Event Knowledge in Sentence Processing: A New Dataset for the Evaluation of Argument Typicality

Paolo Vassallo, Emmanuele Chersoni, Enrico Santus, Alessandro Lenci, Philippe Blache

Event Knowledge in Sentence Processing: A New Dataset for the Evaluation of Argument Typicality

Paolo Vassallo*, Emmanuele Chersoni†, Enrico Santusˆ, Alessandro Lenci*, Philippe Blache†
University of Pisa*, Aix-Marseille University†, Massachusetts Institute of Technologyˆ
paolovassa@virgilio.it, emmanuelechersoni@gmail.com
esantus@mit.edu, alessandro.lenci@unipi.it, blache@lpl-aix.fr

Abstract

In the NLP literature, the *thematic fit estimation* task is defined as the task in which a system has to predict how likely a candidate argument (e.g. *cop*) is to fit a given a verb-specific role (e.g. the agent of *to arrest*) (Santus et al., 2017).
Because of the scarcity of benchmark datasets, thematic fit models are currently evaluated by measuring the correlation between their output and human ratings for isolated verb-filler pairs (Sayeed et al., 2016). However, such evaluation does not account for the dynamic nature of argument expectations: there is robust psycholinguistic evidence that human update their predictions on upcoming arguments during sentence processing, depending on the way other verb arguments are filled (Bicknell et al., 2010; Matsuki et al., 2011). Consider, for example, how the expectation for the patient of *to check* would change if we use *journalist* or *mechanic* as agents.
In this paper we introduce DTFit (Dynamic Thematic Fit), a dataset of human ratings for verb-role fillers in a given event context, with the aim of providing a rigorous benchmark for context-sensitive argument typicality modeling. The dataset accounts for the plausibility of patient, instrument and location roles, given the agent and the predicate.

Keywords: thematic fit modeling, distributional semantics, argument expectations, computational psycholinguistics, sentence processing, linguistic resources

1. Introduction

The psycholinguistic literature of the last two decades has brought extensive evidence for the cognitive relevance of the notion of *thematic fit*, that is to say the degree to which a given lemma fits in a given verb-specific role. A number of studies reported behavioral effects proving that, during on-line sentence processing, hearing a verb induces human subjects to activate expectations about nouns typically filling its thematic roles, and argument nouns in turn activate expectations about their typical predicates and typical co-arguments (McRae et al., 1998; Ferretti et al., 2001; McRae et al., 2005; McRae and Matsuki, 2009; Hare et al., 2009). These findings have been explained by researchers in the light of a *Generalized Event Knowledge* contained in the human semantic memory, which includes information about events and their participants (see Figure 1 for a summary of the priming effects). Such knowledge is activated by lexical cues in the sentences, and it is exploited by human subjects to anticipate the upcoming linguistic input (McRae and Matsuki, 2009).

More recent studies by Bicknell et al. (2010) and Matsuki et al. (2011) showed that verb argument expectations depend on the way other arguments are filled, and they are dynamically updated while the sentence is processed. For example, given the verb *to check*, if *journalist* is the filler of the agent role, then we can expect *spelling* or *report* to be very likely patient fillers. For the same verb, if the agent is *mechanic*, the most likely fillers will be things such as *brakes* and *engine*. Bicknell et al. (2010) presented a self-paced reading and an Event Related Potential (ERP) experiment where they compared sentence pairs differing only for their agents: their results show that sentences with a typicality relation between the agent and the patient are read faster by human and evoke smaller N400

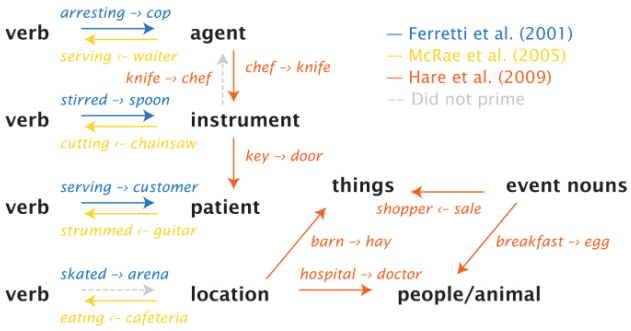

Figure 1: Summary of the experiments on event-based priming, from McRae and Matsuki (2009). The arrows between verb and roles indicate the direction of priming, from the prime to the target.

components[1]. Matsuki et al. (2011) set up a similar experiment with the self-paced reading and the eye-tracking paradigm, but focusing on the typicality relation between instruments and patients (e.g. *She used the shampoo to wash her hair* vs. *She used the shampoo to wash her car*). Coherently, they found significantly faster reading times for patient nouns that were more predictable given a predicate-instrument pair. Moreover, they reported shorter first fixation and gaze duration times for the patient in the eye-tracking experiment.

[1]N400 components are negative ERP deflections that peak around 400 milliseconds after the presentation of a stimulus word. The amplitude of the component elicited by a word has been found to be in an inverse relationship with its cloze probability (Kutas and Hillyard, 1984), and thus it has been considered as an index of the difficulty of integrating the meaning of a word in a given semantic context (Baggio et al., 2012).

The phenomenon of the thematic fit has recently raised interest also in the NLP community. Several studies have developed methods for the automatic estimation of the compatibility between candidate arguments and verb roles, generally adopting an evaluation based on the correlation between system predictions and human ratings. However, most of such work did not take into account the dynamic aspect of the phenomenon, i.e. the fact that the plausibility of arguments changes as the other roles are filled. The main reason behind such limitation is that the current gold standards mostly consist of simple ratings of verb-argument pairs in isolation, and do not take into account how the typicality scores change in function of the other event participants. In the present contribution, we precisely aim to address this issue by introducing the **DTFit dataset** (Dynamic Thematic Fit), a resource that has been built by specifically asking human subjects to produce plausible fillers for verb roles. Similarly to the previous literature, we collect data for the following roles: *patient, instrument* and *locations*, given the agent and the predicate. Our event tuples describe typical and atypical events, differing by just one argument (either the patient, the instrument or the location), and they are associated with human judgements collected in a Crowdflower task. Currently, we are still expanding the dataset, as we started the collection of judgements for new sets of tuples including new instruments and locations. Another planned expansion will regard the *time* role.

The paper is organized as follows. First, we illustrate the methodology and the criteria that guided the data collection, providing some statistical information about the dataset. Then, we will describe two approaches for the evaluation of thematic fit models on DTFit, showing the usefulness of our dataset.

2. Related Work

The thematic fit task, in the last decade, has been typically addressed by means of Distributional Semantic Models (DSMs). To the best of our knowledge, Erk et al. (2010) were the first authors to introduce the evaluation of a thematic fit model in terms of the correlation with human-elicited ratings. The authors used a syntax-based DSM to compute the plausibility of each verb role-filler pair as the similarity between a candidate filler and previously attested fillers for the same role. Finally, they measured the correlation of the system scores with a gold standard consisting of the human judgments collected by McRae et al. (1998) and Padó (2007).

One of the most influential frameworks for thematic fit modeling was the Distributional Memory by Baroni and Lenci (2010) (DM), which is also based on a syntax-based DSM. In the approach adopted by the authors, for each verb-specific role a *prototype vector* is built by averaging the syntax-based vectors of the most typical role fillers. The higher the cosine similarity of a noun with a role prototype, the higher its plausibility as a filler for that role.

Despite its simplicity, the method by Baroni and Lenci (2010) proved to be extremely effective, and inspired several extensions. Sayeed et al. (2015), for example, tried to improve the prototype representation by using vector features based on semantic roles, instead of syntactic dependencies. Moreover, they were the first to test to evaluate the plausibility of the fillers for roles other than the agent and the patient one, by introducing in the literature the Ferretti datasets for instruments and locations (Ferretti et al., 2001). Some other works aimed at addressing the problem of verb polysemy, either by obtaining different prototypes for the different senses through the hierarchical clustering of the fillers (Greenberg et al., 2015), or by testing similarity metrics based on a weighted feature overlap between the dimensions of the vectors (Santus et al., 2017).

It should be pointed out that all the above-mentioned works compare the scores of their systems with human rating for role-filler pairs in isolation: for example, given the patient role of the verb *to cut*, the rating quantify how good is *meat* as a filler. But as we anticipated above, the fitness of a filler depends also on the general event context: if we knew that the agent in the *cut*-event is a *government*, we would probably expect patients like the *taxes*, the *spending*, the *aids* etc. This aspect of dynamic update of the expectation on the fillers, at the present state, has received relatively little attention in the literature.

One of the few proposals addressing the dynamic update was given by Lenci (2011), who extended the original DM model (Baroni and Lenci, 2010) to account for the composition and update of argument expectations. Lenci tested an additive and a multiplicative model of vector composition (Mitchell and Lapata, 2010) to model the agent-related change in the expectations on the patient filler, by using a dataset derived from the sentences of the Bicknell experiment (Bicknell et al., 2010). The Bicknell sentences were turned into subject-verb-object triplets, such as *journalist-check-spelling*, and a binary classification task was set up for the evaluation. More concretely, the system had to measure the plausibility of a patient for pairs of triplets that differ only for the agent noun. It is important to notice that all triplets presented plausible patients with respect to the given predicates: only the agent made the patients of the respective triples more or less plausible. Therefore, the triples in each pair were found either in the *typical* or in the *atypical* condition, as in the following example:

(1) a. *journalist-check-spelling* (typical)
    b. *mechanic-check-spelling* (atypical)

The goal, for each pair, was to identify the triple describing the most typical situation and in the end a global accuracy score was computed for each model.

Another system that was tested on the task of the argument expectation was the neural network architecture by Tilk et al. (2016), which was trained to generate probability distributions over selectional preferences for each thematic role. The authors used the Bicknell dataset as a benchmark, obtaining performances comparable to the multiplicative model by Lenci (2011). The same dataset was finally used by Chersoni et al. (2017), who have implemented some variations of Lenci (2011)'s system to demonstrate that DSMs benefit from structural information (i.e. syntactic information, to be intended as opposite of bag-of-words and bag-of-arguments hypotheses) when composing and updating thematic fit expectations.

To sum up, only a few studies so far have addressed the

problem of dynamic argument expectations, and the Bicknell triplets are currently the only available standard for testing their models.

3. The DTFit Dataset: The Data Collection Procedure

The only benchmark for the task of the argument expectation update, the Bicknell dataset, is limited in the sense that it allows evaluation only in terms of binary choice, i.e. it tests systems just on the capability of recognizing which argument combinations out of two is more typical, and it includes only agent and patient fillers. On the other hand, traditional thematic fit datasets include a wider variety of roles and more fillers for each role, also allowing researchers to perform an evaluation in terms of correlation (i.e., typicality is conceived as a score in a continuum rather than as a binary choice). However, such datasets only consist of verb-specific role-filler pairs and do not take the event context into account. Ideally, the DTFit dataset should combine the qualities of both resources.

In the sentence processing literature, several findings related to argument typicality have been shown to involve both aspects: the update of the expectation based on the event context (the saturation of an role can make a potential filler of another role more or less likely) and the priming relations (see the summary in Figure 1) between the events and the fillers of a wide variety of roles (McRae and Matsuki, 2009; Bicknell et al., 2010; Matsuki et al., 2011; Paczynski and Kuperberg, 2012).

Since our resource has the the advantages of both evaluation strategies (the information on event typicality and a more complex event context on the one hand, human ratings on multiple fillers for a given role on the other hand), we believe it will a useful benchmark for linking distributional models of event knowledge and experimental results.

3.1. Agents and Patients

To start our data collection, we parsed the corpus of image descriptions introduced by (Young et al., 2014). We decided to use this corpus as we wanted to have human-generated descriptions of typical visual scenes (i.e. images taken from Flickr). Then we have extracted from the corpus a list of verb-patient pairs, and we have selected 329 pairs for which it seemed intuitive to imagine a typical agent for the given scenario. For each pair, we produced a typical agent. Then, we created a another set of triples by replacing the original patient of each triple, in order to obtain corresponding atypical combinations (examples in Table 1).

agent	verb	patient	condition
mason	build	house	typical
mason	build	snowman	atypical
cook	clean	fish	typical
cook	clean	window	atypical

Table 1: Examples of triples produced starting from the pairs *build house* and *clean fish*.

In a second phase, we set up two Crowdflower task to obtain typicality ratings both for our agent-verb pairs and for our triples. Collecting judgements for agents and verbs alone was necessary, of course, to check that the perceived typicality of the event was depending on the noun filling the patient role.

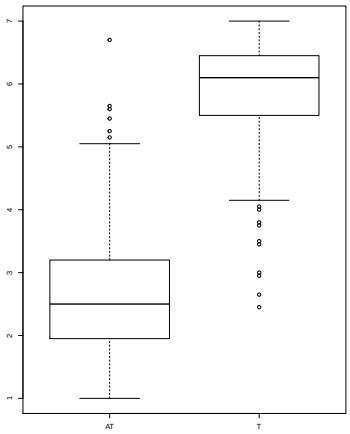

Figure 2: Comparison between the ratings for the atypical (AT, on the left) and the typical triples (T, on the right) in the Patients dataset.

As for the first task, we created two sets of 160 and 159 pairs, respectively. Each subset was rated by a group of 20 native speakers of British or American English. The subjects had to answer questions in the form *How common is for a **mason** to **build** something?*, by assigning a score between 1 (very uncommon, very atypical) and 7 (very common, very typical).

The second task was also taken by groups of 20 native speakers of British and American English. The triples were splitted in four subsets of 168, 168, 161 and 160 items, respectively, equally divided between typical and atypical ones. In this case, the questions had the form *How common is for a **mason** to **build** a **house**?*, and the subject had to provide an answer by using a seven-level Likert scale, as in the previous task.

As a check, we introduced 8 synonymy question for each test set, with the goal of filtering out the answers provided by trolls or non-attentive users. All the questions had the form *can **x** and **y** mean the same thing?* (e.g. *can "help" and "entertain" mean the same thing?*), and they were randomly presented to the subjects while taking the test. The responses of the subjects having less than a 70% accuracy in answering these questions were automatically excluded.

With this strategy, we obtained typicality ratings for all our 657 triples, so we had to check the two conditions differ significantly. We compared the scores for typical and atypical conditions with the Wilcoxon rank sum test, and the test confirmed that the ratings for the former are significantly higher ($W = 106186.5, p < 2.2e-16$; see the boxplots in Figure 2).

3.2. Instruments and Locations

As we already mentioned in the Introduction, processing advantages were not found only for typical agent-patient combinations, but also for other roles, e.g. Matsuki et al.

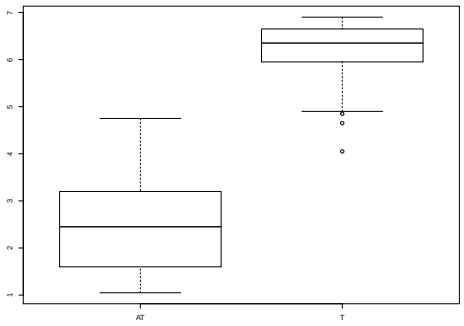

Figure 3: Comparison between the ratings for the atypical (AT, on the left) and the typical triples (T, on the right) in the Instruments dataset.

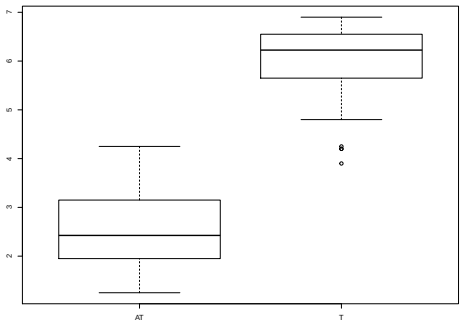

Figure 4: Comparison between the ratings for the atypical (AT, on the left) and the typical triples (T, on the right) in the Locations dataset.

(2011) found them also for sentences in which the patient was more predictable given a verb and an instrument.
From our dataset of agents, verbs and patients, we have thus selected two subsets of 50 triples, for which it was easy to imagine, respectively, typical *instruments* and typical *locations*. For each subset, we generated 100 quadruples by adding either a typical or an atypical argument to each triple (examples are shown in Table 2).

triple	**argument**	**role**	**condition**
mason mix cement	trowel	instrument	typical
mason mix cement	spoon	instrument	atypical
student drink beer	pub	location	typical
student drink beer	classroom	location	atypical

Table 2: Examples of quadruples produced by adding instruments and locations to the dataset triples *mason mix cement* and *student drink beer*.

We asked our subjects to rate the quadruples in the two dataset splits. The question, for each experimental item, was built according to the following pattern:

- how common is for a **agent** to use a **instrument** to **verb** a **patient**? (e.g. how common is for a *mason* to use a *trowel* to *mix cement*?)
- how common is for a **agent** to **verb** a **patient** in a **location**? (e.g. how common is for a *student* to *drink beer* in a *pub*?)

Also for these datasets, the test was taken by 20 native speakers of British or American English and synonymy questions were presented to the subjects as a check, as we previously described. The Wilcoxon rank sum test finally revealed significant differences between typical and atypical condition both for the Instruments ($W = 5, p < 2.2e-16$) and for the Locations dataset ($W = 6.5, p < 2.2e-16$).

3.3. Dataset Description

The current version of the dataset consists of three files:

- 656 triples of agents, verbs and patients (*Patients* dataset);
- 100 quadruples of agents, verbs, patients and locations (*Locations* dataset);
- 100 quadruples of agents, verbs, patients and instruments (*Instruments* dataset).

4. Evaluation Strategies

Previous studies on distributional models for thematic fit evaluated the systems either by measuring the correlation with human judgements or in a classification task. With our dataset, both evaluation strategies are possible:

- given a triple or a verb-filler pair, a system has to output a typicality/probability score and the performance will be evaluated as the correlation between the output scores and the human ratings that we collected in our Crowdflower tasks. Our resource allows to evaluate the typicality for verb-role fillers in isolation (e.g. the ratings for agent-verb and verb-patient combinations will be also made available in the future), but also to compose the expectations for an argument given a verb and the filler noun of another role (reflected by the rating of the entire triple);
- for each pair of triples sharing the agent and the verb, the system has to identify the triple with the higher score. Notice that this typicality has been shown to correspond to a processing advantage of the typical triples over the atypical ones, in terms of shorter reading times and of reduced amplitudes of the N400 component (Bicknell et al., 2010).

4.1. Baselines

In order to verify the quality of the dataset, we tested it by means of two DSMs derived from the approaches to thematic fit estimation by Lenci (2011) and Santus et al. (2017).
Both start by creating a prototype of the filler for a given verb-specific role by summing the distributional vectors of its typical fillers, updating the prototype on the basis of the information coming from the nouns saturating the other roles, and finally calculating the vector cosine with a candidate word. The latter adopts the same principle,

but it uses APSyn as similarity measure between the prototype and the candidate filler.[2] Both the systems were tested on two DSMs, namely the well known Distributional Memory (DM, Baroni and Lenci (2010)) and a dependency based DSM built from the co-occurrences extracted from the British National Corpus (Leech, 1992) and from the Wacky corpus (Baroni et al., 2009).

4.2. Experiments

DSMs. We have implemented two DSMs based on syntactic dependencies. One is based on the data of Distributional Memory (DM, Baroni and Lenci (2010)) and it includes co-occurrences between 30,490 target words and the same words in some syntactic relation with the target (since our space is just a slice of the original DM tensor, contexts have the form *dependency:word*). The other DSM was similarly built on the co-occurrences between 30,063 target words (we have selected the 30K most frequent nouns and verbs in our corpora, plus the words in the datasets) and the same words in some syntactic context.
This latter model, that we called **DEPS**, is a purely dependency-based model, in the sense that all the contexts have been automatically extracted as a syntactic co-occurrence between words. **DM**, on the other hand, has been enhanced with some manually-selected lexical patterns (e.g. *is-a, such-as* etc.).
TASKS AND EVALUATION. We measured the thematic fit for the three parts of our dataset:

1. the fitness of *Patients*, given the agents and the predicates (e.g. predict how likely is *toenail* as patient of *woman paint*);

2. the fitness of *Instruments*, given the agents, the predicates and the patients (e.g. predict how likely is *tray* as instrument of *waiter deliver drink*);

3. the fitness of *Locations*, given the agents, the predicates and the patients (e.g. predict how likely is *pub* as location of *student drink beer*).

The performances are evaluated in terms of both correlation of the scores and binary classification (i.e. typical tuples should get higher thematic fit scores than atypical ones, therefore we measure the accuracy of a system in assigning higher values to typical tuples). The first evaluation consists, concretely, in assessing the Spearman correlation between the scores delivered by our systems and the human ratings (see Section 3.1 and 3.2). The second evaluation consists in measuring the Accuracy of a each system in assigning a higher thematic fit score to typical tuples. This means that, for each dataset pair of tuples sharing the same verb and all the arguments but one, we score a hit each time the thematic fit score of the typical tuple is higher than the one of the corresponding atypical tuple (e.g. the score of *student drink beer pub* should be higher than *student drink beer classroom*).

[2]In their original paper, Santus and colleagues have also filtered the vectors according to certain syntactic relations, demonstrating that some relations contribute more than others to the identification of the similarity between the prototype and the candidate filler. We have however ignored this filtering step in our re-implementation.

Semantic Role	Syntactic Relation
Agent	Subject
Patient	Direct Object
Instrument	Complement introduced by *with*
Location	Complement introduced by *in,on,at*

Table 3: Summary of the syntactic relations that we used to select the typical role fillers.

PROTOTYPE. Following the method introduced by Baroni and Lenci (2010) and adapted by Santus et al. (2017), we measured the thematic fit as the similarity between the candidate filler and a prototype.
The prototype is either the sum or the multiplication between the sub-prototypes, which are vectors containing the sum of the distributional vectors of the most typical fillers for a role, given either the predicate or another argument. These role fillers are identified by means of a syntactic relation, which is used as an approximation of a deeper semantic role (the role-dependency mapping is summarized in Table 3): given a target word and a role, the k typical fillers are those with the highest PLMI association score (Evert, 2004) with the corresponding syntactic relation.[3] As in Baroni and Lenci (2010), we set $k = 20$ for all our models.
As an example, consider the computation of thematic fit for a triple like *mechanic check engine*:

- calculate the prototype of the most typical patient of *to check*, we select the 20 most typical objects of the verbs and sum their vectors;

- for the agent *mechanic*, we create another prototype by summing the vectors of the 20 most typical objects co-occurring with such an argument;

- the two prototypes are combined by either vector addition (**Add**) or vector pointwise multiplication (**Mult**);

- the resulting prototype is fed to the similarity measure, which calculates how similar it is to the candidate filler (in our case, *engine*).

In dataset 1. we have two "partial" prototypes to be combined, in 2. and 3. we have three of them. In other words, each additional argument introduces new information about the role to be predicted, and this information is encoded by means of a new prototype.
SIMILARITY MEASURES. The similarity measures adopted as thematic fit predictors are vector cosine, which is a standard metric for Distributional Models (Turney and Pantel, 2010), and APSyn Santus et al. (2017), which calculates the sum of the inverse of the average rank for each of the top N intersected features between two target vectors. As a value for this parameter, we present the results for $N = 2000$: this parameter value is a common choice in the previous literature and, also in this case, it gave the most stable performances across settings.

[3]Notice that the two models, **DM** and **DEPS**, use different labels to encode the relations, with different granularity.

DSM	*Measure*	*Patients*		*Locations*		*Instruments*	
		Add	*Mult*	*Add*	*Mult*	*Add*	*Mult*
DM	*Cosine*	**0.315**	0.29	**0.2**	0.17	**0.2**	0.11
	APSyn	0.27	0.29	0.17	0.13	0.127	0.146
DEPS	*Cosine*	0.287	0.22	0.17	0.12	**0.105**	0.05
	APSyn	0.33	**0.36**	0.13	**0.278**	0.04	0.09

Table 4: Spearman Correlation. In bold the best results by dataset and DSM; in bold and underlined the best scores by dataset.

Matrix	*Measure*	*Patient*		*Location*		*Instrument*	
		Add	*Mult*	*Add*	*Mult*	*Add*	*Mult*
DM	*Cosine*	67.97%	**69.28%**	61.22%	**63.26%**	**60%**	55.5%
	APSyn	65.03%	68.3%	57.14%	59.18%	55.5%	55.5%
DEPS	*Cosine*	68.67%	62.34%	54%	44%	**60%**	54%
	APSyn	66.77%	**73.4%**	62%	**66%**	50%	52%

Table 5: Accuracy in the binary classification task. In bold the best results by dataset and DSM; in bold and underlined the best scores by dataset.

4.3. Results and Analysis

Table 4 shows the results for the evaluation in terms of Spearman correlation. At a glance, it is clear that the correlation scores of our models in all settings are very low, proving that the task is a difficult one for DSMs. In particular, for the Instruments dataset no model achieve a correlation score above 0.2. This could be due to the fact that Instruments are often not expressed in event descriptions, and this could have led to the creation of more sparse prototype vectors for this dataset.

Concerning the performance, two models seem to perform more consistently: the additive models based on DM and vector cosine, and the multiplicative models based on DEPS and APSyn. The latter ones seem to take advantage from the multiplication operation, which sets to zero all the dimensions that are not shared by all sub-prototypes, and provides a similarity estimation based only on the dimensions that are "relevant" for all the other arguments.

As for the results for the binary classification task, they are shown in Table 5. Again, DM with cosine and addition and DEPS with APSyn and multiplication seem to perform more consistently than the others. Even in this case, the lowest performances overall are reported on the Instrument dataset, which confirms itself as the most difficult to model. If we consider the two evaluation tasks together, it is clear that thematic fit estimation is a complex task for DSMs: for Instruments and Locations tuples, the correlation values with human judgements are extremely low and in the classification task no model manages to do significantly better than random guessing. [4] Future research on this topic might try to address the problem with more sophisticated approaches, i.e. neural network modeling.

5. Conclusion

In this contribution we have introduced DTFit, a new dataset for the evaluation of thematic fit estimation. The dataset has been designed having in mind the dynamic nature of the phenomenon, with the specific goal of providing a resource that allows for the evaluation of context-sensitive argument typicality. We used our dataset to test two different models, which has been shown in the previous literature to perform very well in the traditional evaluation settings for the thematic fit task. The results showed that our dataset is a challenging benchmark for classic syntax-based DSMs, and probably more sophisticated approaches will be required to improve modeling performances.

In the end, we are convinced that thematic fit modeling is an important task for bridging the gap between computational models and experimental results, and that the notion of distributional similarity can be used to model phenomena related to argument expectations (i.e. reduced reading times, or reduced N400 amplitudes for predictable arguments). We hope that our resource will turn out to be a useful tool for the research in computational psycholinguistics going in this direction.

[4] Verified with the Chi-Square test: for the best classifier, $p > 0.1$.

6. Acknowledgements

Emmanuele Chersoni's research is supported by the A*MIDEX grant (n. ANR-11-IDEX-0001-02) funded by the French Government "Investissements d'Avenir" program.

7. Bibliographical References

Baggio, G., Van Lambalgen, M., and Hagoort, P. (2012). The Processing Consequences of Compositionality. In *The Oxford Handbook of Compositionality*, pages 655–672. Oxford University Press.

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

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.

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.

Chersoni, E., Santus, E., Blache, P., and Lenci, A. (2017). Is Structure Necessary for Modeling Argument Expectations in Distributional Semantics? In *Proceedings of IWCS*.

Erk, K., Padó, S., and Padó, U. (2010). A Flexible, Corpus-driven Model of Regular and Inverse Selectional Preferences. *Computational Linguistics*, 36(4):723–763.

Evert, S. (2004). *The Statistics of Word Cooccurrences: Word Pairs and Collocations*. Ph.D. thesis.

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 *Proceedings of HLT-NAACL*, pages 21–31.

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

Kutas, M. and Hillyard, S. A. (1984). Brain Potentials during Reading Reflect Word Expectancy and Semantic Association. *Nature*, 307(5947):161.

Leech, G. N. (1992). 100 million words of english: the british national corpus (bnc).

Lenci, A. (2011). Composing and Updating Verb Argument Expectations: A Distributional Semantic Model. In *Proceedings of the ACL Workshop on Cognitive Modeling and Computational Linguistics*, pages 58–66.

Matsuki, K., Chow, T., Hare, M., Elman, J. L., Scheepers, C., and McRae, K. (2011). Event-based Plausibility Immediately Influences On-line Language Comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 37(4):913.

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 On-line Sentence Comprehension. *Journal of Memory and Language*, 38(3):283–312.

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.

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. *Journal of 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*. Ph.D. thesis.

Santus, E., Chersoni, E., Lenci, A., and Blache, P. (2017). Measuring Thematic Fit with Distributional Feature Overlap. In *Proceedings of EMNLP*.

Sayeed, A., Demberg, V., and Shkadzko, P. (2015). An Exploration of Semantic Features in an Unsupervised Thematic Fit Evaluation Framework. *Italian Journal of Computational Linguistics*, 1(1).

Sayeed, A., Greenberg, C., and Demberg, V. (2016). Thematic Fit Evaluation: An Aspect of Selectional Preferences. In *Proceedings of ACL Workshop on Evaluating Vector Space Representations for NLP*.

Tilk, O., Demberg, V., Sayeed, A., Klakow, D., and Thater, S. (2016). Event Participant Modelling with Neural Networks. In *Proceedings of EMNLP*.

Turney, P. D. and Pantel, P. (2010). From Frequency to Meaning: Vector Space Models of Semantics. *Journal of artificial intelligence research*, 37:141–188.

Young, P., Lai, A., Hodosh, M., and Hockenmaier, J. (2014). From Image Descriptions to Visual Denotations: New Similarity Metrics for Semantic Inference over Event Descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78.