Memory-Based Acquisition of Argument Structures and its Application to Implicit Role Detection

Christian Chiarcos and **Niko Schenk**
Applied Computational Linguistics Lab
Goethe University Frankfurt am Main, Germany
{chiarcos,n.schenk}@em.uni-frankfurt.de

Abstract

We propose a generic, memory-based approach for the detection of implicit semantic roles. While state-of-the-art methods for this task combine hand-crafted rules with specialized and costly lexical resources, our models use large corpora with automated annotations for *explicit* semantic roles only to capture the distribution of predicates and their associated roles. We show that memory-based learning can increase the recognition rate of implicit roles beyond the state-of-the-art.

1 Introduction

Automated implicit semantic role labeling (iSRL) has emerged as a novel area of interest in the recent years. In contrast to traditional SRL, which aims to detect events (e.g., verbal or nominal predicates) together with their associated semantic roles (*agent*, *theme*, *recipient*, etc.) as overtly realized in the current sentence, iSRL extends this analysis with locally *unexpressed* linguistic items. Hence, iSRL requires to broaden the scope beyond isolated sentences to the surrounding discourse. As an illustration, consider the following example from Roth and Frank (2013):

> El Salvador is now the only Latin American country which still has troops in [Iraq]. Nicaragua, Honduras and the Dominican Republic have withdrawn their troops [∅].

In the second sentence, a standard SRL parser would ideally identify *withdraw* as the main verbal predicate. In its thematic relation to the other words within the same sentence, all countries serve as the overtly expressed (explicit) agents, and are thus labeled as arguments A0.[1] Semantically, they are the action performers, whereas *troops* would carry the patient role A1 as the entity which undergoes the action of being withdrawn. However, given these explicit role annotations for A0 and A1 in the second sentence, the standard system would definitely fail to infer the underlying, linguistically unexpressed, i.e., non-overt realization of an *implicit* argument of *withdraw* (denoted by [∅]) about source information. Its corresponding realization is associated with *Iraq* in the preceding sentence, which is outside of the scope of any standard SRL parser. The resulting implicit role has the label A2.

Many role realizations are suppressed on the surface level. The automated detection of such implicit roles and their fillers, which are also called *null instantiations* (NIs) (Fillmore, 1986; Ruppenhofer, 2005), is a challenging task. Yet, if uncovered, NIs provide highly beneficial 'supplementary' information which in turn can be incorporated into practical, downstream NLU applications, like automated text summarization, recognizing textual entailment or question answering.

Current issues in iSRL Corpus data with manually annotated implicit roles is extremely sparse and hard to obtain, and annotation efforts have emerged only recently; cf. Ruppenhofer et al. (2010), Gerber and Chai (2012), and also Feizabadi and Padó (2015) for an attempt to enlarge the number of annotation instances by combination of scarce resources. As a result, most state-of-the-art iSRL systems cannot be trained in a supervised setting and thus integrate custom, rule-based components to detect NIs (we elaborate on related work in Section 2). To this end, a predicate's overt roles are matched against a predefined predicate-specific template. Informally, all roles found in the template but not in the text are regarded as null instantiations. Such pattern-based methods perform satisfactorily, yet there are drawbacks:
(1) They are inflexible and absolute according to

[1] For details on all PropBank labels used in our study, see Palmer et al. (2005).

Proceedings of the SIGDIAL 2015 Conference, pages 178–187,
Prague, Czech Republic, 2-4 September 2015. ©2015 Association for Computational Linguistics

their type, in that they assume that all candidate NIs are equally likely to be missing, which is unrealistic given the variety of different linguistic contexts in which predicates co-occur with their semantic roles.

(2) They are expensive in that they require handcrafted, idiosyncratic rules (Ruppenhofer et al., 2011) and rich background knowledge in the form of language-specific lexical resources, such as FrameNet (Baker et al., 1998), PropBank (Palmer et al., 2005) or NomBank (Meyers et al., 2004). Dictionaries providing information about each predicate and status of the individual roles (e.g., whether they can serve as implicit elements or not) are costly, and for most other languages not available to the same extent as for English.

(3) Most earlier studies heuristically restrict implicit arguments to *core* roles[2] only (Tonelli and Delmonte, 2010; Silberer and Frank, 2012), but this is problematic as it ignores the fact that implicit non-core roles also provide valid and valuable information. Our approach remains agnostic regarding the role inventory, and can address both core and non-core arguments. Yet, in accordance with the limited evaluation data and in line with earlier literature, we had to restrict ourselves to evaluate NI predictions for core arguments only.

Our contribution We propose a novel, generic approach to infer information about implicit roles which does not rely on the availability of manually annotated gold data. Our focus is exclusively on *NI role identification*, i.e., per-predicate detection of the missing implicit semantic role(s) given their overtly expressed explicit role(s) (without finding filler elements) as we believe that it serves as a crucial preprocessing step and still bears great potential for improvement. We treat NI identification *separately* from the resolution of their fillers, also because not all NIs are resolvable from the context. In order to facilitate a more flexible mechanism, we propose to condition on the presence of other roles, and primarily argue that NI detection should be **probabilistic instead of rule-based**. More specifically, we predict implicit arguments using large corpora from which we build a background knowledge base of predicates, co-occurring (explicit) roles and their probabilities. With such a **memory-based** approach, we generalize over large quantities of explicit roles to find evidence for implicit information in a mildly supervised manner. Our proposed models are largely domain independent, include a sense distinction for predicates, and are not bound to a specific release of a hand-maintained dictionary. Our approach is portable across languages in that training data can be created using projected SRL annotations. Unlike most earlier approaches, we employ a generic role set which is based on PropBank/NomBank rather than FrameNet: The PropBank format comprises a relatively small role inventory which is better suited to obtain statistical generalizations than the great variety of highly specific FrameNet roles. While FrameNet roles seem to be more fine-grained, their greater number arises mostly from predicate-specific semantic roles, whose specific semantics can be recovered from PropBank annotations by pairing semantic roles with the predicate.

Yet another motivation of our work is related to the recent development of AMR parsing (Banarescu et al., 2013, Abstract Meaning Representation) which aims at modeling the semantic representation of a sentence while abstracting from syntactic idiosyncrasies. This particular appraoch makes extensive use of the PropBank-style framesets, as well, and would greatly benefit from the integration of information on implicit roles.

The paper is structured as follows: Section 2 outlines related work in which we exclusively focus on how previous research has handled the sole identification of NIs. Sect. 3 describes our approach to probabilistic NI detection; Sect. 4 presents two experiments and their evaluation; Sect. 5 concludes our work.

2 Related Work

In the context of the 2010 SemEval Shared Task on *Linking Events and Their Participants in Discourse*[3] on implicit argument resolution, Ruppenhofer et al. (2010) have released a data set of fiction novels with manual NI role annotations for diverse predicates. The data has been referred to by various researchers in the community for direct or indirect evaluation of their results. The NIs in the data set are further subdivided into two categories: Definite NIs (DNIs) are locally unexpressed arguments which can be resolved to elements in the proceeding or following discourse;

[2] *Core* roles are obligatory arguments of a predicate. Informally, *non-core* roles are optional arguments often realized as adjuncts or modifiers.

[3] http://semeval2.fbk.eu/semeval2.php

Indefinite NIs (INIs) are elements for which no antecedent can be identified in the surrounding context.[4] Also, the evaluation data comes in two flavors: a base format which is compliant with the FrameNet paradigm and a CoNLL-based PropBank format. Previous research has exclusively focused on the former.

Chen et al. (2010) present an extension of an existing FrameNet-style parser (SEMAFOR) to handle implicit elements in text. The identification of NIs is guided by the assumption that, whenever the traditional SRL parser returns the default label involved in a non-saturated analysis for a sentence, an implicit role has to be found in the context instead. Additional FrameNet-specific heuristics are employed in which, e.g., the presence of one particular role in a frame makes the identification of another implicit role redundant.[5]

Tonelli and Delmonte (2010, VENSES++) present a deep semantic approach to NI resolution whose system-specific output is mapped to FrameNet valency patterns. For the detection of NIs, they assume that these are always core arguments, i.e., non-omissible roles in the interaction with a specific predicate. It is unclear how different predicate senses are handled by their approach. Moreover, not all types of NIs can be detected, resulting in a low overall recall of identified NIs, also having drawbacks for nouns. Again using FrameNet-specific modeling assumptions, their work has been significantly refined in Tonelli and Delmonte (2011).

Despite their good performance in the overall task, Silberer and Frank (2012, S&F) give a rather vague explanation regarding NI identification in text. Using a FrameNet API, the authors restrict their analysis only to the core roles by excluding "conceptually redundant" roles without further elaboration.

Laparra and Rigau (2013) propose a deterministic algorithm to detect NIs on grounds of discourse coherence: It predicts an NI for a predicate if the corresponding role has been explicitly realized for the same predicate in the preceding discourse but is currently unfilled. Their approach is promising but ignorant of INIs.

Earlier, Laparra and Rigau (2012, L&R) introduce a statistical approach to identifying NIs similar to ours in that they rely on frequencies from overt arguments to predict implicit arguments. For each predicate template (frame), their algorithm computes all Frame Element patterns, i.e., all co-occurring overt roles and their frequencies. For NI identification a given predicate and its overtly expressed roles are matched against the most frequent pattern not violated by the explicit arguments. Roles of the pattern which are not overtly expressed in the text are predicted as missing NIs. Even though their approach outperforms all previous results in terms of NI detection, Laparra and Rigau (2012) only estimate the *raw* frequencies from a very limited training corpus, raising the question whether all patterns are actually sufficiently robust. Also, the authors disregard all the valuable less frequent patterns and limit their analysis to only a subtype of NI instances which are resolvable from the context.

Finally, Gerber and Chai (2012) describe a supervised model for implicit argument resolution on the NomBank corpus which—unlike the previous literature—follows the PropBank annotation format. However, NI detection is still done by dictionary lookup, and the analysis is limited to only a small set of predicates with only one unambiguous sense. Again limiting NIs to only core roles, the authors empirically demonstrate that this simplification accounts for 8% of the overall error rate of their system.

3 Experimental Setup

3.1 Memory-Based Learning

Memory-based learning for NLP (Daelemans and van den Bosch, 2009) is a lazy learning technique which keeps a record of training instances in the form of a background knowledge base (BKB). Classification compares new items directly to the stored items in the BKB via a distance metric. In semantics, the method has been applied by, e.g., Peñas and Hovy (2010) for semantic enrichment, and Chiarcos (2012) to infer (implicit markers for) discourse relations. Here, we adopt its methodology to identify null-instantiated argument roles in text. More precisely, we setup a BKB of probablistic predicate-role co-occurrences and estimate thresholds which serve as a trigger for the prediction of an implicit role (a slight modification of the distance metric). We elaborate on this methodology in Section 4.

[4] The average F-score annotator agreement for frame assignments is about .75 (Ruppenhofer et al., 2010).

[5] Cf. *CoreSet* and *Exludes* relationship in FrameNet.

3.2 Data & Preprocessing

We train our model on a subset of the *WaCkypedia_EN*[6] corpus (Baroni et al., 2009). The data set provides a 2008 Wikipedia dump from which we extracted the tokens and sentences. We have further divided the dump into pieces of growing size (cumulatively by 100 sentences) and applied MATE[7] (Björkelund et al., 2009) for the automatic detection of semantic roles to the varying portions and annotated them with SRL information. For each sentence, MATE identifies the predicates and all of its associated core and noncore arguments.[8] MATE has been used in previous research on implicit elements in text (Roth and Frank, 2013) and provides semantic roles with a sense disambiguation for both verbal and nominal predicates. The resulting output is based on the PropBank format.

3.3 Model Generation

We build a probablistic model from annotated predicate-role co-occurrences as follows:

1. For every sentence, record all distinct predicate instances and their associated roles.
2. For every predicate instance, sort the role labels lexicographically (not the role fillers), disregarding their sequential order. (We thus obtain a normalized template of role co-occurrences for each frame instantiation.)
3. Compute the frequencies for all templates associated with the same predicate.
4. By relative frequency estimation, derive all conditional probabilities of the form:

$$P(r|R, \textsc{predicate})$$

with $\mathcal{R}$ being the role inventory of the SRL parser, $R \subseteq \mathcal{R}$ a (sub)set of explicitly realized semantic roles, and $r \in \mathcal{R} \setminus R$ an arbitrary semantic role. When we try to gather information on null instantiated roles, r is typically an unrealized role label. The PREDICATE consists of the lemma of the corresponding verb or noun, optionally followed by sense number (if predicates are sense-disambiguated) and its part of speech (V/N), e.g., PLAY.01.N.

Paradigm		#Roles			#Overt
		Overt	DNI	INI	#DNI+#INI
Train	FrameNet	2,526	303	277	4.36
	PropBank	1,027	125	101	4.52
Test	FrameNet	3,141	349	361	4.42
	PropBank	1,332	167	85	5.28

Table 1: Label distribution of the SemEval 2010 data set for overt and null instantiated arguments for both the FrameNet (all roles and parts of speech) and the PropBank version (only core roles for nouns and verbs).

We build models from SRL data in PropBank format, both manually and automatically annotated. We experiment with models for two different styles of predicates: *Sense-ignorant* or ***SI model***s represent predicates by lemma and part of speech (PLAY.N), *sense-disambiguated* or ***SD model***s represent predicates by lemma, sense number and part of speech (PLAY.01.N, PLAY.02.N, etc.).

3.4 Annotated Data

In accordance with previous iSRL studies, we evaluate our model on the SemEval data set (Ruppenhofer et al., 2010). However, to the best of our knowledge, this is the first study to focus on the PropBank version of this data set. It has been derived semi-automatically from the FrameNet base format using hand-crafted mapping rules (as part of the data set) for both verbs and nouns. For example, a conversion for the predicate *fear* in FrameNet's EXPERIENCER_FOCUS frame is defined as *fear.01* (its first sense) with the roles EXPERIENCER and CONTENT mapped to PropBank labels A0 and A1, respectively. In accordance with the mapping patterns, the resulting distribution of NIs varies slightly from the base format. Table 1 shows the label distribution of overt roles, DNIs, INIs for both the FrameNet and PropBank versions, respectively. Some information is lost while the general proportions remain similar to the base format. This is also due to the fact that for some parts of speech (e.g., for adjectives) no mappings are defined, even though some of them are annotated with NI information in the FrameNet version. Moreover, mapping rules exist *only for core roles* A0-A4 (agent, patient, ...). As a consequence, we restrict our analysis to these five (unique) roles, even though our models described in this work incorporate probabilistic information for *all possible roles* in $\mathcal{R}$, i.e., A0-A4, but also for *non-core* (modifier) roles, such as AM-TEMP (temporal), AM-LOC (location), etc.

[6] http://wacky.sslmit.unibo.it/doku.php?id=corpora

[7] `http://code.google.com/p/mate-tools/`

[8] In order to minimize the noise in the data, we attempted to resplit unrealistically long sentences ($>$ 90 tokens) by means of the Stanford Core NLP module (Manning et al., 2014). All resulting splits $>$ 70 tokens were rejected.

Role	Verbs		Nouns	
	Overt	NIs	Overt	NIs
A0	40	**45**	24	23
A1	83	39	29	**33**
A2	3	11	10	6
A3	-	7	-	1
A4	-	24	-	-
totals:	126	126	63	63

Table 2: Label distributions of all roles in both data sets from Experiment 1; majority NI classes in bold.

4 Experiments

4.1 Experiment 1

To evaluate the general usefulness of our memory-based approach to detect implicit roles, we set up a simplified framework for predicates with exactly *one overt argument and one NI* annotated in the SemEval data (for all verbs and all nouns and from both the train and test files to obtain a reasonably large sample; no differentiation of DNIs and INIs). This pattern accounts for 189 instances—roughly 9% of the data samples in the SemEval set. We divided the instances into two subsets based on the predicate's part of speech. The label distributions over overt and null instantiated roles for both verbal and nominal predicates are given in Table 2.

4.1.1 Task Description

Predict the role of the single missing NI (A0–A4) for each given predicate instance.

4.1.2 Predicting Null Instantiations

We trained one sense-disambiguated (*SD*) gold model for verbs (*PB*) and one for nouns (*NB*) according to Sect. 3.3 on the complete PropBank and the complete NomBank, respectively. This was compared with 30 separate *SD* and *SI* models on varying portions of the automatically annotated *WaCkypedia_EN* dump: These were trained on the first k sentences each, in order to make their prediction quality comparable, while k ranges from 50 sentences for the smallest model to $k = 10$ million for the largest model ($\approx \frac{1}{5}$ of the whole corpus). For NI role prediction, we return n_i, i.e., the maximally probable unrealized semantic role given the overt argument o_j plus the predicate:

$$n_i = \arg\max_{n \in \mathcal{R} \setminus R} P(n | o_j, \text{PREDICATE}),$$

where $R = \{o_j\}$, the predicate's single explicit role and $\mathcal{R} = \{A0..A4\} \supset R$, the role inventory.

4.1.3 Results & Evaluation

The prediction accuracies for verbal and nominal predicates are illustrated in Figure 1. Although the number of instances in the data sets is small, some general trends are clearly visible. Our major findings are:

By increasing the number of training sentences the performance of the *SD* and the *SI*-based classification models steadily increases as well. The trend is the same for both verbs and for nouns, even though training in the nominal domain requires more data to obtain similarly good results. More precisely, models trained on only 50k sentences already have an adequate performance on test data for verbs ($\approx$76% with the *SD* model). To reach a similar performance on nouns, we need to increase the training size roughly by a factor of 5.

Likewise, the performance of the *SD* models is better in general than the one of the *SI* models throughout all models analyzing verbal predicates, but only marginally better for nouns.

Both the *SD* and the *SI* models outperform the majority class baseline for both parts of speech.[9]

Also, with 800k sentences for nouns and only 50k sentences for verbs, both *SD* model types reach accuracies equal to or greater than the supervised *PB* and *NB* (gold) models which have been trained on the complete PropBank and NomBank corpus including sense distinctions, respectively.

The classification accuracies for the *SD* models reach their saturated maxima for verbs at around 91.27% (115/126) with 6 million training sentences and 85.71% (54/63) with 2.85 million sentences for nouns. For verbs, a χ^2 test confirms a significant ($p < .01$) improvement of our best model over the *PB* gold model. On the sparse evaluation data for nouns, the improvement over the *NB* gold model is, however, not significant.

Taken together, the improvements confirm that memory-based learning over mass data of automatically annotated (explicit) semantic roles can actually outperform gold models constructed from corpora with manual SRL annotations, even if the tools for automated mass annotation were trained on the very same corpora used to build the gold models (PropBank, NomBank). Also, the experiment demonstrated the feasibility of predicting implicit roles solely using information about the distribution of explicit roles. For the artificially

[9] 35.71% with only 1k training sentences (verbs), 52.38% with 50k sentences (nouns).

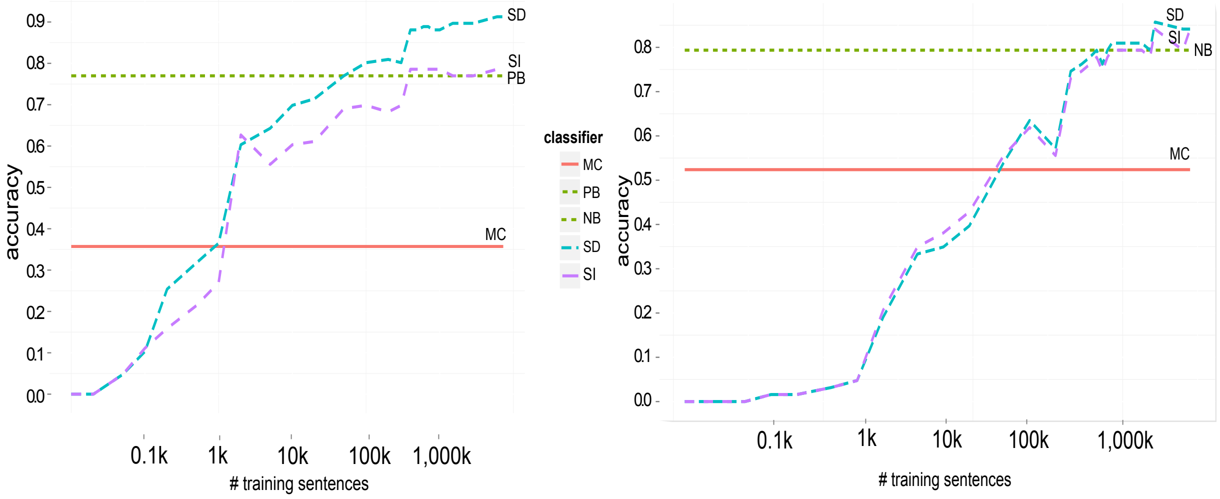

Figure 1: Prediction accuracies for verbal (left figure) and nominal predicates (right figure) from Experiment 1. Majority class (*MC*) baselines in red, PropBank (*PB*) and NomBank (*NB*) gold models in green. The log-scaled x-axis only refers to the *SD* and *SI* models and indicates first k sentences used for training.

simplified NI patterns in Experiment 1, already small portions of automatically annotated SRL data are sufficient to yield adequate results for both types (DNIs and INIs). Sense disambiguation of predicates generally increases the performance.[10]

4.2 Experiment 2

The setup from the previous experiment is by far too simplistic compared to a real linguistic scenario. Usually, a predicate can have an arbitrary number of overt arguments, and similarly the number of missing NIs varies. To tackle this problem, we take the original train and test split (744 vs. 929 unrestricted frame instances of the form: any combination of overt roles vs. any combination of NI roles per predicate). Again, we do not draw a distinction between DNIs and INIs, but treat them generally as NIs. Table 3 shows the distribution of the different NI role patterns in the test data.

4.2.1 Task Description

Given a predicate and its overtly expressed arguments (ranging from any combination of A0 to A4 or none), predict the correct set of null instantiations (which can also be empty or contain up to five different implicit elements).

NI Pattern	Freq	NI Pattern	Freq
-	706	A0 A2	7
A1	86	A1 A2	6
A0	51	A3	5
A2	35	A1 A4	3
A4	18	A0 A1 A2	1
A0 A1	11		

Table 3: The 929 NI role patterns from the test set sorted by their number of occurrence. Most of the predicates are saturated and do not seek an implicit argument. Only one predicate instance has three implicit roles.

4.2.2 Predicting Null Instantiations

We distinguish two main types of classifiers: *supervised classifiers* are directly obtained from NI annotations in the SemEval training data, *mildly supervised classifiers* instead use only information about (automatically obtained) explicitly realized semantic roles in a given corpus, *hybrid classifiers* combine both sources of information. We estimated all parameters optimizing F-measure on the train section of the SemEval data set. Their performance is evaluated on its test section. We aim to demonstrate that mildly supervised classifiers are capable of predicting implicit roles, and to study whether NI annotations can be used to improve their performance.

Baseline: Given the diversity of possible patterns, it is hard to decide how a suitable and competitive baseline should be defined: predicting the majority class means not to predict anything. So, instead, we predict implicit argument roles randomly, but in a way that emulates their frequency distribution in the SemEval data (cf. Tab. 3), i.e., predict

[10] A simple error analysis of the misclassified noun instances revealed that classification on the test data suffers from sparsity issues: In the portions of the *WaCkypedia_EN* that we used for model building, three predicates were not attested (twice *murder.01* and once *murderer.01*). This has a considerable impact on test results.

Classifier	A	B_1	B_2	C_0	C_1	C_2	C_3	C_4	$C4_{n,v}$	$C4_{n,v,B1}$	$C4_{n,v,B2}$
Precision	*0.149*	0.848	**0.853**	*0.368*	*0.378*	0.398	0.400	0.400	0.423	0.561	0.582
Recall	*0.075*	*0.155*	*0.206*	**0.861**	0.851	0.837	0.837	0.837	0.782	0.615	0.814
F_1 Score	*0.100*	*0.262*	*0.332*	*0.516*	*0.523*	*0.540*	*0.541*	*0.541*	*0.549*	0.589	**0.679**

Table 4: Precision, recall and F_1 scores for all classifiers introduced in Experiment 2. Scores are compared row-wise to the best-performing classifier $C4_{n,v,B2}$. A significant improvement over a cell entry with $p < .05$ is indicated in *italics*.

no NIs with a probability of 76.0% (706/929), A1 with 38.6% (86/929), etc. The baseline scores are averaged over 100 runs of this random 'classifier', further referred to as A.

Supervised classifier: Supervised classifiers, as understood here, are classifiers that use the information obtained from manual NI annotations. We set up *two* predictors B_1 and B_2 tuned on the SemEval training set: B_1 is obtained by counting for each predicate its *most frequent NI role pattern*. For instance, for *seem.02*—once annotated with implicit A1, but twice without implicit arguments—B_1 would predict an empty set of NIs. B_2 is similar to B_1 but conditions NI role patterns not only on the predicate, but also on its explicit arguments.[11] For prediction, these classifiers consult the most frequent NI pattern observed for a predicate (B_2: plus its overt arguments). If a test predicate is unknown (i.e., not present in the training data), we predict the majority class (empty set) for NI.

Mildly supervised classifier: Mildly supervised classifiers do not take any NI annotation into account. Instead, they rely on explicitly realized semantic roles observed in a corpus, but use explicit NI annotations only to estimate prediction thresholds. We describe an extension of our prediction method from Exp. 1 and present eight parameter-based classification algorithms for our best-performing *SD* model from Exp. 1, trained on 6 million sentences.

We define prediction for classifier $\mathbf{C_0}$ as follows: Given a predicate PREDICATE, the role inventory $\mathcal{R} = \{A0..A4\}$, its (possibly empty) set of overt roles $R \subseteq \mathcal{R}$ and a fixed, predicate-independent threshold t_0. We start by optimizing threshold t_0 on all predicate instances with *no* given overt argument. If there is *no* overt role and an unrealized role $n_i \in \mathcal{R}$ for which it is true that $P(n_i|\text{PREDICATE}) > t_0$, then predict n_i as an implicit role. If there is an overt role $o_j \in R$ and an unrealized role $n_i \in \mathcal{R} \setminus R$ for which it is true that $P(n_i|o_j,\text{PREDICATE}) > t_0$, then predict n_i as an implicit role. Note that C_0 requires that this condition to hold for *one* o_j, not all explicit arguments of the predicate instance (logical disjunction).

We refine this classifier by introducing an additional parameter that accounts for the group of overtly realized frames with exactly *one* overt argument, i.e., $\mathbf{C_1}$ predicts n_i if $P(n_i|o_j,\text{PREDICATE}) > t_1$; for all other configurations the procedure is the same as in C_0, i.e., the threshold t_0 is applied.

Classifiers $\mathbf{C_2}$, $\mathbf{C_3}$ and $\mathbf{C_4}$ extend C_1 accordingly and introduce additional thresholds t_2, t_3, t_4 for the respective number of overt arguments. For example, $\mathbf{C_3}$ predicts n_i if $P(n_i|o_{j_1}, o_{j_2}, o_{j_3},\text{PREDICATE}) > t_3$, for configurations with less arguments, it relies on C_2, etc. Our general intuition here is to see whether the increasing number of specialized parameters for increasingly marginal groups of frames is justified by the improvements we achieve in this way.

A final classifier $\mathbf{C4_{n,v}}$ extends C_4 by distinguishing verbal and nominal predicates, yielding a total of ten parameters $t_{0_n}..t_{4_n}, t_{0_v}..t_{0_n}$.

Hybrid classifier: To explore to what extent explicit NI annotations improve the classification results, we combine the best-performing and most elaborate mildly supervised classifier $C4_{n,v}$ with the supervised classifiers B_1 and B_2: For predicates encountered in the training data, $\mathbf{C4_{n,v,B_1}}$ (resp., $\mathbf{C4_{n,v,B_2}}$) uses B_1 (resp., B_2) to predict the most frequent pattern observed for the predicate; for unknown predicates, apply the threshold-based procedure of $C4_{n,v}$.

4.2.3 Results & Evaluation

Table 4 contains the evaluation scores for the individual parameter-based classifiers. All classifiers demonstrate significant improvements over the random baseline. Also the mildly supervised

[11] Specifically, we extract finer-grained patterns, e.g., *evening.01*[A1] → {}=2, {A2}=3, where a predicate is associated with its overt role(s) (left side of the arrow). The corresponding implicit role patterns and their number of occurrence is shown to the right.

classifiers outperform the supervised algorithms in terms of F_1 score and recall. However, detecting NIs by the supervised classifiers is very accurate in terms of high precision. Classifier B_2 outperforms B_1 as a result of directly incorporating additional information about the overt arguments.

Concerning our parameter-based classifiers, the main observations are: First, the overall performance (F_1 score) increases from C_0 to C_4 (yet not significantly). Secondly, with more parameters, recall decreases while precision increases. We can observe, however, that improvements from C_2 to C_4 are marginal, at best, due to the sparsity of predicates with two or more overt arguments. Similar problems related to data sparsity have been reported in Chen et al. (2010). Results for C_3 and C_4 are identical, as no predicate with more than three overt arguments occurred in the test data. Encoding the distinction between verbal and nominal predicates into the classifier again slightly increases the performance.

A combination of the high-precision supervised classifiers and the best performing mildly supervised algorithm yields a significant boost in performance (Tab. 4, last two columns). The optimal parameter values for all classifiers $C_{4_{n,v}}$ estimated on the train section of the SemEval data set are given in Table 5.

Noun thresholds	$t_{C_{0_n}}$	$t_{C_{1_n}}$	$t_{C_{2_n}}$	$t_{C_{3_n}}$	$t_{C_{4_n}}$
Values	0.35	0.10	0.20	0.35	0.45
Verb thresholds	$t_{C_{0_v}}$	$t_{C_{1_v}}$	$t_{C_{2_v}}$	$t_{C_{3_v}}$	$t_{C_{4_v}}$
Values	0.05	0.25	0.25	0.30	0.20

Table 5: Optimal parameter values for the thresholds in all $C_{4_{n,v}}$ classifiers estimated on the train section of the SemEval data set.

In Table 6, we report the performance of our best classifier $C_{4_{n,v,B2}}$ with detailed label scores. Its overall NI recognition rate of 0.81 (recall) outperforms the state-of-the-art in implicit role identification: cf. L&P (0.66), SEMAFOR (0.63), S&F (0.58), T&D (0.54), VENSES++ (0.08).[12]

Summarizing our results, Exp. 2 has shown that combining supervised and mildly supervised strategies to NI detection achieves the best results on the SemEval test set. Concerning the mildly supervised, parameter-based classifiers, it has proven beneficial to incorporate a maximum of available information on overtly expressed arguments in order to determine implicit roles. Our best-performing classifier achieves NI recognition rate beyond state-of-the-art.

Roles	A0	A1	A2	A3	A4
# Labels	70	107	49	5	21
Precision	0.675	0.578	0.432	0.400	0.791
Recall	0.800	0.897	0.653	0.400	0.905
F_1 Score	0.732	0.703	0.520	0.400	0.844

Table 6: Evaluation of $C_{4_{n,v,B2}}$ for all 252 implicit roles.

Interestingly, memory-based learning offers the capability to detect both DNIs (resolvable from context), as well as INIs (not resolvable from context), simply by learning patterns from local explicit role realizations. Subsequent experiments should extend this approach to distinguish between the two types, as well, which we have treated equivalently in our settings. First promising experiments in this direction are being conducted in Chiarcos and Schenk (2015).

5 Summary and Outlook

We have presented a novel, statistical method to infer evidence for implicit roles from their explicit realizations in large amounts of automatically annotated SRL data. We conclude that—especially when annotated training data is sparse—memory-based approaches to implicit role detection seem highly promising. With a much greater degree of flexibility, they offer an alternative solution to static rule-/template-based methods.

Despite its simplicity, we demonstrated the suitability of our approach: It is competitive with state-of-the-art systems in terms of the overall recognition rate, however, still suffers in precision of the respective null instantiated arguments. Thus, directions for future research should consider integrating additional contextual features, and would benefit from the *complete* role inventory of our models (including non-core roles). In this extended setting, we would like to experiment with other machine learning approaches to assess whether the accuracy of the detected NIs can be increased. Also, we plan to apply the memory-based strategy described in this paper to NI *resolution* (on top their detection), and in this context, examine more closely the characteristic (possibly contrastive) distributions of DNIs and INIs.

[12] Note that only an indirect comparison of these scores is possible due to the aforementioned difference between data formats and also because none of the other systems report precision scores for their pattern-based NI detection systems.

References

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet Project. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics - Volume 1*, ACL '98, pages 86–90, Stroudsburg, PA, USA. Association for Computational Linguistics.

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. Abstract Meaning Representation for Sembanking. Proc. Linguistic Annotation Workshop.

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

Anders Björkelund, Love Hafdell, and Pierre Nugues. 2009. Multilingual Semantic Role Labeling. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL 2009): Shared Task*, pages 43–48, Boulder, Colorado, June. Association for Computational Linguistics.

Desai Chen, Nathan Schneider, Dipanjan Das, and Noah A. Smith. 2010. SEMAFOR: Frame Argument Resolution with Log-linear Models. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, SemEval '10, pages 264–267, Stroudsburg, PA, USA. Association for Computational Linguistics.

Christian Chiarcos and Niko Schenk. 2015. (accepted) Towards the Unsupervised Acquisition of Implicit Semantic Roles. In *Recent Advances in Natural Language Processing, RANLP 2015, September, 2015, Hissar, Bulgaria.*

Christian Chiarcos. 2012. Towards the Unsupervised Acquisition of Discourse Relations. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers - Volume 2*, ACL '12, pages 213–217, Stroudsburg, PA, USA. Association for Computational Linguistics.

Walter Daelemans and Antal van den Bosch. 2009. *Memory-Based Language Processing*. Cambridge University Press, New York, NY, USA, 1st edition.

Parvin Sadat Feizabadi and Sebastian Padó. 2015. Combining Seemingly Incompatible Corpora for Implicit Semantic Role Labeling. In *Proceedings of STARSEM*, pages 40–50, Denver, CO.

Charles J. Fillmore. 1986. Pragmatically Controlled Zero Anaphora. In *Proceedings of Berkeley Linguistics Society*, pages 95–107, Berkeley, CA.

Matthew Gerber and Joyce Chai. 2012. Semantic Role Labeling of Implicit Arguments for Nominal Predicates. *Comput. Linguist.*, 38(4):755–798, December.

Egoitz Laparra and German Rigau. 2012. Exploiting Explicit Annotations and Semantic Types for Implicit Argument Resolution. In *Sixth IEEE International Conference on Semantic Computing, ICSC 2012.*, Palermo, Italy. IEEE Computer Society.

Egoitz Laparra and German Rigau. 2013. ImpAr: A Deterministic Algorithm for Implicit Semantic Role Labelling. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1180–1189. Association for Computational Linguistics.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60.

Adam Meyers, Ruth Reeves, Catherine Macleod, Rachel Szekely, Veronika Zielinska, Brian Young, and Ralph Grishman. 2004. The NomBank Project: An Interim Report. In A. Meyers, editor, *HLT-NAACL 2004 Workshop: Frontiers in Corpus Annotation*, pages 24–31, Boston, Massachusetts, USA, May 2 - May 7. Association for Computational Linguistics.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The Proposition Bank: An Annotated Corpus of Semantic Roles. *Comput. Linguist.*, 31(1):71–106, March.

Anselmo Peñas and Eduard Hovy. 2010. Semantic Enrichment of Text with Background Knowledge. In *Proceedings of the NAACL HLT 2010 First International Workshop on Formalisms and Methodology for Learning by Reading*, FAM-LbR '10, pages 15–23, Stroudsburg, PA, USA. Association for Computational Linguistics.

Michael Roth and Anette Frank. 2013. Automatically Identifying Implicit Arguments to Improve Argument Linking and Coherence Modeling. In *Proceedings of the Second Joint Conference on Lexical and Computational Semantics (*SEM)*, pages 306–316, Atlanta, Georgia, USA, June. Association for Computational Linguistics.

Josef Ruppenhofer, Caroline Sporleder, Roser Morante, Collin Baker, and Martha Palmer. 2010. SemEval-2010 Task 10: Linking Events and Their Participants in Discourse. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, SemEval '10, pages 45–50, Stroudsburg, PA, USA. Association for Computational Linguistics.

Josef Ruppenhofer, Philip Gorinski, and Caroline Sporleder. 2011. In Search of Missing Arguments: A Linguistic Approach. In Galia Angelova, Kalina Bontcheva, Ruslan Mitkov, and Nicolas Nicolov, editors, *RANLP*, pages 331–338. RANLP 2011 Organising Committee.

Josef Ruppenhofer. 2005. Regularities in Null Instantiation. Ms, University of Colorado.

Carina Silberer and Anette Frank. 2012. Casting Implicit Role Linking as an Anaphora Resolution Task. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics - Volume 1: Proceedings of the Main Conference and the Shared Task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, SemEval '12, pages 1–10, Stroudsburg, PA, USA. Association for Computational Linguistics.

Sara Tonelli and Rodolfo Delmonte. 2010. VENSES++: Adapting a Deep Semantic Processing System to the Identification of Null Instantiations. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 296–299. Association for Computational Linguistics.

Sara Tonelli and Rodolfo Delmonte. 2011. Desperately Seeking Implicit Arguments in Text. In *Proceedings of the ACL 2011 Workshop on Relational Models of Semantics*, pages 54–62. Association for Computational Linguistics.