, (, :, {, /, }, size, domain*	*maj., army, battalion, colonel, lieutenant, troops, regiment, infantry, confederate, cavalry*
	4	0.01	*france, adriatic, mediterranean, spain, croatia, czech, portuguese, german, italy, baltic*	*sen., hawkins, elijah, malcolm, john, smith, lord, peyton, county, sheriff*

Table 7.1.: Words with highest and lowest activations for PCA dimensions (summed GloVe vectors). The 'Var' column gives the ratio of the variability in the data that is explained by each PCA dimension.

	Dim	Var	Dominant category (-)	Dominance (-)	Dominant category (+)	Dominance (+)
Hurricanes	0	0.70	*hurricane names*	6/10	*interpunction, particles*	3/10
	1	0.03	*weather terms*	7/10	*geography (US)*	6/10
	2	0.03	*geography/directions*	10/10	*currency/money-related*	7/10
	3	0.02	*change-related*	8/10	*units, interpunction*	9/10
	4	0.02	*damage-related*	10/10	*geography (Asia)*	10/10
Concert tours	0	0.76	*numbers*	2/10	*interpunction, particles*	4/10
	1	0.03	*countries/continents*	10/10	*music-related*	10/10
	2	0.01	*economy-related*	6/10	*months*	7/10
	3	0.01	*interpunction*	8/10	*geography (North-West Europe)*	6/10
	4	0.01	*music-related*	10/10	*'misfortune'-related*	8/10
Battles	0	0.68	*foreign words*	9/10	*interpunction, particles*	3/10
	1	0.02	*(saint) names*	10/10	*conflicts in the Middle East*	10/10
	2	0.02	*navy/air force-related*	8/10	*king/emporer-related*	8/10
	3	0.01	*interpunction*	8/10	*army-related*	9/10
	4	0.01	*geography (Europe)*	10/10	*names/titles*	9/10

Table 7.2.: Interpretation of word categories in Table 7.1

'orchestra' in dimension 4). Another interesting dimension is what we call the 'bad luck' dimension (dimension 4, positive activations) which contains words like 'accident', 'died', and 'injured'. While these words do not seem directly related to concert tours, a manual inspection of the concert event descriptions makes clear where they come from: many concert tours are somehow related to some life event of one of the band members. For example, in the entry for the 'Unity Tour' by The Jacksons[2], we find "The tour also marked the first time the brothers have toured as the Jacksons without brother Michael, who died in June 2009".

Finally, for battles we find a dimension for European geography (dimension 4, negative activations), and also two military-related dimensions. The words in the first of these (dimension 2, negative activations) seem to be more related to air forces and navies ('marine', 'harbor', 'aircraft') whereas the words in dimension 4 (positive activations) seem to be more army-related ('army', 'battalion', 'infantry', 'cavalry'). There is also a dimension that seems related to recent (religious) conflicts in the Middle East (dimension 1, positive activations), activating words such as 'airstrikes', 'al-qaeda', 'iraqi', and 'taliban'. Finally, an interesting pattern is that many positively activated words in dimension 2 refer to rulers and emperors from different periods in history: 'ptolemy' (Ptolemaic Egypt, from the 4th century BCE), 'claudius' (Roman emperor), 'charlemagne' (early middle ages). Other words in the same dimension also seem vaguely related to this theme, e.g. 'visigoths', 'constantinople'.

An interesting general pattern is that the first PCA dimension of every event space explains a very large portion of the variation in that space, but does not yield very clear semantic patterns. For example, for all three event types, the tokens that are positively activated by the first dimension contain interpunction (comma, period), particle words (''s', 'n't') and the words 'new', 'first', and 'time'. The groups of negatively activated words in these dimensions are somewhat more coherent, but do not form a clear semantic group: for battles, the only thing that the

[2] https://en.wikipedia.org/wiki/Unity_Tour

words have in common is that they are from foreign languages (Chinese, Russian [?], Romanian, French); for concert tours, there seems to be no pattern at all. For hurricanes, we find a more interpretable pattern: most of the negatively activated words for the first dimension are names or alternative names[3] of hurricanes (typhoons) in the Pacific ocean. Although our analysis is admittedly very informal and somewhat speculative, these observations seem in line with the finding by Bullinaria and Levy (2012) that the first 100 principal components in a dimensionality-reduced count-based space "tend to be contaminated by aspects other than lexical semantics" (p. 898). In our case (if our analysis is correct), only the first component is 'bad', but this could be explained by the fact that our summed GloVe space only had 300 dimensions to begin with.

As an extra experiment, we also looked at the activations of event vectors themselves in the G'_s space. While this did not always yield clear results, we did find a few interesting patterns. For example, the hurricanes that have high activations for the 'Asia-dimension' are all typhoons (and hence happened in Asia), and the hurricanes with high activations on the 'U.S. dimension' all happened in America. However, it is unclear what exactly the relationship is between description words that have very high or very low activations on a particular dimension and the event vectors with extreme scores on that dimension. For example, while for battles there is a clear 'conflicts in the Middle East' dimension, most battles with high values for that dimension do not seem to be related to this. A list of events activated by each dimension is given in the appendix (Table A.6).

[3]Hurricanes sometimes have different names in different countries, yielding descriptions such as 'Hurricane X, also known in country Y as Z ...'

8. Conclusion

This thesis investigated the referential properties of distributional representations of named events. Our work has three main contributions: (i) we proposed methods of creating distributional representations for named events; (ii) we tested what referential information these representations encode; and (iii) we proposed a method for qualitatively analyzing event spaces. Additionally, we designed a theoretical framework for interpreting our findings. In this chapter, we will synthesize our findings and suggest directions for future research.

8.1. Synthesis

Our theoretical framework (see the schema in Figure 2.1 on p. 19) consists of three components: Language (the way that we talk about the world), Meaning (our mental representation of the world), and World (the physical world as we perceive it). In our experiments, we use two types of data: information from Wikipedia infoboxes, which we see as an approximation of World, and textual data (i.e., all of Wikipedia, and in particularly event descriptions), which we see as an approximation of Language. Moreover, we take Meaning to be an abstract representation of World which consists of event entities that are connected to semantic roles through frames.

The theoretical question that this thesis has focused on is how language use (Language) about named events reflects our conceptualization of the world of these events (Meaning), and, indirectly, what the events are like in the actual world. We have approached this question computationally using two steps. First, we created distributional representations for events, which can be thought of as a numerical version of Language. Next, we investigated the link between Language and Meaning by trying to predict referential attributes from these distributional representations.

We described the first step of our computational work in Chapters 4 and 5. In Chapter 4, we modeled language use about named events by creating distributional representations of encyclopedic descriptions of these events. We used two methods for doing this: taking pre-trained GloVe-embeddings (Pennington et al. 2014) for the content words in the event descriptions and then summing these to create a composed representation, and extracting sentence embeddings from BERT (Devlin et al. 2018), a recent neural language model. Because of the newness of BERT, there is no canonical way yet of extracting sentence representations from the pre-trained model; the approach that would turn out to be most fruitful was using hidden layer activations corresponding to individual word tokens and combining these by averaging them.

In Chapter 5, we modeled language use about named events using a different approach: we produced representations of the co-occurrence contexts of the names of the events. We did this in two different ways. First, we computed count-based representations of event names in the Wikipedia corpus. Second, we retrieved vectors for named events from pre-trained distributional models. These models were (i) a Word2Vec (skipgram) model (Mikolov et al. 2013)

with representations of named entities in the Freebase database, and (ii) the Wikipedia2Vec model (Yamada et al. 2018), which provides representations of many named entities that have a Wikipedia page. Wikipedia2Vec embeddings are based not only on how the entity names are used in a text corpus, but also on additional information from the graph structure of Wikipedia; despite being not purely distributional, we included them in our study to see if they would perform better than the other two approaches.

The second step of our work was described in Chapter 6. This chapter defined classification problems based on the referential attributes (which we interpret as describing the MEANING component of these events) that we extracted for the named events in our dataset. For each of our classification problems, we trained machine learning models (a linear SVM and a simple feedforward neural network) and evaluated their performance using a cross-validation setup. Our results were very positive: our models outperformed a simple baseline model (which makes random predictions based on the distribution of classes in the training data) for all attributes. However, performance varied across attributes and representation types. In general, our results for description representations were better than those for event name representations, but this is probably due to the limited size of the event name datasets.

In Chapter 7, we took a first step towards explaining how our prediction models work by performing a qualitative analysis of the space of (GloVe-Summed) event description representations. We did this by applying PCA (Principal Component Analysis), and analyzing which words are 'activated' by (i.e., have high absolute values for) the dimensions in the transformed space. Even though our analysis was limited to only the first five principal components, we found many semantically coherent activation patterns; for example, there are PCA dimensions that seem to encode geographical information (e.g., the 'Asia' dimension for hurricanes) or information about event participants (e.g., the 'emperor' dimension for battles). While much remains unclear about how the event space 'works', our analysis shows that at least certain types of referential information are implicitly encoded in the space.

8.2. Future work

We see several directions for future extensions of our research, particularly related to (i) predicting semantic information from distributional representations and (ii) analyzing event spaces. In this section, we will briefly discuss both of these directions.

First, instead of predicting different sets of referential properties for each of the event types in our dataset, it would be interesting to make predictions for an event space containing different types of events. For example, one could try to predict temporal attributes (e.g. 'year', 'duration'), to test whether time is encoded in the same way for different event types. Other possibilities would be learning to discriminate between events and other entities (e.g., 'Italy', 'Barack Obama') and concepts (e.g. 'country', 'president'), or predicting event kinds from event instances (i.e., learning to relate 'Battle of Waterloo' to 'battle'). For the latter idea, an interesting type of events to consider could be recurring events, such as sport events (e.g. the 'Tokyo Olympic Summer Games' as an instance of 'Olympic Games'), festivals, and elections.

We also see additional possibilities for analyzing event spaces. Our analysis so far (in Chapter 7) only applies to GloVe-Summed description representations, but similar analyses could be

done for other kinds of spaces. We already performed preliminary experiments with count-based spaces, but an analysis of the first few principal components did not yield interpretable results. A possible alternative way of finding (PCA) dimensions that encode referentially-relevant information would be to test which dimensions contribute to classifying referential attributes.[1] Another direction would be to make use of the fact that the dimensions of count-based space are interpretable (i.e., they encode co-occurrence counts with a particular context word), and investigate which context words are relevant for predicting referential properties. A possible method of doing this would be using SVM coefficient weights (cf. Chang and Lin 2008). It would be especially interesting to investigate how much attribute prediction relies on 'superficial' contextual cues (e.g., year numbers for time attributes or country names for place attributes) and to what extent it uses more subtle information.

[1] An algorithm for selecting dimensions for optimizing a distributional space for a particular task already exists (see `https://github.com/akb89/entropix`) and could be adapted to our problem.

A. Supplementary tables and figures

A.1. Prediction accuracy

See Tables A.1 through A.5.

Note on missing values: In the results for count vectors (Table A.3), some scores are missing for binary hurricane attributes. This is because, due to the small size of the available dataset, in some cross-validation splits the training data contained samples from only a single class, meaning that the model could not be trained. We decided to exclude attributes for which this happened at least once.

A.2. Dimension analysis

See Table A.6.

ACCURACY													
	BL	**SVM**						**MLP**					
		GloVe	BERT					GloVe	BERT				
			CLS		Tokens				CLS		Tokens		
		Gs	*Bpp*	*Bps*	*Bm5*	*Bm9*	*Bm12*	*Gs*	*Bpp*	*Bps*	*Bm5*	*Bm9*	*Bm12*
H (b.)													
ArNS	0.75	0.99	0.96	0.94	0.99	0.99	0.99	0.98	0.96	0.93	0.99	0.98	0.99
ArWE	0.51	0.99	0.96	0.94	0.98	0.98	0.98	0.98	0.95	0.93	0.98	0.97	0.98
Ca	0.64	0.89	0.87	0.82	0.91	0.89	0.90	0.88	0.85	0.79	0.91	0.89	0.90
Da	0.50	0.70	0.67	0.65	0.69	0.71	0.71	0.70	0.65	0.64	0.71	0.71	0.71
Du	0.49	0.74	0.65	0.63	0.71	0.72	0.72	0.72	0.63	0.62	0.73	0.72	0.73
Fa	0.50	0.69	0.66	0.63	0.68	0.69	0.71	0.70	0.63	0.62	0.70	0.70	0.71
Pr	0.49	0.80	0.73	0.72	0.83	0.80	0.81	0.81	0.72	0.69	0.83	0.81	0.82
Wi	0.50	0.78	0.71	0.69	0.80	0.78	0.78	0.79	0.71	0.66	0.80	0.78	0.78
Ye	0.50	0.72	0.79	0.73	0.84	0.82	0.84	0.73	0.78	0.70	0.84	0.82	0.85
H (m.)													
Ar	0.40	0.98	0.94	0.89	0.98	0.96	0.97	0.98	0.93	0.90	0.98	0.96	0.97
Ca	0.17	0.44	0.35	0.32	0.47	0.41	0.42	0.43	0.32	0.28	0.47	0.42	0.42
Da	0.25	0.41	0.37	0.36	0.41	0.40	0.41	0.42	0.35	0.37	0.43	0.43	0.41
Du	0.26	0.44	0.36	0.34	0.44	0.43	0.43	0.45	0.34	0.35	0.45	0.44	0.43
Fa	0.25	0.41	0.35	0.35	0.40	0.38	0.38	0.43	0.33	0.32	0.43	0.43	0.42
Pr	0.25	0.54	0.44	0.42	0.54	0.51	0.51	0.55	0.42	0.39	0.54	0.52	0.52
Wi	0.25	0.57	0.47	0.42	0.59	0.55	0.55	0.58	0.45	0.39	0.59	0.56	0.56
Ye	0.26	0.48	0.56	0.47	0.61	0.57	0.62	0.49	0.55	0.44	0.62	0.59	0.62
C (b.)													
Du	0.51	0.65	0.64	0.59	0.67	0.67	0.68	0.66	0.58	0.56	0.68	0.69	0.69
Le	0.52	0.70	0.66	0.62	0.68	0.68	0.69	0.71	0.62	0.60	0.70	0.70	0.70
Ye	0.50	0.73	0.85	0.80	0.85	0.83	0.87	0.74	0.85	0.79	0.86	0.85	0.87
C (m.)													
Du	0.25	0.40	0.37	0.34	0.43	0.41	0.42	0.40	0.34	0.32	0.44	0.43	0.44
Le	0.26	0.42	0.38	0.35	0.41	0.42	0.40	0.43	0.35	0.33	0.43	0.44	0.43
Ye	0.25	0.49	0.66	0.55	0.68	0.62	0.69	0.49	0.65	0.52	0.68	0.63	0.70
B (b.)													
ArNS	0.88	0.97	0.96	0.95	0.98	0.97	0.98	0.97	0.95	0.94	0.98	0.97	0.98
ArWE	0.55	0.92	0.90	0.88	0.93	0.91	0.93	0.92	0.89	0.87	0.95	0.93	0.94
Be	0.58	0.74	0.73	0.72	0.74	0.73	0.75	0.74	0.73	0.72	0.75	0.75	0.74
InF	0.74	0.93	0.91	0.89	0.93	0.92	0.93	0.94	0.91	0.89	0.94	0.93	0.93
InS	0.88	0.96	0.94	0.94	0.96	0.95	0.96	0.96	0.94	0.94	0.96	0.96	0.96
InUS	0.76	0.95	0.94	0.93	0.95	0.95	0.95	0.95	0.93	0.92	0.95	0.95	0.95
StR	0.50	0.57	0.53	0.54	0.56	0.55	0.58	0.58	0.53	0.56	0.59	0.59	0.58
StT	0.51	0.72	0.69	0.67	0.72	0.70	0.75	0.74	0.68	0.64	0.76	0.75	0.75
Ye	0.50	0.88	0.95	0.90	0.96	0.94	0.95	0.88	0.94	0.90	0.96	0.95	0.96
B (m.)													
Ar	0.48	0.90	0.86	0.83	0.92	0.89	0.90	0.91	0.83	0.83	0.93	0.91	0.92
Be	0.31	0.50	0.49	0.48	0.50	0.48	0.49	0.51	0.49	0.47	0.50	0.50	0.52
StR	0.25	0.30	0.27	0.28	0.30	0.28	0.28	0.32	0.26	0.29	0.31	0.29	0.29
StT	0.25	0.42	0.38	0.34	0.41	0.41	0.41	0.44	0.35	0.33	0.45	0.45	0.45
Ye	0.25	0.79	0.90	0.81	0.91	0.88	0.91	0.79	0.88	0.80	0.90	0.89	0.92

Legend. *Events/attributes:* H=Hurricanes, C=Concert tours, B=Battles; (b.)=binary, (m.)=multi-class; Ar=Area, Be=Belligerents, Ca=Category, Da=Damage, Du=Duration, Fa=Fatalities, InF/InS/InUS=Involves France/Spain/US, Le=Legs, Pr=Air pressure, StR/StT = Strength Ratio/Total, Wi=Wind speed, Ye=Year (see Table 3.1). *Embeddings:* Gs=GloVe-Summed, Bpp=Bert-Pooled-Paragraph, Bps=Bert-Pooled-Sentences, Bm{5, 9, 12}=Bert-Mean-{5, 9, 12}. BERT representations are grouped by whether they use pooled representations from the CLS token ('CLS'), or compose the representations from individual work tokens ('Tokens').

Table A.1.: Prediction accuracy results

F1													
	BL	**SVM**						**MLP**					
		GloVe	BERT					GloVe	BERT				
			CLS		Tokens				CLS		Tokens		
		Gs	*Bpp*	*Bps*	*Bm5*	*Bm9*	*Bm12*	*Gs*	*Bpp*	*Bps*	*Bm5*	*Bm9*	*Bm12*
H (b.)													
ArNS	0.49	0.98	0.92	0.87	0.99	0.97	0.97	0.97	0.90	0.85	0.98	0.96	0.97
ArWE	0.51	0.99	0.96	0.94	0.98	0.98	0.98	0.98	0.95	0.93	0.98	0.97	0.98
Ca	0.50	0.84	0.81	0.74	0.88	0.85	0.85	0.84	0.77	0.67	0.87	0.85	0.86
Da	0.50	0.70	0.67	0.65	0.69	0.71	0.71	0.70	0.65	0.64	0.71	0.71	0.71
Du	0.49	0.74	0.65	0.63	0.71	0.72	0.72	0.72	0.63	0.62	0.73	0.72	0.73
Fa	0.50	0.69	0.66	0.62	0.68	0.69	0.71	0.70	0.63	0.62	0.70	0.70	0.71
Pr	0.49	0.80	0.73	0.72	0.83	0.80	0.80	0.81	0.72	0.69	0.82	0.80	0.82
Wi	0.50	0.78	0.71	0.69	0.80	0.78	0.78	0.79	0.71	0.66	0.80	0.78	0.78
Ye	0.50	0.72	0.79	0.73	0.84	0.82	0.84	0.73	0.77	0.70	0.83	0.82	0.85
H (m.)													
Ar	0.24	0.84	0.70	0.66	0.79	0.75	0.79	0.81	0.67	0.65	0.77	0.74	0.77
Ca	0.14	0.39	0.31	0.29	0.45	0.39	0.41	0.37	0.21	0.20	0.40	0.33	0.38
Da	0.25	0.41	0.37	0.36	0.41	0.40	0.41	0.41	0.34	0.36	0.42	0.42	0.41
Du	0.26	0.43	0.35	0.34	0.44	0.42	0.42	0.44	0.33	0.35	0.44	0.43	0.43
Fa	0.25	0.40	0.35	0.34	0.40	0.38	0.38	0.40	0.32	0.32	0.41	0.42	0.41
Pr	0.25	0.54	0.43	0.41	0.53	0.51	0.50	0.53	0.41	0.38	0.53	0.51	0.52
Wi	0.25	0.57	0.47	0.42	0.59	0.55	0.55	0.57	0.45	0.39	0.59	0.56	0.56
Ye	0.26	0.48	0.56	0.47	0.62	0.58	0.63	0.49	0.55	0.44	0.62	0.60	0.63
C (b.)													
Du	0.51	0.65	0.64	0.59	0.67	0.67	0.68	0.66	0.58	0.56	0.68	0.69	0.69
Le	0.50	0.67	0.63	0.56	0.66	0.66	0.67	0.70	0.55	0.53	0.68	0.68	0.69
Ye	0.50	0.73	0.85	0.80	0.85	0.83	0.87	0.74	0.85	0.79	0.86	0.85	0.87
C (m.)													
Du	0.25	0.39	0.37	0.34	0.43	0.40	0.41	0.40	0.33	0.32	0.44	0.43	0.44
Le	0.25	0.40	0.35	0.30	0.40	0.41	0.39	0.42	0.28	0.27	0.42	0.43	0.42
Ye	0.25	0.49	0.67	0.55	0.69	0.62	0.69	0.48	0.65	0.52	0.68	0.63	0.71
B (b.)													
ArNS	0.50	0.83	0.78	0.64	0.90	0.86	0.89	0.87	0.64	0.62	0.91	0.86	0.90
ArWE	0.50	0.91	0.88	0.86	0.93	0.91	0.92	0.92	0.88	0.86	0.94	0.92	0.93
Be	0.50	0.61	0.60	0.54	0.67	0.65	0.66	0.65	0.60	0.58	0.67	0.66	0.66
InF	0.50	0.85	0.80	0.72	0.87	0.84	0.85	0.87	0.79	0.73	0.87	0.85	0.85
InS	0.50	0.76	0.63	0.57	0.82	0.78	0.76	0.81	0.57	0.56	0.83	0.79	0.81
InUS	0.50	0.89	0.86	0.84	0.90	0.89	0.90	0.89	0.86	0.84	0.91	0.90	0.90
StR	0.50	0.57	0.53	0.54	0.56	0.55	0.58	0.58	0.53	0.56	0.59	0.59	0.58
StT	0.51	0.72	0.69	0.67	0.72	0.70	0.75	0.74	0.68	0.64	0.76	0.75	0.75
Ye	0.50	0.88	0.95	0.90	0.96	0.94	0.95	0.88	0.94	0.90	0.96	0.95	0.96
B (m.)													
Ar	0.25	0.80	0.72	0.57	0.87	0.81	0.86	0.84	0.52	0.56	0.87	0.82	0.85
Be	0.25	0.34	0.34	0.29	0.40	0.38	0.39	0.37	0.34	0.32	0.39	0.39	0.41
StR	0.25	0.30	0.27	0.28	0.29	0.28	0.28	0.32	0.24*	0.29	0.31	0.28	0.29
StT	0.25	0.42	0.38	0.34	0.41	0.41	0.41	0.43	0.35	0.32	0.45	0.45	0.45
Ye	0.25	0.79	0.90	0.81	0.91	0.88	0.91	0.79	0.88	0.80	0.90	0.89	0.92

Legend: see Table A.1.

Table A.2.: Prediction results (F1 scores)

	ACC						F1						N
	BL	SVM					BL	SVM					
		Raw2	Raw5	PPMI2	PPMI5	Gs		Co2	Co5	PPMI2	PPMI5	Gs	
H (b.)													
ArNS	-	-	-	-	-	-	-	-	-	-	-	-	49
ArWE	0.77	0.85	0.83	0.85	0.86	0.89	0.51	0.67	0.57	0.46*	0.47*	0.75	48
Ca	-	-	-	-	-	-	-	-	-	-	-	-	50
Da	-	-	-	-	-	-	-	-	-	-	-	-	51
Du	0.65	0.77	0.76	0.78	0.80	0.79	0.48	0.62	0.53	0.44*	0.45*	0.59	51
Fa	-	-	-	-	-	-	-	-	-	-	-	-	51
Pr	0.81	0.85	0.83	0.88	0.88	0.85	0.53	0.54	0.48*	0.47*	0.47*	0.53*	51
Wi	0.82	0.85	0.85	0.88	0.88	0.85	0.53	0.59	0.50*	0.47*	0.47*	0.51*	51
Ye	0.50	0.75	0.78	0.48*	0.70	0.61	0.50	0.74	0.78	0.41*	0.70	0.60	51
H (m.)													
Ar	0.78	0.84	0.86	0.87	0.87	0.86	0.34	0.38	0.40	0.31*	0.31*	0.41	47
Ca	0.31	0.37	0.41	0.38	0.39	0.35	0.18	0.20	0.20	0.14*	0.16*	0.18*	50
Da	0.82	0.88	0.88	0.90	0.90	0.88	0.31	0.34	0.35	0.32	0.32	0.31*	51
Du	0.41	0.45	0.52	0.59	0.59	0.48	0.24	0.24*	0.26	0.19*	0.19*	0.22*	51
Fa	0.50	0.63	0.69	0.72	0.71	0.62	0.21	0.34	0.36	0.27	0.22	0.29	51
Pr	0.40	0.39*	0.42	0.43	0.46	0.48	0.23	0.22*	0.24	0.16*	0.20*	0.27	51
Wi	0.45	0.45*	0.44*	0.56	0.58	0.54	0.23	0.27	0.19*	0.18*	0.19*	0.30	51
Ye	0.26	0.46	0.49	0.20*	0.39	0.23*	0.25	0.47	0.49	0.11*	0.32	0.22*	51
C (b.)													
Du	0.52	0.57	0.65	0.51*	0.65	0.60	0.49	0.52	0.60	0.41*	0.50	0.52	74
Le	0.49	0.57	0.59	0.48*	0.58	0.62	0.49	0.55	0.58	0.39*	0.56	0.62	69
Ye	0.51	0.58	0.75	0.54	0.75	0.76	0.50	0.58	0.74	0.36*	0.74	0.75	76
C (m.)													
Du	0.28	0.33	0.34	0.39	0.38	0.29	0.25	0.31	0.33	0.21*	0.25*	0.24*	74
Le	0.27	0.35	0.33	0.33	0.31	0.26*	0.25	0.32	0.30	0.20*	0.19*	0.23*	69
Ye	0.25	0.34	0.44	0.28	0.42	0.42	0.23	0.32	0.39	0.14*	0.31	0.37	76
B (b.)													
ArNS	0.90	0.94	0.94	0.94	0.94	0.95	0.51	0.50*	0.65	0.49*	0.49*	0.70	*369*
ArWE	0.53	0.78	0.81	0.61	0.91	0.90	0.51	0.76	0.81	0.42*	0.90	0.90	*369*
Be	0.49	0.63	0.66	0.55	0.71	0.66	0.49	0.63	0.66	0.41*	0.71	0.66	*420*
InF	0.69	0.81	0.89	0.82	0.89	0.89	0.50	0.61	0.82	0.52	0.77	0.80	*420*
InS	0.91	0.95	0.95	0.96	0.96	0.95	0.50	0.49*	0.66	0.60	0.62	0.54	*420*
InUS	0.66	0.83	0.89	0.79	0.91	0.91	0.50	0.71	0.83	0.50*	0.85	0.86	*420*
StR	0.55	0.61	0.58	0.70	0.67	0.60	0.49	0.45*	0.49*	0.46*	0.42*	0.47*	*102*
StT	0.62	0.72	0.72	0.75	0.75	0.76	0.50	0.55	0.59	0.43*	0.44*	0.57	*102*
Ye	0.51	0.82	0.90	0.59	0.96	0.88	0.49	0.81	0.90	0.42*	0.96	0.88	*428*
B (m.)													
Ar	0.47	0.70	0.78	0.56	0.86	0.88	0.25	0.44	0.47	0.19*	0.44	0.67	*369*
Be	0.27	0.36	0.39	0.29	0.49	0.43	0.25	0.33	0.35	0.16*	0.39	0.37	*420*
StR	0.29	0.27*	0.34	0.42	0.41	0.32	0.25	0.23*	0.27	0.15*	0.20*	0.21*	*102*
StT	0.35	0.46	0.52	0.50	0.53	0.60	0.24	0.33	0.36	0.17*	0.25	0.39	*102*
Ye	0.27	0.62	0.78	0.32	0.91	0.78	0.26	0.59	0.76	0.18*	0.89	0.76	*428*

Legend. '-' indicate missing scores. *Embeddings:* Raw2/Raw5=raw counts, window size 2 or 5; PPMI2/PPMI5=PPMI-weighted counts, window size 2 or 5; Gs=Summed GloVe vectors. N=number of events with a particular attribute for which count vectors are available.

Table A.3.: Results for count-based event name vectors

	ACC					F1					N
	BL	SVM		MLP		BL	SVM		MLP		
		Fr	Gs	Fr	Gs		Fr	Gs	Fr	Gs	
Hurricanes											
(binary)											
Area (NS)	0.66	0.98	0.97	0.99	0.88	0.49	0.96	0.95	0.98	0.83	153
Area (WE)	0.52	0.97	0.95	0.95	0.94	0.50	0.97	0.95	0.94	0.94	148
Category	0.84	0.91	0.92	0.91	0.89	0.50	0.54	0.58	0.53	0.58	155
Damage	0.68	0.85	0.75	0.87	0.75	0.52	0.74	0.52*	0.79	0.62	143
Duration	0.55	0.66	0.72	0.67	0.69	0.51	0.60	0.67	0.62	0.66	157
Fatalaities	0.57	0.82	0.79	0.83	0.75	0.49	0.76	0.72	0.78	0.68	157
Pressure	0.67	0.84	0.85	0.84	0.80	0.50	0.68	0.71	0.68	0.68	157
Wind speed	0.65	0.78	0.79	0.78	0.78	0.50	0.62	0.64	0.61	0.65	157
Year	0.50	0.96	0.71	0.94	0.70	0.50	0.96	0.71	0.94	0.70	157
(multi)											
Area	0.41	0.95	0.92	0.95	0.87	0.24	0.84	0.67	0.81	0.67	147
Category	0.25	0.32	0.35	0.33	0.32	0.15	0.17	0.23	0.15*	0.22	155
Damage	0.38	0.57	0.48	0.61	0.47	0.24	0.35	0.25	0.37	0.26	143
Duration	0.30	0.38	0.40	0.38	0.37	0.27	0.32	0.35	0.31	0.33	157
Fatalaties	0.30	0.55	0.41	0.53	0.45	0.24	0.43	0.33	0.42	0.35	157
Pressure	0.35	0.49	0.48	0.48	0.45	0.24	0.37	0.35	0.34	0.38	157
Wind speed	0.32	0.44	0.49	0.44	0.47	0.24	0.32	0.38	0.32	0.37	157
Year	0.34	0.84	0.53	0.83	0.52	0.25	0.72	0.36	0.72	0.34	157
Concert tours											
(binary)											
Duration	0.56	0.62	0.62	0.63	0.61	0.50	0.53	0.46*	0.55	0.52	146
Legs	0.50	0.61	0.66	0.62	0.62	0.49	0.60	0.64	0.60	0.61	139
Year	0.54	0.88	0.64	0.88	0.62	0.51	0.87	0.57	0.87	0.59	149
(multi)											
Duration	0.29	0.33	0.29*	0.36	0.28*	0.25	0.25*	0.22*	0.26	0.22*	146
Legs	0.25	0.38	0.37	0.40	0.37	0.23	0.31	0.31	0.33	0.32	139
Year	0.33	0.74	0.40	0.76	0.43	0.32	0.75	0.38	0.76	0.42	149
Battles											
(binary)											
Area (NS)	0.84	0.95	0.94	0.96	0.93	0.49	0.80	0.71	0.81	0.72	*501*
Area (WE)	0.50	0.92	0.92	0.92	0.92	0.50	0.92	0.92	0.92	0.92	*501*
Belligerents	0.51	0.59	0.67	0.66	0.69	0.49	0.57	0.65	0.63	0.68	*578*
Involved France	0.77	0.90	0.90	0.91	0.90	0.50	0.75	0.74	0.77	0.76	*578*
Involved Spain	0.92	0.96	0.96	0.96	0.96	0.50	0.54	0.58	0.54	0.66	*578*
Involved US	0.54	0.90	0.89	0.91	0.88	0.49	0.89	0.87	0.91	0.87	*578*
Strength (Ratio)	0.50	0.46*	0.48*	0.49*	0.53	0.50	0.46*	0.48*	0.48*	0.53	*151*
Strength (Total)	0.51	0.69	0.76	0.71	0.77	0.50	0.69	0.75	0.71	0.76	*151*
Year	0.51	0.89	0.87	0.90	0.87	0.49	0.89	0.86	0.89	0.87	*586*
(multi)											
Area	0.43	0.88	0.87	0.88	0.88	0.25	0.70	0.65	0.67	0.72	*501*
Belligerents	0.27	0.38	0.44	0.40	0.46	0.25	0.34	0.38	0.35	0.40	*578*
Strength (Ratio)	0.27	0.24*	0.23*	0.24*	0.25*	0.26	0.22*	0.20*	0.20*	0.24*	*151*
Strength (Total)	0.24	0.42	0.46	0.43	0.49	0.23	0.37	0.43	0.38	0.46	*151*
Year	0.26	0.82	0.81	0.81	0.80	0.25	0.80	0.80	0.79	0.79	*586*

Legend. *Embeddings:* Fr=Freebase word2vec vectors, Gs=Summed GloVe vectors. N=number of events with a particular attribute for which Freebase vectors are available.

Table A.4.: Results for word2vec Freebase vectors

	ACC					F1					N
	BL	SVM		MLP		BL	SVM		MLP		
		Wi	Gs	Wi	Gs		Wi	Gs	Wi	Gs	
Hurricanes											
(binary)											
Area (NS)	0.79	1.00	0.99	1.00	0.98	0.50	0.99	0.98	0.99	0.95	792
Area (WE)	0.52	0.99	0.99	1.00	0.98	0.50	0.99	0.99	1.00	0.98	761
Category	0.68	0.82	0.90	0.84	0.88	0.51	0.69	0.84	0.75	0.81	793
Damage	0.50	0.78	0.68	0.78	0.68	0.50	0.78	0.68	0.78	0.68	730
Duration	0.50	0.71	0.73	0.73	0.73	0.50	0.71	0.73	0.73	0.73	817
Fatalities	0.50	0.74	0.70	0.76	0.70	0.49	0.74	0.70	0.76	0.70	803
Pressure	0.51	0.80	0.79	0.82	0.80	0.51	0.80	0.79	0.82	0.80	802
Wind speed	0.50	0.76	0.80	0.78	0.80	0.50	0.76	0.79	0.78	0.80	813
Year	0.51	0.86	0.69	0.87	0.69	0.51	0.86	0.69	0.87	0.69	817
(multi)											
Area	0.43	0.99	0.98	0.99	0.97	0.24	0.90	0.79	0.92	0.76	756
Category	0.17	0.36	0.43	0.39	0.42	0.13	0.26	0.36	0.28	0.33	793
Damage	0.26	0.51	0.39	0.53	0.40	0.25	0.50	0.39	0.52	0.38	730
Duration	0.25	0.40	0.44	0.42	0.45	0.25	0.39	0.43	0.41	0.44	817
Fatalaties	0.25	0.46	0.41	0.48	0.43	0.25	0.46	0.40	0.47	0.42	803
Pressure	0.26	0.49	0.54	0.52	0.56	0.26	0.48	0.53	0.50	0.54	802
Wind speed	0.26	0.49	0.58	0.51	0.57	0.25	0.49	0.58	0.51	0.57	813
Year	0.25	0.70	0.43	0.73	0.45	0.25	0.71	0.43	0.74	0.45	817
Concert tours											
(binary)											
Duration	0.51	0.65	0.63	0.67	0.65	0.51	0.65	0.62	0.67	0.65	978
Legs	0.50	0.66	0.67	0.68	0.67	0.50	0.65	0.67	0.68	0.67	843
Year	0.50	0.90	0.74	0.91	0.74	0.50	0.90	0.74	0.91	0.74	988
(multi)											
Duration	0.25	0.40	0.36	0.44	0.37	0.25	0.40	0.36	0.44	0.38	978
Legs	0.26	0.41	0.37	0.44	0.39	0.25	0.40	0.35	0.43	0.39	843
Year	0.26	0.76	0.47	0.79	0.48	0.25	0.75	0.46	0.78	0.48	988
Battles											
(binary)											
Area (NS)	0.88	0.99	0.97	0.99	0.97	0.50	0.96	0.84	0.97	0.87	*2775*
Area (WE)	0.54	0.96	0.93	0.97	0.93	0.50	0.95	0.92	0.96	0.92	*2775*
Belligerents	0.56	0.74	0.72	0.75	0.73	0.50	0.66	0.62	0.69	0.65	*4405*
Involved France	0.73	0.94	0.93	0.95	0.93	0.50	0.89	0.86	0.90	0.88	*4405*
Involved Spain	0.87	0.96	0.95	0.97	0.96	0.50	0.83	0.75	0.87	0.81	*4405*
Involved US	0.73	0.96	0.95	0.97	0.94	0.50	0.93	0.90	0.94	0.89	*4405*
Strength (Ratio)	0.50	0.59	0.58	0.60	0.59	0.50	0.59	0.57	0.59	0.58	*989*
Strength (Total)	0.51	0.76	0.73	0.79	0.74	0.51	0.76	0.73	0.78	0.73	*989*
Year	0.50	0.98	0.88	0.99	0.88	0.50	0.98	0.88	0.99	0.88	*4473*
(multi)											
Area	0.48	0.95	0.91	0.96	0.91	0.25	0.95	0.82	0.95	0.84	*2775*
Belligerents	0.30	0.50	0.48	0.52	0.48	0.25	0.39	0.35	0.43	0.38	*4405*
Strength (Ratio)	0.25	0.28	0.30	0.31	0.30	0.25	0.28	0.29	0.31	0.30	*989*
Strength (Total)	0.25	0.49	0.43	0.51	0.45	0.25	0.48	0.42	0.50	0.44	*989*
Year	0.25	0.98	0.78	0.98	0.79	0.25	0.98	0.78	0.98	0.78	*4473*

Legend. *Embeddings:* Wi= Wikipedia2Vec vectors, Gs=GloVe-Summed vectors. N=number of events with a particular attribute for which count vectors are available.

Table A.5.: Prediction results for Wikipedia2Vec vectors

	Dim	Var	0	1	2
Hurricanes	0	0.70	/wiki/Hurricane_Maria	/wiki/Typhoon_Utor	/wiki/Tropical_Storm_Sonca_(2017)
	1	0.03	/wiki/Cyclone_Rusty	/wiki/Cyclone_Amara	/wiki/Tropical_Storm_Etau_(2009)
	2	0.03	/wiki/Hurricane_Hanna	/wiki/Tropical_Storm_Claudette_(1979)	/wiki/Hurricane_Agnes
	3	0.02	/wiki/Typhoon_Ken_(1982)	/wiki/Hurricane_Keith	/wiki/Typhoon_Bess_(1974)
	4	0.02	/wiki/Typhoon_Morakot	/wiki/Hurricane_Blanca_(2015)	/wiki/Cyclone_Nargis
Concert tours	0	0.76	/wiki/Theatre_of_Madness_Tour	/wiki/The_Ozzman_Cometh_Tour	/wiki/The_Ultimate_Sin_Tour
	1	0.03	/wiki/The_Sun_Comes_Out_World_Tour	/wiki/The_Ride_Tour	/wiki/Blue_Moon_World_Tour
	2	0.01	/wiki/Couldn%27t_Stand_the_Weather_Tour	/wiki/Number_Ones,_Up_Close_and_Personal	/wiki/Loud_Tour
	3	0.01	/wiki/TP.3_Reloaded	/wiki/Born_This_Way_Ball	/wiki/Unfinished_Business_(Jay-Z_and_R._Kelly_album)
	4	0.01	/wiki/Barefoot_at_the_Symphony_Tour	/wiki/Vespertine_World_Tour	/wiki/The_Onyx_Hotel_Tour
Battles	0	0.68	/wiki/Battle_of_Khresili	/wiki/Battle_of_Wabho	/wiki/Battle_of_the_Hotels
	1	0.02	/wiki/Battle_of_Steenbergen_(1583)	/wiki/Battle_of_Grand_Pr%C3%A9	/wiki/Battle_of_Annan
	2	0.02	/wiki/Battle_of_Fairfax_Court_House_(June_1863)	/wiki/Battle_of_Sattelberg	/wiki/First_Battle_of_Charleston_Harbor
	3	0.01	/wiki/First_Battle_of_Passchendaele	/wiki/Battle_of_Nicopolis_(1798)	/wiki/Battles_of_Kawanakajima
	4	0.01	/wiki/Battle_of_Ponta_Delgada	/wiki/First_Battle_of_Ypres	/wiki/Battle_of_the_Basque_Roads

	Dim	Var	-3	-2	-1
Hurricanes	0	0.70	/wiki/Hurricane_Bill_(2009)	/wiki/Typhoon_Kinna_(1991)	/wiki/Cyclone_Alan
	1	0.03	/wiki/1900_Galveston_hurricane	/wiki/Tropical_Storm_Fay_(2008)	/wiki/Hurricane_Jeanne
	2	0.03	/wiki/Typhoon_Kinna_(1991)	/wiki/Typhoon_Morakot	/wiki/Cyclone_Nargis
	3	0.02	/wiki/Cyclone_Vance	/wiki/Cyclone_Gamede	/wiki/1949_Florida_hurricane
	4	0.02	/wiki/Typhoon_Vicente	/wiki/Typhoon_Nabi	/wiki/Typhoon_Rammasun
Concert tours	0	0.76	/wiki/Red_Hot_Chili_Peppers_1983_Tour	/wiki/Speak_of_the_Devil_Tour	/wiki/A_Big_Night_in_with_Darren_Hayes_Tour
	1	0.03	/wiki/Volta_Tour	/wiki/No_Prayer_on_the_Road	/wiki/Red_Hot_Chili_Peppers_1983_Tour
	2	0.01	/wiki/R40_Live_Tour	/wiki/Damage,_Inc._Tour	/wiki/Speak_of_the_Devil_Tour
	3	0.01	/wiki/Don%27t_Believe_the_Truth_Tour	/wiki/Be_Here_Now_Tour	/wiki/The_Ride_Tour
	4	0.01	/wiki/The_Uplift_Mofo_Party_Tour	/wiki/Black_%26_Blue_Tour	/wiki/Merry_Mayhem_Tour
Battles	0	0.68	/wiki/Battle_of_Monmouth	/wiki/Battle_of_Ngh%C4%A9a_L%E1%BB%99_(1951)	/wiki/Second_Battle_of_Kom%C3%A1rom_(1849)
	1	0.02	/wiki/Battle_of_Long_Kh%C3%A1nh	/wiki/Battle_of_Coral%E2%80%93Balmoral	/wiki/Battle_of_Zahleh
	2	0.02	/wiki/Battle_of_the_Colline_Gate_(82_BC)	/wiki/Battle_of_the_Camel	/wiki/Battle_of_Gallipoli_(1312)
	3	0.01	/wiki/Second_Battle_of_Auburn	/wiki/Battle_of_Lookout_Mountain	/wiki/Battle_of_Monck%27s_Corner
	4	0.01	/wiki/Battle_of_Alamance	/wiki/Battle_of_Langside	/wiki/Battle_of_Chancellorsville

Note: Events are represented here using (partial) Wikipedia URLs. The full URL for each event can be obtained by prefixing `https://en.wikipedia.org`.

Table A.6.: Event vectors activated by PCA dimensions

Bibliography

Abzianidze, Lasha, Johannes Bjerva, Kiliang Evang, Hessel Haagsma, Rik van Noord, Pierre Ludmann, Duc-Duy Nguyen, and Johan Bos (2017). "The Parallel Meaning Bank: Towards a Multilingual Corpus of Translations Annotated with Compositional Meaning Representations". In: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 242–247.

Baker, Collin F., Charles J. Fillmore, and John B. Lowe (1998). "The Berkeley FrameNet Project". In: *Proceedings of the 17th International Conference on Computational Linguistics - Volume 1*. COLING '98. Montreal, Quebec, Canada, pp. 86–90.

Baroni, Marco (2013). "Composition in Distributional Semantics". In: *Language and Linguistics Compass* 7 (10), pp. 511–522.

Baroni, Marco (2019). "Linguistic generalization and compositionality in modern artificial neural networks". In: *CoRR* abs/1904.00157. arXiv: 1904.00157. URL: http://arxiv.org/abs/1904.00157.

Baroni, Marco and Alessandro Lenci (2010). "Distributional Memory: A General Framework for Corpus-Based Semantics". In: *Computational Linguistics* 36 (4), pp. 673–721.

Baroni, Marco and Robereto Zamparelli (2010). "Nouns are vectors, adjectives are matrices: Representing adjective-noun constructions in semantic space". In: *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pp. 1183–1193.

Baroni, Marco, Raffaella Bernardi, Ngoc-Quynh Do, and Chung-chieh Shan (2012). "Entailment above the word level in distributional semantics". In: *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 23–32.

Baroni, Marco, Georgiana Dinu, and Germán Kruszewski (2014a). "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)*. Vol. 1, pp. 238–247.

Baroni, Marco, Raffaella Bernardi, and Roberto Zamparelli (2014b). "Frege in Space: A Program for Compositional Distributional Semantics". In: *Linguistic Issues in Language Technology* 9, pp. 241–346.

Bengio, Yoshua, Réjean Ducharme, Pascal Vincent, and Christian Jauvin (2003). "A Neural Probabilistic Language Model". In: *Journal of Machine Learning Research* 3, pp. 1137–1155.

Bennett, Jonathan (2002). "What Events Are". In: *The Blackwell Guide to Metaphysics*. Ed. by Richard M. Gale. Online version, https://www.earlymoderntexts.com/assets/jfb/events.pdf. Blackwell, pp. 43–65.

Bird, Steven, Edward Loper, and Ewan Klein (2009). *Natural Language Processing with Python*. O'Reilly Media Inc.

Boleda, Gemma and Aurélie Herbelot (2017). "Formal Distributional Semantics: Introduction to the Special Issue". In: *Computational Linguistics* 42 (4), pp. 619–635.

Bollacker, Kurt, Colin Evans, Pavreen Paritosh, Tim Sturge, and Jamie Taylor (2008). "Freebase: A Collaboratively Created Graph Database For Structuring Human Knowledge". In: *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data*.

Bos, Johan and Malvina Nissim (2008). "Combining Discourse Representation Theory with FrameNet". In: *Frames, corpora and knowledge representation*. Ed. by R. Rossini Favretti. Bologna: Bononia University Press.

Bullinaria, John A. and Joseph P. Levy (2012). *Extracting semantic representations from word co-occurrence statistics: stop-lists, stemming, and SVD*.

Casati, Roberto and Achille Varzi (2015). "Events". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Winter 2015. Metaphysics Research Lab, Stanford University.

Chang, Yin-Weng and Chih-Jen Lin (2008). "Feature Ranking Using Linear SVM". In: *JMLR: Workshop and Conference Proceedings* 3: WCCI2008 workshop on causality, pp. 53–64.

Chomsky, Noam (2013). "Notes on Denotation and Denoting". In: *From Grammar To Meaning: The Spontaneous Logicality of Language*. Ed. by Carlo Cechetto and Ivano Caponigro. Cambridge: Cambridge University Press, pp. 38–45.

Davidson, Donald (1967). "The logical form of action sentences". In: *Essays in honor of Carl G. Hempel*. Springer, pp. 216–234.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (2018). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *CoRR* abs/1810.04805. arXiv: 1810.04805. URL: http://arxiv.org/abs/1810.04805.

Erk, Katrin (2013). "Towards a semantics for distributional representations". In: *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013)–Long Papers*, pp. 95–106.

Erk, Katrin (2016). "What do you know about an alligator by the company that it keeps?" In: *Semantics and Pragmatics* 9, pp. 1–63.

Erk, Katrin and Sebastian Padó (2008). "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.

Erk, Katrin and Sebastian Padó (2010). "Exemplar-Based Models for Word Meaning In Context". In: *Proceedings of the ACL 2010 Conference Short Papers*, pp. 92–97.

Fillmore, Charles J. (1977). "Scenes-and-frames semantics". In: *Linguistic Structures Processing*. Ed. by Antonio Zampolli. North Holland Publishing, pp. 55–88.

Făgărăşan, Luana, Eva Maria Vecchi, and Stephen Clark (2015). "From distributional semantics to feature norms: grounding semantic models in human perceptual data". In: *Proceedings of the 11th International Conference on Computational Semantics*, pp. 52–57.

Gulordava, Kristina, Piotr Bojanowski, Edouard Grave, Tal Linzen, and Marco Baroni (2018). "Colorless Green Recurrent Networks Dream Hierarchically". In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 1195–1205.

Gupta, Abhijeet, Gemma Boleda, Marco Baroni, and Sebastian Padó (2015). "Distributional vectors encode referential attributes". In: *Proceedings of the 2015 Conference on Emperical Methods in Natural Language Processing*, pp. 12–21.

Harris, Zelig (1954). "Distributional structure". In: *Word* 10 (2-3), pp. 146–162.

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

Herbelot, Aurélie and Eva Maria Vecchi (2015). "Building a shared world: Mapping distributional to model-theoretic semantic spaces". In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 22–32.

Hermann, Karl Moritz, Edward Grefenstette, and Phil Blunsom (2013). ""Not not bad" is not "bad": A distributional account of negation". In: *Proceedings of the Workshop on Continuous Vector Space Models and their Compositionality*, pp. 74–82.

Johns, Brendan T. and Michael N. Jones (2012). "Perceptual inference through global lexical similarity". In: *Topics in Cognitive Science* 4 (1), pp. 103–120.

Kim, Jaegwon (1966). "On the Psycho-Physical Identity Theory". In: *American Philosophical Quarterly* 3 (3), pp. 227–235.

Kingma, Diederik P. and Jimmy Lei Ba (2014). "Adam: A Method for Stochastic Optimization". In: *CoRR* abs/1412.6980. ArXiv preprint, http://arxiv.org/abs/1412.6980. arXiv: 1412.6980.

Kuzmenko, Elizaveta and Aurélie Herbelot (2019). "Distributional semantics in the real world: building word vector representations from a truth-theoretic model". In: *Proceedings of the 19th Conference on Computational Semantics - Short Papers*, pp. 16–23.

Landauer, Thomas K. and Susan T. Dumais (1997). "A Solution to Plato's Problem: The Latent Semantic Analysis Theory of Acquisition, Induction and Representation of Knowledge". In: *Psychological Review* 104 (2).

Landauer, Thomas K., D. Laham, Bob Rehder, and M.E. Schreiner (1997). "How well can passage meaning be derived without using word order: A comparison of Latent Semantic Analysis and humans". In: *Proceedings of the 19th Annual Conference of the Cognitive Science Society*, pp. 412–417.

Lazaridou, Angeliki, Marco Marelli, and Marco Baroni (2017). "Multimodal Word Meaning Induction From Minimal Exposure to Natural Text". In: *Cognitive Science* 41, pp. 677–705.

Lenci, Alessandro (2008). "Distributional semantics in linguistic and cognitive research". In: *Italian Journal of Linguistics* 20 (1), pp. 1–31.

Liu, Nelson F., Matt Gardner, Yonatan Belinkov, Matthew Peters, and Noah A. Smith (2019). "Linguistic Knowledge and Transferability of Contextual Representations". In: *CoRR* abs/1903.08855. arXiv: 1903.08855. URL: http://arxiv.org/abs/1903.08855.

McRae, Ken, George S. Cree, Mark S. Seidenberg, and Chris McNorgan (2005). "Semantic feature production norms for a large set of living and nonliving things". In: *Behavior research methods* 37 (4), pp. 547–559.

Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean (2013). "Efficient Estimation of Word Representations in Vector Space". In: *CoRR* abs/1301.3781. http://arxiv.org/abs/1301.3781. arXiv: 1301.3781.

Miller, George A. (1995). "WordNet: A Lexical Database for English". In: *Communications of the ACM* 38 (11).

Mitchell, Jeff and Mirella Lapata (2008). "Vector-based Models of Semantic Composition". In: *Proceedings of ACL-08: HLT*, pp. 236–244.

Moltmann, Friederike (2018). "Natural Language and its Ontology". In: *Metaphysics and Cognitive Science*. Ed. by A. Goldman and B. McLaughlin. Online postprint, http://friederike-moltmann.com/uploads/Language%20and%20Ontology-OUP.pdf.

Montague, Richard (1973). "The Proper Treatment of Quantification in Ordinary English". In: *Approaches to Natural Language*. Ed. by Patrick Suppes, Julius Moravcsik, and Jaakko Hintikka. Boston/Dordrecht: Reidel, pp. 221–242.

Padó, Sebastian and Mirella Lapata (2007). "Dependency-Based Construction of Semantic Space Models". In: *Computational Linguistics* 33 (2), pp. 161–199.

Parsons, Terence (1980). "Modifiers and Quantifiers in Natural Language". In: *Canadian Journal of Philosophy* 10, pp. 29–60.

Paszke, Adam, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer (2017). "Automatic differentiation in PyTorch". In: *31st Conference on Neural Information Processing Systems (NIPS 2017)*.

Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay (2011). "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12, pp. 2825–2830.

Pennington, Jeffrey, Richard Socher, and Christopher D. Manning (2014). "GloVe: Global Vectors for Word Representation". In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543.

Peters, Matthew, Sebastian Ruder, and Noah A. Smith (2019). "To Tune or Not to Tune? Adapting Pretrained Representations to Diverse Tasks". In: *CoRR* abs/1903.05987. arXiv: 1903.05987. URL: http://arxiv.org/abs/1903.05987.

Peters, Matthew E., Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer (2018). "Deep contextualized word representations". In: *CoRR* abs/1802.05365. arXiv: 1802.05365. URL: http://arxiv.org/abs/1802.05365.

Rajpurkar, Pranav, Jian Zhang, Konstantin Lopyrev, and Percy Liang (2016). "SQuAD: 100, 000+ Questions for Machine Comprehension of Text". In: *CoRR* abs/1606.05250. arXiv: 1606.05250. URL: http://arxiv.org/abs/1606.05250.

Shin, Jamin, Andrea Madotto, and Pascale Fung (2018). "Interpreting Word Embeddings with Eigenvector Analysis". In: *32nd Conference on Neural Information Processing Systems (NIPS 2018)*.

Srivastava, Nitish, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov (2014). "Dropout: A Simple Way to Prevent Neural Networks from Overfitting". In: *J. Mach. Learn. Res.* 15.1, pp. 1929–1958. ISSN: 1532-4435. URL: http://dl.acm.org/citation.cfm?id=2627435.2670313.

Strubell, Emma, Ananya Ganesh, and Andrew McCallum (2019). "Energy and Policy Considerations for Deep Learning in NLP". In: *CoRR* abs/1906.02243. arXiv: 1906.02243. URL: http://arxiv.org/abs/1906.02243.

Thater, Stefan, Hagen Fürstenau, and Manfred Pinkal (2010). "Contextualizing Semantic Representations Using Syntactically Enriched Vector Models". In: *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pp. 948–957.

Turney, Peter D. and Patrick Pantel (2010). "From Frequency to Meaning: Vector Space Models of Semantics". In: *Journal of Artificial Intelligence Research* 37, pp. 141–188.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin (2017). "Attention is All you Need". In: *Advances in Neural Information Processing Systems 30*, pp. 5998–6008. URL: http://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf.

Wang, Alex, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman (2018). "GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding". In: *CoRR* abs/1804.07461. arXiv: 1804.07461. URL: http://arxiv.org/abs/1804.07461.

Yamada, Ikuya, Hiroyuki Shindo, Hideaki Takeda, and Yoshiyasu Takefuji (2016). "Joint Learning of the Embeddings of Words and Entities for Named Entity Disambiguation". In: *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Processing (CoNLL)*, pp. 250–259.

Yamada, Ikuya, Akari Asai, Hiroyuki Shindo, Hideaki Takeda, and Yoshiyasu Takefuji (2018). "Wikipedia2Vec: An Optimized Tool for Learning Embeddings of Words and Entities from Wikipedia". In: *CoRR* abs/1812.06280. arXiv: 1812.06280. URL: http://arxiv.org/abs/1812.06280.