Predicate-specific Annotations for Implicit Role Binding: Corpus Annotation, Data Analysis and Evaluation Experiments

Tatjana Moor Michael Roth Anette Frank
Department of Computational Linguistics, Heidelberg University
{moor,mroth,frank}@cl.uni-heidelberg.de

Abstract

Current research on linking implicit roles in discourse is severely hampered by the lack of sufficient training resources, especially in the verbal domain: learning algorithms require higher-volume annotations for specific predicates in order to derive valid generalizations, and a larger volume of annotations is crucial for insightful evaluation and comparison of alternative models for role linking.

We present a corpus of predicate-specific annotations for verbs in the FrameNet paradigm that are aligned with PropBank and VerbNet. A qualitative data analysis leads to observations regarding implicit role realization that can guide further annotation efforts. Experiments using role linking annotations for five predicates demonstrate high performance for these target predicates. Using our additional data in the SemEval task, we obtain overall performance gains of 2-4 points F_1-score.

1 Introduction

Automatic annotation of semantic predicate-argument structure (PAS) is an important subtask to be solved for high-quality information access and natural language understanding. Semantic role labeling (SRL) has made tremendous progress in addressing this task, using supervised and recently also semi- and unsupervised methods (Palmer et al., 2010).

Traditional SRL is restricted to the local syntactic domain. In discourse interpretation, however, we typically find locally unrealized argument roles that are contextually bound to antecedents beyond their local structure. Thus, by using strictly local methods, we are far from capturing the full potential offered by semantic argument structure (Fillmore and Baker, 2001; Burchardt et al., 2005).

The task of resolving the reference of implicit arguments has been addressed in previous work: Gerber and Chai (2012) address the task in the nominal domain by learning a model from manually annotated data following the NomBank paradigm (Meyers et al., 2004). In contrast, Ruppenhofer et al. (2010) follow the FrameNet paradigm, which is not restricted to nominal predicates. However, their data set suffers from considerable sparsity with respect to annotation instances per predicate (cf. Section 2).

Our contribution addresses the problem of sparse training resources for implicit role binding by providing a higher volume of predicate-specific annotations for non-local role binding, using OntoNotes (Weischedel et al., 2011) as underlying corpus. A qualitative analysis of the produced annotations leads to a number of hypotheses on implicit role realization. Using the extended set of annotations, we perform experiments to measure their impact, using a state-of-the-art system for implicit role binding.

2 Motivation and Related Work

The main motivation for this work relates to the SemEval 2010 Task 10 on implicit role linking[1] and the problem of data sparsity that became evident by the poor performance of the participating systems, at 1% F_1-score (Tonelli and Delmonte, 2010; Chen et al., 2010).[2] Later systems could only marginally improve

[1] http://www.coli.uni-saarland.de/projects/semeval2010_FG

[2] The data set provides 245/259 instances of resolvable implicit roles for training/testing. All cases of implicit roles (580/710) are distributed over 317/452 frame types and a small overall number of frame instances (1,370/1,703 training/testing).

on these results, with performance up to 19% F_1-score due to improved recognition of resolvable implicit roles, heuristic data acquisition, and variations in model properties.[3] Gerber and Chai (2010, 2012), working on a related task following the NomBank/PropBank paradigm, achieved higher performance of 50% F_1-score, using as training data a substantial amount of annotations for 10 noun predicates.

3 Corpus and Annotation

3.1 Corpus

While there is a rich source of annotated sentences in the FrameNet paradigm, contextualized FrameNet annotations are restricted in coverage. As we target high-frequency annotations for specific verbs, and in order to make annotations available for a corpus that is widely used, we chose OntoNotes (V.4.0) (Weischedel et al., 2011) as underlying corpus. OntoNotes contains semantic role annotations following the PropBank annotation style (Palmer et al., 2005). We map these annotations to FrameNet using the mapping specified in the SemEval 2010 Task 10 data, which is based on SemLink (Loper et al., 2007).

3.2 Selection of Annotation Targets

Our goal was to produce a high volume of annotations for specific verb predicates, ideally reaching a margin of 100-200 instances involving locally unfilled argument roles (cf. Gerber and Chai (2010)). In order to make the task feasible for the annotators, we selected predicates and frame readings that are relatively easy to discriminate, so that the annotators can concentrate on the role linking task.

We applied a number of further selection criteria to make the resulting annotations as useful as possible: (i) We excluded light verbs, as they are not well covered in FrameNet, and typically involve difficult sense distinctions. (ii) We only chose predicates (and senses) that are covered in VerbNet, PropBank and FrameNet, according to the Unified Verb Index.[4] This ensures that the corpus can also be used for experimentation using the VerbNet or PropBank paradigm. Finally, (iii) for the selected candidate predicates and readings, we investigated the FrameNet annotation data base to determine whether the annotated frames involve a critical number of non-instantiated roles that can be resolved from discourse context.[5] In case we found little or no such cases for the candidate reading, the predicate was not chosen.

The list of predicates that resulted from this selection process is given in Table 1. They exhibit varying numbers of core roles (2-7), frame ambiguity (1-7), and different syntactic properties.

3.3 Annotation Process and Categories

Data preparation. We extracted annotation instances for the selected target predicates from the OntoNotes corpus. The resulting corpus consists of overall 1.992 instances. Each annotation instance was embedded within its full document context. The average document length is 612 words.

Annotation Categories. Our annotation maily follows the SemEval task guidelines for role linking (Ruppenhofer et al., 2010, 2012), with the exception that we differentiate between non-instantiated (NI) roles that are *resolvable* vs. *non-resolvable* within discourse instead of classifying them as 'definite (DNI)' vs. 'indefinite (INI)'. This distinction makes the task of linking NIs much clearer, as definite null-instantiations may or may not be resolvable within the discourse context.[6]

Two examples of NI occurrences are given below: (1) illustrates a (resolvable) DNI: the implicit role's referent is anaphorically bound within the prior discourse. In (2) the non-instantiated role can only be interpreted existentially within the given discourse (non-resolvable, INI).

[3] See Tonelli and Delmonte (2011); Ruppenhofer et al. (2011); Silberer and Frank (2012); Laparra and Rigau (2012).

[4] `http://verbs.colorado.edu/verb-index`

[5] Even though FrameNet annotations are out of context, non-realized core roles are marked for definite vs. indefinite interpretation.

[6] In fact, only 80.9%/74.2% of all DNIs in the SemEval training/test corpus are linked within discourse.

(1) (s3) Nearly 200 American agents went to [Yemen]$_{Source}$ right after the attack on the ”Cole”.
(s9) They$_{Theme}$ are *leaving* frustrated. (*Source*: resolvable, DNI)

(2) I$_{Donor}$ tried to *give* as good as I got. (*Recepient*, non-resolvable, INI; *Theme*, non-resolvable, INI)

The annotation consists of three sub-tasks that are applied to null-instantiated core roles (NIs) only: (a) classifying each NI as a *resolvable or non-resolvable null instantiation*; (b) distinguishing between *Lexical* and *Constructional Licensing* of each NI;[7] and (c) *linking resolvable DNIs* to the closest antecedent mention within the preceding context.[8]

Before proceeding to these decisions, the annotator determines whether the mapped frame corresponds to the actual predicate meaning in the given context. If not, it is marked as ’NgFNS’ (no genuine FN sense). We also flag roles whose filler does not correspond to the role definition given in FrameNet (e.g., roles categorized as ’Physical object’ that are filled by an abstract concept). For each predicate, we record the chosen frame as well as the mapping to the corresponding readings in PropBank and VerbNet.

Calibration of Annotation Quality. The annotation was performed by two annotators, both students of Computational Linguistics. Both of them studied the SemEval guidelines and used the first 50 sentences of the SemEval corpus as a trial corpus, in order to validate their understanding of the guidelines.

We performed two calibration experiments, in which we measured Kappa (Cohen, 1960) for the assignment of role interpretation type (resolvable vs. non-resolvable role), and percentage of agreement on the chosen antecedent for resolvable roles.

(I) **Agreement with SemEval:** After initial training, we measured IAA between our main annotator and the SemEval gold annotations for sentences 51-100 of the SemEval data set. For *interpretation type* (resolvable/non-resolvable classification) we achieved a Kappa of 0.77. For NI-linking, we measured 71.43% agreement (15 out of 21 resolvable roles were correctly linked).

(II) **Agreement between Annotators:** We determined agreement between both annotators on all annotation instances pertaining to the predicate *give*. For *interpretation type*, the annotators achieved a Kappa value of 0.94. For *linking*, the annotators agreed on the marked antecedent in 85.7% of all cases (48 out of 56 cases).

After this calibration phase, the annotation was done independently by the two annotators.

4 Data Analysis

Table 1 gives an overview of the annotations we obtained. Overall we annotated 630 NI occurrences for genuine frame meanings, distributed over 438 verb occurrences (i.e., 1.44 NIs/verb).[9] We observe great variation in the number of NI occurrences for the different predicates (e.g., *leave* vs. *pay*). We find a predominance of non-resolvable over resolvable role classifications (61.6% vs. 38.4%). 78% of the resolvable NIs are realized within a window of 3 sentences,[10] as opposed to 69.6% in the SemEval and 90% in Gerber&Chai’s data. This can be explained by variation in text genre and target categories.

Considering the distribution of NI-realizations and the properties of the corresponding predicates, we note some tendencies that seem worth investigating on a larger scale, as potential factors determining null instantiation of roles. Predicates with low frame ambiguity rate (*pay, bring, give*) tend to have a higher NI-realization rate than frames with a higher ambiguity rate (*leave, put*). A higher number of core roles of the target frame tends to go along with a higher NI-realization potential (*bring, pay*).

[7] As the SemEval guidelines for lexical and constructional licensing are not very explicit, and given these annotations are not required for evaluating system annotations, we do not report details about this part of the annotation.

[8] If the antecedent is not found in the preceding context, we also inspect the following discourse.

[9] 204 out of all 834 NIs (24.5%) do not pertain to genuine frame readings. These were held out from further data statistics.

[10] 9.9% of the fillers were found in the following discourse.

verb	frame	verb	core	frame	NI occurrences								
		occ.	roles	ambig.	all	other	frame	frame	NIs per	resolvable		non-resolv.	
						reading	reading	occ.	frame	abs.	%	abs.	%
give	GIVING	524	3	1	218	63	155	144	1.08	62	40.0	93	60.0
put	PLACING	427	4	3	39	17	22	22	1.00	10	45.5	12	54.5
leave	DEPARTING	354	2	7	70	30	40	39	1.03	25	62.5	15	37.5
bring	BRINGING	351	7	2	103	38	65	45	1.44	28	43.1	37	56.9
pay	COMMERCE_PAY	336	5	1	404	56	348	188	1.85	117	33.6	231	66.4
all		1992	–	–	834	204	630	438	–	242	–	388	–

Table 1: Annotated predicates and data analysis: Implicit role interpretation and linking.

give: GIVING		all	**Donor**	**Recepient**	**Theme**		
Interpretation	resolvable	40.0	**25.2**	13.5	1.3		
	non-resolvable	60.0	19.4	**37.4**	3.2		
put: PLACING		all	**Agent**	**Cause**	**Theme**	**Goal**	
Interpretation	resolvable	45.5	**40.9**	4.6	0.0	0.0	
	non-resolvable	54.4	**45.5**	9.1	0.0	0.0	
leave: DEPARTING		all	**Theme**	**Source**			
Interpretation	resolvable	62.5	0.0	**62.5**			
	non-resolvable	37.5	7.5	30.0			
bring: BRINGING		all	**Agent**	**Goal**	**Source**	**Carrier**	
Interpretation	resolvable	43.1	9.3	16.9	16.9	0.0	
	non-resolvable	56.9	21.5	1.5	**23.1**	10.8	
pay: COMMERCE_PAY		all	**Buyer**	**Seller**	**Goods**	**Money**	**Rate**
Interpretation	resolvable	33.6	6.6	**14.6**	9.2	2.9	0.3
	non-resolvable	66.4	2.5	**25.3**	17.2	10.9	10.3

Table 2: Distribution of resolvable vs. non-resolvable NI roles over predicate roles (in percent).

Further data statistics for particular predicates and which roles they omit under the different NI-interpretations are given in Table 2. Typically, we find NI-realization concentrated on one or two roles for a given predicate. Yet, these are observations on a small number of predicates that need substantiation by further data annotation, with systematic exploration of other determining properties, such as role meaning or perspectivation (*pay/sell*; *leave/arrive*) and the influence of constructional licensing.

5 Evaluation Experiments

We evaluate the impact of predicate-specific annotations for classification using two scenarios: (**CV**) we examine the linking performance of models trained and tested on the same predicate by adopting the 10-fold Cross-Validation scenario used by Gerber and Chai (2012) (G&C).[11] (**SemEval**) Secondly, we examine the direct effect of using our annotations as additional training data for linking NIs in the SemEval 2010 task on implicit role binding. We use the state-of-the-art system and best performing feature set described in Silberer and Frank (2012) (S&F) to make direct comparisons to previous results.

CV. As positive training and test samples for this scenario, we use all annotated (and resolvable) NIs and randomly split them into 10 folds. Negative training samples (i.e., incorrect NI fillers) are automat-

[11] Note that this is not a direct comparison as both the annotation paradigm and the data sets are different.

	Cross-Validation			SemEval 2010 test set					
verb	precision	recall	F_1	**training data**	FS	NgFNS	precision	recall	F_1
give	48.8	33.3	39.6	**S&F'12**					
put	33.3	20.0	25.0	no additional data	+	n.a.	25.6	25.1	25.3
leave	48.3	**56.0**	**51.9**	+ best heuristic data	+	n.a.	30.8	25.1	27.7
bring	**72.7**	27.6	40.0	**this paper**					
pay	35.4	20.0	25.6	+ our annotations	–	–	21.7	21.2	21.5
—	—	—	—	+ our annotations	+	–	33.3	22.0	26.5
average	47.7	31.4	36.4	+ our annotations	+	+	**34.3**	**26.3**	**29.8**

Table 3: Results for both evaluations (all figures are percentages). FS indicates whether feature selection was applied. NgFNS indicates the use of frame annotations that do not match the contextual meaning.

ically added by extracting constituents that overtly fill a role according to the semantic annotations in the OntoNotes gold standard. We only consider phrases of type NPB, S, VP, SBAR and SG within the current and the two preceding sentences as potential fillers.[12]

SemEval. This setting is identical with the linking evaluation in S&F. Like them, we (optionally) apply an additional step of feature selection (±FS) on the SemEval training data to select a feature subset that generalizes best across data sets, i.e., the fully annotated novel from the shared task and our predicate-specific annotations based on OntoNotes. We further compare models trained w/ and w/o non-genuine frame annotations (±NgFNS). As in the CV setting, we assume that all resolvable NIs are known and only the correct fillers are unknown. Thus, our results are not comparable to those of participants of the *full* SemEval task, who solved two further sub-tasks. Instead we compare to the NI linking results in S&F, with models trained on the SemEval data and using additional heuristically labelled data.

Table 3 summarizes our results for both settings. They are not strictly comparable due to varying properties, i.a., the number of available annotations. The **CV** results show that few annotations can be sufficient to achieve a high linking precision and f-score (up to 72.7 P, 51.9 F_1). However, this is highly dependent on the target predicate (cf. *bring* vs. *pay*). Overall, the results exhibit a similar variance and lie within the same range as those reported by G&C. Even though the numbers are not directly comparable, they generally indicate a similar difficulty of linking implicit arguments across lexical predicate types.

In the **SemEval** setting, we obtain improved precision and recall over S&F's results (± additional heuristic data, cf. Silberer&Frank, 2012)) when linking NIs using our additional training data and feature selection. Using our full additional data set (+NgFNS) we obtain higher performance compared to S&F's best setting with heuristically labelled data, yielding highest scores of 34.3% precision and 26.3% recall. The resulting F_1-score of 29.76% lies 2.1 points above the best model of S&F, whose full system also achieved state-of-the-art performance on the *full* SemEval task.

6 Conclusions

We presented an annotation effort for implicit role linking targeting five verb predicates. The FrameNet annotations are mapped to PropBank and VerbNet and will be available for the community. Annotations follow the SemEval guidelines and were quality-controlled. We annotated 630 NI realizations for the intended predicate senses. Our experiments show that even a moderate amount of annotations per predicate yield substantial performance gains of 2.1-4.5 points F_1-score. Our data set complements the SemEval corpus in terms of text genre and Gerber&Chai's data set in terms of category and explicit annotation for interpretation type. Due to higher-volume predicate-specific annotations, it enables more insightful evaluation and comparison between different models, including comparison across frameworks. In future work, we plan to extend the annotation to further predicates using, i.a., active learning techniques.

[12] This corresponds to the *SentWin* setting in Silberer and Frank (2012) and is motivated by the fact that most NI fillers both in the SemEval training data and in our annotations are located within a span of the current and two preceding sentences.

References

Burchardt, A., A. Frank, and M. Pinkal (2005). Building Text Meaning Representations from Contextually Related Frames – A Case Study. In *Proceedings of the 6th International Workshop on Computational Semantics*, IWCS-6, Tilburg, The Netherlands, pp. 66–77.

Chen, D., N. Schneider, D. Das, and N. A. Smith (2010, July). SEMAFOR: Frame Argument Resolution with Log-Linear Models. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, Uppsala, Sweden, pp. 264–267.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement 20*(1), 37–46.

Fillmore, C. J. and C. F. Baker (2001, June). Frame Semantics for Text Understanding. In *Proceedings of the NAACL 2001 Workshop on WordNet and Other Lexical Resources*, Pittsburgh.

Gerber, M. and J. Chai (2010). Beyond NomBank: A Study of Implicit Arguments for Nominal Predicates. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, Uppsala, Sweden, pp. 1583–1592.

Gerber, M. and J. Y. Chai (2012). Semantic role labeling of implicit arguments for nominal predicates. *Computational Linguistics 38*(4), 755–798.

Laparra, E. and G. Rigau (2012). Exploiting Explicit Annotations and Semantic Types for Implicit Argument Resolution. In *6th IEEE International Conference on Semantic Computing (ICSC'12)*, Palermo, Italy.

Loper, E., S. Yi, and M. Palmer (2007). Combining Lexical Resources: Mapping between PropBank and VerbNet. In *Proceedings of the 7th International Workshop on Computational Semantics*, Tilburg, the Netherlands.

Meyers, A., R. Reeves, C. Macleod, R. Szekely, V. Zielinska, B. Young, and R. Grishman (2004). Annotating Noun Argument Structure for NomBank. In *Proceedings of the 4th International Conference on Language Resources and Evaluation*, LREC-2004, Lisbon, Portugal, pp. 803–806.

Palmer, M., D. Gildea, and P. Kingsbury (2005). The Proposition Bank: An Annotated Corpus of Semantic Roles. *Computational Linguistics 31*(1), 71–106.

Palmer, M., D. Gildea, and N. Xue (2010). *Semantic Role Labeling*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.

Ruppenhofer, J., P. Gorinski, and C. Sporleder (2011). In Search of Missing Arguments: A Linguistic Approach. In *Proceedings of RANLP*, Hissar, Bulgaria, pp. 331–338.

Ruppenhofer, J., R. Lee-Goldman, C. Sporleder, and R. Morante (2012). Beyond sentence-level semantic role labeling: linking argument structures in discourse. *Language Resources and Evaluation*. DOI: 10.1007/s10579-012-9201-4.

Ruppenhofer, J., C. Sporleder, R. Morante, C. Baker, and M. Palmer (2010). SemEval-2010 Task 10: Linking Events and Their Participants in Discourse. In *Proceedings of the 5th International Workshop on Semantic Evaluations*, Uppsala, Sweden, pp. 45–50.

Silberer, C. and A. Frank (2012). Casting Implicit Role Linking as an Anaphora Resolution Task. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics*, Montréal, Canada, pp. 1–10.

Tonelli, S. and R. Delmonte (2010). VENSES++: Adapting a Deep Semantic Processing System to the Identification of Null Instantiations. In *Proceedings of the 5th International Workshop on Semantic Evaluations*, Uppsala, Sweden, pp. 296–299.

Tonelli, S. and R. Delmonte (2011). Desperately Seeking Implicit Arguments in Text. In *Proceedings of the ACL 2011 Workshop on Relational Models of Semantics*, Portland, pp. 54–62.

Weischedel, R., E. Hovy, M. Palmer, M. Marcus, R. Blevin, S. Pradhan, L. Ramshaw, and N. Xue (2011). OntoNotes: A Large Training Corpus for Enhanced Processing. In J. Olive, C. Christianson, and J. McCary (Eds.), *Handbook of Natural Language Processing and Machine Translation*. Springer.