SEMAFOR: Frame Argument Resolution with Log-Linear Models

Desai Chen Nathan Schneider Dipanjan Das Noah A. Smith
School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213, USA
{desaic@andrew,dipanjan@cs,nschneid@cs,nasmith@cs}.cmu.edu

Abstract

This paper describes the SEMAFOR system's performance in the SemEval 2010 task on linking events and their participants in discourse. Our entry is based upon SEMAFOR 1.0 (Das et al., 2010a), a frame-semantic probabilistic parser built from log-linear models. The extended system models *null instantiations*, including non-local argument reference. Performance is evaluated on the task data with and without gold-standard overt arguments. In both settings, it fares the best of the submitted systems with respect to recall and F_1.

1 Introduction

The theory of frame semantics (Fillmore, 1982) holds that meaning is largely structured by holistic units of knowledge, called *frames*. Each frame encodes a conventionalized gestalt event or scenario, often with conceptual dependents (participants, props, or attributes) filling roles to elaborate the specific instance of the frame. In the FrameNet lexicon (Fillmore et al., 2003), each frame defines **core roles** tightly coupled with the particular meaning of the frame, as well as more generic **non-core** roles (Ruppenhofer et al., 2006). Frames can be evoked with linguistic predicates, known as **lexical units (LUs)**; role fillers can be expressed overtly and linked to the frame via (morpho)syntactic constructions. However, a great deal of conceptually-relevant content is left unexpressed or is not explicitly linked to the frame via linguistic conventions; rather, it is expected that the listener will be able to infer the appropriate relationships pragmatically. Certain types of implicit content and implicit reference are formalized in the theory of **null instantiations (NIs)** (Fillmore, 1986; Ruppenhofer, 2005). A complete frame-semantic analysis of text thus incorporates covert *and* overt predicate-argument information.

In this paper, we describe a system for frame-semantic analysis, evaluated on a semantic role labeling task for explicit and implicit arguments (§2). Extending the SEMAFOR 1.0 frame-semantic parser (Das et al., 2010a; outlined in §3), we detect null instantiations via a simple two-stage pipeline: the first stage predicts *whether* a given role is null-instantiated, and the second stage (§4) predicts *how* it is null-instantiated, if it is not overt. We report performance on the SemEval 2010 test set under the full-SRL and NI-only conditions.

2 Data

The SemEval 2007 task on frame-semantic parsing (Baker et al., 2007) provided a small (about 50,000 words and 2,000 sentences) dataset of news text, travel guides, and bureaucratic accounts of weapons stockpiles. Sentences in this dataset were fully annotated with frames and their arguments. The SemEval 2010 task (Ruppenhofer et al., 2010) adds annotated data in the fiction domain: parts of two Sherlock Holmes stories by Arthur Conan Doyle. The SemEval 2010 **training set** consists of the SemEval 2007 data plus one document from the new domain. This document has about 7800 words in 438 sentences; it has 1492 annotated frame instances, including 3169 (overt and null-instantiated) argument annotations. The **test set** consists of two chapters from another story: Chapter 13 contains about 4000 words, 249 sentences, and 791 frames; Chapter 14 contains about 5000 words, 276 sentences, and 941 frames (see also Table 3). Figure 1 shows two annotated test sentences. All data released for the 2010 task include part-of-speech tags, lemmas, and phrase-structure trees from a parser, with head annotations for constituents.

3 Argument identification

Our starting point is SEMAFOR 1.0 (Das et al., 2010a), a discriminative probabilistic frame-semantic parsing model that operates in three stages: (a) rule-based target selection, (b) probabilistic disambiguation that resolves each target to a FrameNet frame, and (c) joint selection of text spans to fill the roles of each target through a second probabilistic model.[1]

[1] Das et al. (2010a) report the performance of this system on the complete SemEval 2007 task at 46.49% F_1.

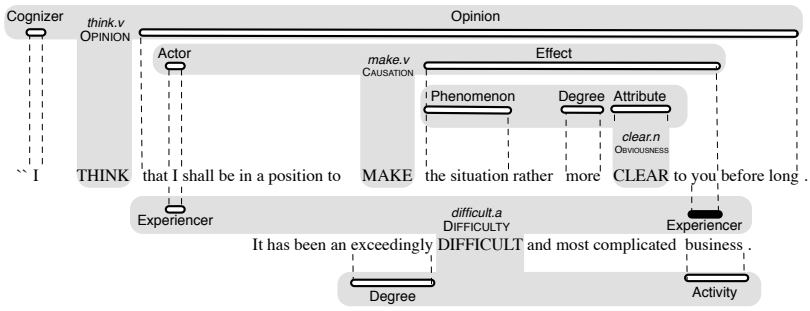

Figure 1. Two consecutive sentences in the test set, with frame-semantic annotations. Shaded regions represent frames: they include the target word in the sentence, the corresponding frame name and lexical unit, and arguments. Horizontal bars mark gold argument spans—white bars are gold annotations and black bars show mistakes of our NI-only system.

	Chapter 13			Chapter 14		
Training Data	**Prec.**	**Rec.**	F_1	**Prec.**	**Rec.**	F_1
SemEval 2010 data (includes SemEval 2007 data)	0.69	0.50	0.58	0.66	0.48	0.56
SemEval 2007 data + 50% new, in-domain data	0.68	0.47	0.55	0.66	0.45	0.54
SemEval 2007 data only	0.67	0.41	0.50	0.64	0.40	0.50

Table 1. Overt argument labeling performance.

Stage (c), known as **argument identification** or SRL, is most relevant here. In this step, the system takes the target (frame-evoking) phrase t and corresponding frame type f predicted by the previous stages, and independently fills each role of f with a word or phrase from the sentence, or the symbol OTHER to indicate that the role has no (local) overt argument. Features used to inform this decision include aspects of the syntactic dependency parse (e.g. the path in the parse from the target to the argument); voice; word overlap of the argument with respect to the target; and part-of-speech tags within and around the argument. SEMAFOR as described in (Das et al., 2010a) does not distinguish between different types of null instantiations or find non-local referents. Given perfect input to stage (c), the system achieved 68.5% F_1 on the SemEval 2007 data (exact match, evaluating overt arguments only). The only difference in our use of SEMAFOR's argument identification module is in preprocessing the training data: we use dependency parses transformed from the head-augmented phrase-structure parses in the task data.

Table 1 shows the performance of our argument identification model on this task's test data. The SRL systems compared in (Ruppenhofer et al., 2010) all achieved precision in the mid 60% range, but SEMAFOR achieved substantially higher recall, F_1, and label accuracy on this subtask. (The table also shows how performance of our model degrades when half or all of the new data are not used for training; the 9% difference in recall suggests the importance of in-domain training data.)

4 Null instantiation detection

In this subtask, which follows the argument identification subtask (§3), our system seeks to characterize non-overt core roles given gold standard local frame-argument annotations. Consider the following passage from the test data:

> "That's lucky for him—in fact, it's lucky for all of you, since you are all on the wrong side of the law in this matter. I am not sure that as a conscientious detective [Authorities my] first duty is not to arrest [Suspect the whole household]. [DNI Charges ∅]

The frame we are interested in, ARREST, has four core roles, two of which (Authorities and Suspect) have overt (local) arguments. The third core role, Charges, is annotated as having anaphoric or **definite null instantiation (DNI)**. "Definite" means that the discourse implies a specific referent that should be recoverable from context, without marking that referent linguistically. Some DNIs in the data are linked to phrases in syntactically non-local positions, such as in another sentence (see Figure 1). This one is not (though our model incorrectly labels *this matter* from the previous sentence as a DNI referent for this role). The fourth core role, Offense, is not annotated as a null instantiation because it belongs to the same **CoreSet** as Charges—which is to say they are relevant in a similar way to the frame as a whole (both pertain to the rationale for the arrest) and only one is typically expressed.[2] We will use the term **masked** to refer to any non-overt core role which does not need to be specified as null-instantiated due to a structural connection to another role in its frame.

The typology of NIs given in Ruppenhofer (2005) and employed in the annotation distinguishes anaphoric/definite NIs from existential or **indefinite null instantiations (INIs)**. Rather than having a specific referent accessible in the discourse, INIs are left vague or deemphasized, as in

[2] If the FrameNet lexicon marks a pair of roles within a frame as being in a CoreSet or Excludes relationship, then filling one of them satisfies the requirement that the other be (expressly or implicitly) present in the use of the frame.

	Training Data	Chapter 13 Prec.	Chapter 13 Rec.	Chapter 13 F_1	Chapter 14 Prec.	Chapter 14 Rec.	Chapter 14 F_1
NI-only	SemEval 2010 new: 100%	0.40	0.64	0.50	0.53	0.60	0.56
	SemEval 2010 new: 75%	0.66	0.37	0.50	0.70	0.37	0.48
	SemEval 2010 new: 50%	0.73	0.38	0.51	0.75	0.35	0.48
Full	All	0.35	0.55	0.43	0.56	0.49	0.52

Table 2. Performance on the full task and the NI-only task. The NI model was trained on the new SemEval 2010 document, "The Tiger of San Pedro" (data from the 2007 task was excluded because none of the null instantiations in that data had annotated referents).

Gold \ *Predicted*	overt	DNI	INI	masked	inc.	total
overt	**2068 (1630)**	5	362	327	0	2762
DNI	64	**12 (3)**	182	90	0	348
INI	41	2	**214**	96	0	353
masked	73	0	240	**1394**	0	1707
inc.	12	2	55	2	**0**	71
total	2258	21	1053	1909	0	**3688 correct**

Table 3. Instantiation type confusion matrix for the full model (argument identification plus NI detection). Parenthesized numbers count the predictions of the correct type which also predicted the same (argument or referent) span. On the NI-only task, our system has a similar distribution of NI detection errors.

the thing(s) eaten in the sentence *We ate.*

The problem can be decomposed into two steps: (a) *classifying* each null instantiation as definite, indefinite, or masked; and (b) *resolving* the DNIs, which entails finding referents in the non-local context. Instead, our model makes a single NI prediction for any role that received no local argument (OTHER) in the argument identification phase (§3), thereby combining classification and resolution.[3]

4.1 Model

Our model for this subtask is analogous to the argument identification model: it chooses one from among many possible fillers for each role. However, whereas the argument identification model considers parse constituents as potential local fillers (which might constitute an overt argument within the sentence) along with a special category, OTHER, here the set of candidate fillers consists of phrases from outside the sentence, along with special categories INI or MASKED. When selected, a non-local phrase will be interpreted as a non-local argument and labeled as a DNI referent.

These non-local candidate fillers are handled differently from candidates within the sentence considered in the argument identification model: they are selected using more restrictive criteria, and are associated with a different set of features.

Restricted search space for DNI referents. We consider nouns, pronouns, and noun phrases from the previous three sentences as candidate DNI referents. This narrows the search space considerably to make learning tractable, but at a cost: many gold DNI referents will not even be considered. In the training data, there are about 250 DNI instances with explicit referents; their distribution is chaotic.[4] Judging by the training data, our heuristics thus limit oracle recall to about 20% of DNIs.[5]

Modified feature set. Since it is not obvious how to calculate a syntactic path between two words in different sentences, we replaced dependency path features with simpler features derived from FrameNet's lexicographic exemplar annotations. For each candidate span, we use two types of features to model the affinity between the head word and the role. The first indicates whether the head word is used as a filler for this role in at least one of the lexicographic exemplars. The second encodes the maximum distributional similarity to any word heading a filler of that role in the exemplars.[6] In practice, we found that these features received negligible weight and had virtually no effect on performance, possibly due to data sparseness. An additional change in the feature set is that ordering/distance features (Das et al., 2010b, p. 13) were replaced with a feature indicating the number of *sentences* away the candidate is from the target.[7] Otherwise, the null identifica-

[3] Investigation of separate modeling is left to future work.

[4] 91 DNI referents are found no more than three sentences prior; another 90 are in the same sentence as the target. 20 DNIs have referents which are not noun phrases. Six appear *after* the sentence containing its frame target; 28 appear at least 25 sentences prior. 60 have no referent.

[5] Our system ignores DNIs with no referent or with a referent in the same sentence as the target. Experiments with variants on these assumptions show that the larger the search space (i.e. the more candidate DNI referents are under consideration), the worse the trained model performs at distinguishing NIs from non-NIs (though DNI vs. INI precision improves). This suggests that data sparseness is hindering our system's ability to learn useful generalizations about NIs.

[6] Distributional similarity scores are obtained from D. Lin's Proximity-based Thesaurus (`http://webdocs.cs.ualberta.ca/~lindek/Downloads/sims.lsp.gz`) and quantized into binary features for intervals: $[0, .03)$, $[.03, .06)$, $[.06, .08)$, $[.08, \infty)$.

[7] All of the new features are instantiated in three forms:

tion model uses the same features as the argument identification model.

The theory of null instantiations holds that the grammaticality of lexically-licensed NI for a role in a given frame depends on the LU: for example, the verbs *buy* and *sell* share the same frame but differ as to whether the Buyer or Seller role may be lexically null-instantiated. Our model's feature set is rich enough to capture this in a soft way, with lexicalized features that fire, e.g., when the Seller role is null-instantiated and the target is *buy*. Moreover, (Ruppenhofer, 2005) hypothesizes that each role has a strong preference for one interpretation (INI or DNI) when it is lexically null-instantiated, regardless of LU. This, too, is modeled in our feature set. In theory these trends should be learnable given sufficient data, though it is doubtful that there are enough examples of null instantiations in the currently available dataset for this learning to take place.

4.2 Evaluation

We trained the model on the non-overt arguments in the new SemEval 2010 training document, which has 580 null instantiations—303 DNIs and 277 INIs.[8,9] Then we used the task scoring procedure to evaluate the NI detection subtask in isolation (given gold-standard overt arguments) as well as the full task (when this module is combined in a pipeline with argument identification). Results are shown in Table 2.[10]

Table 3 provides a breakdown of our system's predictions on the test data by instantiation type: overt local arguments, DNIs, INIs, and the MASKED category (marking the role as redundant or irrelevant for the particular use of the frame, given the other arguments). It also shows counts for **incorporated** ("inc.") roles, which are filled by the frame-evoking target, e.g. *clear* in Figure 1.[11] This table shows that the system is reasonably effective at discriminating NIs from masked roles, but DNI identification suffers from low recall and INI identification from low precision. Data sparseness is likely the biggest obstacle here. To put this in perspective, there are over 20,000 training examples of overt arguments, but fewer than 600 examples of null instantiations, two thirds of which do not have referents. Without an order of magnitude more NI data (at least), it is unlikely that a supervised learner could generalize well enough to recognize on new data null instantiations of the over 7000 roles in the lexicon.

5 Conclusion

We have described a system that implements a clean probabilistic model of frame-semantic structure, considering overt arguments as well as various forms of null instantion of roles. The system was evaluated on SemEval 2010 data, with mixed success at detecting null instantiations. We believe in-domain data sparseness is the predominant factor limiting the robustness of our supervised model.

Acknowledgments

This work was supported by DARPA grant NBCH-1080004 and computational resources provided by Yahoo. We thank the task organizers for providing data and conducting the evaluation, and two reviewers for their comments.

References

C. Baker, M. Ellsworth, and K. Erk. 2007. SemEval-2007 Task 19: Frame Semantic Structure Extraction. In *Proc. of SemEval*.

D. Das, N. Schneider, D. Chen, and N. A. Smith. 2010a. Probabilistic frame-semantic parsing. In *Proc. of NAACL-HLT*.

D. Das, N. Schneider, D. Chen, and N. A. Smith. 2010b. SEMAFOR 1.0: A probabilistic frame-semantic parser. Technical Report CMU-LTI-10-001, Carnegie Mellon University.

C. J. Fillmore, C. R. Johnson, and M. R.L. Petruck. 2003. Background to FrameNet. *International Journal of Lexicography*, 16(3).

C. J. Fillmore. 1982. Frame semantics. In *Linguistics in the Morning Calm*, pages 111–137. Hanshin Publishing Co., Seoul, South Korea.

C. J. Fillmore. 1986. Pragmatically controlled zero anaphora. In *Proc. of Berkeley Linguistics Society*, pages 95–107, Berkeley, CA.

J. Ruppenhofer, M. Ellsworth, M. R.L. Petruck, C. R. Johnson, and J. Scheffczyk. 2006. FrameNet II: extended theory and practice.

J. Ruppenhofer, C. Sporleder, R. Morante, C. Baker, and M. Palmer. 2010. SemEval-2010 Task 10: Linking Events and Their Participants in Discourse. In *Proc. of SemEval*.

J. Ruppenhofer. 2005. Regularities in null instantiation.

one specific to the frame and the role, one specific to the role name only, and one to learn the overall bias of the data.

[8]For feature engineering we held out the last 25% of sentences from the new training document as development data, retraining on the full training set for final evaluation.

[9]We used Nils Reiter's FrameNet API, version 0.4 (`http://www.cl.uni-heidelberg.de/trac/FrameNetAPI`) in processing the data.

[10]The other system participating in the NI-only subtask had much lower NI recall of 8% (Ruppenhofer et al., 2010).

[11]We do not predict any DNIs without referents or incorporated roles, though the evaluation script gives us credit when we predict INI for these cases.