Thematic fit evaluation: an aspect of selectional preferences

Asad Sayeed, Clayton Greenberg, and Vera Demberg
Computer Science / Computational Linguistics and Phonetics
Saarland University
66117 Saarbrücken, Germany
{asayeed,claytong,vera}@coli.uni-saarland.de

Abstract

In this paper, we discuss the *human thematic fit judgement correlation* task in the context of real-valued vector space word representations. Thematic fit is the extent to which an argument fulfils the selectional preference of a verb given a role: for example, how well "cake" fulfils the patient role of "cut". In recent work, systems have been evaluated on this task by finding the correlations of their output judgements with human-collected judgement data. This task is a representation-independent way of evaluating models that can be applied whenever a system score can be generated, and it is applicable wherever predicate-argument relations are significant to performance in end-user tasks. Significant progress has been made on this cognitive modeling task, leaving considerable space for future, more comprehensive types of evaluation.

1 Introduction

In this paper, we discuss a way of evaluating real-valued semantic representations: *human thematic fit judgement correlations*. This evaluation method permits us to model the relationship between the construction of these semantic representation spaces and the cognitive decision-making process that goes into predicate-argument compositionality in human language users. We focus here on verb-noun compositionality as a special case of thematic fit judgement evaluation.

A verb typically evokes expectations regarding the participants in the event that the verb describes. By generalizing over different verbs, we can create a scheme of *thematic roles*, which characterize different ways to be a participant. Schemes vary, but most contain *agent*, *patient*, *instrument*, and *location* (Aarts, 1997). The verb "cut" creates an expectation, among others, for a *patient* role that is to be fulfilled by something that is cuttable. This role-specific expectation is called the *patient selectional preference* of "cut". The noun "cake" fulfils the patient selectional preference of "cut", "form" less so. As such, we can see that selectional preferences are likely to be graded.

We define *thematic fit* to be the extent to which a noun fulfils the selectional preference of a verb given a role. This can be quantified in *thematic fit ratings*, human judgements that apply to combinations of verb, role, and noun[1]. One of the goals of this type of evaluation is both for cognitive modeling and for future application. From a cognitive modeling perspective, thematic fit judgements offer a window into the decision-making process of language users in assigning semantic representations to complex expressions. Psycholinguistic work has shown that these introspective judgements map well to underlying processing notions (Padó et al., 2009; Vandekerckhove et al., 2009).

One of our goals in developing this type of evaluation is to provide another method of testing systems designed for applications in which predicate-argument relations may have a significant effect on performance, especially in user interaction. This particularly applies in tasks where non-local dependencies have semantic relevance, for example, such as in judging the plausibility of a candidate coreferent from elsewhere in the discourse. Such applications include statistical sentence generation in spoken dialog contexts, where systems must make plausible lexical choices in context. This is particularly important as dialog systems grow steadily less task-specific. Indeed, applications that depends on predicting or generating match-

[1] Sometimes roles can be fulfilled by clausal arguments, which we leave for the future.

Proceedings of the 1st Workshop on Evaluating Vector Space Representations for NLP, pages 99–105,
Berlin, Germany, August 12, 2016.

ing predicate-argument pairs in a human-plausible way, such as question-answering, summarization, or machine translation, may benefit from this form of thematic fit evaluation.

Both from the cognitive modeling perspective and from the applications perspective, there is still significant work to be done in constructing models, including distributional representations. We thus need to determine whether and how we can find judgements that are a suitable gold standard for evaluating automatic systems. We seek in this paper to shed some light on the aspects of this problem relevant to vector-space word representation and to highlight the evaluation data currently available for this task.

This task differs from other ways of evaluating word representations because it focuses partly on the psychological plausibility of models of predicate-argument function application. Analogy task evaluations, for example, involve comparisons of word representations that are similar in their parts of speech (Mikolov et al., 2013b). Here we are evaluating relations between words that are "counterparts" of one another and that exist overall in complementary distribution to one another. There are other forms of evaluation that attempt to replicate role assignments or predict more plausible role-fillers given observed text data (Van de Cruys, 2014), but this does not directly capture human biases as to plausibility: infrequent predicate-argument combinations can nevertheless have high human ratings. Consequently, we view this task as a useful contribution to the family of evaluations that would test different aspects of general-purpose word representations.

2 Existing datasets

The first datasets of human judgements were obtained in the context of a larger scientific discussion on human sentence processing. In particular, McRae et al. (1998) proposed incremental evaluation of thematic fit for the arguments in potential parses as a method of parse comparison. Human judgements of thematic fit were needed for incorporation into this model.

McRae et al. (1997) solicited thematic fit ratings on a scale from 1 (least common) to 7 (most common) using "How common is it for a {*snake, nurse, monster, baby, cat*} to *frighten* someone/something?" (for agents) and "How common is it for a {*snake, nurse, monster, baby, cat*} to *be frightened by* someone/something?" (for patients). A small sample of scores from this dataset is given in Table 1. Each (*role-filler, verb, role*) triple received ratings from 37 different participants. The 37 ratings for each triple were averaged to generate a final thematic fit score. The verbs were all transitive, thus allowing an agent rating and patient rating for each verb-noun pair. As shown, many nouns were chosen such that they fit at least one role very well. This meant that some verb-roles in this dataset have no poorly-fitting role-fillers, e.g., patients of "accept" and "agents of "admire". This had strong ramifications for the "difficulty" of this dataset for correlation with automatic systems because extreme differences in human judgements are much easier to model than fine-grained ones.

verb	role-filler	agent	patient
accept	friend	6.1	5.8
accept	student	5.9	5.3
accept	teenager	5.5	4.1
accept	neighbor	5.4	4.4
accept	award	1.1	6.6
admire	groupie	6.9	1.9
admire	fan	6.8	1.7
admire	disciple	5.6	4.1
admire	athlete	4.8	6.4
admire	actress	4.6	6.4

Table 1: Sample of McRae et al. (1997) ratings.

MST98, a 200 item subset of the McRae et al. (1997) dataset created for McRae et al. (1998), has two animate role-fillers for each verb. The first was a good agent and a poor patient, and the other a poor agent and a good patient. The ratings were still well-distributed, but these conditions made correlation with automatic systems easier.

Ferretti et al. (2001) created a dataset of 248 instrument ratings (**F-Inst**) and a dataset of 274 location ratings (**F-Loc**) using questions of the form "How common is it for someone to use each of the following to perform the action of *stirring*?" (instruments) and "How common is it for someone to *skate* in each of the following locations?". 40 participants supplied ratings on a seven point scale.

Ken McRae, Michael Spivey-Knowlton, Maryellen MacDonald, Mike Tanenhaus, Neal Pearlmutter and Ulrike Padó compiled a master list of thematic fit judgements from Pearlmutter and MacDonald (1992), Trueswell et al. (1994),

McRae et al. (1997), a replication of Binder et al. (2001) [Experiment B], and follow-up studies of Binder et al. (2001) [Experiment C]. These studies had slightly different requirements for the kinds of verbs and nouns used and significant overlap in stimuli due to collaboration. This represents the largest to-date dataset of agent-patient thematic fit ratings (1,444 single-word verb/noun judgements), referenced herein as **MSTNN**.

Padó (2007) created a new dataset of 414 agent and patient ratings (**P07**) to be included in a sentence processing model. The verbs were chosen based on their frequencies in the Penn Treebank and FrameNet. Role-fillers were selected to give a wide distribution of scores within each verb. The final dataset contains fine-grained distinctions from FrameNet, which many systems map to familiar agent and patient roles. Judgements were obtained on a seven point scale using questions of the form "How common is it for an *analyst* to *tell* [something]?" (subject) and "How common is it for an *analyst* to be *told*?" (object).

Finally, Greenberg et al. (2015a) created a dataset of 720 patient ratings (**GDS-all**) that were designed to be different from the others in two ways. First, they changed the format of the judgement elicitation question, since they believed that asking how common/typical something is would lead the participants to consider frequency of occurrence rather than semantic plausibility. Instead, they asked participants how much they agreed on a 1-7 scale with statements such as "*cream* is something that is *whipped*". This dataset was constructed to vary word frequency and verb polysemy systematically; the experimental subset of the dataset contained frequency-matched monosemous verbs (**GDS-mono**) and polysemous verbs (**GDS-poly**). Synonymous pairs of nouns (one frequent and one infrequent) were chosen to fit a frequent sense, an infrequent sense (for polysemous verbs only), or no senses per verb.

3 Evaluation approaches

The dominant approach in recent work in thematic fit evaluation has been, given a verb/role/noun combination, to use the vector space to construct a prototype filler of the given role for the given verb, and then to compare the given noun to that prototype (Baroni and Lenci, 2010). The prototype fillers are constructed by averaging some number of "typical" (e.g., most common by frequency or by some information statistic) role-fillers for that verb—the verb's vector is not itself directly used in the comparison. Most recent work instead varies in the construction of the vector space and the use of the space to build the prototype.

The importance of the vector space A semantic model should recognize that cutting a cake with an improbable item like a sword is still highly plausible, even if cakes and swords rarely appear in the same genres or discourses; that is, it should recognize that swords and knives (more typically used to cut cakes) are both cutting-instruments, even if their typical genre contexts are different.

Because of their indirect relationship to probability, real-valued vector spaces have produced the most successful recent high-coverage models for the thematic fit judgement correlation task. Even if cakes and swords may rarely appear in the same discourses, swords and knives sometimes may. A robust vector space allows the representation of unseen indirect associations between these items. In order to understand the progress made on the thematic fit question, we therefore look at a sample of recent attempts at exploring the feature space and the handling of the vector space as a whole.

Comparing recent results In table 2, we sample results from recent vector-space modeling efforts in the literature in order to understand the progress made. The table contains:

BL2010 Results from the TypeDM system of Baroni and Lenci (2010). This space is constructed from counts of rule-selected dependency tree snippets taken from a large web crawl corpus, adjusted via local mutual information (LMI) but is otherwise unsupervised. The approach they take generates a vector space above a 100 million dimensions. The top 20 typical role-fillers by LMI are chosen for prototype construction. Some of the datasets presented were only created and tested later by Sayeed et al. (2015) (*) and Greenberg et al. (2015a) (**).

BDK2014 Tests of word embedding spaces from Baroni et al. (2014), constructed via word2vec (Mikolov et al., 2013a). These are the best systems reported in their paper. The selection of typical role-fillers for constructing the prototype role-filler comes from TypeDM, which is not consulted for the vectors themselves.

Dataset	BL2010	BDK2014	GSD2015	GDS2015	SDS2015-avg	SDS2015-swap
P07	28	41	50	-	59	48
MST98	51	27	-	-	-	-
MSTNN	33*	-	36	-	34	25
F-Loc	23*	-	29	-	21	19
F-Inst	36*	-	42	-	39	45
GDS-all	53**	-	-	55	51	50
GDS-mono	41**	-	-	43	-	-
GDS-poly	66**	-	-	67	-	-

Table 2: Spearman's ρ values ($\times 100$) for different datasets with results collected from different evaluation attempts. All models evaluated have coverage higher than 95% over all datasets.

GSD2015 The overall best-performing system from Greenberg et al. (2015b), which is TypeDM from BL2010 with a hierarchical clustering algorithm that automatically clusters the typical role-fillers into verb senses relative to the role. For example, "cut" has multiple senses relative to its patient role, in one of which "budget" may be typical, while in another sense "cake" may be typical.

GSD2015 The overall best-performing system from Greenberg et al. (2015a). This is the same TypeDM system with hierarchical clustering as in GSD2015, but applied to a new set of ratings intended to detect the role of verb polysemy in human decision-making about role-fillers.

SDS2015-avg Sayeed et al. (2015) explore the contribution of semantics-specific features by using a semantic role labeling (SRL) tool to label a corpus similar to that of BL2010 and constructing a similar high-dimensional vector space. In this case, they average the results of their system, SDDM, with TypeDM and find that SRL-derived features make an additional contribution to the correlation with human ratings. Prototypes are constructed using typical role-fillers from the new corpus, weighted, like TypeDM, by LMI.

SDS2015-swap This is similar to SDS2015-avg, but instead, the typical role-fillers of SDDM are used to retrieve the vectors of TypeDM for prototype construction.

It should be emphasized that each of these papers tested a number of parameters, and some of them (Baroni and Lenci, 2010; Baroni et al., 2014) used vector-space representations over a number of tasks. Baroni et al. (2014) found that trained, general-purpose word embeddings—BDK2014—systematically outperform count-based representations on most of these tasks. However, they also found that the thematic fit correlation task was one of the few for which the same word embedding spaces underperform. We confirm this by observing that every system in Table 2 dramatically outperforms BDK2014.

One hint from this overview as to why trained word embedding spaces underperform on this task is that the best performing systems involve very large numbers of linguistically-interpretable dimensions (features)[2]. SDS2015-avg involves the combination of two different systems with high-dimensional spaces, and it demonstrates top performance on the high-frequency agent-patient dataset of Padó (2007) and competitive performance on the remainder of evaluated datasets. SDS2015-swap, on the other hand, involves the use of one high-dimensional space with the typical role-filler selection of another one, and performs comparatively poorly on all datasets except for instrument roles. Note that the typical role-fillers are themselves chosen by the magnitudes of their (LMI-adjusted) frequency dimensions in the vector space itself, relative to their dependency relationships with the given verb, as per the evaluation procedure of Baroni and Lenci (2010). In other words, not only do many meaningful dimensions seem to matter in comparing the vectors, the selection of vectors is itself tightly dependent on the model's own magnitudes.

What these early results in thematic fit evaluation suggest is that, more so than many other kinds

[2]Baroni and Lenci provide a reduction to 5000-dimensions via random indexing (Kanerva et al., 2000) on their web site derived from TypeDM that performs competitively. Most high-performing general-purpose trained word embeddings, including those in (Baroni et al., 2014), have a much smaller dimensionality, and they tend not to be trained from linguistically-rich feature sets.

of lexical-semantic tasks, thematic fit modeling is particularly sensitive to linguistic detail and interpretability of the vector space.

4 Future directions

In the process of proposing this evaluation task, we have presented in this paper an overview of the issues involved in vector-space approaches to human thematic fit judgement correlation. Thematic fit modeling via real-valued vector-space word representations has made recent and significant progress. But in the interest of building evaluations that truly elucidate the cognitive underpinnings of human semantic "decision-making" in a potentially application-relevant way, there are a number of areas in which such evaluations could be strengthened. We present some suggestions here:

Balanced datasets In order to investigate the apparent relationship between the linguistic interpretability of the vector space dimensions and the correlations with human judgements, we need more evaluation data sets balanced for fine-grained linguistic features. The data collected in Greenberg et al. (2015a) is a step in this direction, as it was used to investigate the relationship between polysemy, frequency, and thematic fit, and so it was balanced between polysemy and frequency. However, a thematic role like location—on which all systems reported here perform poorly—could be similarly investigated by collecting data balanced by, for example, the preposition that typically indicates the location relation ("in the kitchen" vs. "on the bus").

Compositionality Both the currently available thematic fit judgements and the vector spaces used to evaluate them are not designed around compositionality, as they have very limited flexibility in combining the subspaces defined by typical role-filler prototypes (Lenci, 2011). Language users may have the intuition that cutting a budget and cutting a cake are both highly plausible scenarios. However, if we were to introduce an agent role-filler such as "child", the human ratings may be quite different, as children are not typical budget-cutters. The thematic fit evaluation tasks of the future will have to consider compositionality more systematically, possibly by taking domain and genre into account.

Perceptuomotor knowledge A crucial question in the use of distributional representations for thematic fit evaluation is the extent to which the distributional hypothesis really applies to predicting predicate-argument relations. Humans presumably have access to world-knowledge that is beyond the mere texts that they have consumed in their lifetimes. While there is evidence from psycholinguistic experimentation that both forms of knowledge are involved in the neural processing of linguistic input (Amsel et al., 2015), the boundary between world-knowledge and distributional knowledge is not at all clear. However, thematic fit judgement data represents the output of the complete system. An area for future work would be to see whether the distinction between these two types of knowledge (such as image data or explicitly-specified logical features) can be incorporated into the evaluation itself. However, the single rating approach has its own advantages, in that we expect an optimal vector-space (or other) representation will also include the means by which to combine these forms of linguistic knowledge.

Rating consistency 240 items, containing the most frequent verbs from the **MSTNN** dataset, were deliberately included in the **GDS-all** dataset, in order to evaluate consistency of judgements between annotators, especially when the elicitation method varied. There was a significant positive correlation between the two sets of ratings, Pearson's $r(238)$ 95% CI [0.68, 0.80], $p < 2.2 \times 10^{-16}$. The residuals appeared normal with homogeneous variances, and the Spearman's ρ was 0.75. This high correlation provides a possible upper-bound on computational estimators of thematic fit. The fact that it is well above the state of the art for any dataset and estimator configuration suggests that there is still substantial room for development for this task.

Acknowledgments

This research was funded by the German Research Foundation (DFG) as part of SFB 1102: "Information Density and Linguistic Encoding" as well as the Cluster of Excellence "Multimodal Computing and Interaction" (MMCI). Also, the authors wish to thank the two anonymous reviewers whose valuable ideas contributed to this paper.

References

Bas Aarts. 1997. *English syntax and argumentation*. St. Martin's Press, New York.

Ben D Amsel, Katherine A DeLong, and Marta Kutas. 2015. Close, but no garlic: Perceptuomotor and event knowledge activation during language comprehension. *Journal of memory and language* 82:118–132.

Marco Baroni, Georgiana Dinu, and Germán Kruszewski. 2014. Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Baltimore, Maryland, pages 238–247.

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

Katherine S. Binder, Susan A. Duffy, and Keith Rayner. 2001. The effects of thematic fit and discourse context on syntactic ambiguity resolution. *Journal of Memory and Language* 44(2):297–324.

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 the 6th Workshop on Cognitive Modeling and Computational Linguistics*. Association for Computational Linguistics, Denver, Colorado, pages 48–57.

Clayton Greenberg, Asad Sayeed, and Vera Demberg. 2015b. Improving unsupervised vector-space thematic fit evaluation via role-filler prototype clustering. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Denver, Colorado, pages 21–31.

Pentti Kanerva, Jan Kristofersson, and Anders Holst. 2000. Random indexing of text samples for latent semantic analysis. In *Proceedings of the 22nd annual conference of the cognitive science society*. Citeseer, volume 1036.

Alessandro Lenci. 2011. Composing and updating verb argument expectations: A distributional semantic model. In *Proceedings of the 2nd Workshop on Cognitive Modeling and Computational Linguistics*. Association for Computational Linguistics, Portland, Oregon, USA, pages 58–66.

Ken McRae, Todd R. Ferretti, and Liane Amyote. 1997. Thematic roles as verb-specific concepts. *Language and cognitive processes* 12(2-3):137–176.

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

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013a. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*. pages 3111–3119.

Tomas Mikolov, Wen tau Yih, and Geoffrey Zweig. 2013b. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2013)*. Association for Computational Linguistics.

Ulrike Padó. 2007. *The integration of syntax and semantic plausibility in a wide-coverage model of human sentence processing*. Ph.D. thesis, Saarland University.

Ulrike Padó, Matthew W. Crocker, and Frank Keller. 2009. A probabilistic model of semantic plausibility in sentence processing. *Cognitive Science* 33(5):794–838.

Neal J. Pearlmutter and Maryellen C. MacDonald. 1992. Plausibility and syntactic ambiguity resolution. In *Proceedings of the 14th Annual Conference of the Cognitive Science Society*. Lawrence Erlbaum Associates, Inc., Hillsdale, New Jersey, pages 498–503.

Asad Sayeed, Vera Demberg, and Pavel Shkadzko. 2015. An exploration of semantic features in an unsupervised thematic fit evaluation frame-

work. *Italian Journal of Computational Linguistics* 1(1).

John C. Trueswell, Michael K. Tanenhaus, and Susan M. Garnsey. 1994. Semantic influences on parsing: Use of thematic role information in syntactic ambiguity resolution. *Journal of memory and language* 33(3):285–318.

Tim Van de Cruys. 2014. A neural network approach to selectional preference acquisition. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. pages 26–35.

Bram Vandekerckhove, Dominiek Sandra, and Walter Daelemans. 2009. A robust and extensible exemplar-based model of thematic fit. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*. Association for Computational Linguistics, Athens, Greece, pages 826–834.