PISA: A Measure of Preference in Selection of Arguments to Model Verb Argument Recoverability

Giulia Cappelli
Scuola Normale Superiore
giulia.cappelli@sns.it

Alessandro Lenci
CoLing Lab
University of Pisa
alessandro.lenci@unipi.it

Abstract

Our paper offers a computational model of the semantic recoverability of verb arguments, tested in particular on direct objects and Instruments. Our fully distributional model is intended to improve on older taxonomy-based models, which require a lexicon in addition to the training corpus. We computed the selectional preferences of 99 transitive verbs and 173 Instrument verbs as the mean value of the pairwise cosine similarity between their arguments (a weighted mean between all the arguments, or an unweighted mean with the topmost k arguments). Results show that our model can predict the recoverability of objects and Instruments, providing a similar result to that of taxonomy-based models but at a much cheaper computational cost.

1 Introduction

Verb meaning, together with pragmatic and discourse-related factors, has long been identified as playing a major role in determining argument optionality (Levin, 1993). Here, we specifically focus on the **semantic recoverability** of arguments, which has often been claimed to be the key determinant of object omission (Jespersen, 1927; Hopper and Thompson, 1980; Levin, 1993; Resnik, 1993, 1996; Conklin et al., 2004; Medina, 2007; Glass, 2020). Resnik (1993, 1996) link argument recoverability to selectional preferences by means of an experiment we detail in Section 2, and also show that it is correlated with plausibility and typicality judgments provided by human subjects.
The relation linking a verb to its optional argument is a grammatical function in Resnik (1993, 1996), such as "subject" or "direct object", but it may also be a semantic role, such as "Instrument" or "Patient". The choice between the two depends on computational requirements rather than on theoretical constraints.
Let us consider some examples. Recoverability affects the grammaticality of the sentences in (1), since the Patient of an eating event (as in (1a)) is easily recoverable as a member of the class of edibles, but there are no such strict constraints on what one can make (as in (1b)).

(1) a. John ate $\varnothing_{\text{object}}$.
b. *John made $\varnothing_{\text{object}}$.

The same applies to prepositional phrases filling the Instrument role, albeit in a more nuanced fashion. Koenig et al. (2002, 2003, 2007) have shown that verbs describing actions may be divided into two classes based on whether they semantically require an Instrument (as in (2a)) or they merely allow it (as in (2b)). Require-Instrument verbs select for a smaller range of Instruments than Allow-Instrument verbs: For instance, a beheading event is very likely to involve a heavy bladed tool like a sword, while a killing event may happen with a much larger set of tools or even without any. While both sentences in (2) are grammatically acceptable, given that Instruments are syntactic adjuncts (Rissman and Rawlins, 2017), it is much easier to infer the type of instrument used in the event depicted in (2a) than the one in (2b), given the high recoverability of the tool used in the former from verb meaning.

(2) a. John beheaded the prisoner $\varnothing_{\text{Instrument}}$.
b. John killed the prisoner $\varnothing_{\text{Instrument}}$.

Instruments are an interesting object of study for a number of reasons. First of all, they are arguably underesearched with respect to recoverability if compared to direct objects, and to our knowledge there are no computational models of Instrument recoverability. Moreover, this semantic role is usu-

*Proceedings of the Ninth Joint Conference on Lexical and Computational Semantics (*SEM)*, pages 131–136
Barcelona, Spain (Online), December 12–13, 2020

ally overtly realized with prepositional phrases in English and typologically similar languages, making it very different from direct objects and a computationally challenging problem to tackle.

The first attempt to formalize the requirement that a verbal dependent be recoverable in order to be omitted was provided by Resnik (1993, 1996). Resnik's result is a probabilistic information-theoretic model of selectional constraints relying on a manually-built lexicon. In this paper, we propose a fully distributional model of argument recoverability. The crucial idea is that the more mutually similar the arguments selected by a verb are, the more recoverable they are from the meaning of the verb alone. This intuition builds upon the hypotheses and results by Resnik (1993, 1996); Koenig et al. (2007), adding a distributional dimension that those works lacked. We will show that this solution reproduces Resnik's findings about direct objects, addresses the drawbacks of Resnik's model, and is robust enough to predict the recoverability of Instruments. A model such as ours can find a wide range of applications, both in Natural Language Processing (e.g., in semantic role labeling and word sense disambiguation), in linguistic research (e.g., language acquisition and processing), and even in real-life technology (e.g., to provide robots with common-sense knowledge).

2 Related work

Given a verb-relation pair (such as the verb-object relation) Resnik (1993, 1996) formulate the *selectional preference strength* (SPS) of the verb with respect to the possible fillers in the given role (see Eq. 1) as the Kullback-Leibler divergence between the (posterior) distribution of WordNet synsets for the given verb–relation pair and the (prior) distribution of synsets participating in the given relation over all verbs in the corpus. In 1, "classes" refers to the taxonomic classes used in WordNet:

$$SPS_{v,r} = \sum_{c \in classes} p(c|v,r)\, log\frac{p(c|v,r)}{p(c|r)} \quad (1)$$

This yields higher SPS scores for verb-relation pairs admitting only a restricted range of arguments.

Resnik also defines the *selectional association* (SA) of a verb-relation-class triple as the ratio of the SPS for that class and the overall SPS of the verb-relation pair (see Eq. 2), and the SA of a verb-relation-argument triple as the highest verb-relation-class SA among those computed for each WordNet class the argument belongs to:

$$SA_{v,r,c} = \frac{p(c|v,r)\, log\frac{p(c|v,r)}{p(c|r)}}{SPS_{v,r}} \quad (2)$$

Resnik's work inspired more taxonomy-based models of SA over the years (Grishman and Sterling, 1992; Abe and Li, 1996; Ciaramita and Johnson, 2000; Clark and Weir, 2001; Alishahi and Stevenson, 2007; Padó et al., 2009), but no further refinements of the SPS itself.

Distributional Semantic Models (DSMs) (Lenci, 2018) instead tackle the main drawback of taxonomy-based models, i.e. the need for a manually-built lexicon, by requiring no other resource than the corpus they are trained on. They rely on different strategies to compute SA, such as clustering (Pereira et al., 1993), Support Vector Machines (Bergsma et al., 2008), and hybrid approaches (Schulte im Walde et al., 2008).

Erk (2007) and Erk et al. (2010) provide a cognitively plausible distributional model that proves particularly relevant for our purposes. Given a verb-relation pair, it computes the plausibility of a potential argument of the pair (i.e., the SA of the triple) via the weighted similarity between that argument and the exemplar arguments stored in the model as vectors, as shown in Eq. 3:

$$SA_{v,r}(a_0) = \sum_{a \in args(v,r)} wt_{v,r}(a)\, sim(a_0, a) \quad (3)$$

3 PISA: a novel measure of Preference In Selection of Arguments

We introduce **PISA**, our own distributional measure of **Preference In Selection of Arguments**, which we use to model argument recoverability in the spirit of Resnik's SPS. It stems from the intuition that *the vector-based SPS of a given verb-relation pair should be positively correlated with the distributional similarity of their arguments*. The simplest way to capture this notion is by computing the **semantic density** of the verb-relation pair as the mean value of the pairwise cosine similarity between the arguments of the pair. In order to take into account the fact that some arguments are more associated with a given verb-relation pair than others, it is possible to compute a **weighted** measure of semantic density. This is tantamount to averaging Erk et al. (2010)'s SA in Eq. 3 over n arguments of a given verb-relation pair:

$$PISA_{v,r} = \frac{1}{n}\sum_{i=1}^{n} SA_{v,r}(a_i) \tag{4}$$

As in previous literature, relations in our model may be syntactic ones or semantic roles, depending on their availability in a corpus. We used only one similarity measure, cosine.

3.1 Weighted models

We assigned a weight to each argument in our equation based on the 5 weight functions below. The weight functions UNI, FRQ and IDF are taken from Erk et al. (2010):

UNI assumes a uniform distribution:

$$wt_{v,r}(a) = 1 \tag{5}$$

FRQ is the co-occurrence frequency of a given argument with the verb-relation pair:

$$wt_{v,r}(a) = freq(a, v, r) \tag{6}$$

IDF is inspired to the well-known Inverse Document Frequency weighting scheme, which assigns higher scores to arguments occurring with fewer verb-relation pairs:

$$wt_{v,r}(a) = \log \frac{|v, r|}{|v, r : a \in v, r|} \tag{7}$$

LMI is the Local Mutual Information of the argument and a given verb-relation pair:

$$wt_{v,r}(a) = f(a, v, r) \log_2 \frac{p(a, v, r)}{p(a)p(v, r)} \tag{8}$$

ENT is the entropy of the argument of a given verb-relation pair:

$$wt_{v,r}(a) = -\sum_{a \in args(v,r)} p(a) \log_2 p(a) \tag{9}$$

with $p(x) = \frac{f(x)}{\sum_{a \in A} f(a)}$ where A is the complete set of arguments extracted. We entered in the equation only the verbs of our interest.

3.2 Unweighted models

In addition to the weighted models, we created unweighted models taking into consideration only the top/bottom k argument nouns for each verb-relation pair, sorted based on the FRQ, IDF, LMI and ENT weighting functions. In particular we considered the top/bottom 300 nouns for the direct object relation and the top/bottom 20 for the Instrument semantic role. The parameters were determined empirically depending on the fact that our transitive verbs occur with a large number of direct objects, while our Instrument-verbs occur with a much smaller set of Instruments. Both top-k and bottom-k models were computed as not all weighting functions are directly proportional to PISA (e.g. a high IDF means that the nouns is quite selective with respect to what verbs it appears with, whereas high entropy means the opposite).

4 Experimental settings

The datasets and the scripts we used to run our model are freely available on GitHub[1].

4.1 Datasets

We tested our model on two datasets:

- 99 transitive verbs (50 recoverable-object + 49 non-recoverable-object), comprising Resnik's original 34-verb dataset, 35 recoverable-object verbs from Levin (1993), and 30 non-recoverable-object verbs we sampled among high-frequency transitive verbs.
- 173 Instrument verbs (116 recoverable-Instrument + 57 non-recoverable-Instrument), taken from Koenig et al. (2007).

4.2 Extraction of verb arguments

We extracted the arguments participating in the verb-relation pairs of our interest from ukWaC, a 2-billion token part-of-speech tagged and lemmatized corpus of English (Ferraresi et al., 2008). We limited the extraction to the head nouns of the phrases involved in a direct object or Instrument relation with each verb, excluding determiners and modifiers (e.g., *sword* instead of *a big rusty sword*). We mapped the Instrument role to PPs headed by *with* and having an Artifact as a noun argument[2], following Erk et al. (2010) in considering syntactic relations as noisy approximations of semantic roles. Since we are only interested in implicit argument alternations (Ex. (3)), we discarded sentences having an Artifact subject to avoid including inchoative/causative alternations (Ex. (4)) and instrument alternations (Ex. (5)) in our computation.

[1] https://github.com/ellepannitto/PISA

[2] As defined in WordNet 3.0 (Miller, 1995)

SVD	w2v	w2vf
synt.c1000	CBOW.w10	SGNS.synt.c1000
synt.c500	CBOW.w2	SGNS.synt.c500
w10	SGNS.w10	SGNS.w10
w2	SGNS.w2	SGNS.w2

Table 1: Tested embedding types (w2v = word2vec; w2vf = word2vecf).

(3) a. John broke the vase with a hammer.
b. John broke the vase.

(4) a. John broke the vase with a hammer.
b. The vase broke.

(5) a. John broke the vase with a hammer.
b. The hammer broke the vase.

4.3 Word embeddings

As for the vector representation of arguments, we a variety of 300-dimensional embeddings trained on a concatenation of ukWaC and a 2018-dump of English Wikipedia. The embeddings we tested include both SVD reduced count-based DSMs and neural embeddings created via `word2vec`[3] (Mikolov et al., 2013), testing both SGNS and CBOW models, and `word2vecf` (Levy and Goldberg, 2014). We used both window-based and syntax-based contexts, and we tested different window sizes (2 or 10) for `word2vec` and SVD models, for a total of 12 models (Table 1).

5 Results and discussion

5.1 Resnik's SPS

We tested the hypothesis that recoverable-argument verbs have higher SPS scores than non-recoverable argument verbs, by means of Mann-Whitney U tests. We replicated Resnik's experiment by calculating the SPS scores for our transitive and Instrument verbs, based on the distribution of their arguments in ukWaC. The results are consistent with our and Resnik's hypothesis.

The mean score for transitive verbs is 4.27 for recoverable-object verbs and 1.89 for non-recoverable-object verbs. The difference between the two groups is significant ($U = 264$, $n_1 = 50$, $n_2 = 49$, $P < .001$). Similarly, the mean score for Instrument verbs is 4.72 for recoverable-Instrument verbs and 3.60 for non-recoverable-Instrument verbs, and their difference is significant too ($U = 4646$, $n_1 = 116$, $n_2 = 57$, $P < .001$).

[3] https://code.google.com/archive/p/word2vec/

		weighted	top k	bot k
UNI	SVD	***	-	-
	w2v	***	-	-
	w2vf	** (***)	-	-
FRQ	SVD	***	** (***)	ns
	w2v	***	***	ns
	w2vf	***	** (***)	ns
IDF	SVD	***	** (ns)	ns (***)
	w2v	***	*** (ns)	***
	w2vf	** (***)	ns	ns
LMI	SVD	*** (**)	** (ns)	ns (**)
	w2v	***	* (ns)	*
	w2vf	*** (*)	* (ns)	* (**)
ENT	SVD	*** (*)	ns (***)	ns
	w2v	*** (**)	***	ns
	w2vf	*** (**)	* (ns)	*

Table 2: Mann-Whitney U tests comparing recoverable- and non-recoverable-argument verbs (significance levels). Whenever transitive and Instrument-verb results are different, the former are on the left and the latter on the right of the same cell

5.2 PISA

In Tables 2 to 4, we collapsed the results from the 12 distributional models into three types (see Table 1), since there are no within-group notable differences. For each group, we report the worst score. Significance levels are given as follows: *** $p \leq 0.001$, ** $p \leq 0.01$, * $p \leq 0.05$, *ns* $p > 0.05$.
As shown in Table 2, PISA can reliably separate the two groups of recoverable- and non-recoverable-argument verbs based on Mann-Whitney U tests comparing the mean score of the two groups. Looking closer at the results, it appears that the weighted versions of PISA yield highly significant results overall, while the versions using the top k nouns yield varying results depending on both the weights and the distributional spaces. The most notable pattern within the top-k models is that the word2vec spaces lead to consistently significant results, and `FRQ` appears to be the best-performing weight.
The same significance pattern observed for the Mann-Whitney U scores is found, with slight differences, in the Spearman correlations (Tables 3 and 4). In this case, we considered how much the ranking of our set of verbs based on PISA is similar to the ranking yielded by Resnik's SPS, since PISA is intended to improve on Resnik's methodology, while building on the same theoretical premises. Once again, the best scores come from the weighted models (especially those weighted with entropy), and the `FRQ` models are the best within the top-k group.

		weighted	top300	bot300
	SVD	.832***	-	-
UNI	w2v	.851***	-	-
	w2vf	.250*	-	-
	SVD	.854***	.341***	-.041 ns
FRQ	w2v	.835***	.712***	-.024 ns
	w2vf	.743***	-.368***	-.090 ns
	SVD	.750***	-.328***	.211 ns
IDF	w2v	.818***	-.388***	.457***
	w2vf	.256*	-.154 ns	.164 ns
	SVD	.791***	-.385***	-.092 ns
LMI	w2v	.711***	-.135 ns	.129 ns
	w2vf	.667***	-.092 ns	.091 ns
	SVD	-.905***	.163 ns	.111 ns
ENT	w2v	-.908***	.579***	.134 ns
	w2vf	-.911***	.254*	.320**

Table 3: Spearman correlations between PISA and Resnik scores for transitive verbs.

		weighted	top20	bot20
	SVD	.404***	-	-
UNI	w2v	.244***	-	-
	w2vf	.105 ns	-	-
	SVD	.283***	.481***	-.025 ns
FRQ	w2v	.179*	.519***	-.005 ns
	w2vf	.127 ns	.326***	.037 ns
	SVD	.384***	.005 ns	.135 ns
IDF	w2v	.242***	.09 ns	.265***
	w2vf	.082 ns	.176*	.03 ns
	SVD	.170*	.152*	-.011 ns
LMI	w2v	.134 ns	.134 ns	-.065 ns
	w2vf	.077 ns	.266***	-.013 ns
	SVD	-.885***	.118 ns	.003 ns
ENT	w2v	-.920***	.256***	.088 ns
	w2vf	-.928***	.031 ns	.334***

Table 4: Spearman correlations between PISA and Resnik scores for Instrument verbs.

6 Conclusions and future work

In this paper, we presented a novel distributional measure of semantic selectivity (PISA) and used it to quantify direct object and Instrument recoverability. Most notably, PISA achieves a degree of precision comparable with Resnik's SPS but at a much cheaper computational cost, since it does not require a taxonomy or other lexical resources external to the training corpus.

Considering the full picture provided in Tables 2 to 4, the question arises of which is the best choice amongst the various PISA variants. We believe there is no univocal answer. The simplest solution would be to choose a `UNI` weighted model, since it does requires neither an additional step to compute the weight nor an assessment of the optimal value of *k* to compute the top-*k* models. If one cares for their results to resemble Resnik's in the verb ranking, the most conservative solution would be to choose an `ENT` weighted model. If one works with a very large set of verbs, each occurring with many different arguments, the computationally most parsimonious solution would be to choose a `UNI` or better a `FRQ` top-*k* model.

It will be interesting to use our new measure to predict the recoverability of arguments participating in other syntactic relations or semantic roles, to see whether it generalizes to them and to what extent.

Acknowledgments

We would like to thank Ludovica Pannitto for helping us with the computational implementation of our model, Najoung Kim for her contribution to the shaping of the ideas hereby presented, and the anonymous reviewers for their comments and suggestions.

References

Naoki Abe and Hang Li. 1996. Learning Word Association Norms Using Tree Cut Pair Models.

Afra Alishahi and Suzanne Stevenson. 2007. A cognitive model for the representation and acquisition of verb selectional preferences. In *Proceedings of the Workshop on Cognitive Aspects of Computational Language Acquisition*, pages 41–48, Prague, Czech Republic. Association for Computational Linguistics.

Shane Bergsma, Dekang Lin, and Randy Goebel. 2008. Discriminative Learning of Selectional Preference from Unlabeled Text. In *Proceedings of the 2008*

Conference on Empirical Methods in Natural Language Processing, pages 59–68. Association for Computational Linguistics.

Massimiliano Ciaramita and Mark Johnson. 2000. Explaining away ambiguity: Learning verb selectional preference with Bayesian networks. In *COLING 2000 Volume 1: The 18th International Conference on Computational Linguistics*.

Stephen Clark and David Weir. 2001. Class-based probability estimation using a semantic hierarchy. In *Second Meeting of the North American Chapter of the Association for Computational Linguistics*.

Kathy Conklin, Jean-Pierre Koenig, and Gail Mauner. 2004. The role of specificity in the lexical encoding of participants. 90(1-3):221–230.

Katrin Erk. 2007. A simple, similarity-based model for selectional preferences. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 216–223, Prague, Czech Republic. Association for Computational Linguistics.

Katrin Erk, Sebastian Padó, and Ulrike Padó. 2010. A Flexible, Corpus-Driven Model of Regular and Inverse Selectional Preferences. 36(4):723–763.

Adriano Ferraresi, Eros Zanchetta, Marco Baroni, and Silvia Bernardini. 2008. Introducing and evaluating ukwac, a very large web-derived corpus of english. In *In Proceedings of the 4th Web as Corpus Workshop (WAC-4*.

Lelia Glass. 2020. Verbs describing routines facilitate object omission in English. 5(1):44.

Ralph Grishman and John Sterling. 1992. Acquisition of Selectional Patterns. In *COLING 1992 Volume 2: The 15th International Conference on Computational Linguistics*.

Paul J. Hopper and Sandra A. Thompson. 1980. Transitivity in Grammar and Discourse. 56(2):251.

Otto Jespersen. 1927. *A Modern English Grammar on Historical Principles. Part III. Syntax*, volume 2. Routledge.

Jean-Pierre Koenig, Gail Mauner, and Breton Bienvenue. 2002. Class Specificity and the Lexical Encoding of Participant Information. 81(1-3):224–235.

Jean-Pierre Koenig, Gail Mauner, and Breton Bienvenue. 2003. Arguments for adjuncts. *Cognition*, 89(2):67–103.

Jean-Pierre Koenig, Gail Mauner, Breton Bienvenue, and Kathy Conklin. 2007. What with? The Anatomy of a (Proto)-Role. 25(2):175–220.

Alessandro Lenci. 2018. Distributional Models of Word Meaning. *Annual Review of Linguistics*, 4:151–171.

Beth Levin. 1993. *English Verb Classes and Alternations: A Preliminary Investigation*. University of Chicago Press.

Omer Levy and Yoav Goldberg. 2014. Dependency-based word embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 302–308.

Tamara Nicol Medina. 2007. Learning which verbs allow object omission: Verb semantic selectivity and the implicit object construction.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

George A. Miller. 1995. WordNet: A lexical database for English. *Communications of the ACM*, 38(11):39–41.

Ulrike Padó, Matthew W. Crocker, and Frank Keller. 2009. A Probabilistic Model of Semantic Plausibility in Sentence Processing. 33(5):794–838.

Fernando Pereira, Naftali Tishby, and Lillian Lee. 1993. Distributional clustering of English words. In *Proceedings of the 31st Annual Meeting on Association for Computational Linguistics* -, pages 183–190. Association for Computational Linguistics.

Philip Resnik. 1993. *Selection and Information: A Class-Based Approach to Lexical Relationships*. IRCS Technical Reports Series. University of Pennsylvania.

Philip Resnik. 1996. Selectional constraints: An information-theoretic model and its computational realization. 61(1-2):127–159.

Lilia Rissman and Kyle Rawlins. 2017. Ingredients of Instrumental Meaning. 34(3):507–537.

Sabine Schulte im Walde, Christian Hying, Christian Scheible, and Helmut Schmid. 2008. Combining EM Training and the MDL Principle for an Automatic Verb Classification Incorporating Selectional Preferences. In *Proceedings of ACL-08: HLT*, pages 496–504. Association for Computational Linguistics.