Measuring Thematic Fit with Distributional Feature Overlap

Enrico Santus[1], **Emmanuele Chersoni**[2], **Alessandro Lenci**[3] and **Philippe Blache**[2]
enrico_santus@sutd.edu.sg
emmanuelechersoni@gmail.com
alessandro.lenci@unipi.it
philippe.blache@univ-amu.fr
[1] Singapore University of Technology and Design
[2] Aix-Marseille University
[3] University of Pisa

Abstract

In this paper, we introduce a new distributional method for modeling predicate-argument thematic fit judgments. We use a syntax-based DSM to build a prototypical representation of verb-specific roles: for every verb, we extract the most salient *second order contexts* for each of its roles (i.e. the most salient dimensions of *typical role fillers*), and then we compute the thematic fit as a *weighted overlap* between the top features of candidate fillers and role prototypes. Our experiments show that our method consistently outperforms a baseline re-implementing a state-of-the-art system, and achieves better or comparable results to those reported in the literature for the other unsupervised systems. Moreover, it provides an explicit representation of the features characterizing verb-specific semantic roles.

1 Introduction

Several psycholinguistic studies in the last two decades have brought extensive evidence that humans activate a rich array of event knowledge during sentence processing: verbs (e.g. *arrest*) activate expectations about their typical arguments (e.g. *cop, thief*) (McRae et al., 1998; Altmann and Kamide, 1999; Ferretti et al., 2001; McRae et al., 2005; Hare et al., 2009; Matsuki et al., 2011), and nouns activate other nouns typically co-occurring in the same events (Kamide et al., 2003; Bicknell et al., 2010). Subjects are able to determine the plausibility of a noun for a given argument role and quickly use this knowledge to anticipate upcoming linguistic input (McRae and Matsuki, 2009). This phenomenon is referred to in the literature as *thematic fit*. Thematic fit estimation has been extensively used in sentence comprehension studies on constraint-based models, mainly as a predictor variable allowing to disambiguate between possible structural analyses.[1] More in general, thematic fit is considered as a key factor in a variety of studies concerned with structural ambiguity (Vandekerckhove et al., 2009).

Starting from the work of Erk et al. (2010), several distributional semantic methods have been proposed to compute the extent to which nouns fulfill the requirements of verb-specific thematic roles, and their performances have been evaluated against human-generated judgments (Baroni and Lenci, 2010; Lenci, 2011; Sayeed and Demberg, 2014; Sayeed et al., 2015, 2016; Greenberg et al., 2015a,b). Most research on thematic fit estimation has focused on *count-based* vector representations (as distinguished from *prediction-based* vectors).[2] Indeed, in their comparison between high-dimensional explicit vectors and low-dimensional neural embeddings, Baroni et al. (2014) found that thematic fit estimation is the only benchmark on which prediction models are lagging behind state-of-the-art performance. This is consistent with Sayeed et al. (2016)'s observation that "thematic fit modeling is particularly sensitive to linguistic detail and interpretability of the vector space".

The present work sets itself among the unsupervised approaches to thematic fit estimation. By relying on explicit and interpretable *count-based* vector representations, we propose a simple, cognitively-inspired, and efficient thematic fit model using information extracted from dependency-parsed corpora. The key features of our proposal are *a*) prototypical representations of verb-specific thematic roles, based on feature weighting and filtering of *second order contexts*

[1] For an overview on constraint-based models, see MacDonald and Seidenberg (2006).

[2] We adopt the terminology from Baroni et al. (2014).

Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 659–669
Copenhagen, Denmark, September 7–11, 2017.

(i.e. contexts that are salient for many of the typical fillers of a given verb-specific thematic role), and *b*) a similarity measure which computes the *Weighted Overlap* (WO) between prototypes and candidate fillers.[3]

2 Related Work

Erk et al. (2010) were, at the best of our knowledge, the first authors to measure the correlation between human-elicited thematic fit ratings and the scores assigned by a syntax-based Distributional Semantic Model (DSM). More specifically, their gold standard consisted of the human judgments collected by McRae et al. (1998) and Padó (2007). The plausibility of each verb-filler pair was computed as the similarity between new candidate nouns and previously attested exemplars for each specific verb-role pairing (as already proposed in Erk (2007)).

Baroni and Lenci (2010) evaluated their Distributional Memory (henceforth DM)[4] framework on the same datasets, adopting an approach to the task that has become dominant in the literature: for each verb role, they built a prototype vector by averaging the dependency-based vectors of its most typical fillers. The higher the similarity of a noun with a role prototype, the higher its plausibility as a filler for that role. Lenci (2011) has later extended the model to account for the dynamic update of the expectations on an argument, depending on how another role is filled. By using the same DM tensor, this study tested an additive and a multiplicative model (Mitchell and Lapata, 2010) to compose and update the expectations on the patient filler of the subject-verb-object triples of the Bicknell dataset (Bicknell et al., 2010).

The thematic fit models proposed by Sayeed and Demberg (2014) and Sayeed et al. (2015) are similar to Baroni and Lenci's, but their DSMs were built by using the roles assigned by the SENNA semantic role labeler (Collobert et al., 2011) to define the feature space. These authors argued that the prototype-based method with dependencies works well when applied to the agent and to the patient role (which are almost always syntactically realized as subjects and objects), but that it might be problematic to apply it to different roles, such as instruments and locations, as the construction of the prototype would have to rely on prepositional complements as typical fillers, and the meaning of prepositions can be ambiguous. Comparing their results with Baroni and Lenci (2010), the authors showed that their system outperforms the syntax-based model DepDM and almost matches the scores of the best performing TypeDM, which uses hand-crafted rules. Moreover, they were the first to evaluate thematic role plausibility for roles other than agent and patient, as they computed the scores also for the instruments and for the locations of the Ferretti datasets (Ferretti et al., 2001).

Greenberg et al. (2015a,b) further developed the TypeDM and the role-based models, investigating the effects of verb polysemy on human thematic fit judgments and introducing a hierarchical agglomerative clustering algorithm into the prototype creation process. Their goal was to cluster together typical fillers into multiple prototypes, corresponding to different verb senses, and their results showed constant improvements of the performance of the DM-based model.

Finally, Tilk et al. (2016) presented two neural network architectures for generating probability distributions over selectional preferences for each thematic role. Their models took advantage of supervised training on two role-labeled corpora to optimize the distributional representation for thematic fit modeling, and managed to obtain significant improvements over the other systems on almost all the evaluation datasets. They also evaluated their model on the task of composing and updating verb argument expectations, obtaining a performance comparable to Lenci (2011).

3 Methodology

As pointed out by Sayeed et al. (2016), most works on unsupervised thematic fit estimation vary in the method adopted for constructing the prototypes. The semantic role prototype is usually a vector, obtained by averaging the most typical fillers, and plausibility of new fillers depends on their similarity to the prototype, assessed by means of vector cosine (the standard similarity measure for DSMs; see Turney and Pantel (2010)).

Its merits notwithstanding, we argue that this method is not optimal for characterizing roles. Distributional vectors are typically built as out-of-

[3] Code: https://github.com/esantus/Thematic_Fit

[4] In this paper, we will make reference to two different models of DM: DepDM and TypeDM. DepDM counts the frequency of dependency links between words (e.g. *read, obj, book*), while TypeDM uses the variety of surface forms that express the link between words, rather than the link itself.

context representations, and they conflate different senses. By building the prototype as the centroid of a cluster of vectors and measuring then the thematic fit with vector cosine, the plausibility score is inevitably affected by many contexts that are irrelevant for the specific verb-argument combination.[5] This is likely to be one of the main reasons behind the difficulties of modeling roles other than agent and patient with syntax-based DSMs. We claim that improving the prototype representation might lead to a better characterization of thematic roles, and to a better treatment of polysemy.

When a verb and an argument are composed, humans are intuitively able to select only the part of the potential meaning of the words that is relevant for the concept being expressed (e.g. in *The player hit the ball*, humans would certainly exclude from the meaning of *ball* semantic dimensions that are strictly related to its dancing sense). In other words, not all the features of the semantic representations are active, and the composition process makes some features more 'prominent', while moving others to the background.[6]

Although we are not aware of experimental works specifically dedicated to verb-argument composition, a similar idea has been supported in studies on conceptual combinations (Hampton, 1997, 2007): when a head and a modifier are combined, their interaction affects the saliency of the features in the original concepts. For example, in *racing car*, the most salient properties would be those related to SPEED, whereas in *family car* SPACE properties would probably be more prominent. Yeh and Barsalou (2006) used a property priming experiment to show how the concept features activated during language comprehension vary across the background situations described by the sentence they occur in. When concepts are combined in a sentence, the features that are relevant for the specific combination are activated and are then easier to verify for human subjects.

The same could be true for linguistically-derived properties of lexical meaning: Simmons et al. (2008) brought neuroimaging evidence of the early activation of word association areas during property generation tasks, and Santos et al. (2011) showed that word associates are often among the properties generated for a given concept. Such findings suggest that, while we combine concepts, both embodied simulations and word distributions influence property salience (Barsalou et al., 2008).

Our model makes the following assumptions:

- the composition between a verb role representation and an argument shares the same cognitive mechanism underlying conceptual combinations;
- at least part of semantic representations is derived from, and/or mirrored in, linguistic data.[7] Consistently, the process of selecting the relevant features of the concepts being composed corresponds to modify the salience of the dimensions of distributional vectors;
- thematic fit computation is carried out on the basis of the activation and selection of salient features of a verb thematic role prototype and of the candidate argument filler vectors.

We rely on syntax-based DSMs, using dependency relations to approximate verb-specific roles and to identify their most typical fillers: for agents/patients, we extract the most frequent subjects/objects, for instruments we use the prepositional complements introduced by *with*, and for locations those introduced by either *on, at* or *in*.

Assuming that the linguistic features of distributional vectors correspond to the properties of conceptual composition processes, a candidate filler can be represented as a sorted distributional vector of the filler term, in which the most salient contexts occupy the top positions. Similarly, the abstract representation of a verb-specific role is a sorted prototype-vector, whose features derive from the sum of the most typical filler vectors for that verb-specific role.

Differently from Baroni and Lenci, the core and novel aspect of our proposal, described in the following subsections, is that we do not simply measure the correlation between all the features of candidate and prototype vectors (as vector cosine would do on unsorted vectors), but rather we *rank* and *filter* the features, computing the *weighted overlap* with a rank-based similarity measure inspired by $APSyn$, a recent proposal by Santus

[5] For an overview on the limitations of vector cosine, see: Li and Han (2013); Dinu et al. (2015); Schnabel et al. (2015); Faruqui et al. (2016); Santus et al. (2016a).

[6] An early proposal going in this direction is the predication theory by Kintsch (2001), which exploited Latent Semantic Analysis to select only the vector features that are appropriate for predicate-argument composition.

[7] See also the so-called 'strong version' of the Distributional Hypothesis (Miller and Charles, 1991; Lenci, 2008).

et al. (2016a,b,c) which has shown interesting results in synonymy detection and similarity estimation. As we will show in the next sections, the new metric assigns high scores to candidate fillers sharing many salient contexts with the verb-specific role prototype.

3.1 Typical Fillers

The first step of our method consists in identifying the typical fillers of a verb-specific role. Following Baroni and Lenci (2010), we weighted the raw co-occurrences between verbs, syntactic relations and fillers in the TypeDM tensor of DM with Positive Local Mutual Information (PLMI; Evert (2004)).

Given the co-occurrence count O_{vrf} of the verb v, a syntactic relation r and the filler f, we computed the expected count E_{vrf} under the assumption of statistical independence:

$$PLMI(v,r,f) = log\left(\frac{O_{v,r,f}}{E_{v,r,f}}\right) * O_{v,r,f} \quad (1)$$

From the ranked list of *(v,r,f)* tuples, for each slot, we selected as typical fillers the top k lexemes with the highest PLMI scores (see examples in Table 1, *Typical Fillers* column). In our experiments, we report results for $k = \{10, 30, 50\}$.

3.2 Role Prototype Vectors

To represent the typical fillers, the candidate fillers and the verb-specific role prototypes (which are obtained by summing their typical filler vectors), we built a syntax-based DSM. This includes *relation:word* contexts, like *sbj:dog, obj:apple, etc.*.

Contexts were weighted with Positive Pointwise Mutual Information (PPMI; Church and Hanks (1990), Bullinaria and Levy (2012), Levy et al. (2015)). Given a context c and a word w, the PPMI is defined as follows:

$$PPMI(w,c) = max(PMI(w,c), 0) \quad (2)$$

$$PMI(w,c) = log\left(\frac{P(w,c)}{P(w)P(c)}\right) = log\left(\frac{|w,c|D}{|w||c|}\right) \quad (3)$$

where *w* is the target word, *c* is the given context, *P(w,c)* is the probability of co-occurrence, and *D* is the collection of observed word-context pairs.[8]

The context c of the prototype vector P representing a thematic role has a value corresponding to the sum of the values of c for each of the k typical fillers used to build P. The contexts of P are then sorted according to their weight. Desirably, the highest-ranking contexts for a role prototype will be those that are more strongly associated with many of its typical fillers. Such *second order contexts* correspond to the most salient features of the verb-specific thematic role, as they are salient for many role fillers (some examples are reported in Table 1, *Top Second Order Contexts* column).

In summary, we built centroid vectors for our verb-specific thematic roles by means of *second order contexts*, which are first order dependency-based contexts of the most typical fillers of a verb-specific role. Since we are interested only in the most salient contexts, we ranked the centroid contexts according to their PPMI score, and we took the resulting rank as a distributional characterization of the thematic roles.

3.3 Filtering the Contexts

Filtering the prototype dimensions according to syntactic criteria might be useful to improve our role representations. It is, indeed, reasonable to hypothesize that predicates co-occurring with the typical patients of a verb are more relevant for the characterization of its patient role than – let's say – prepositional complements, as they correspond to other actions that are typically performed on the same patients.

Imagine that *apple, pizza, cake* etc. are among the most salient fillers for the OBJ slot of *to eat*, and that *OBJ-1:slice-v, OBJ-1:devour-v, SBJ:kid-n, INSTRUMENT:fork-n, LOCATION:table-n* are some of the most salient contexts of the prototype.[9] Things that are typically *sliced* and/or *devoured* are more likely to be good fillers for the patient role *to eat* than things that are simply located on a *table* or that are patients of actions performed by *kids*. To test this hypothesis, we evaluated the performance of the system in three different settings, each of which selecting:

[8] A variant of this DSM weighted with PLMI (which is simply the PPMI multiplied by the word-context frequency) was also built, but because of its lower and inconsistent performance we will not discuss it further. Santus et al. (2016c) previously showed that their rank-based measure performs worse on PLMI-weighted vectors, as they are biased towards frequent contexts.

[9] Our DSM also makes use of inverse syntactic dependencies: *target SYN-1 context* means that *target* is linked to *context* by the dependency relation *SYN* (e.g. *meal OBJ-1 devour* means that *meal* is *OBJ* of *devour*).

	Typical Fillers	Top Second Order Contexts
subject: cure-v	treatment-n, drug-n, resin-n, doctor-n, surgery-n, medicine-n, therapy-n, antibiotic-n, dose-n, operation-n, water-n...	obj-1:prescribe-v, sbj-1:prescribe-v, sbj-1:prevent-v, sbj-1:contraindicate-v, [...]
object: abandon-v	plan-n, idea-n, project-n, attempt-n, position-n, principle-n, policy-n, ship-n, practice-n, hope-n, fort-n, claim-n...	obj-1:revive-v, obj-1:defend-v, obj-1:renounce-v, obj-1:espouse-v, sbj-1:entail-v...
instrument: eat-v	bread-n, hand-n, spoon-n, sauce-n, relish-n, fork-n, finger-n, meal-n, knife-n, friend-n, chopstick-n, rice-n, food-n...	obj-1:flavour-v, obj-1:taste-v, obj-1:spoon-v, sbj-1:taste-v, obj-1:slice-v in:bowl-n...
location: walk-v	in:direction-n, at:time, at:pace-n, on:path-n, at:night, on:side-n, at:end, on:beach-n, on:leg, in:area, in:way...	obj-1:wander-v, obj-1:stroll-v, obj-1:litter-v, obj-1:sweep-v, sbj-1:slope-v, obj-1:tread-v...

Table 1: Typical fillers and top *second order contexts* for several verb-specific roles.

- only predicates in a subject/object relation (**SO** setting);
- only prepositional complements (**PREP** setting);
- both of them (**ALL** setting).

3.4 Computing the Thematic Fit

Our hypothesis is that fillers whose salience-ranked vector has a large overlap with the prototype representation should have a high thematic fit. Such overlap should take into account not only the number of shared features, but also their respective ranks in the salience-ranked vectors.

When the prototype has been computed and the candidate filler vector has also been sorted, we can measure the *Weighted Overlap* by adapting $APSyn$ (Santus et al., 2016a,b,c) to our needs:

$$WO(w_x, w_y) = \sum_{\forall f \epsilon (x_{[1:N]} \cap y_{[1:N]})} \frac{1}{avg(r_x(f), r_y(f))} \quad (4)$$

where for every feature f in the intersection between the top N features of the sorted vectors x, $x_{[1:N]}$, and y, $y_{[1:N]}$, we sum 1 divided by the average rank of the shared feature in x and y, $r_x(f)$ and $r_y(f)$ (N is a tunable parameter).

This measure assigns the maximum score to vectors sharing exactly the same dimensions, in the same salience ranking. The lower the rank of a shared context in the sorted vector, the smaller its contribution to the thematic fit score. If the feature set intersection is empty, the score will be 0.

Differently from cosine similarity, which conflates multiple senses, measuring the *Weighted Overlap* between prototype and candidate filler can improve the estimation of the thematic fit by favoring the appropriate word senses: for example, for a verb-argument pair like *embrace-v–communism-n*, *communism-n* is likely to intersect and to increase the saliency (through the average rank) only of the second-order features of *embrace-v* referring to its abstract sense.

Data	Our system	BL2010	SD2014	G2015	T2016
Padó	96	100	99	100	99
McRae	100	95	96	95	96
Instr.	100	93	94	93	94
Loc.	96	99	100	99	100

Table 2: Dataset coverage (%) for all systems.

4 Experiments

Datasets. We tested our method on three popular datasets for thematic fit estimation, namely McRae et al. (1998), Ferretti et al. (2001) and Padó (2007). All the datasets contain human plausibility judgments for verb-role-filler triples. McRae and Padó include scores for agent and patient roles, whereas Ferretti includes instruments and locations (see Table 2 for the coverage of each system for the datasets).

Metrics. Performance is evaluated as the Spearman correlation between the scores of the systems and the human plausibility judgments.

Fillers. In order to make our results more comparable with previous studies, the typical fillers for each verb role were extracted from the TypeDM tensor of the Distributional Memory framework (see Section 3.1).[10] Those were the same fillers used by Baroni and Lenci (2010) and Greenberg et al. (2015b).

DSM. Distributional information is derived from the concatenation of two corpora: the British National Corpus (Leech, 1992) and Ukwac (Baroni et al., 2009). Both were parsed with the Maltparser (Nivre and Hall, 2005). From this concatenation, we built a dependency-based DSMs, weighted with PPMI, containing 20,145 targets (i.e. nouns and verbs with frequency above 1000) and 94,860 contexts. The syntactic relations taken into account were: *sbj, sbj-1, obj, obj-1, at-1, in-1, on-1, with-1*.

Settings. To prove our hypotheses and verify the consistency of the system, we tested a large range of settings, varying:

[10] http://clic.cimec.unitn.it/dm/

Weight	N	# Fillers	Padó			Mcrae			Ferretti - Instruments			Ferretti - Locations		
			ALL	SO	PREP	ALL	SO	PREP	ALL	SO	PREP	ALL	SO	PREP
PPMI	2000	10	0.43	0.45	0.26	0.25	0.27	0.19	0.43	0.41	0.46	0.25	0.27	0.28
		30	0.47	0.49	0.33	0.26	0.28	0.22	0.42	0.41	**0.50**	0.28	0.31	0.37
		50	0.46	**0.50**	0.35	0.27	**0.29**	0.24	0.39	0.38	0.47	0.28	0.32	**0.39**
Vector Cosine (Baseline)		10	0.43			0.25			0.42			0.29		
		30	0.47			0.26			0.41			0.32		
		50	0.48			0.26			0.38			0.31		
State of the Art														
Baroni and Lenci (2010)			0.53			0.33			0.36			0.23		
Sayeed and Demberg (2014)			0.56			0.27			0.28			0.13		
Greenberg et al. (2015)			0.53			0.36			0.42			0.29		
Tilk et al. (2016)			0.52			0.38			0.45			0.44		

Table 3: Results for Padó, McRae and Ferretti, Instruments and Locations, with WO computed on PPMI matrix, varying the number of fillers (i.e. 10, 30 and 50) and the types of dependency contexts (i.e. ALL, SO and PREP). The best results of our system are in bold. A baseline reimplementing Baroni and Lenci (2010) – with 10, 30 and 50 fillers – and state of the art results from previous literature are reported for comparison.

- the number of fillers used to build the prototype, with the most typical values in the literature ranging between 10 and 50. We report the results for 10, 30 and 50 fillers
- the types of the dependency relations used for calculating the overlap: we report results for the SO, PREP and ALL settings;
- the value of N, that is the number of top contexts that we take into account when computing the weighted overlap. Table 3 reports the scores for our best setting, while the performances for other values of N are discussed in the Section 5.

Baseline and State of the Art. As a baseline, we use the thematic fit model by Baroni and Lenci (2010), with no ranking of the features of the prototypes and with vector cosine as a similarity metric.[11] Results are reported for 10, 30 and 50 fillers. For reference, we also report the results of state-of-the-art models, both the unsupervised (Baroni and Lenci, 2010; Sayeed and Demberg, 2014; Greenberg et al., 2015b) and the supervised ones (Tilk et al., 2016).

5 Results

Table 3 describes the performance of the best setting (weight: PPMI; N=2000). In the first three rows, the table shows the scores obtained by our system varying the types of dependency contexts (i.e. ALL, SO, PREP) and the number of fillers considered for the prototype (i.e. 10, 30 and 50). The other rows respectively show i) the scores obtained by calculating the vector cosine between the role prototype vector (i.e. the vector obtained by summing the most typical fillers, with no salience ranking of the dimensions) and the candidate filler vector and ii) the scores reported in the literature for the best unsupervised and supervised models.

At a glance, our best scores always outperform the reimplementation of Baroni and Lenci, being mostly competitive with the state of the art models. More precisely, for agents and patients the performance is close to the reported scores for DM, when only predicates are used in the WO calculation, as hypothesized in Section 3.3. The neural network of Tilk and colleagues retains a significant advantage on our models only for the McRae dataset. Our system, however, shows a remarkable improvements on the Ferretti's datasets, and specifically on Ferretti-Instruments, when only complements are used (see Section 3.3), outperforming even the supervised and more complex model by Tilk et al. (2016), which has access to semantic roles information. Compared to the other unsupervised models, our system has a statistically significant advantage over Baroni and Lenci (2010) on the locations dataset and over Sayeed and Demberg (2014) on the locations and on the instruments dataset ($p < 0.05$).[12]

At the best of our knowledge, the result for the

[11] This baseline is equivalent to the approach of Baroni and Lenci (2010), except for the fact that it is applied on a standard dependency-based DSM and not on TypeDM, which combines dependency links and handcrafted lexico-syntactic patterns: see Section 2.

[12] p-values computed with Fisher's r-to-z transformation.

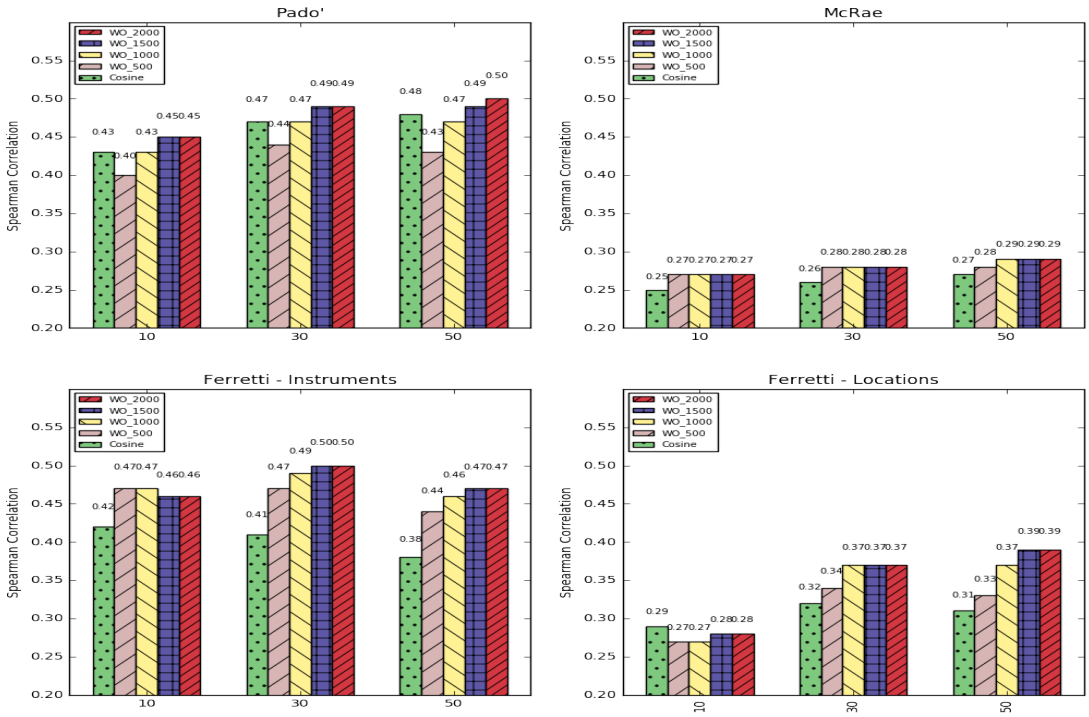

Figure 1: Results for Padó, McRae and Ferretti, Instruments and Locations, with WO (respectively SO and PREP) computed on PPMI matrix, varying the number of fillers (i.e. 10, 30 and 50) and the value of N (i.e. 500, 1000, 1500 and 2000). A baseline reimplementing Baroni and Lenci (2010) – with 10, 30 and 50 fillers – is also reported in every test for comparison.

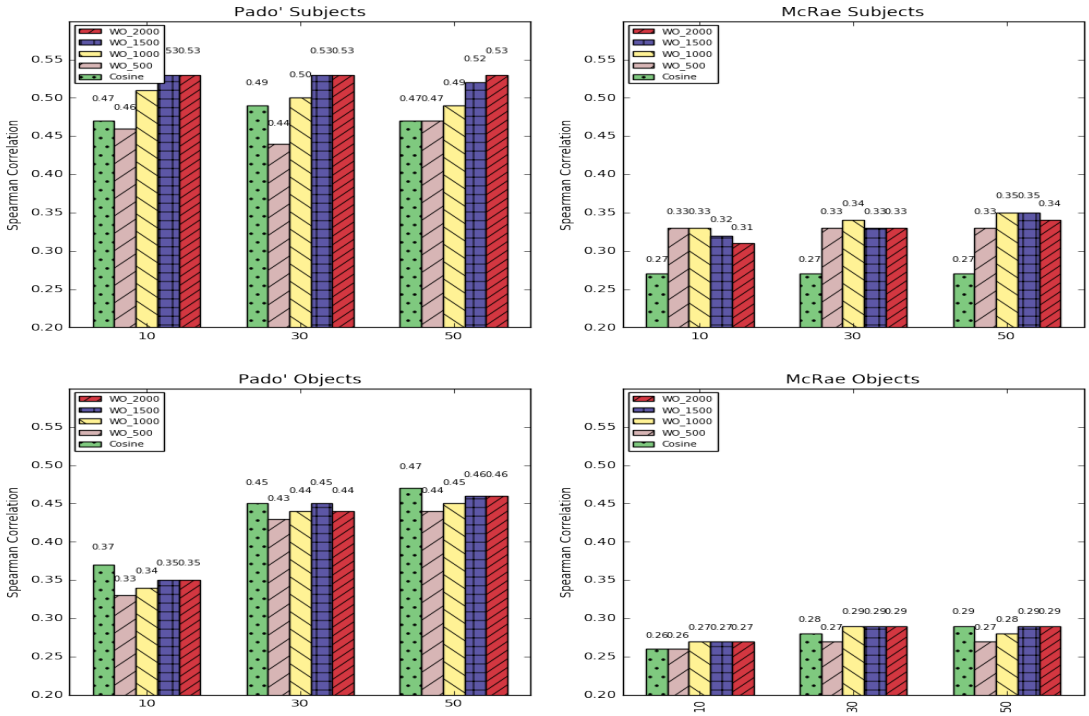

Figure 2: Results for the agent and patient roles in Padó and McRae, with WO (SO) computed on PPMI matrix, varying the number of fillers (i.e. 10, 30 and 50) and the value of N (i.e. 500, 1000, 1500 and 2000). A baseline reimplementing Baroni and Lenci (2010) – with 10, 30 and 50 fillers – is also reported in every test for comparison.

		BEST 35				WORST 35			
	Metric	**Avg. Gold**	**Overlap**	**Syntax**	**Lexemes**	**Avg. Gold**	**Overlap**	**Syntax**	**Lexemes**
McRae (k=50, Pred)	Cos	4.90	3	14 sbj, 21 obj	3 sentence, 2 devour, 2 scratch...	4.75	26	24 sbj, 11 obj	2 consider, 2 entertain, 2 scrub
	WO 2000	4.20	4	17 sbj, 18 obj	2 haunt	4.15	26	23 sbj, 12 obj	2 admire, 2 arrest, 2 consider, 2 entertain
Padó (k=50, Pred)	Cos	4.07	10	12 sbj, 23 obj	4 advise, 4 eat, 4 embarrass	4.77	16	17 sbj, 18 obj	9 tell, 7 kill, 4 see
	WO 2000	4.35	10	21 sbj, 14 obj	3 confuse, 3 hear, 3 promise, 3 raise	4.68	16	15 sbj, 20 obj	7 resent, 5 increase, 4 hear, 4 see
Ferretti - Instruments (k=30, Compl)	Cos	4.53	16	35 with	3 hung, 3 eat, 3 teach	4.51	22	35 with	4 repair, 3 teach, 3 inflate
	WO 2000	5.06	15	35 with	3 dig, 3 hunt	4.49	22	35 with	3 repair, 2 attract, 2 dig, 2 draw, 2 drink...
Ferretti - Locations (k=50, Compl)	Cos	5.15	11	35 on/at/in	3 draw, 3 rescue	4.72	23	35 on/at/in	3 run, 2 wait, 2 wash, 2 shower...
	WO 2000	4.97	11	35 on/at/in	3 browse, 3 eat, 3 mingle, 3 rescue	4.47	23	35 on/at/in	3 run, 2 draw, 2 exercise, 2 shower, 2 wait...

Table 4: Average gold values, number of items listed for both metrics, and distribution of syntactic and lexical forms among the 35 best and worst correlated items for every measure in the given datasets.

instruments is the best reported until now in the literature. This is particularly interesting because – as pointed out by Sayeed and Demberg (2014) – instruments and locations are difficult to model for a dependency-based system, given the ambiguity of prepositional phrases (e.g. *with* does not only encode instruments, but it can also encode other roles, such as in *I ate a pizza with Mark*). We think this is the main reason behind the different trend observed for the Instruments datasets with respect to the number of the fillers (see Table 3 and Figure 1). Unlike all the other datasets, instrument prototypes built with more fillers tend to be more noisy and therefore to pull down both the vector cosine and WO performance (this is partially true also for locations, where the performances – for cosine and WO with a lower number of contexts – drop with more than 30 fillers: see Figure 1). Systems based on semantic role labeling have an advantage in this sense, as they do not have to deal with prepositional ambiguity.

Our results show that, by weighting and filtering the features of the role prototype, dependency-based approaches can be successful in modeling roles other than agent and patient, eventually dealing also with the ambiguity of prepositional phrases.

Settings. Apart from the above-mentioned exceptions, the best scores are obtained building the prototypes with a higher number of fillers, typically with 50, and calculating the WO only with a syntactically-filtered set of contexts. More specifically, Padó and McRae benefit from the calculation of WO using only second order subject-object predicates (i.e. SO), while Ferretti-Instruments and Ferretti-Locations benefit from the exclusive use of prepositional complements (i.e. PREP). On the other hand, the opposite setting (e.g. SO for Ferretti-Instruments and Ferretti-Locations and PREP for Padó and McRae) leads to much lower scores, whereas the full vectors (i.e. ALL) tend to have a stable-but-not-excellent performances on all datasets.

As briefly mentioned above, in our experiments, we tested both PPMI and PLMI as weighting measures. Table 3 only reports PPMI scores because it performs more regularly than PLMI, whose behaviour is often unpredictable.

A parameter that has an impact on the performance of our system is the value of N, which is the number of *second order contexts* that are considered when calculating the WO. We have noticed that the performance of WO is directly related to the growth of N, and this can be noticed in Figure 1, where WO is plotted for the different values of N with every combination of dataset and number of fillers. For space reasons, the plot only contains the performance for the best type of *second order contexts* for each dataset (i.e. SO for Padó and McRae and COMP for Ferretti-Locations and Ferretti-Instruments). As it can be seen in Figure 1, the scores of WO tend to grow with the growth of N in all datasets. Interestingly, they are largely above the competitive baseline in most of the cases, the only exceptions being Padó (where a large N is necessary to outperform the baseline) and Ferretti-Locations with 10 fillers (prepositional ambiguity might have caused the introduction of noisy fillers among the top ones).

Agent & Patient. In order to further evaluate our system, we have split Padó and McRae datasets into agent and patient subsets. Figure 2 describes the performance of WO and vector cosine baseline while varying N and the number of fillers. The plot shows a clearly better performance of WO for the agent role (i.e. *subject*), especially when N is equal or over 1000 (note that the value of N has little impact in the agent subset of the McRae dataset). Such advantage, however, is reduced for the patient role (i.e. *object*). This is particularly interesting because we do not observe large drops in performance for the vector cosine

between agent and patient role (except for Padó, $k = 10$). The drop is particularly noticeable in Padó, a dataset which has several non-constraining verbs (especially for the patient role: a similar observation was also made by Tilk et al. (2016)). As the constraints on the typical fillers of such verbs are very loose, we hypothesize that it is more difficult to find a set of salient features that are shared by many typical fillers. Therefore, estimations based on the whole vectors turn out to be more reliable. This can be confirmed by looking at the worst correlated words reported in *Lexemes* column, in Table 4.

5.1 Error Analysis

We performed an error analysis to verify – for the best settings of WO in each dataset – the correlation between vector cosine and WO scores (see Table 5), and the peculiarities of the entries with the strongest and the weakest correlation (see Table 4).

We found that WO and vector cosine always have a high correlation (i.e. above 0.80), with the highest correlations reported for McRae and Ferretti-Instruments. Looking at Table 4 we can also observe that:

- the average gold value of the 35 most (4.65) and least (4.56) correlated items does not substantially differ from the average gold value calculated on the full datasets (4.31), meaning that the distribution of likely and unlikely fillers among the best and worst correlated items is similar to the one in the datasets (i.e. no bias can be identified);
- both measures have difficulties on the same test items (probably because of loose semantic constraints), but report their best performances on different pairs (see *Overlap* and *Lexemes* columns);
- syntactically, vector cosine correlates better with objects, while WO is more balanced between objects and subjects, often showing a preference for the latter (see the distribution in *Syntax* column).

6 Conclusions

In this paper, we have introduced an unsupervised distributional method for modeling predicate-argument thematic fit judgments which works purely on syntactic information.

Dataset	Correlation
McRae	0.88
Padó	0.81
Ferretti - Instruments	0.90
Ferretti - Locations	0.83

Table 5: Correlation between WO and vector cosine in WO best settings for all datasets

The method, inspired by cognitive and psycholinguistic findings, consists in: i) extracting and filtering the most salient *second order contexts* for each verb-specific role, i.e. the most salient semantic dimensions of *typical verb-specific role fillers*; and then ii) estimating the thematic fit as a *weighted overlap* between the top features of the candidate fillers and of the prototypes. Once tested on some popular datasets of thematic fit judgments, our method consistently outperforms a baseline re-implementing the thematic fit model of Baroni and Lenci (2010) and proves to be competitive with state of the art models. It even registered the best performance on the Ferretti-Instruments dataset and it is the second best on the Ferretti-Locations, which were known to be particularly hard to model for dependency-based approaches.

Our method is simple, economic and efficient, it works purely on syntactic dependencies (so it does not require a role-labeled corpus) and achieves good results even with no supervised training. Finally, it offers linguistically and cognitively grounded insights on the process of prototype creation and contextual feature salience, preparing the ground for further speculations and optimizations. For example, future work might aim at identifying strategies for tuning the parameter N to account for the different degrees of selectivity of each verb-specific role. Another possible extension would be the inclusion of a mechanism for updating the role prototypes depending on how the other roles are filled, which would be the key for a more realistic and dynamic model of thematic fit expectations (Lenci, 2011).

Acknowledgments

We would like to thank the anonymous reviewers for their helpful suggestions.

This work has been carried out thanks to the support of the A*MIDEX grant (nANR-11-IDEX-0001-02) funded by the French Government “Investissements d’Avenir” program.

References

Gerry T.M Altmann and Yuki Kamide. 1999. Incremental Interpretation at Verbs: Restricting the Domain of Subsequent Reference . *Cognition* 73(3):247 – 264.

Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. 2009. The WaCky wide web: a Collection of Very Large Linguistically Processed Web-crawled Corpora. *Language Resources and Evaluation* 43(3):209–226.

Marco Baroni, Georgiana Dinu, and Germán Kruszewski. 2014. Dont Count, Predict! A Systematic Comparison of Context-counting vs. Context-predicting Semantic Vectors. In *Proceedings of ACL*. volume 1.

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

Lawrence W Barsalou, Ava Santos, W Kyle Simmons, and Christine D Wilson. 2008. Language and Simulation in Conceptual Processing. *Symbols, embodiment, and meaning* pages 245–283.

Klinton Bicknell, Jeffrey L Elman, Mary Hare, Ken McRae, and Marta Kutas. 2010. Effects of Event Knowledge in Processing Verbal Arguments. *Journal of Memory and Language* 63(4):489–505.

John A. Bullinaria and Joseph P. Levy. 2012. Extracting Semantic Representations from Word Co-occurrence Statistics: Stop-lists, Stemming, and SVD. *Behavior Research Methods* 44(3):890–907.

Kenneth Ward Church and Patrick Hanks. 1990. Word Association Norms, Mutual Information, and Lexicography. *Computational Linguistics* 16(1):22–29.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural Language Processing (Almost) from Scratch. *Journal of Machine Learning Research* 12:2493–2537.

Georgiana Dinu, Angeliki Lazaridou, and Marco Baroni. 2015. Improving Zero-shot Learning by Mitigating the Hubness Problem. In *Proceedings of ICLR*.

Katrin Erk. 2007. A Simple, Similarity-based Model for Selectional Preferences. In *Proceedings of ACL*.

Katrin Erk, Sebastian Padó, and Ulrike Padó. 2010. A flexible, corpus-driven model of regular and inverse selectional preferences. *Computational Linguistics* 36:723–763.

Stefan Evert. 2004. *The Statistics of Word Cooccurrences: Word Pairs and Collocations*. Ph.D. thesis.

Manaal Faruqui, Yulia Tsvetkov, Pushpendre Rastogi, and Chris Dyer. 2016. Problems With Evaluation of Word Embeddings Using Word Similarity Tasks. In *Proceedings of ACL Workshop on Evaluating Vector Space Representations for NLP*.

Todd R. Ferretti, Ken McRae, and Andrea Hatherell. 2001. Integrating Verbs, Situation Schemas, and Thematic Role Concepts . *Journal of Memory and Language* 44(4):516 – 547.

Clayton Greenberg, Vera Demberg, and Asad Sayeed. 2015a. Verb Polysemy and Frequency Effects in Thematic Fit Modeling. In *Proceedings of NAACL Workshop on Cognitive Modeling and Computational Linguistics*.

Clayton Greenberg, Asad B. Sayeed, and Vera Demberg. 2015b. Improving Unsupervised Vector-space Thematic Fit Evaluation via Role-filler Prototype Clustering. In *Proceedings of HLT-NAACL*.

James A Hampton. 1997. Conceptual Combination. *Knowledge, Concepts, and Categories* pages 133–159.

James A. Hampton. 2007. Typicality, Graded Membership, and Vagueness. *Cognitive Science* 31:355–384.

Mary Hare, Michael Jones, Caroline Thomson, Sarah Kelly, and Ken McRae. 2009. Activating Event Knowledge. *Cognition* 111 2:151–67.

Yuki Kamide, Gerry T.M Altmann, and Sarah L Haywood. 2003. The Time-course of Prediction in Incremental Sentence Processing: Evidence from Anticipatory Eye Movements . *Journal of Memory and Language* 49(1):133 – 156.

Walter Kintsch. 2001. Predication. *Cognitive Science* 25:173–202.

Geoffrey Leech. 1992. 100 Million Words of English: the British National Corpus (BNC). *Language Research* 28(1):1–13.

Alessandro Lenci. 2008. Distributional Semantics in Linguistic and Cognitive Research. *Italian Journal of Linguistics* 20(1):1–31.

Alessandro Lenci. 2011. Composing and Updating Verb Argument Expectations: A Distributional Semantic Model. In *Proceedings of ACL Workshop on Cognitive Modeling and Computational Linguistics*.

Omer Levy, Yoav Goldberg, and Ido Dagan. 2015. Improving Distributional Similarity with Lessons Learned from Word Embeddings. *TACL* 3:211–225.

Baoli Li and Liping Han. 2013. Distance Weighted Cosine Similarity Measure for Text Classification. In *Intelligent Data Engineering and Automated Learning*. Springer Berlin Heidelberg, Berlin, Heidelberg, pages 611–618.

Maryellen C MacDonald and Mark S Seidenberg. 2006. Constraint Satisfaction Accounts of Lexical and Sentence Comprehension. *Handbook of psycholinguistics* 2:581–611.

Kazunaga Matsuki, Tracy Chow, Mary Hare, Jeffrey L Elman, Christoph Scheepers, and Ken McRae. 2011. Event-based Plausibility Immediately Influences On-line Language Comprehension. *Journal of experimental psychology. Learning, memory, and cognition* 37 4:913–34.

Ken McRae, Mary Hare, Jeffrey L. Elman, and Todd Ferretti. 2005. A basis for generating expectancies for verbs from nouns. *Memory & Cognition* 33(7):1174–1184.

Ken McRae and Kazunaga Matsuki. 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.

Ken McRae, Michael J. Spivey-Knowlton, and Michael K. Tanenhaus. 1998. Modeling the Influence of Thematic Fit (and Other Constraints) in Online Sentence Comprehension. *Journal of Memory and Language* 38(3):283–312.

George A. Miller and Walter G. Charles. 1991. Contextual Correlates of Semantic Similarity. *Language and Cognitive Processes* 6(1):1–28.

Jeff Mitchell and Mirella Lapata. 2010. Composition in Distributional Models of Semantics. *Cognitive science* 34(8):1388–1429.

Joakim Nivre and Johan Hall. 2005. Maltparser: A Language-independent System for Data-driven Dependency Parsing. In *Proceedings of Workshop on Treebanks and Linguistic Theories*. pages 13–95.

Ulrike Padó. 2007. *The Integration of Syntax and Semantic Plausibility in a Wide-coverage Model of Human Sentence Processing*. Ph.D. thesis.

Ava Santos, Sergio E. Chaigneau, W. Kyle Simmons, and Lawrence W. Barsalou. 2011. Property Generation Reflects Word Association and Situated Simulation. *Language and Cognition* 3(1):83119.

Enrico Santus, Emmanuele Chersoni, Alessandro Lenci, Chu-Ren Huang, and Philippe Blache. 2016a. Testing APSyn against Vector Cosine on Similarity Estimation. In *Proceedings of PACLIC*.

Enrico Santus, Tin-Shing Chiu, Qin Lu, Alessandro Lenci, and Chu-Ren Huang. 2016b. Unsupervised Measure of Word Similarity: how to Outperform Co-occurrence and Vector Cosine in VSMs. In *Proceedings of the AAAI*. AAAI Press, pages 4260–4261.

Enrico Santus, Tin-Shing Chiu, Qin Lu, Alessandro Lenci, and Chu-Ren Huang. 2016c. What a Nerd! Beating Students and Vector Cosine in the ESL and TOEFL Datasets. In *Proceedings of LREC*.

Asad Sayeed and Vera Demberg. 2014. Combining Unsupervised Syntactic and Semantic Models of Thematic Fit. In *Proceedings of CLIC*.

Asad Sayeed, Vera Demberg, and Pavel Shkadzko. 2015. An Exploration of Semantic Features in an Unsupervised Thematic Fit Evaluation Framework. In *Italian Journal of Linguistics*.

Asad Sayeed, Clayton Greenberg, and Vera Demberg. 2016. Thematic Fit Evaluation: an Aspect of Selectional Preferences. In *Proceedings of ACL Workshop for Evaluating Vector Space Representations for NLP*.

Tobias Schnabel, Igor Labutov, David M. Mimno, and Thorsten Joachims. 2015. Evaluation Methods for Unsupervised Word Embeddings. In *Proceedings of EMNLP*.

W Kyle Simmons, Stephan B Hamann, Carla L Harenski, Xiaoping P Hu, and Lawrence W Barsalou. 2008. fMRI Evidence for Word Association and Situated Simulation in Conceptual Processing. *Journal of Physiology* 102 1-3:106–19.

Ottokar Tilk, Vera Demberg, Asad B. Sayeed, Dietrich Klakow, and Stefan Thater. 2016. Event Participant Modelling with Neural Networks. In *Proceedings of EMNLP*.

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

Bram Vandekerckhove, Dominiek Sandra, and Walter Daelemans. 2009. A Robust and Extensible Exemplar-Based Model of Thematic Fit. In *Proceedings of EACL*.

Wenchi Yeh and Lawrence W Barsalou. 2006. The Situated Nature of Concepts. *The American Journal of Psychology* 119:3:349–84.