Combining Seemingly Incompatible Corpora for Implicit Semantic Role Labeling

Parvin Sadat Feizabadi and **Sebastian Padó**
Institut für Maschinelle Sprachverarbeitung
University of Stuttgart, Germany
{parvin.feizabadi,pado}@ims.uni-stuttgart.de

Abstract

Implicit semantic role labeling, the task of retrieving locally unrealized arguments from wider discourse context, is a knowledge-intensive task. At the same time, the annotated corpora that exist are all small and scattered across different annotation frameworks, genres, and classes of predicates. Previous work has treated these corpora as incompatible with one another, and has concentrated on optimizing the exploitation of single corpora. In this paper, we show that *corpus combination* is effective after all when the differences between corpora are bridged with domain adaptation methods. When we combine the SemEval-2010 Task 10 and Gerber and Chai noun corpora, we obtain substantially improved performance on both corpora, for all roles and parts of speech. We also present new insights into the properties of the implicit semantic role labeling task.

1 Introduction

Semantic role labeling (SRL) is the task of identifying semantic arguments of predicates in text. It is an important step in text analysis and has applications in information extraction (Christensen et al., 2010), question answering (Shen and Lapata, 2007; Moreda et al., 2011) and machine translation (Wu and Fung, 2009; Xiong et al., 2012) . A large body of work exists on algorithms for SRL (Gildea and Jurafsky, 2002; Srikumar and Roth, 2011). Their success is closely connected to the availability of two large, hand-constructed semantic role resources, FrameNet (Fillmore et al., 2003) and PropBank (Palmer et al., 2005). They used to concentrate on *overt* semantic roles, that is, semantic roles that are realized within the local syntactic structure of the predicate.

Recent years have seen a broadening of the focus in SRL to *implicit semantic roles*, that is all roles that remain locally unrealized but can be retrieved in the (typically prior) context (Ruppenhofer et al., 2010). In the following example annotated with PropBank roles (cf. Section 2), the target predicate *come* has two roles, a locally realized one (A1, the entity in motion), it, and an implicit role mentioned in the previous sentence (A4, the goal):

> Well, sir, it's [$_{\text{A4}}$ this lonely, silent house] and the queer thing in the kitchen . ... I thought [$_{\text{A1}}$ it] had **come** again.

Implicit SRL is useful to complete predicates' argument structures for inference (Mirkin et al., 2010) and paraphrasing (Roth and Frank, 2013), or to assess the coherence of discourse (Burchardt et al., 2005). It however requires (even) more training data than traditional SRL. One reason is that potential arguments come from the whole text rather than just the sentence. Another one is that most of the powerful syntactic features that are a staple in traditional SRL are unavailable across sentence boundaries. Unfortunately, existing corpora for implicit SRL are quite small: The task requires full-text annotation, which is time-consuming and pushes semantic role frameworks to their limits (Palmer and Sporleder, 2010). It is also hard to do consistently, and can only be crowdsourced in limited settings (Feizabadi and Padó, 2014). Thus, even though multiple systems for implicit SRL exist (among others, Tonelli and Delmonte (2011), Laparra and Rigau (2012), Silberer

*Proceedings of the Fourth Joint Conference on Lexical and Computational Semantics (*SEM 2015)*, pages 40–50,
Denver, Colorado, June 4–5, 2015.

and Frank (2012)), results are still relatively poor.

In this paper, we focus on the fact that the corpora that exist for implicit SRL differ not only in the semantic role frameworks used (FrameNet vs. PropBank), but also in genre (newswire vs. novels), and classes of annotated predicates (verbs vs. nouns). As a result, they are generally regarded as incompatible, and previous work has concentrated on getting most out of individual corpora, or spending annotation effort on focused extensions of these corpora. Instead, we will follow the intuition that the performance of implicit SRL can be improved significantly by *combining corpora*, using simple domain adaptation techniques to bridge the differences between them. We combine the two largest datasets for implicit SRL, the SemEval-2010 Task 10 dataset (Ruppenhofer et al., 2010) and the Gerber and Chai dataset (Gerber and Chai, 2012). This combination achieves improvements across all target and semantic roles despite the differences in genre, domain, and parts of speech. Our analyses indicates that the properties of the implicit SRL task – where syntactic features play a relatively minor role compared to semantic and discourse features – are responsible for this picture, and mean that models can actually profit from complementarity between combined corpora.

Plan of the paper. Section 2 summarizes the resource and model situation in SRL. Section 3 defines a simple system for implicit SRL that uses domain adaptation. Sections 4 and 5 report experiments and provide analysis. Section 6 concludes.

2 Traditional and Implicit SRL

This section first describes existing resources for traditional and implicit SRL (frameworks and corpora). Then it outlines the state of the art in modeling.

2.1 Frameworks for Semantic Roles

Almost all contemporary work on SRL is based on one of two frameworks: FrameNet and PropBank.

FrameNet is a dictionary and corpus annotated in the Frame Semantics paradigm (Fillmore et al., 2003). In Frame Semantics, the meaning of predicates (verbs, nouns, or adjectives) is conveyed by frames, conceptual structures which represent situations and define salient entities. Semantic roles describe these salient entities and are therefore located at the level of frames. E.g., the verb *approach* is analyzed as an instance of the frame ARRIVING, with the roles THEME, SOURCE, GOAL:

> [$_{\text{Theme}}$ He] was **approaching** [$_{\text{Source}}$ from behind and slightly to the right of Sharpe].

Frame Semantics also offers an analysis of unrealized roles, called Null Instantiations, that distinguishes three classes. Indefinite non-instantiations (INIs) are interpreted generically. Constructional non-instantiations (CNI) include, e.g., passives. Finally, definite non-instantiations (DNIs) have a specific interpretation and often refer to expressions in the context. DNIs correspond to the pre-theoretic concept of implicit roles. The FrameNet corpus, however, does not annotate the antecedents of DNIs, so it cannot be used directly as training data for implicit SRL.

PropBank The second major framework for semantic role annotation is PropBank (Palmer et al., 2005). It defines a set of general semantic roles named ARG0-ARG5 of which ARG0 and ARG1 are interpreted as proto-agent and proto-patient (Dowty, 1991), respectively. The higher-numbered roles receive more predicate-specific interpretations. These "core" roles are complemented by adjunct roles such as MNR (manner) or TMP (time). For example,

> Jim Unruh ... said [$_{\text{A1}}$ he] is **approaching** [$_{\text{A2}}$ next year] [$_{\text{MNR}}$ with caution].

PropBank has annotated the WSJ part of the Penn Treebank, i.e., newswire text, exhaustively with semantic roles. While it originally concentrated on verbs, the NomBank project (Meyers et al., 2004) extended the annotation scheme to nouns. PropBank does not have a specific taxonomy of null instantiations like FrameNet, but it can nevertheless be used equally for implicit role annotation.

2.2 Annotated Corpora for Implicit SRL

FrameNet and PropBank are both very large corpora, covering tens of thousands of instances. Corpora with implicit role annotation are generally much smaller; the main corpora are summarized in Table 1.

Ruppenhofer et al. Arguably the first corpus with a substantial set of annotations for implicit roles was

Corpus	Scheme	POS	Genre	# predicates	# instances	# implicit roles
Ruppenhofer et al. (2010)	FrameNet	V, N	Novels	801	1575	245
Gerber & Chai (2012)	PropBank	N	Newswire	10	1253	1172
Moor et al. (2013)	FrameNet	V	Newswire	5	1992	242
Feizabadi & Padó (2014)	FrameNet	V	Novels	10	384	363

Table 1: Size of available English corpora with implicit semantic role annotation

created for SemEval 2010 Task 10 (Ruppenhofer et al., 2010). This dataset covers a number of chapters from Arthur Conan Doyle short stories and provides full-text annotation of both explicit and implicit semantic roles. The texts were annotated manually with FrameNet roles. This dataset is a de-facto standard benchmark for implicit SRL.

Gerber and Chai. A study by Gerber and Chai (2012) investigated implicit arguments of NomBank nominalizations. They extended a part of the PropBank corpus with implicit roles for 10 nominal predicates, of which they annotated all instances.

Further Corpora with Implicit Role Annotation. Moor et al. (2013) created a corpus with all annotated instances for five verbs with the goal of focused improvement of implicit SRL. Feizabadi & Padó (2014) investigated the use of crowdsourcing to create annotations for implicit roles. Both corpora are more restricted in size and scope than the first two.

2.3 Models for Semantic Role Labeling

Traditional SRL. A broad range of models have been proposed for "traditional", i.e., local SRL (Palmer et al., 2010). The task can be seen as a sequence of two classification tasks, predicate disambiguation and role labeling. Earlier models modeled them in a pipeline architecture, but recent works demonstrates the benefits of joint inference (Srikumar and Roth, 2011; Das et al., 2014). SRL models have drawn on a wide variety of features from two main groups: syntactic features describing the structural relation between predicate and argument candidate, and semantic features describing role and candidate. A general observation is that SRL models are lexically specific to a substantial degree, i.e., do not generalize very well between predicates, so that the availability of annotations remains a bottleneck.

Implicit SRL was formulated by SemEval 2010 Task 10 in two versions. The "full task" includes identification of all (explicit or implicit) semantic roles of the target predicate. The "null instantiation task" is the subtask of the full task concerned only with the identification and labeling of antecedents for implicit roles. It assumes that predicates and overt roles are already available. We follow the lead of almost all models for implicit SRL on the null instantiation task. Structurally, it can be approached similarly to role identification in traditional SRL.

The first systems on large-coverage implicit SRL adopted traditional SRL modeling techniques (Chen et al., 2010; Tonelli and Delmonte, 2010). but struggled with the scarcity of training data for the complex task. Work since then has concentrated on tapping into novel knowledge and data sources. There are three main directions. The first one is knowledge about *semantic types*. This includes Ruppenhofer et al. (2011) who extract semantic types for null instantiations from FrameNet and Laparra and Rigau (2012) who learn distributions over semantic types for each role from explicit role annotations in FrameNet. Similarly, Roth and Frank (2013) retrieve overt instances of implicit roles from comparable corpora. The second direction is *discourse level knowledge*. Laparra and Rigau (2013) and Gorinski et al. (2013) treat implicit SRL as a task similar to anaphor resolution, which motivates the use of several features of discourse such as distance and salience. A third set of studies concentrated on simply obtaining *more annotated instances*. Silberer and Frank (2012) use an entity-based coreference resolution model to automatically extended the training set. Moor et al. (2013) and Feizabadi and Padó (2014) manually construct focused corpora (cf. Section 2.2).

3 Combining Corpora for Implicit SRL

3.1 Rationale and Challenges

Despite the progress made by on implicit SRL, as discussed in the previous section, *data sparsity* remains the main bottleneck. This has two main reasons.

First, the set of constitutents included in the search for each role is very large, potentially including the whole discourse. To address this problem, implicit SRL systems typically concentrate on a window of n sentences, typically the sentence with the predicate and its preceding discourse. Second, the powerful class of syntactic features becomes largely unavailable beyond sentence boundaries.

This situation calls for large, richly annotated corpora. Unfortunately, the annotation effort that has been expended on implicit role has been distributed over a number of different corpora, all of which are fairly small (cf. Section 2.2). The question that we are asking in this paper is: *Can data from existing corpora be combined rather than spending annotation effort on yet another corpus?*

We will consider the combination of the standard benchmark, the SemEval 2010 Task 10 dataset (Ruppenhofer et al., 2010) (henceforth SEMEVAL), with the corpus with the largest number of implicit roles, the Gerber and Chai (2012) corpus (henceforth GERBERCHAI). The main challenge in this endeavour is that these corpora have very different properties (cf. Table 1). Consequently, a number of challenges arise for data combination. Below we discuss them, our expectations, and our strategies to address them.

Challenge: Differences in Role Framework. SEMEVAL was annotated with FrameNet roles, while GERBERCHAI was annotated with PropBank roles. While semi-automatic conversion schemes now exist in both directions, we decided to adopt the PropBank paradigm, working on the basis of the semi-automatically converted SEMEVAL annotation provided by the task organizers. The reasons are twofold: (a), we believe that, in parallel to results on traditional SRL, PropBank roles should be generally easier to label than FrameNet roles; (b), this effect should be particularly pronounced when facing sparse data problems, as is the case here.

Challenge: Differences in Parts of Speech. SEMEVAL covers both verbal and nominal predicates, while GERBERCHAI contains only nominal predicates (cf. Table 1). Given the absence of syntactic features from implicit SRL, we believe that this is not a huge impediment. We will, however, evaluate on a per-POS basis to test this assumption.

Challenge: Differences in Genre/Domain. Also, SEMEVAL is based on novels dealing with everyday affairs, while GERBERCHAI consists of newswire text focusing on finance and politics. It is well known that the performance of NLP models degrades when applied across domains and genres. This holds for traditional SRL (Carreras and Màrquez, 2005) and is likely to extend to the implicit variant. For this reason, we believe that it is crucial to apply domain adaptation methods to ensure that reasonable generalizations can be learned. See Section 3.3 for details.

3.2 A Simple Implicit SRL System

We now describe the simple classification-based system for implicit SRL that we will use in our experiments. Like many systems from the literature, it focuses on the "null instantiation" step (cf. Section 2.2) – i.e., we assume that overtly realized roles are already available. The architecture of our system is inspired by the system by Laparra and Rigau (2012) which is among the best-performing systems on SEMEVAL.

Our system decomposes the task into two steps: (1), Determining a set of implicit roles that should be identified in context; (2) Determining the antecedents of these missing roles. For the first step, we extract the *predominant role set* (i.e., most frequently realized set) for each predicate by searching the predicate in a large corpus, OntoNotes (Hovy et al., 2006). We assume that all instances of the predicate realize these roles and select the subset that is not realized overtly for inclusion in the second step.

We phrase the second step as binary classification. The items to be classified are triples ⟨target predicate, implicit role, candidate realization⟩. The set of candidate realizations is defined as all constituents from the target predicate's sentence and the two prior sentences which do not fill an explicit role for the target. We employ a Naive Bayes classifier that can deal relatively well with sparse data.[1] We use 10 features, shown in Table 1 which attempt to capture relevant syntacto-semantic and the discourse features.

3.3 Domain Adaptation

The standard assumption in machine leaning is that data are independent and identically distributed, that is, drawn from the same underlying population. This

[1] We also experimented with other classifiers including SVMs, but did not achieve better results.

Name	Description
Expected roles	Set of roles required by the target predicates (based on PropBank and NomBank). This feature serves as a delexicalized target representation.
Semantic Type	Semantic type of the candidate realization's head word (WordNet supersenses) or, if pronoun, of the next content words in the coreference chain
Word Frequency	Lemma frequency of the candidate filler's head word
POS	Part of Speech of candidate realization's head word
Constituent type	The constituent type of the candidate filler, e.g. NP, PP, VP, etc.
Distance	Distance between candidate realization and target predicate (in sentences)
Salience	Whether the candidate realization's head word is included in a non-singleton coreference chain
Previous Role	Whether the candidate realization has overtly realized any semantic role in the dataset
Same Role	Whether the candidate realization has realized the implicit role as an overt role in the dataset
Role Percentage	The percentage with which the candidate realization has realized the implicit role

Table 2: Feature Set (above: syntacto-semantic features; below: discourse features)

assumption is violated if the test data differs substantially from the training data, and consequently the performance of models learned on the training data suffers on the test data. Since this situation arises frequently, the field of *domain adaptation* has developed (Jiang, 2008). In our application, SEMEVAL and GERBERCHAI can be understood as two domains.

We adopt Daumé's (2007) simple but effective *feature augmentation* method which makes use of some training data in both source and target domain. Each feature is stored in three variants: a general version, a source version and a target version. Each of the two domains (source and target) activates two versions, the general one and its specific one, which can also be given a Bayesian interpretation (Finkel and Manning, 2009). In this manner, the model balances global and domain-specific trends against each other. As an example, the "expected roles" feature (cf. Table 2), which is shaped by subcategorization, is a likely candidate for changess across domains, due to sense shifts. In contrast, we would not expect the part-of-speech features of realization candidates to undergo major changes across domains.

4 Experiment 1

We present three experiments. Experiment 1 extends the SEMEVAL training data with out-of-domain data from GERBERCHAI and evaluates on SEMEVAL. Experiment 2 swaps the setup, extending the GERBERCHAI dataset with SEMEVAL data and evaluating on GERBERCHAI. Experiment 3 aims at providing a better understanding of these observations.

4.1 Experimental Setup

Design. In this experiment, we evaluate our approach on the SEMEVAL dataset (SEMEVAL is the target domain and GERBERCHAI is the source domain). Since there is an established split of SEMEVAL into training and test parts, we simply use the test part for evaluation, and designate the SEMEVAL training part as well as GERBERCHAI for training.

We compare four experimental scenarios (cf. Table 3): (1) The standard "in-domain" setup that only uses SEMEVAL, as assumed by most studies on the dataset. (2) A pure "out-of-domain" setup where we use only GERBERCHAI as training data. Of course, there is reason to believe that this strategy will perform quite poorly. (3) A simple "concatenation" setup where we train on the union of GERBERCHAI and the SEMEVAL training corpus. (4) The feature augmentation setting where we train on the combined corpus, but apply Daumé's (2007) learning method.

Preprocessing. SEMEVAL comes pre-parsed with the Collins (Collins, 1997) parser. We parsed GERBERCHAI with the same parser, ignoring the Penn Treebank gold trees. Since all datasets are manually annotated with semantic roles, no overt SRL is necessary. Coreference information, which we require for some features, is available from manual annotation in the SEMEVAL test part, but not for the other datasets. We computed coreference chains with the Stanford CoreNLP tools (Manning et al., 2014).

Evaluation. We evaluate implicit role predictions with precision, recall, and F_1 score, following the official SemEval 2010 Task 10 guidelines. Note that

Training Set	Pr.	Rec.	F_1
(1) SEMEVAL train (in-domain)	0.10	0.20	0.13
(2) GERBERCHAI (out-of-domain)	0.12	0.08	0.10
(3) SEMEVAL train + GERBERCHAI, concat.	0.11	0.19	0.14
(4) SEMEVAL train + GERBERCHAI, feature augmentation	**0.13**	**0.30**	**0.18**
Laparra and Rigau (2013)	0.12	0.16	0.14

Table 3: Evaluation of implicit SRL (PropBank roles) on the SEMEVAL test set

% of GERBERCHAI	Pr.	Rec.	F_1
0	0.10	0.20	0.13
5	0.13	0.29	0.17
10	**0.14**	**0.31**	**0.19**
15	0.13	0.31	0.18
20	0.13	0.30	0.18
100	0.13	0.30	0.18

Table 4: Results on SEMEVAL test, training on SEMEVAL train plus varying amounts of data from GERBERCHAI

according to the guidelines, the true positives include all predictions that match the gold span indirectly through a (manually annotated) coreference chain.

Baseline. All previous studies on the SEMEVAL dataset used the FrameNet annotation, and without access to the actual predictions we cannot directly compare our predictions to theirs. We are grateful to Laparra and Rigau who agreed to share the predictions of their 2013 model with us, which is, at the time of writing, the system with the second-best reported scores. We converted the predictions into the PropBank format, using the FrameNet-to-PropBank mapping provided by the task organizers.

Upper bound. Implicit SRL systems typically trade off recall against precision by restricting the search space. Our system uses two heuristics: It restricts search to the current and two preceding sentences and to the predominant role set (cf. Section 3.2). The upper bound in recall on SEMEVAL test that can still be achieved in this setting is 60.1%.

4.2 Results

Table 3 shows the results of the four experimental conditions defined above and the comparison system, the converted Laparra and Rigau (2013). Our system, trained in-domain (1), achieves a performance comparable to Laparra and Rigau, albeit with a different precision-recall trade-off. Not surprisingly, pure out-of-domain training (2) does not perform well either. Simple data concatenation (3) leads to a minimal numeric improvement, but indicates that the datasets are indeed rather different.

We see a substantial improvement in performance when feature augmentation (4) is used. There is not only a major improvement in recall (+10 percentage points) but also a smaller improvement in precision (+3 points). We tested the difference to the in-domain model (1) for significance with bootstrap resampling (Efron and Tibshirani, 1993) and found it to be higly significant ($p < 0.01$). In sum, we see an improvement of 5% F-Score, despite the differences between the corpora, when feature augmentation is used. Notably, we achieve a high recall, despite the upper bound imposed by the filtering heuristics.

Unfortunately, it is rather difficult to pinpoint individual instances whose improvements can be interpreted in a linguistically meaningful way. A comparative feature ablation study for models (1) and (4) showed that discourse features such as Previous Role (cf. Table 2) are among the most important features in (4), while they are almost useless in (1). This indicates that discourse-level features particularly profit from the inclusion of out-of-domain data.

Analysis by Amount of Out-of-Domain Data. Since GERBERCHAI is about ten times as large as the SEMEVAL training set, we wondered whether the out-of-domain GERBERCHAI data is "overwhelming" the SEMEVAL data. Keeping the SEMEVAL test set for evaluation, we combined SEMEVAL train with subsets of GERBERCHAI in increments of 5% of the total number of predicates. The results, shown in Table 4, show that almost the complete benefit of the GERBERCHAI data is already present when we add 5% of GERBERCHAI, and we achieve the optimal result by adding 10%. The results are marginally higher than when we add the complete GERBERCHAI (difference not significant). Our take away is that, in contrast to the proposal by Moor et al. (2013), we do not require many annotations for each predicate: the results are best when the in-domain and out-of-domain corpora have about the same size.

	Verbal predicates			Nominal predicates		
Training Set	Pr.	Rec.	F_1	Pr.	Rec.	F_1
(1) SEMEVAL train ("in-domain")	0.11	0.20	0.14	0.10	0.21	0.14
(2) GERBERCHAI train ("out-of-domain")	0.09	0.12	0.10	0.07	0.11	0.09
(3) SEMEVAL train + GERBERCHAI, concat.	0.11	0.18	0.13	0.11	0.21	0.14
(4) SEMEVAL train + GERBERCHAI, feature aug.	0.13	**0.30**	**0.18**	**0.14**	**0.32**	**0.20**
Laparra and Rigau (2013)	**0.15**	0.20	0.17	0.09	0.11	0.09

Table 5: Evaluation of implicit SRL (PropBank roles) on the SEMEVAL test set, by target part of speech

	A0			A1			A2		
Training Set	Pr.	Rec.	F_1	Pr.	Rec.	F_1	Pr.	Rec.	F_1
(1) SEMEVAL train ("in-domain")	0.19	0.29	0.23	0.09	0.26	0.13	0.06	0.10	0.07
(2) GERBERCHAI ("out-of-domain")	0.19	0.34	0.24	0.03	0.06	0.03	0.0	0.0	0.0
(3) SEMEVAL train + GERBERCHAI, concat.	0.23	0.34	0.27	0.08	0.22	0.11	0.0	0.0	0.0
(4) SEMEVAL train + GERBERCHAI, feature aug.	**0.24**	**0.42**	**0.31**	**0.11**	**0.37**	**0.17**	0.09	**0.24**	0.13
Laparra and Rigau (2013)	0.21	0.28	0.24	0.10	0.13	0.11	**0.13**	0.19	**0.15**

Table 6: Evaluation of implicit SRL (PropBank roles) on the SEMEVAL test set, by role

Analysis by Predicate POS. Since GERBERCHAI contains only noun targets, we could hypothesize that its inclusion improves results in SEMEVAL specifically for nominal predicates. To test this hypothesis, we evaluated verbal and nominal predicates separately. The results in Table 5 are actually comparable across parts-of-speech. Even though the benefit is somewhat smaller for verbs, there is still a substantial improvement (+4.1% F_1 for verbs; +5.9% F_1 for nouns). In contrast, studies on traditional SRL found only small (albeit consistent) improvements for extending training sets with instances of targets with different parts-of-speech (Li et al., 2009).

We believe that this is the case because implicit SRL, as discussed in Section 2.3, can rely less on syntactic features but must make predictions on the basis of semantic and discourse features, which are more comparable across target parts of speech. Consider these two examples – one verbal and one nominal predicate – of implicit A0 roles. Both occur in the same sentences as their predicates, but outside their syntactic domains:

> SEMEVAL: The wagonette was **paid off** ... while [$_{A0}$ we] started walking.
> GERBERCHAI: His ... house ... is up for **sale** to pay for [$_{A0}$ his] lawyers.

While the role realizations are quite different structurally (subject vs. posessive), they are similar on the semantics and discourse levels: both are pronouns referring to agent-like entities and are realized in the immediately following discourse.

Analysis by Role. Finally, we performed an evaluation by individual semantic roles, shown in Table 6, to assess to what extent differences in role distribution between SEMEVAL and GERBERCHAI influence the improvements. We concentrate on A0 through A2, since A3 and A4 are so infrequent in SEMEVAL that evaluation results are not reliable.

Not surprisingly, we see the overall best results for A0, followed by A1 and A2. The improvement for combining corpora correlates with the overall performance: +7% F_1 for A0, +4% for A1, +6% for A2. The overall pattern of a major boost to recall and a minor one to precision are also stable across roles. Thus, corpus combination seems to benefit all roles as well. A notable observation is the inability of the naive out-of-domain models (2) and (3) to correctly predict any A2 roles. The reason is that for the nominal targets in GERBERCHAI, A2 is an *incorporated role*, that is, realized by the predicate itself. This pattern hardly occurs in SEMEVAL. Interestingly, the domain adaptation model (4) manages to extract relevant information from GERBERCHAI. Nevertheless, the fact that (4) is still worse than Laparra & Rigau (2013), which is trained just in-domain, indicates that more informative features for A2 are also necessary.

Training Set	A0 Pr.	A0 Rec.	A0 F_1	A1 Pr.	A1 Rec.	A1 F_1
(1) GERBERCHAI ("in-domain")	0.15	0.10	0.12	0.18	0.23	0.16
(4) SEMEVAL + GERBERCHAI, feature augmentation	**0.19**	**0.13**	**0.15**	**0.26**	**0.35**	**0.30**

Table 7: Evaluation of implicit SRL (PropBank roles) on the SEMEVAL test set, by role

Training Set	Pr.	Rec.	F_1
(1) GERBERCHAI (in-domain)	0.16	0.10	0.12
(2) SEMEVAL (out-of-domain)	0.11	0.06	0.07
(3) SEMEVAL + GERBERCHAI, concat.	0.16	0.09	0.11
(4) SEMEVAL + GERBERCHAI, feature augmentation	**0.24**	**0.18**	**0.21**
Upper bound: Gerber & Chai (2012)	0.58	0.44	0.50

Table 8: Evaluation of implicit SRL (PropBank roles) on GERBERCHAI (3-fold CV)

5 Experiment 2

In Experiment 2, we use a combination of GERBERCHAI and the complete SEMEVAL for training and evaluate on GERBERCHAI. The main question is whether the addition of the (much smaller) SEMEVAL corpus to GERBERCHAI can improve performance.

We consider the same four conditions as in Experiment 1. To obtain reliable results, we split GERBERCHAI into three equal-sized parts and report averages over three cross-validation runs where we always use two thirds for training and one third for testing. Evaluation also is performed as before, with the exception that in the absence of manually annotated coreference chains, we only count direct matches as true positives. The upper bound for recall on this dataset (using the same 3-sentence window and predominant role set) is rather low, at 44%, which reflects the structural tendency of nominalizations to realize few roles locally.

Unfortunately, we do not have a directly comparable competitor, since Laparra and Rigau did not run their system on GERBERCHAI data. The results obtained by Gerber and Chai (2012) are not directly comparable, since their approach was hand-tailored towards nominal implicit SRL in the newswire domain. It incorporates a large number of detailed linguistic resources (Penn Treebank, Penn Discourse Bank, NomBank, FrameNet) and assumes gold standard information on all levels. We therefore see this system as an upper bound rather than as a competitor.

The results are shown in Table 8. The overall patterns are very similar to Experiment 1: out-of-domain training (2) works worse than in-domain training (1), and simple concatenation (3) does not improve over in-domain training. With feature augmentation, however, we see a significant improvement of 9% in precision, recall and F_1. The difference is highly significant at $p< 0.01$. This confirms the effectiveness of corpus combination, despite the small size of the added SEMEVAL dataset compared to GERBERCHAI. It is also clear, however, that the results are much worse than the upper bound set by Gerber and Chai.

Table 7 subdivides the results by semantic roles for (1), as the in-domain baseline, and (4), as the best model. Again, we see improvements for both A0 and A1, both regarding precision and recall. Interestingly, the improvements as well as the performance for A1 exceed those for A0 – a difference to the SEMEVAL results, where we found the best results for A0.

6 Experiment 3

In Experiments 1 and 2, we have found an improvement for including out-of-domain data. However, it is unclear so far whether the improvements are simply due to the increased amount of training data, or to the training data becoming more *varied*. To distinguish between these two hypotheses, Experiment 3 keeps the total size of the training set constant and varies the proportions of the two source corpora, SEMEVAL and GERBERCHAI, in 10% increments, from 100% SEMEVAL to 100% GERBERCHAI. The size of the training set is limited by the smaller one of the training sets (SEMEVAL, cf. Table 1).

As before, we apply feature augmentation and train models, which we now evaluate on both the SEMEVAL and GERBERCHAI test sets. If the improvements we have seen before are solely due to the larger size of the training sets, we expect to see the highest performance for the 100% in-domain training

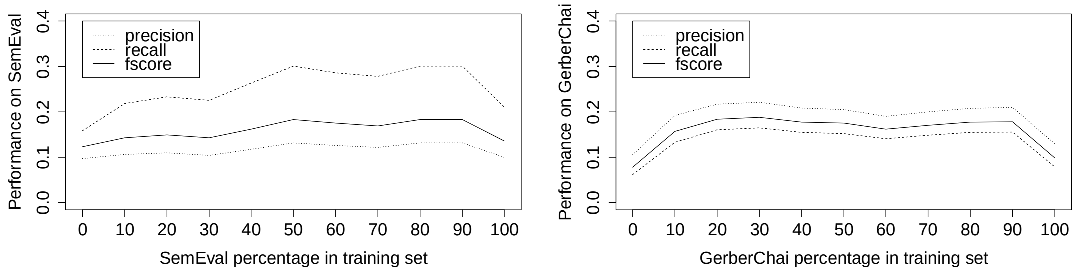

Figure 1: Evaluation of models trained on a constant-size training set with changing composition

set, and decreasing performance with more out-of-domain data. If however variety is important, we expect to see a maximum somewhere between the two extremes, at the point where there is enough out-of-domain training data to introduce variety but not enough to overwhelm the in-domain data.

Figure 1 shows the results. On both test sets, we do *not* see the best result for 100% in-domain data – there is a substantial improvement moving from 100% to 90% in-domain data (from 0.13 to 0.18 F-Score on SemEval and from 0.10 to 0.18 on GerberChai). On the SemEval test set, the result for 90% is the (tied) best result. We see minor variation until roughly the 50-50 split and then a mild degradation to the cases where the GerberChai training data dominates, consistent with Experiment 1. On the GerberChai test set, we see a more symmetrical picture, with relatively constant performance for almost all mixtures. We see degradation for the both "pure" (100%) training sets, but still better performance for in-domain than for out-of-domain (100% GerberChai: 0.10; 100% SemEval: 0.08).[2]

Overall, the results are compatible with the second, but not the first hypothesis: the models do seem to profit from the combination of different corpora even when this does not involve larger training sets.

7 Conclusion

This paper has reviewed the state-of-the-art in implicit semantic role labeling (SRL) where scarcity of training data is the major bottleneck. We have argued that rather than annotating new datasets, researchers should gauge the potential for *combining existing corpora*, even if they are very different at first glance.

We have presented experiments on two standard corpora, the SemEval 2010 Task 10 corpus (novels) and Gerber and Chai's nominalization corpus (newswire). They demonstrate that systems trained on either corpus can benefit substantially from combination with the other one. More specifically, we find that (a) domain adaptation techniques are helpful to bridge the differences between corpora; (b) improvements from corpus combination apply surprisingly uniformly to different roles and different parts of speech; (c) improvements can be obtained from relatively small amounts of "out-of-domain" data.

Further analyses have indicated that it is indeed the *complementarity* of the corpora, rather than the addition of training data, which is responsible for the improvement. This suggests that rather than annotating as many instances as possible, researchers should concentrate on annotating instances that are as *varied* as possible, similar to uncertainty sampling in active learning (Lewis and Gale, 1994). In future work, we will experiment with combining more than two corpora to test the scalability of the present approach.

An open question is to what extent the benefits that we see for implicit SRL generalize to other tasks. We believe that two factors combine to give us the present picture: the first one is the set of properties of implicit SRL as a task where semantic and discourse features play important roles. The second one is simply the low baseline performance; overall better models are presumably harder to improve.

Acknowledgments. We would like thank Roman Klinger for helpful suggestions.

[2] Note that these numbers do not match Experiment 2, since the training set in this experiment is much smaller.

References

Aljoscha Burchardt, Anette Frank, and Manfred Pinkal. 2005. Building text meaning representations from contextually related frames – a case study. In *Proceedings of the International Conference on Computational Semantics*, pages 66–77, Tilburg, Netherlands.

Xavier Carreras and Lluís Màrquez. 2005. Introduction to the CoNLL-2005 Shared Task: Semantic Role Labeling. In *Proceedings of the Conference on Computational Natural Language Learning 2005*, pages 152–164, Ann Arbor, MI.

Desai Chen, Nathan Schneider, Dipanjan Das, and Noah A. Smith. 2010. SEMAFOR: Frame argument resolution with log-linear models. In *Proceedings of the International Workshop on Semantic Evaluation*, pages 264–267, Uppsala, Sweden.

Janara Christensen, Mausam, Stephen Soderland, and Oren Etzioni. 2010. Semantic role labeling for open information extraction. In *Proceedings of the NAACL HLT 2010 First International Workshop on Formalisms and Methodology for Learning by Reading*, pages 52–60, Los Angeles, CA.

Michael Collins. 1997. Three generative, lexicalised models for statistical parsing. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics*, pages 16–23, Madrid, Spain.

Dipanjan Das, Desai Chen, André FT Martins, Nathan Schneider, and Noah A Smith. 2014. Frame-semantic parsing. *Computational Linguistics*, 40(1):9–56.

Hal Daumé III. 2007. Frustratingly easy domain adaptation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 256–263, Prague, Czech Republic.

David Dowty. 1991. Thematic Proto-Roles and Argument Selection. *Language*, 67:547–619.

Bradley Efron and Robert J. Tibshirani. 1993. *An Introduction to the Bootstrap*. Chapman and Hall, New York.

Parvin Sadat Feizabadi and Sebastian Padó. 2014. Crowdsourcing annotation of non-local semantic roles. In *Proceedings of the Annual Meeting of the European Chapter of the Association for Computational Linguistics*, pages 226–230, Gothenburg, Sweden.

Charles J. Fillmore, Christopher R. Johnson, and Miriam R.L. Petruck. 2003. Background to FrameNet. *International Journal of Lexicography*, 16(3):235–250.

Jenny Rose Finkel and Christopher D Manning. 2009. Hierarchical Bayesian Domain Adaptation. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 602–610, Boulder, Colorado, USA.

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

Daniel Gildea and Daniel Jurafsky. 2002. Automatic labeling of semantic roles. *Computational Linguistics*, 28(3):245–288.

Philip Gorinski, Josef Ruppenhofer, and Caroline Sporleder. 2013. Towards weakly supervised resolution of null instantiations. In *Proceedings of the International Conference on Computational Semantics*, pages 119–130, Potsdam, Germany.

Eduard H Hovy, Mitch Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2006. OntoNotes: The 90% Solution. In *Proceedings of the Joint Human Language Technology Conference and Annual Meeting of the North American Chapter of the Association for Computational Linguistics*, pages 57–60, New York City, NY.

Jing Jiang. 2008. A literature survey on domain adaptation of statistical classifiers. Technical report, Department of Computer Science, University of Illinois at Urbana-Champaign.

Egoitz Laparra and German Rigau. 2012. Exploiting explicit annotations and semantic types for implicit argument resolution. In *Proceedings of the International Conference on Semantic Computing*, pages 75–78, Palermo, Italy.

Egoitz Laparra and German Rigau. 2013. Sources of evidence for implicit argument resolution. In *Proceedings of the International Conference on Computational Semantics*, pages 155–166, Potsdam, Germany.

David D. Lewis and William A. Gale. 1994. A sequential algorithm for training text classifiers. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3–12, Dublin, Ireland.

Junhui Li, Guodong Zhou, Hai Zhao, Qiaoming Zhu, and Peide Qian. 2009. Improving nominal SRL in Chinese language with verbal SRL information and automatic predicate recognition. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1280–1288, Singapore.

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 the Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60, Baltimore, MD.

Adam Meyers, Ruth Reeves, Catherine Macleod, Rachel Szekely, Veronika Zielinska, Brian Young, and Ralph Grishman. 2004. The NomBank project: An interim report. In *Proceedings of the HLT-NAACL 2004 workshop on Frontiers in Corpus Annotation*, pages 24–31, Boston, MA, USA.

Shachar Mirkin, Ido Dagan, and Sebastian Padó. 2010. Assessing the Role of Discourse References in Entailment Inference. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1209–1219, Uppsala, Sweden.

Tatjana Moor, Michael Roth, and Anette Frank. 2013. Predicate-specific annotations for implicit role binding: Corpus annotation, data analysis and evaluation experiments. In *Proceedings of the International Conference on Computational Semantics*, pages 369–375, Potsdam, Germany.

Paloma Moreda, Hector Llorens, Estela Saquete, and Manuel Palomar. 2011. Combining semantic information in question answering systems. *Information Processing & Management*, 47(6):870–885.

Alexis Palmer and Caroline Sporleder. 2010. Evaluating FrameNet-style semantic parsing: the role of coverage gaps in FrameNet. In *Proceedings of the International Conference on Computational Linguistics*, pages 928–936, Beijing, China.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The proposition bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–106.

Martha Palmer, Dan Gildea, and Nianwen Xue. 2010. *Semantic Role Labeling*. Morgan & Claypool.

Michael Roth and Anette Frank. 2013. Automatically identifying implicit arguments to improve argument linking and coherence modeling. In *Proceedings of the Joint Conference on Lexical and Computational Semantics*, pages 306–316, Atlanta, GA.

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 International Workshop on Semantic Evaluation*, pages 45–50, Uppsala, Sweden.

Josef Ruppenhofer, Philip Gorinski, and Caroline Sporleder. 2011. In search of missing arguments: A linguistic approach. In *Proceedings of the Conference on Recent Advances in Natural Language Processing*, pages 331–338, Hissar, Bulgaria.

Dan Shen and Mirella Lapata. 2007. Using semantic roles to improve question answering. In *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Conference on Computational Natural Language Learning*, pages 12–21, Prague, Czech Republic.

Carina Silberer and Anette Frank. 2012. Casting implicit role linking as an anaphora resolution task. In *Proceedings of the Joint Conference on Lexical and Computational Semantics*, pages 1–10, Jeju Island, South Korea.

Vivek Srikumar and Dan Roth. 2011. A joint model for extended semantic role labeling. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 129–139, Edinburgh, United Kingdom.

Sara Tonelli and Rodolfo Delmonte. 2010. VENSES++: Adapting a deep semantic processing system to the identification of null instantiations. In *Proceedings of the International Workshop on Semantic Evaluation*, pages 296–299, Uppsala, Sweden.

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, Portland, OR, USA.

Dekai Wu and Pascale Fung. 2009. Semantic roles for SMT: A hybrid two-pass model. In *Proceedings of the Annual Meeting of the North American Chapter of the Association for Computational Linguistics*, pages 13–16, Boulder, CO.

Deyi Xiong, Min Zhang, and Haizhou Li. 2012. Modeling the translation of predicate-argument structure for SMT. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 902–911, Jeju Island, South Korea.