Natural Language Engineering (2019), **25**, pp. 483–502
doi:10.1017/S1351324919000214

CAMBRIDGE UNIVERSITY PRESS

ARTICLE

A structured distributional model of sentence meaning and processing

E. Chersoni[1]*, E. Santus[2], L. Pannitto[3], A. Lenci[3], P. Blache[4], and C.-R. Huang[1]

[1]Department of Chinese and Bilingual Studies, Hong Kong Polytechnic University, Hong Kong, China, [2]Computer Science and Artificial Intelligence Lab, MIT, Cambridge (MA), United States, [3]Department of Philology, Literature and Linguistics, University of Pisa, Pisa, Italy and [4]Laboratoire Parole et Langage, Aix-Marseille University, France
*Corresponding author. Email: emmanuelechersoni@gmail.com

Abstract

Most compositional distributional semantic models represent sentence meaning with a single vector. In this paper, we propose a structured distributional model (SDM) that combines word embeddings with formal semantics and is based on the assumption that sentences represent events and situations. The semantic representation of a sentence is a formal structure derived from discourse representation theory and containing distributional vectors. This structure is dynamically and incrementally built by integrating knowledge about events and their typical participants, as they are activated by lexical items. Event knowledge is modelled as a graph extracted from parsed corpora and encoding roles and relationships between participants that are represented as distributional vectors. SDM is grounded on extensive psycholinguistic research showing that generalized knowledge about events stored in semantic memory plays a key role in sentence comprehension. We evaluate SDM on two recently introduced compositionality data sets, and our results show that combining a simple compositional model with event knowledge constantly improves performances, even with dif ferent types of word embeddings.

Keywords: distributional semantics; event knowledge; discourse representation theory; word embeddings; sentence processing

1. Sentence meaning in vector spaces

While for decades sentence meaning has been represented in terms of complex formal structures, the most recent trend in computational semantics is to model semantic representations with dense distributional vectors (aka *embeddings*). As a matter of fact, distributional semantics has become one of the most influential approaches to lexical meaning, because of the important theoretical and computational advantages of representing words with continuous vectors, such as automatically learning lexical representations from natural language corpora and multimodal data, assessing semantic similarity in terms of the distance between the vectors, and dealing with the inherently gradient and fuzzy nature of meaning (Erk 2012; Lenci 2018*a*).

Over the years, intense research has tried to address the question of how to project the strengths of vector models of meaning beyond word level, to phrases and sentences. The mainstream approach in distributional semantics assumes the representation of sentence meaning to be a vector, exactly like lexical items. Early approaches simply used pointwise vector operations (such as addition or multiplication) to combine word vectors to form phrase or sentence vectors (Mitchell and Lapata 2010), and in several tasks they still represent a non-trivial baseline to beat (Rimell *et al.* 2016). More recent contributions can be essentially divided into two separate trends. The former attempts to model 'Fregean compositionality' in vector space, and aims at finding progressively more sophisticated compositional operations to derive sentence representations from the vectors of the words composing them (Baroni *et al.* 2013; Paperno *et al.* 2014). In the latter trend, dense

vectors for sentences are learned as a whole, in a similar way to neural word embeddings (Mikolov *et al.* 2013; Levy and Goldberg 2014): for example, the encoder–decoder models of works like Kiros *et al.* (2015) and Hill *et al.* (2016) are trained to predict, given a sentence vector, the vectors of the surrounding sentences.

Representing sentences with vectors appears to be unrivalled from the applicative point of view, and has indeed important advantages such as the possibility of measuring similarity between sentences with their embeddings, as it is customary at the lexical level, which is then exploited in tasks like automatic paraphrasing and captioning, question-answering, etc. Recently, probing tasks have been proposed to test what kind of syntactic and semantic information is encoded in sentence embeddings (Ettinger *et al.* 2016; Adi *et al.* 2017; Conneau *et al.* 2018; Zhu *et al.* 2018). In particular, Zhu *et al.* (2018) show that current models are not able to discriminate between different syntactic realization of semantic roles and fail to recognize that *Lilly loves Imogen* is more similar to its passive counterpart than to *Imogen loves Lilly*. Moreover, it is difficult to recover information about the component words from sentence embeddings (Adi *et al.* 2017; Conneau *et al.* 2018). The semantic representations built with tensor product in the question-answering system by Palangi *et al.* (2018) have been claimed to be grammatically interpretable as well. However, the complexity of the semantic information brought by sentences and the difficulty to interpret the embeddings raise doubts about the general theoretical and empirical validity of the 'sentence-meaning-as-vector' approach.

In this paper, we propose a structured distributional model (SDM) of sentence meaning that combines word embeddings with formal semantics and is based on the assumption that sentences represent events and situations. These are regarded as inherently complex semantic objects, involving multiple entities that interact with different roles (e.g. agents, patients and locations). The semantic representation of a sentence is a formal structure inspired by discourse representation theory (DRT) (Kamp 2013) and containing distributional vectors. This structure is dynamically and incrementally built by integrating knowledge about events and their typical participants, as they are activated by lexical items. Event knowledge is modelled as a graph extracted from parsed corpora and encoding roles and relationships between participants that are represented as distributional vectors. The semantic representations of SDM retain the advantages of embeddings (e.g. learnability and gradability), but also contain directly interpretable formal structures, differently from classical vector-based approaches.

SDM is grounded on extensive psycholinguistic research showing that generalized knowledge about events stored in semantic memory plays a key role in sentence comprehension (McRae and Matsuki 2009). On the other hand, it is also close to recent attempts to look for a 'division of labour' between formal and vector semantics, representing sentences with logical forms enriched with distributional representations of lexical items (Beltagy *et al.* 2016; Boleda and Herbelot 2016; McNally 2017). Like SDM, McNally and Boleda (2017) propose to introduce embeddings within DRT semantic representations. At the same time, differently from these other approaches, SDM consists of formal structures that integrate word embeddings with a distributional representation of activated event knowledge, which is then dynamically integrated during semantic composition.

The contribution of this paper is twofold. First, we introduce SDM as a cognitively inspired distributional model of sentence meaning, based on a structured formalization of semantic representations and contextual event knowledge (Section 2). Secondly, we show that the event knowledge used by SDM in the construction of sentence meaning representations leads to improvements over other state-of-the-art models in compositionality tasks. In Section 3, SDM is tested on two different benchmarks: the first is RELPRON (Rimell *et al.* 2016), a popular data set for the similarity estimation between compositional distributional representations; the second is DTFit (Vassallo *et al.* 2018), a data set created to model an important aspect of sentence meaning, that is the typicality of the described event or situation, which has been shown to have important processing consequences for language comprehension.

https://doi.org/10.1017/S1351324919000214 Published online by Cambridge University Press

2. Dynamic composition with embeddings and event knowledge

SDM rests on the assumption that natural language comprehension involves the *dynamic construction of semantic representations, as mental characterization of the events or situations described in sentences*. We use the term 'dynamic' in the sense of dynamic semantic frameworks like DRT, to refer to a bidirectional relationship between linguistic meaning and context (see also Heim 1983):

The meaning of an expression depends on the context in which it is used, and its content is itself defined as a *context-change potential*, which affects and determines the interpretation of the following expressions.

The content of an expression E used in a context C depends on C, but – once the content has been determined – it will contribute to update C to a new context C', which will help fixing the content of the next expression. Similarly to DRT, SDM integrates word embeddings in a dynamic process to construct the semantic representations of sentences. Contextual knowledge is represented in distributional terms and affects the interpretation of following expressions, which in turn cue new information that updates the current context.[a]

Context is a highly multifaceted notion that includes several types of factors guiding and influencing language comprehension: information about the communicative settings, preceding discourse, general presuppositions and knowledge about the world, etc. In DRT, Kamp (2016) has introduced the notion of *articulated context* to model different sources of contextual information that intervene in the dynamic construction of semantic representations. In this paper, we focus on the contribution of a specific type of contextual information, which we refer to as *Generalized Event Knowledge* (GEK). This is knowledge about events and situations that we have experienced under different modalities, including the linguistic input (McRae and Matsuki 2009), and is generalized because it contains information about prototypical event structures.

In linguistics, the Generative Lexicon theory (Pustejovsky 1995) argues that the lexical entries of nouns also contain information about events that are crucial to define their meaning (e.g. *read* for *book*). Psycholinguistic studies in the last two decades have brought extensive evidence that the array of event knowledge activated during sentence processing is extremely rich: verbs (e.g. *arrest*) activate expectations about typical arguments (e.g. *cop* and *thief*) and vice versa (McRae *et al.* 1998; Ferretti *et al.* 2001; McRae *et al.* 2005), and similarly nouns activate other nouns typically co-occurring as participants in the same events (*key, door*) (Hare *et al.* 2009). The influence of argument structure relations on how words are neurally processed is also an important field of study in cognitive neuroscience (Thompson and Meltzer-Asscher 2014; Meltzer-Asscher *et al.* 2015; Williams *et al.* 2017).

Stored event knowledge has relevant processing consequences. Neurocognitive research showed that the brain is constantly engaged in making predictions to anticipate future events (Bar 2009; Clark 2013). Language comprehension, in turn, has been characterized as a largely predictive process (Kuperberg and Jaeger 2015). Predictions are memory-based, and experiences about events and their participants are used to generate expectations about the upcoming linguistic input, thereby minimizing the processing effort (Elman 2014; McRae and Matsuki 2009). For instance, argument combinations that are more 'coherent' with the event scenarios activated by the previous words are read faster in self-paced reading tasks and elicited smaller N400 amplitudes in Event Related Potentials (ERP) experiments (Bicknell *et al.* 2010; Matsuki *et al.* 2011; Paczynski and Kuperberg 2012; Metusalem *et al.* 2012).[b]

[a] An early work on a distributional model of lexical expectations in context is Washtell (2010), but its focus was more on word sense disambiguation than on representing sentence meaning.

[b] Event-related potentials are the electrophysiological response of the brain to a stimulus. In the sentence processing literature, the ERPs are recorded for each stimulus word, and the N400, one of the most studied ones, is a negative-going deflection appearing 400 ms after the presentation of the word. A common interpretation of the N400 assumes that the wave amplitude is proportional to the difficulty of semantic unification (Baggio and Hagoort 2011).

https://doi.org/10.1017/S1351324919000214 Published online by Cambridge University Press

Elman (2009; 2014) has proposed a general interpretation of these experimental results in the light of the Words-as-Cues framework. According to this theory, words are arranged in the mental lexicon as a sort of network of mutual expectations, and listeners rely on pre-stored representations of events and common situations to try to identify the one that a speaker is more likely to communicate. As new input words are processed, they are quickly integrated in a data structure containing a dynamic representation of the sentence content, until some events are recognized as the 'best candidates' for explaining the cues (i.e. the words) observed in the linguistic input. It is important to stress that, in such a view, the meaning of complex units such as phrases and sentences is not always built by composing lexical meanings, as the representation of typical events might be already stored and retrieved as a whole in semantic memory. Participants often occurring together become active when the representation of one of them is activated (see also Bar et al. (2007) on the relation between associative processing and predictions).

SDM aims at integrating the core aspects of dynamic formal semantics and the evidence on the role of event knowledge for language processing into a general model for compositional semantic representations that relies on two major assumptions:

- Lexical items are represented as embeddings within a network of relations encoding knowledge about events and typical participants, which corresponds to what we have termed above GEK.
- The *semantic representation* (SR) of a sentence (or even larger stretches of linguistic input, such as discourse) is a formal structure that dynamically combines the information cued by lexical items.

Like in Chersoni *et al.* (2017), the model is inspired by Memory, Unification and Control (MUC), proposed by Hagoort (Hagoort 2013, 2016) as a general model for the neurobiology of language. MUC incorporates three main functional components: (i) *Memory* corresponds to knowledge stored in long-term memory; (ii) *Unification* refers to the process of combining the units stored in *Memory* to create larger structures, with contributions from the context; and (iii) *Control* is responsible for relating language to joint action and social interaction. Similarly, our model distinguishes between a component storing event knowledge, in the form of a *Distributional Event Graph* (DEG, Section 2.1), and a *meaning composition function* that integrates information activated from lexical items and incrementally builds the SR (Section 2.2).

2.1 The distributional event graph

The Distributional Event Graph represents the event knowledge stored in long-term memory with information extracted from parsed corpora. We assume a very broad notion of *event*, as an *n*-ary relation between entities. Accordingly, an event can be a complex situation involving multiple participants, such as *The student reads a book in the library*, but also the association between an entity and a property expressed by the noun phrase *heavy book*. This notion of event corresponds to what psychologist call *situation knowledge* or *thematic associations* (Binder 2016). As McRae and Matsuki (2009) argue, GEK is acquired from both sensorimotor experience (e.g. watching or playing football matches) and linguistic experience (e.g. reading about football matches). DEG can thus be regarded as a model of the GEK derived from the linguistic input.

Events are extracted from parsed sentences, using syntactic relations as an approximation of deeper semantic roles (e.g. the subject relation for the agent and the direct object relation for the patient). In the present paper, we use dependency parses, as it is customary in distributional semantics, but nothing in SDM hinges on the choice of the syntactic representation. Given a verb

https://doi.org/10.1017/S1351324919000214 Published online by Cambridge University Press

Table 1. The five nearest paradigmatic and syntagmatic neighbours for the lexical item book, extracted from DEG

Paradigmatic neighbours	Syntagmatic neighbours
essay, story, novel, author, biography	publish, write, read, child, series

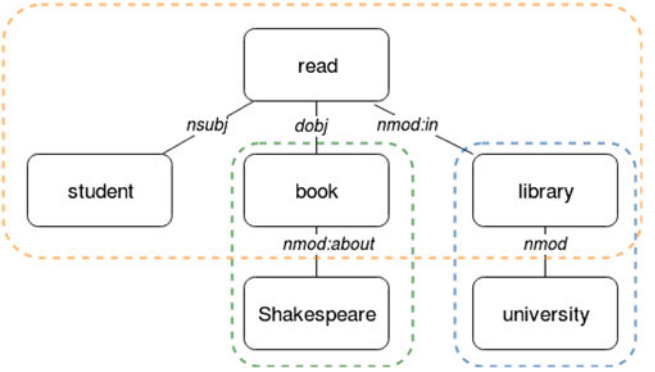

Figure 1. Reduced version of the parsing for the sentence *The student is reading the book about Shakespeare in the university library*. Three events are identified, each represented with a dotted box.

or a noun head, all its syntactic dependents are grouped together.[c] More schematic events are also generated by abstracting from one or more event participants for every recorded instance. Since we expect each participant to be able to trigger the event and consequently any of the other participants, a relation can be created and added to the graph from every subset of each group extracted from a sentence (cf. Figure 1).

The resulting DEG structure is a *weighted hypergraph*, as it contains weighted relations holding between nodes pairs, and a *labelled multigraph*, since the edges are labelled in order to represent specific syntactic relations. The weights σ are derived from co-occurrence statistics and measure the association strengths between event nodes. They are intended as salience scores that identify the most prototypical events associated with an entity (e.g. the typical actions performed by a student). Crucially, the graph nodes are represented as word embeddings. Thus, given a lexical cue w, the information in DEG can be activated along two dimensions during processing (cf. Table 1):

(1) by retrieving the most similar nodes to w (the paradigmatic neighbours), on the basis of their cosine similarity between their vectors and the vector of w;
(2) by retrieving the closest associates of w (the syntagmatic neighbours), using the edge weights.

Figure 2 shows a toy example of DEG. The little boxes with circles in them represent the embedding associated with each node. Edges are labelled with syntactic relations (as a surface approximation of event roles) and weighted with salience scores σ. Each event is a set of co-indexed edges. For example, e_2 corresponds to the event of students reading books in libraries, while e_1 represents a schematic event of students performing some generic action on books (e.g. reading, consulting and studying).

2.2 The meaning composition function

We assume that during sentence comprehension lexical items activate fragments of event knowledge stored in DEG (like in Elman's Words-as-Cues model), which are then dynamically integrated

[c]The extracted graphs are similar to the syntactic joint contexts for verb representation that were proposed by Chersoni *et al.* (2016).

https://doi.org/10.1017/S1351324919000214 Published online by Cambridge University Press

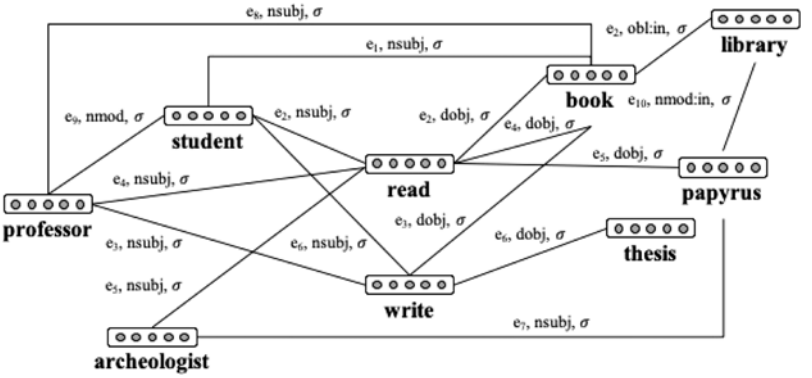

Figure 2. Toy sample of DEG showing several instances of events, each represented by a sequence of co-indexed *e*. The σ are the event salience weights.

in a semantic representation SR. This is a formal structure directly inspired by DRT and consisting of three different yet interacting information tiers:

(1) *universe* (U) – this tier, which we do not discuss further in the present paper, includes the entities mentioned in the sentence (corresponding to the *discourse referents* in DRT). They are typically introduced by noun phrases and provide the targets of anaphoric links.
(2) *linguistic conditions* (LC) – a context-independent tier of meaning that accumulates the embeddings associated with the lexical items. This corresponds to the conditions that in DRT content words add to the discourse referents. The crucial difference is that now such conditions are embeddings.
(3) *active context* (AC) – similarly to the notion of *articulated context* in Kamp (2016), this component consists of several types of contextual information available during sentence processing or activated by lexical items (e.g. information from the current communication setting and general world knowledge). More specifically, we assume that AC contains the embeddings activated from DEG by the single lexemes (or by other contextual elements) and integrated into a semantically coherent structure contributing to the sentence interpretation.

Figure 3 shows an example of SR built from the sentence *The student drinks the coffee* (ignoring the specific contribution of determiners and tense). The universe U contains the discourse referents introduced by the noun phrases, while LC includes the embeddings of the lexical items in the sentence, each linked to the relevant referent (e.g. $\overrightarrow{student} : u$ means that the embedding introduced by *student* is linked to the discourse referent *u*). AC consists of the embeddings activated from DEG and ranked by their salience with respect to the current content in the SR. The elements in AC are grouped by their syntactic relation in DEG, which again we regard here just as a surface approximation of their semantic role (e.g. the items listed under 'obl:*loc*' are a set of possible locations of the event expressed by the sentence). AC makes it possible to enrich the semantic content of the sentence with contextual information, predict other elements of the event and generate expectations about incoming input. For instance, given the AC in Figure 3, we can predict that the student is most likely to be drinking a coffee at the cafeteria and that he/she is drinking it for breakfast or in the morning. The ranking of each element in AC depends on two factors: (i) its degree of activation by the lexical items and (ii) its overall coherence with respect to the information already available in the AC.

A crucial feature of each SR is that LC and AC are also represented with vectors that are incrementally updated with the information activated by lexical items. Let SR_{i-1} be the semantic

https://doi.org/10.1017/S1351324919000214 Published online by Cambridge University Press

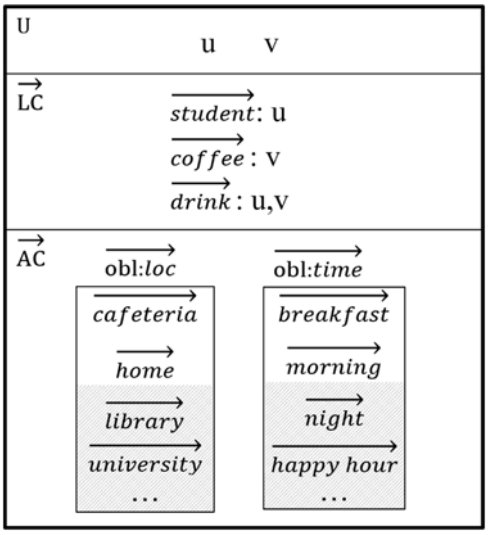

Figure 3. Sample SR for the sentence *The student drinks the coffee*. The sentence activates typical locations and times in which the event could take place.

representation built for the linguistic input $w_1, \ldots, w_{i-1}$. When we process a new pair $\langle w_i, r_i \rangle$ with a lexeme w_i and syntactic role r_i:

(1) LC in SR_{i-1} is updated with the embedding $\overrightarrow{w_i}$;
(2) AC in SR_{i-1} is updated with the embeddings of the syntagmatic neighbours of w_i extracted from DEG.

Figures 4 and 5 exemplify the update of the SR for the subject *The student* with the information is activated by the verb *drink*. The update process is defined as follows:

(1) LC is represented with the vector $\overrightarrow{LC}$ obtained from the linear combination of the embeddings of the words contained in the sentence. Therefore, when $\langle w_i, r_i \rangle$ is processed, the embedding $\overrightarrow{w_i}$ is simply added to $\overrightarrow{LC}$;[d]
(2) for each syntactic role r_i, AC contains a set of ranked lists (one for each processed pair) of embeddings corresponding to the most likely words expected to fill that role. For instance, the AC for the fragment *The student* in Figure 4 contains a list of the embeddings of the most expected direct objects associated with *student*, a list of the embeddings of the most expected locations, etc. Each list of expected role fillers is itself represented with the weighted centroid vector (e.g. $\overrightarrow{dobj}$) of their k most prominent items (with k a model hyperparameter). For instance, setting $k = 2$, the $\overrightarrow{dobj}$ centroid in the AC in Figure 4 is built just from $\overrightarrow{book}$ and $\overrightarrow{research}$; less salient elements (the gray areas in Figures 3–5) are kept in the list of likely direct objects, but at this stage do not contribute to the centroid representing the expected fillers for that role. AC is then updated with the DEG fragment activated by the new lexeme w_i (e.g. the verb *drink*):

- The event knowledge activated by w_i for a given role r_i is ranked according to cosine similarity with the vector $\overrightarrow{r_i}$ available in AC: in our example, the direct objects activated by the verb *drink* (e.g. $\overrightarrow{beer}$ and $\overrightarrow{coffee}$) are ranked according to their cosine similarity to the $\overrightarrow{dobj}$ vector of the AC.

[d] At the same time, the embedding is linked either to a new discourse referent added to U, or to an already available one.

https://doi.org/10.1017/S1351324919000214 Published online by Cambridge University Press

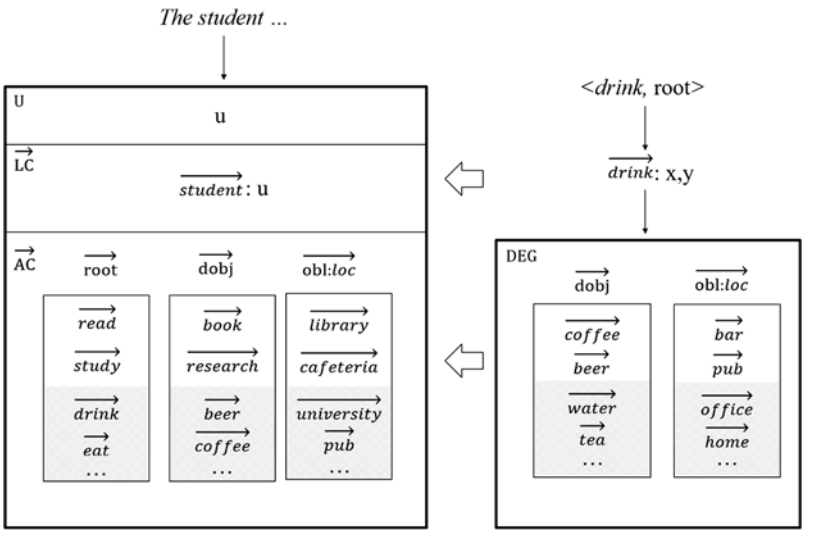

Figure 4. On the left, the SR for *The student*. On the right, the embedding and DEG portion activated by the verb *drink*.

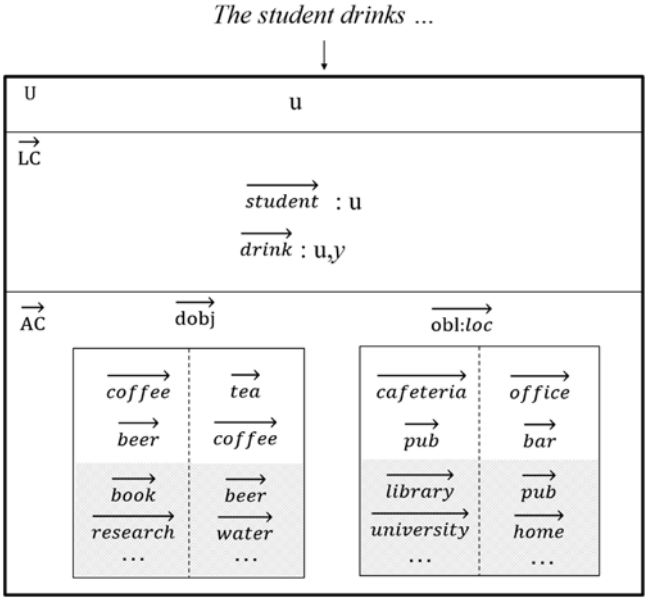

Figure 5. The original semantic representation SR for *The student* . . . is updated with the information activated by the verb, producing the SR for *The student drinks* . . . The new event knowledge is re-ranked with respect to the previous content of AC.

- The ranking process works also in the opposite direction: the newly retrieved information is used to update the centroids in AC. For example, the direct objects activated by the verb *drink* are aggregated into centroids and the corresponding weighted lists in AC are re-ranked according to the cosine similarity with the new centroids, in order to maximize the semantic coherence of the representation. At this point, $\overrightarrow{book}$ and $\overrightarrow{research}$, which are not as salient as $\overrightarrow{coffee}$ and $\overrightarrow{beer}$ in the *drinking* context, are downgraded in the ranked list and are therefore less likely to become part of the $\overrightarrow{dobj}$ centroid at the next step.

The newly retrieved information is now added to the AC: as shown in Figure 5, once the pair ⟨*drink*, *root*⟩ has been fully processed, the AC contains two ranked lists for the *dobj*

https://doi.org/10.1017/S1351324919000214 Published online by Cambridge University Press

role and two ranked lists for the *obl:loc* role, the top k elements of each list will be part of the centroid for their relation in the next step. Finally, the whole AC is represented with the centroid vector $\overrightarrow{AC}$ built out of the role vectors $\overrightarrow{r_1}, \ldots, \overrightarrow{r_n}$ available in AC. The vector $\overrightarrow{AC}$ encodes the integrated event knowledge activated by the linguistic input.

As an example of GEK re-ranking, assume that after processing the subject noun phrase *The student*, the AC of the corresponding SR predicts that the most expected verbs are *read, study, drink*, etc., the most expected associated direct objects are *book, research, beer*, etc., and the most expected locations are *library, cafeteria, university,* etc. (Figure 4). When the main verb *drink* is processed, the corresponding role list is removed by the AC, because that syntactic slot is now overtly filled by this lexeme, whose embedding is then added to the LC. The verb *drink* cues its own event knowledge, for instance, that the most typical objects of drinking are *tea, coffee, beer*, etc., and the most typical locations are *cafeteria, pub, bar*, etc. The information cued by *drink* is re-ranked to promote those items that are most compatible and coherent with the current content of AC (i.e. direct objects and locations that are likely to interact with students). Analogously, the information in the AC is re-ranked to make it more compatible with the GEK cued by *drink* (e.g. the salience of *book* and *research* gets decreased, because they are not similar to the typical direct objects and locations of *drink*). The output of the SR update is shown in Figure 5, whose AC now contains the GEK associated with an event of drinking by a student.

A crucial feature of SR is that it is a much richer representation than the bare linguistic input: the overtly realized arguments in fact activate a broader array of roles than the ones actually appearing in the sentence. As an example of how these unexpressed arguments contribute to the semantic representation of the event, consider a situation in which three different sentences are represented by means of AC, namely *The student writes the thesis*, *The headmaster writes the review* and *The teacher writes the assignment*. Although *teacher* could be judged as closer to *headmaster* than to *student*, and *thesis* as closer to *assignment* than to *review*, taking into account also the typical locations (e.g. a *library* for the first two sentences, a *classroom* for the last one) and writing supports (e.g. a *laptop* in the first two cases, a *blackboard* in the last one) would lead to the first two events being judged as the most similar ones.

In the case of unexpected continuations, the AC will be updated with the new information, though in this case the re-ranking process would probably not change the GEK prominence. Consider the case of an input fragment like *The student plows...*: *student* activates event knowledge as it is shown in Figure 3, but the verb does not belong to the set of expected events given *student*. The verb triggers different direct objects from those already in the AC (e.g. typical objects of *plow* such as *furrow* and *field*). Since the similarity of their centroid with the elements of the direct object list in the AC will be very low, the relative ordering of the ranked list will roughly stay the same, and direct objects pertaining to the plowing situation will coexist with direct objects triggered by *student*. Depending on the continuation of the sentence, then, the elements triggered by *plow* might gain centrality in the representation or remain peripheral.

It is worth noting that the incremental process of the SR update is consistent with the main principles of formal dynamic semantics frameworks like DRT. As we said above, dynamics semantics assumes the meaning of an expression to be a context-change potential that affects the interpretation of the following expressions. Similarly, in our distributional model of sentence representation the AC in SR_{i-1} affects the interpretation of the incoming input w_i, via the GEK re-ranking process.[e]

[e]For a more comprehensive analysis of the relationship between distributional semantics and dynamics semantics, see Lenci (2018*b*).

https://doi.org/10.1017/S1351324919000214 Published online by Cambridge University Press

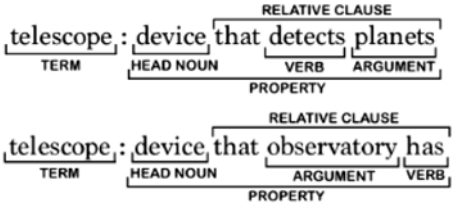

Figure 6. Image from Rimell *et al.* (2016), showing the terminology for terms and properties in RELPRON: subject relative clause top, object relative clause bottom.

3. Experiments

3.1 Data sets and tasks

Our goal is to test SDM in compositionality-related tasks, with a particular focus on the contribution of event knowledge. For the present study, we selected two different data sets: the development set of the RELPRON data set (Rimell *et al.* 2016)[f] and the DTFit data set (Vassallo *et al.* 2018).

RELPRON consists of 518 target–property pairs, where the target is a noun labelled with a syntactic function (either subject or direct object) and the property is a subject or object relative clause providing the definition of the target (Figure 6). Given a model, we produce a compositional representation for each of the properties. In each definition, the *verb*, the *head noun* and the *argument* are composed to obtain a representation of the property. Following the original evaluation in Rimell *et al.* (2016), we tested six different combinations for each composition model: the verb only, the argument only, the head noun and the verb, the head noun and the argument, the verb and the argument and all three of them. For each target, the 518 composed vectors are ranked according to their cosine similarity to the target. Like Rimell *et al.* (2016), we use mean average precision (henceforth MAP) to evaluate our models on RELPRON. Formally, MAP is defined as

$$MAP = \frac{1}{N}\sum_{i=1}^{N} AP(t_i) \tag{1}$$

where N is the number of terms in RELPRON, and $AP(t)$ is the average precision for term t, defined as

$$AP(t) = \frac{1}{P_t}\sum_{k=1}^{M} Prec(k) \times rel(k) \tag{2}$$

where P_t is the number of correct properties for term t in the data set, M is the total number of properties in the data set, $Prec(k)$ is the precision at rank k and $rel(k)$ is a function equal to one if the property at rank k is a correct property for t, and zero otherwise. Intuitively, $AP(t)$ will be 1 if, for the term t, all the correct properties associated to the term are ranked in the top positions, and the value becomes lower when the correct items are ranked farther from the head of the list.

Our second evaluation data set, **DTFit**, has been introduced with the goal of building a new gold standard for the *thematic fit* estimation task (Vassallo *et al.* 2018). Thematic fit is a psycholinguistic notion similar to selectional preferences, the main difference being that the latter involve the satisfaction of constraints on discrete semantic features of the arguments, while thematic fit is a continuous value expressing the degree of compatibility between an argument and a semantic role (McRae *et al.* 1998). Distributional models for thematic fit estimation have been proposed by several authors (Erk 2007; Baroni and Lenci 2010; Erk *et al.* 2010; Lenci 2011; Sayeed *et al.* 2015; Greenberg *et al.* 2015; Santus *et al.* 2017; Tilk *et al.* 2016; Hong *et al.* 2018). While thematic fit data sets typically include human-elicited typicality scores for argument–filler pairs taken in isolation, DTFit includes tuples of arguments of different length, so that the typicality value of an argument

[f]We used the development set of RELPRON in order to compare our results with those published by Rimell *et al.* (2016).

https://doi.org/10.1017/S1351324919000214 Published online by Cambridge University Press

depends on its interaction with the other arguments in the tuple. This makes it possible to model the dynamic aspect of argument typicality, since the expectations on an argument are dynamically updated as the other roles in the sentence are filled. The argument combinations in DTFit describe events associated with crowdsourced scores ranging from 1 (very atypical) to 7 (very typical). The data set items are grouped into typical and atypical pairs that differ only for one argument, and divided into three subsets:

- 795 triplets, each differing only for the **Patient** role:
  - *sergeant*_N *assign*_V *mission*_N (typical)
  - *sergeant*_N *assign*_V *homework*_N (atypical)
- 300 quadruples, each differing only for the **Location** role:
  - *policeman*_N *check*_V *bag*_N *airport*_N (typical)
  - *policeman*_N *check*_V *bag*_N *kitchen*_N (atypical)
- 200 quadruples, each differing only for the **Instrument** role:
  - *painter*_N *decorate*_V *wall*_N *brush*_N (typical)
  - *painter*_N *decorate*_V *wall*_N *scalpel*_N (atypical)

However, the Instrument subset of DTFit was excluded from our current evaluation. After applying the threshold of 5 for storing events in the DEG (cf. Section 3.2.3), we found that the SDM coverage on this subset was too low.

For each tuple in the DTFit data set, the task for our models is to predict the upcoming argument on the basis of the previous ones. Given a model, we build a compositional vector representation for each data set item by excluding the last argument in the tuple, and then we measured the cosine similarity between the resulting vector and the argument vector. Models are evaluated in terms of the Spearman's correlation between the similarity scores and the human ratings.

As suggested by the experimental results of Bicknell *et al.* (2010) and Matsuki *et al.* (2011), the typicality of the described events has important processing consequences: atypical events lead to longer reading times and stronger N400 components, while typical ones are easier to process thanks to the contribution of GEK. Thus, the task of modelling typicality judgements can be seen as closely related to modelling semantic processing complexity.

3.2 Models settings

In this study, we compare the performance of SDM with three baselines. The simple additive model formulated in Mitchell and Lapata (2010), a smoothed additive model and a multilayer Long-Short-Term-Memory (LSTM) neural language model trained against one-hot targets (Zaremba *et al.* 2015).

The additive models (Mitchell and Lapata 2010) have been evaluated on different types of word embeddings. We compared their performances with SDM.[g] Despite their simplicity, previous evaluation studies on several benchmarks showed that such models can be difficult to beat, even for sophisticated compositionality frameworks (Rimell *et al.* 2016; Arora *et al.* 2017; Tian *et al.* 2017).

The embeddings we used in our tests are the WORD2VEC models by Mikolov *et al.* (2013), that is the Skip-Gram with negative sampling (**SG**) and the continuous-bag-of-words (**CBOW**), and the *C-Phrase* model by Pham *et al.* (2015). The latter model incorporates information about

[g]We also tested pointwise multiplicative models, but in our tasks the performances were extremely low, so they were omitted.

https://doi.org/10.1017/S1351324919000214 Published online by Cambridge University Press

syntactic constituents, as the principles of the model training are (i) to group the words together according to the syntactic structure of the sentences and (ii) to optimize simultaneously the context predictions at different levels of the syntactic hierarchy (e.g. given the training sentence *A sad dog is howling in the park*, the context prediction will be optimized for *dog, a dog* and *a sad dog*, that is for all the words that form a syntactic constituent). The performance of C-Phrase is particularly useful to assess the benefits of using vectors that encode directly structural/syntactic information.

We used the same corpora both for training the embeddings and for extracting the syntactic relations for DEG. The training data come from the concatenation of three dependency-parsed corpora: the BNC Leech and Smith (2000), the Ukwac (Baroni *et al.* 2009) and a 2018 dump of the English Wikipedia, for a combined size of approximately 4 billion tokens. The corpora were parsed with Stanford CoreNLP (Manning *et al.* 2014). The hyperparameters of the embeddings were the following for all models: 400 dimensions, a context window of size 10, 10 negative samples and 100 as the minimum word frequency.[h]

3.2.1 Simple additive models

Our additive models, corresponding to an SR consisting of the $\overrightarrow{LC}$ component only, represent the meaning of a sentence *sent* by summing the embeddings of its words:

$$\overrightarrow{sent} = \sum_{w \in sent} \vec{w} \tag{3}$$

The similarity with the targets is measured with the cosine between the target vector and the sentence vector.

3.2.2 Smoothed additive models

These models are a smoothed version of the additive baseline, in which the final representation is simply the sum of the vectors of the words in the sentence, plus the top $k = 5$ nearest neighbour of each word in the sentence.[i] Therefore, the meaning of a sentence *sent* is obtained by

$$\overrightarrow{sent} = \sum_{w \in sent} \left(\vec{w} + \sum_{x \in N_5(w)} \vec{x} \right) \tag{4}$$

where $N_k(w)$ is the set of the k nearest neighbours of w. Compared to the GEK models, the smoothed additive baseline modifies the sentence vector by adding the vectors of related words. Thus, it represents a useful comparison term for understanding the actual added value of the structural aspects of SDM.[j]

3.2.3 The structured distributional models

The SDM introduced in Section 2 consists of a full SR including the linguistic conditions vector $\overrightarrow{LC}$ and the event knowledge vector $\overrightarrow{AC}$. In this section, we detail the hyperparameter setting for the actual implementation of the model.

Distributional event graph We included in the graph only events with a minimum frequency of 5 in the training corpora. The edges of the graph were weighted with *Smoothed Local Mutual*

[h] We tested different values for the dimension hyperparameter, and we noticed that vectors with higher dimensionality lead to constant improvements on the thematic fit data sets. The best results were obtained with 400 dimensions.

[i] We have experimented with $k = 2, 5, 10$ and, although the scores do not significantly differ, this baseline model reports slightly better scores for $k = 5$.

[j] We would like to thank one of the anonymous reviewers for the suggestion.

https://doi.org/10.1017/S1351324919000214 Published online by Cambridge University Press

Information (LMI). Given a triple composed by the words w_1 and w_2, and a syntactic relation s linking them, we computed its weight by using a smoothed version of the Local Mutual Information (Evert 2004):

$$LMI_\alpha(w_1, w_2, s) = f(w_1, w_2, s) * log(\frac{P(w_1, w_2, s)}{P(w_1) * P_\alpha(w_2) * P(s)}) \quad (5)$$

where the smoothed probabilities are defined as follows:

$$P_\alpha(x) = \frac{f(x)^\alpha}{\sum_x f(x)^\alpha} \quad (6)$$

This type of smoothing, with $\alpha = 0.75$, was chosen to mitigate the bias of Mutual Information (MI) statistical association measures towards rare events (Levy *et al.* 2015). While this formula only involves pairs (as only pairs were employed in the experiments), it is easily extensible to more complex tuples of elements.

Re-ranking settings For each word in the data set items, the top 50 associated words were retrieved from DEG. Both for the re-ranking phase and for the construction of the final representation, the event knowledge vectors (i.e. the role vectors $\overrightarrow{r}$ and the AC vector $\overrightarrow{AC}$) are built from the top 20 elements of each weighted list. As detailed in Section 2.2, the ranking process in SDM can be performed in the forward direction and in the backward direction at the same time (i.e. the AC can be used to re-rank newly retrieved information and vice versa, respectively), but for simplicity we only implemented the forward ranking.

Scoring As in SDM the similarity computations with the target words involves two separate vectors, we combined the similarity scores with addition. Thus, given a *target* word in a sentence *sent*, the score for SDM will be computed as

$$score(target, sent) = cos(\overrightarrow{target}, \overrightarrow{LC}(sent)) + cos(\overrightarrow{target}, \overrightarrow{AC}(sent)) \quad (7)$$

In all settings, we assume the model to be aware of the syntactic parse of the test items. In DTFit, word order fully determines the syntactic constituents, as the sentences are always in the *subject verb object [location-obl|instrument-obl]* order. In RELPRON, on the other hand, the item contains information about the relation that is being tested: in the *subject* relative clauses, the properties always show the *verb* followed by the *argument* (e.g. *telescope: device that detects planets*), while in the *object* relative clauses the properties always present the opposite situation (e.g. *telescope: device that observatory has*). In the present experiments, we did not use the predictions on non-expressed arguments to compute $\overrightarrow{AC}$, and we restricted the evaluation to the representation of the target argument. For example, in the DTFit Patients set, $\overrightarrow{AC}(sent)$ only contains the $\overrightarrow{dobj}$ centroid.

3.2.4 LSTM neural language model

We also compared the additive vector baselines and SDM with an LSTM neural network, taking as input WORD2VEC embeddings. For every task, we trained the LSTM on syntactically labelled tuples (extracted from the same training corpora used for the other models), with the objective of predicting the relevant target. In DTFit, for example, for the Location task, in the tuple *student learn history library*, the network is trained to predict the argument *library* given the tuple *student learn history*. Similarly, in RELPRON, for the tuple *engineer patent design*, the LSTM is trained to predict *engineer* in the subject task and *design* in the object task, given *patent design* and *engineer patent*, respectively.

In both DTFit and RELPRON, for each input tuple, we took the top N network predictions (we tested with $N = 3, 5, 10$, and we always obtained the best results with $N = 10$), we averaged their respective word embeddings, and we used the vector cosine between the resulting vector and the embedding of the target reported in the gold standard.

https://doi.org/10.1017/S1351324919000214 Published online by Cambridge University Press

Table 2. Results for the vector addition baseline, smoothed vector addition baseline, LSTM and the SDM on the RELPRON development set (MAP scores). Rows refer to the different word combinations tested in Rimell et al. 2016 (R&al.)

	Word combination	R&al.	SG	CBOW	C-Phrase
Additive	verb	0.18	0.16	0.16	0.13
	arg	0.35	0.33	0.32	0.37
	head noun+verb	0.26	0.26	0.25	0.21
	head noun+arg	0.45	0.44	0.46	0.45
	verb+arg	0.40	0.43	0.36	0.41
	head noun+verb+arg	**0.50**	**0.50**	0.47	0.47
Smoothed	verb	–	0.15	0.16	0.14
	arg	–	0.35	0.33	0.40
	head noun+verb	–	0.24	0.23	0.22
	head noun+arg	–	0.45	0.46	0.49
	verb+arg	–	0.41	0.36	0.41
	head noun+verb+arg	–	**0.49**	0.46	0.47
LSTM	LSTM_10	–	0.10	0.32	–
SDM	verb	–	0.21	0.20	0.19
	arg	–	0.38	0.36	0.41
	head noun+verb	–	0.27	0.28	0.26
	head noun+arg	–	0.50	0.50	0.50
	verb + arg	–	0.41	0.36	0.41
	head noun + verb + arg	–	**0.54**	0.52	**0.54**

The LSTM is composed by (i) an input layer of the same size of the WORD2VEC embeddings (400 dimensions, with dropout=0.1); (ii) a single-layer monodirectional LSTM with l hidden layers (where $l = 2$ when predicting Patients and $l = 3$ when predicting Locations) of the same size of the embeddings; (iii) a linear layer (again with dropout= 0.1) of the same size of the embeddings, which takes in input the average of the hidden layers of the LSTM; and (iv) finally a softmax layer that projects the filler probability distribution over the vocabulary.

4. Results and discussion

4.1 RELPRON

Given the targets and the composed vectors of all the definitions in RELPRON, we assessed the cosine similarity of each pair and computed the MAP scores shown in Table 2. First of all, the Skip-Gram-based models always turn out to be the best performing ones, with rare exceptions, closely followed by the C-Phrase ones. The scores of the additive models are slightly inferior, but very close to those reported by Rimell *et al.* (2016), while the LSTM model lags behind vector addition, improving only when the parameter N is increased. Results seem to confirm the original findings: even with very complex models (in that case, the Lexical Function Model by Paperno et al. 2014), it is difficult to outperform simple vector addition in compositionality tasks.

Interestingly, SDM shows a constant improvement over the simple vector addition equivalents (Table 2), with the only exception of the composition of the verb and the argument. All the results

https://doi.org/10.1017/S1351324919000214 Published online by Cambridge University Press

Table 3. Results for the vector addition baseline, smoothed vector addition baseline, LSTM and the SDM on the Patients and Locations subsets of DTFit. The scores are expressed in terms of Spearman's correlation with the gold standard ratings. The LSTM scores refer to the best configuration, with $N = 10$ and vectors of size 400. The statistical significance of the improvements over the additive baseline is reported as follows: $^{*}p < 0.05$; $^{**}p < 0.01$ (p values computed with Fisher's r-to-z transformation, one-tailed test). ns = non-significant correlation

	Data set	SG	CBOW	C-Phrase
Additive	Patients	**0.63**	0.52	0.60
	Locations	**0.74**	0.70	0.74
Smoothed	Patients	**0.58**	0.51	**0.58**
	Locations	0.74	0.71	**0.76**
LSTM	Patients	ns	0.42	–
	Locations	0.58	**0.60**	–
SDM	Patients	0.65	0.62**	**0.66** *
	Locations	0.75	0.74	**0.76**

for the *headNoun* + *verb* + *arg* composition are, to the best of our knowledge, the best scores reported so far on the data set. Unfortunately, given the relatively small size of RELPRON, the improvement of the GEK models fails to reach significance ($p > 0.1$ for all comparisons between a basic additive model and its respective augmentation with DEG, p values computed with the Wilcoxon rank-sum test). Compared to SDM, the smoothed vector addition baseline seems to be way less consistent (Table 2): for some combinations and for some vector types, adding the nearest neighbours is detrimental. We take these results as supporting the added value of the structured event knowledge and the SR update process in SDM, over the simple enrichment of vector addition with nearest neighbours. Finally, we can notice that the Skip-Gram vectors have again an edge over the competitors, even over the syntactically informed C-Phrase vectors.

4.2 DTFit

At a first glance, the results on DTFit follow a similar pattern (Table 3): the three embedding types perform similarly, although in this case the CBOW vectors perform much worse than the others in the Patients data set. LSTM also largely lags behind all the additive models, showing that thematic fit modelling is not a trivial task for language models, and that more complex neural architectures are required in order to obtain state-of-the-art results (Tilk *et al.* 2016).[k]

The results for SDM again show that including the DEG information leads to improvements in the performances (Table 3). While on the Locations the difference is only marginal, also due to the smaller number of test items, two models out of three showed significantly higher correlations than their respective additive baselines. The increase is particularly noticeable for the CBOW vectors that, in their augmented version, manage to fill the gap with the other models and to achieve a competitive performance. However, it should also be noticed that there is a striking difference between the two subsets of DTFit: while on patients the advantage of the GEK models on both the baselines is clear, on locations the results are almost indistinguishable from those of the smoothed additive baseline, which simply adds the nearest neighbours to the vectors of the words in the sentence. This complies with previous studies on thematic fit modelling with dependency-based distributional models (Sayeed *et al.* 2015; Santus *et al.* 2017). Because of the ambiguous nature

[k]It should also be noticed that our LSTM baseline has been trained on simple syntactic dependencies, while state-of-the-art neural models rely simultaneously on dependencies and semantic role labels (Tilk *et al.* 2016; Hong *et al.* 2018).

https://doi.org/10.1017/S1351324919000214 Published online by Cambridge University Press

Table 4. Spearman's correlations of the vector addition baseline and the SDM based on Skip-Gram on the DTFit patients subset

Dimensions	Additive	SDM
100	0.58	0.63
200	0.58	0.63
300	0.60	0.64
400	0.64	**0.65**

of the prepositions used to identify potential locations, the role vectors used by SDM can be very noisy. Moreover, since most locative complements are optional adjuncts, it is likely that the event knowledge extracted from corpora contain a much smaller number of locations. Therefore, the structural information about locations in DEG is probably less reliable and does not provide any clear advantage compared to additive models.

Concerning the comparison between the different types of embeddings, Skip-Gram still retains an advantage over C-Phrase in its basic version, while it is outperformed when the latter vectors are used in SDM. However, the differences are clearly minimal, suggesting that the structured knowledge encoded in the C-Phrase embeddings is not a plus for the thematic fit task. Concerning this point, it must be mentioned that most of the current models for thematic fit estimation rely on vectors relying either on syntactic information (Baroni and Lenci 2010; Greenberg *et al.* 2015; Santus *et al.* 2017; Chersoni *et al.* 2017) or semantic roles (Sayeed *et al.* 2015; Tilk *et al.* 2016). On the other hand, our results comply with studies like Lapesa and Evert (2017), who reported comparable performance for bag-of-words and dependency-based models on several semantic modelling tasks, thus questioning whether the injection of linguistic structure in the word vectors is actually worth its processing cost. However, this is the first time that such a comparison is carried out on the basis of the DTFit data set, while previous studies proposed slightly different versions of the task and evaluated their systems on different benchmarks.[1] A more extensive and in-depth study is required in order to formulate more conclusive arguments on this issue.

Another constant finding of previous studies on thematic fit modelling was that high-dimensional, count-based vector representations perform generally better than dense word embeddings, to the point that Sayeed *et al.* (2016) stressed the sensitivity of this task to linguistic detail and to the interpretability of the vector space. Therefore, we tested whether vector dimensionality had an impact on task performance (Table 4). Although the observed differences are generally small, we noticed that higher-dimensional vectors are generally better in the DTFit evaluation and, in one case, the differences reach a marginal significance (i.e. the difference between the 100-dimensional and the 400-dimensional basic Skip-Gram model is marginally significant at $p < 0.1$). This point will also deserve future investigation, but it seems plausible that for this task embeddings benefit from higher dimensionality for encoding more information, as it has been suggested by Sayeed and colleagues. However, these advantages do not seem to be related to the injection of linguistic structure directly in the embeddings (i.e. not to the direct use of syntactic contexts for training the vectors), as bag-of-words models perform similarly to – if not better than – a syntactic-based model like C-Phrase. We leave to future research a systematic comparison with sparse count-based models to assess whether interpretable dimensions are advantageous for modelling context-sensitive thematic fit.

[1]In data sets such as McRae *et al.* (1998) and Padó (2007), the verb-filler compatibility is modelled without taking into account the influence of the other fillers. On the other hand, studies on the composition and update of argument expectations generally propose evaluations in terms of classification tasks (Lenci 2011; Chersoni *et al.* 2017) instead of assessing directly the correlation with human judgements.

https://doi.org/10.1017/S1351324919000214 Published online by Cambridge University Press

Table 5. Comparison of the performance of the vector addition baseline, smoothed vector addition baseline and the SDM on the typical items of DTFit Patients (Spearman's correlations). Δ reports the SDM improvements over the basic additive models. Significance is noted with the following notation: $^*p < 0.05$

Model	Additive	Smoothed	SDM	Δ
CBOW	0.18	0.18	0.30	+ 0.12*
SG	0.29	0.24	0.33	+ 0.04
C-Phrase	0.30	0.29	**0.37**	+ 0.07

Table 6. Comparison of the SDM performance (MAP) on the subject and object relative clauses in RELPRON

SDM	SG		CBOW		C-Phrase	
Subset	*sbj*	*obj*	*sbj*	*obj*	*sbj*	*obj*
head noun+verb	0.29	0.31	0.32	0.32	0.29	0.28
head noun+arg	0.54	0.57	0.54	0.56	0.56	0.57
verb+arg	0.45	0.47	0.40	0.43	0.47	0.47
head noun+verb+arg	0.56	0.61	0.58	0.57	0.60	0.58

4.3 Error analysis

One of our basic assumptions about GEK is that semantic memory stores representations of typical events and their participants. Therefore, we expect that integrating GEK into our models might lead to an improvement especially on the typical items of the DTFit data set. A quick test with the correlations revealed that this is actually the case (Table 5): all models showed increased Spearman's correlations on the tuples in the typical condition (and in the larger Patients subset of DTFit, the increase is significant at $p < 0.05$ for the CBOW model), while they remain unchanged or even decrease for the tuples in the atypical conditions. Notice that this is true only for SDM, which is enriched with GEK. On the other hand, the simple addition of the nearest neighbours never leads to improvements, as proved by the low correlation scores of the smoothed additive baseline. As new and larger data sets for compositionality tasks are currently under construction (Vassallo *et al.* 2018), it will be interesting to assess the consistency of these results.

Turning to the RELPRON data set, we noticed that the difference between subject and object relative clauses is particularly relevant for SDM, which generally shows better performances on the latter. Table 6 summarizes the scores component on the two subsets. While relying on syntactic dependencies, SDM also processes properties in linear order: the *verb+arg* model, therefore, works differently when applied to *subject* clauses than to *object* clauses. In the *subject* case, in fact, the verb is found first, and then its expectations are used to re-rank the object ones. In the *object* case, on the other hand, things proceed the opposite way: at first the subject is found, and then its expectations are used to re-rank the verb ones. Therefore, the event knowledge triggered by the verb seems not only less informative than the one triggered by the argument, but it is often detrimental to the composition process.

5. Conclusion

In this contribution, we introduced a SDM that represents sentence meaning with formal structures derived from DRT and including embeddings enriched with event knowledge. This is modelled with a distributional event graph that represents events and their prototypical participants

https://doi.org/10.1017/S1351324919000214 Published online by Cambridge University Press

with distributional vectors linked in a network of syntagmatic relations extracted from parsed corpora. The compositional construction of sentence meaning in SDM is directly inspired by the principles of dynamic semantics. Word embeddings are integrated in a dynamic process to construct the semantic representations of sentences: contextual event knowledge affects the interpretation of following expressions, which cue new information that updates the current context.

Current methods for representing sentence meaning generally lack information about typical events and situation, while SDM rests on the assumption that such information can lead to better compositional representations and to an increased capacity of modelling typicality, which is one striking capacity of the human processing system. This corresponds to the hypothesis by Baggio and Hagoort (2011) that semantic compositionality actually results from a balance between storage and computation: on the one hand, language speakers rely on a wide amount of stored events and scenarios for common, familiar situations; on the other hand, a compositional mechanism is needed to account for our understanding of new and unheard sentences. Processing complexity, as revealed by effects such as the reduced amplitude of the N400 component in ERP experiments, is inversely proportional to the typicality of the described events and situations: the more they are typical, the more they will be coherent with already-stored representations.

We evaluated SDM on two tasks, namely a classical similarity estimation tasks on the target–definition pairs of the RELPRON data set (Rimell *et al.* 2016) and a thematic fit modelling task on the event tuples of the DTFit data set (Vassallo *et al.* 2018). Our results still proved that additive models are quite efficient for compositionality tasks, and that integrating the event information activated by lexical items improves the performance on both the evaluation data sets. Particularly interesting for our evaluation was the performance on the DTFit data set, since this data set has been especially created with the purpose of testing computational models on their capacity to account for human typicality judgements about event participants. The reported scores on the latter data set showed that not only SDM improves over simple and smoothed additive models, but also that the increase in correlation concerns the data set items rated as most typical by human subjects, fulfilling our initial predictions.

Differently from other distributional semantic models tested on the thematic fit task, 'structure' is now externally encoded in a graph, whose nodes are embeddings, and not directly in the dimension of the embeddings themselves. The fact that the best performing word embeddings in our framework are the Skip-Gram ones is somewhat surprising, and against the finding of previous literature in which bag-of-words models were always described as struggling on this task (Baroni *et al.* 2014; Sayeed *et al.* 2016). Given our results, we also suggested that the dimensionality of the embeddings could be an important factor, much more than the choice of training them on syntactic contexts.

References

Adi Y., Kermany E., Belinkov Y., Lavi O. and Goldberg Y. (2017). Fine-grained analysis of sentence embeddings using auxiliary prediction tasks. In *ICLR*.

Arora S., Liang Y. and Ma T. (2017). A simple but tough-to-beat baseline for sentence embeddings. In *ICLR*.

Baggio G. and Hagoort P. (2011). The balance between memory and unification in semantics: A dynamic account of the N400. *Language and Cognitive Processes* **26**(9), 1338–1367.

Bar M. (2009). The proactive brain. *Philosophical Transactions of the Royal Society B* **364**(March), 1235–1243.

Bar M., Aminoff E., Mason M. and Fenske M. (2007). The units of thought. *Hippocampus* **17**(6), 420–428.

Baroni M., Bernardi R. and Zamparelli R. (2013). Frege in space: A program for compositional distributional semantics. *Linguistic Issues in Language Technologies* **9**.

Baroni M., Bernardini S., Ferraresi A. and Zanchetta E. (2009). The WaCky Wide Web: A collection of very large linguistically processed Web-Crawled Corpora. *Language Resources and Evaluation* **43**(3), 209–226.

Baroni M., Dinu G. and Kruszewski G. (2014). Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In *ACL*.

Baroni M. and Lenci A. (2010). Distributional memory: A general framework for corpus-based semantics. *Computational Linguistics* **36**(4), 673–721 .

https://doi.org/10.1017/S1351324919000214 Published online by Cambridge University Press

Beltagy I., Roller S., Cheng P., Erk K. and Mooney R.J. (2016). Representing meaning with a combination of logical and distributional models. *Computational Linguistics* **42**(4), 763–808.

Bicknell K., Elman J.L., Hare M., McRae K. and Kutas M. (2010). Effects of event knowledge in processing verbal arguments. *Journal of Memory and Language* **63**(4), 489–505.

Binder J.R. (2016). In defense of abstract conceptual representations. *Psychonomic Bulletin & Review* **23**, 1096–1108.

Boleda G. and Herbelot A. (2016). Formal distributional semantics: Introduction to the special issue. *Computational Linguistics* **42**(4).

Chersoni E., Lenci A. and Blache P. (2017). Logical metonymy in a distributional model of sentence comprehension. In **SEM*.

Chersoni E., Santus E., Blache P. and Lenci A. (2017). Is structure necessary for modeling argument expectations in distributional semantics? In *IWCS*.

Chersoni E., Santus E., Lenci A., Blache P. and Huang C.-R. (2016). Representing verbs with rich contexts: An evaluation on verb similarity. In *EMNLP*.

Clark A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences* **36**(3), 181–204.

Conneau A., Kruszewski G., Lample G., Barrault L. and Baroni M. (2018). What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In *ACL*.

Elman J.L. (2009). On the meaning of words and Dinosaur bones: Lexical knowledge without a Lexicon. *Cognitive Science* **33**(4), 1–36.

Elman J.L. (2014). Systematicity in the Lexicon: On having your cake and eating it too. In *The Architecture of Cognition: Rethinking Fodor and Pylyshyn's Systematicity Challenge*. Cambridge, MA: The MIT Press. pp. 115–146.

Erk K. (2007). A simple, similarity-based model for selectional preferences. In *ACL*. pp. 635–653.

Erk K. (2012). Vector space models of word meaning and phrase meaning: A survey. *Linguistics and Language Compass* **6**(10).

Erk K., Padó S. and Padó U. (2010). A flexible, corpus-driven model of regular and inverse selectional preferences. *Computational Linguistics* **36**.

Ettinger A., Elgohary A. and Resnik P. (2016). Probing for semantic evidence of composition by means of simple classification tasks. In *Workshop on evaluating vector-space representations for NLP*. ACL.

Evert S. (2004). *The Statistics of Word Cooccurrences Word Pairs and Collocations*. PhD Thesis, University of Stuttgart.

Ferretti T.R., McRae K. and Hatherell A. (2001). Integrating verbs, situation schemas, and thematic role concepts. *Journal of Memory and Language* **44**(4), 516–547.

Greenberg C., Sayeed A.B. and Demberg V. (2015). Improving unsupervised vector-space thematic fit evaluation via role-filler prototype clustering. In *NAACL-HLT*.

Hagoort P. (2013). MUC (Memory, Unification, Control) and beyond. *Frontiers in Psychology* **4**.

Hagoort P. (2016). MUC (Memory, Unification, Control): A model on the neurobiology of language beyond single word processing. In *Neurobiology of Language*, Volume 28. Amsterdam: Elsevier. pp. 339–347.

Hare M., Jones M., Thomson C., Kelly S. and McRae K. (2009). Activating event knowledge. *Cognition* **111**(2), 151–167.

Heim I. (1983). File change semantics and the familiarity theory of definiteness. In *Meaning, Use, and Interpretation of Language*. Berlin: De Gruyter.

Hill F., Cho K. and Korhonen A. (2016). Learning distributed representations of sentences from unlabelled data. In *NAACL-HLT*.

Hong X., Sayeed A. and Demberg V. (2018). Learning distributed event representations with a multi-task approach. In **SEM*.

Kamp H. (2013). *Meaning and the Dynamics of Interpretation: Selected papers by Hans Kamp*. Leiden-Boston: Brill.

Kamp H. (2016). *Entity Representations and Articulated Contexts: An Exploration of the Semantics and Pragmatics of Definite Noun Phrases*. Unpublished manuscript.

Kiros R., Zhu Y., Salakhutdinov R.R., Zemel R., Urtasun R., Torralba A. and Fidler S. (2015). Skip-thought vectors. In *NIPS*.

Kuperberg G.R. and Jaeger T.F. (2015). What do we mean by prediction in language comprehension? *Language Cognition & Neuroscience* **3798**, 32–59.

Lapesa G. and Evert S. (2017). Large-scale evaluation of dependency-based DSMs: Are they worth the effort? In *EACL*.

Leech G.N. and Smith N. (2000). Manual to accompany the British National Corpus (version 2) with improved word-class tagging.

Lenci A. (2011). Composing and updating verb argument expectations: A distributional semantic model. In *ACL Workshop on Cognitive Modeling and Computational Linguistics*.

Lenci A. (2018a). Distributional models of word meaning. *Annual Review of Linguistics* **4**, 151–171.

Lenci A. (2018b). *Dynamic Distributional Semantics*. Unpublished manuscript.

Levy O. and Goldberg Y. (2014). Neural word embedding as implicit matrix factorization. In *NIPS*.

Levy O., Goldberg Y. and Dagan I. (2015). Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics* **3**.

Manning C., Surdeanu M., Bauer J., Finkel J., Bethard S. and McClosky D. (2014). The Stanford CoreNLP Natural Language Processing Toolkit. In *ACL (System Demonstrations)*.

Matsuki K., Chow T., Hare M., Elman J.L., Scheepers C. and McRae K. (2011). Event-based plausibility immediately influences online language comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **37**(4).

McNally L. (2017). Kinds, descriptions of kinds, concepts, and distributions. In *Bridging Formal and Conceptual Semantics. Selected Papers of BRIDGE-14*. DUP.

McNally L. and Boleda G. (2017). Conceptual vs. referential affordance in concept composition. In *Compositionality and Concepts in Linguistics and Psychology*. Berlin: Springer. pp. 245–267.

McRae K., Hare M., Elman J.L. and Ferretti T. (2005). A basis for generating expectancies for verbs from nouns. *Memory & Cognition* **33**(7), 1174–1184.

McRae K. and Matsuki K. (2009). People use their knowledge of common events to understand language, and do so as quickly as possible. *Language and Linguistics Compass* **3**(6), 1417–1429.

McRae K., Spivey-Knowlton M.J. and Tanenhaus M.K. (1998). Modeling the influence of thematic fit (and other constraints) in online sentence comprehension. *Journal of Memory and Language* **38**(3), 283–312.

Meltzer-Asscher A., Mack J.E., Barbieri E. and Thompson C.K. (2015). How the brain processes different dimensions of argument structure complexity: Evidence from fMRI. *Brain and Language* **142**, 65–75.

Metusalem R., Kutas M., Urbach T.P., Hare M., McRae K. and Elman J.L. (2012). Generalized event knowledge activation during online sentence comprehension. *Journal of Memory and Language* **66**(4), 545–567.

Mikolov T., Sutskever I., Chen K., Corrado G.S. and Dean J. (2013). Distributed representations of words and phrases and their compositionality. In *NIPS*.

Mitchell J. and Lapata M. (2010). Composition in distributional models of semantics. *Cognitive Science* **34**(8), 1388–1429.

Paczynski M. and Kuperberg G.R. (2012). Multiple influences of semantic memory on sentence processing: Distinct effects of semantic relatedness on violations of real-world event/state knowledge and animacy selection restrictions. *J Memory and Language* **67**(4), 426–448.

Padó U. (2007). *The Integration of Syntax and Semantic Plausibility in a Wide-coverage Model of Human Sentence Processing*. PhD Thesis, University of Stuttgart.

Palangi H., Smolensky P., He X. and Deng L. (2018). Question-answering with grammatically-interpretable representations. In *AAAI*.

Pham N., Kruszewski G., Lazaridou A., Baroni M. (2015). Jointly optimizing word representations for lexical and sentential tasks with the C-phrase model. In *ACL*.

Paperno D., Pham N.T. and Baroni M. (2014). A practical and linguistically-motivated approach to compositional distributional semantics. In *ACL*, Volume 1.

Pustejovsky J. (1995). *The Generative Lexicon*. Cambridge, MA: MIT Press.

Rimell, L., Maillard J., Polajnar T. and Clark S. (2016). RELPRON: A relative clause evaluation data set for compositional distributional semantics. *Computational Linguistics* **42**(4), 661–701.

Santus E., Chersoni E., Lenci A. and Blache P. (2017). Measuring thematic fit with distributional feature overlap. In *EMNLP*.

Sayeed A., Demberg V. and Shkadzko P. (2015). An exploration of semantic features in an unsupervised thematic fit evaluation framework. *Italian Journal of Linguistics* **1**(1).

Sayeed A., Greenberg C. and Demberg V. (2016). Thematic fit evaluation: An aspect of selectional preferences. In *ACL Workshop for Evaluating Vector Space Representations for NLP*.

Thompson C.K. and Meltzer-Asscher A. (2014). Neurocognitive mechanisms of verb argument structure processing. In *Structuring the Argument: Multidisciplinary Research on Verb Argument Structure*. Amsterdam: John Benjamins.

Tian R., Okazaki N. and Inui K. (2017). The mechanism of additive composition. *Machine Learning* **106**(7), 1083–1130.

Tilk O., Demberg V., Sayeed A.B., Klakow D. and Thater S. (2016). Event participant modelling with neural networks. In *EMNLP*.

Vassallo P., Chersoni E., Santus E., Lenci A. and Blache P. (2018). Event knowledge in sentence processing: A new dataset for the evaluation of argument typicality. In *LREC Workshop on Linguistic and Neurocognitive Resources (LiNCR)*.

Washtell J. (2010). Expectation vectors: A semiotics inspired approach to geometric lexical-semantic representation. In *Workshop on Geometrical Models of Natural Language Semantics*. ACL.

Williams A., Reddigari S. and Pylkkänen L. (2017). Early sensitivity of left perisylvian cortex to relationality in nouns and verbs. *Neuropsychologia* **100**, 131–143.

Zarcone A., Utt J. and Padó S. (2012). Modeling covert event retrieval in logical metonymy: Probabilistic and distributional accounts. In *Proceedings of the NAACL Workshop on Cognitive Modeling and Computational Linguistics*.

Zaremba W., Sutskever I. and Vinyals O. (2015). Recurrent neural network regularization. In ICLR. arXiv:1409.2329.

Zhu X., Li T. and de Melo G. (2018). Exploring semantic properties of sentence embeddings. In *ACL*.

Cite this article: Chersoni E, Santus E, Pannitto L, Lenci A, Blache P and Huang C.-R. A structured distributional model of sentence meaning and processing. *Natural Language Engineering* **25**, 483–502. https://doi.org/10.1017/S1351324919000214