# Frame-Semantic Role Labeling with Heterogeneous Annotations

**Meghana Kshirsagar**[*] **Sam Thomson**[*] **Nathan Schneider**[†]
**Jaime Carbonell**[*] **Noah A. Smith**[*] **Chris Dyer**[*]
[*]School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, USA
[†]School of Informatics, University of Edinburgh, Edinburgh, Scotland, UK

## Abstract

We consider the task of identifying and labeling the semantic arguments of a predicate that evokes a FrameNet frame. This task is challenging because there are only a few thousand fully annotated sentences for supervised training. Our approach augments an existing model with features derived from FrameNet and PropBank and with partially annotated exemplars from FrameNet. We observe a 4% absolute increase in $F_1$ versus the original model.

## 1 Introduction

Paucity of data resources is a challenge for semantic analyses like frame-semantic parsing (Gildea and Jurafsky, 2002; Das et al., 2014) using the FrameNet lexicon (Baker et al., 1998; Fillmore and Baker, 2009).[1] Given a sentence, a frame-semantic parse maps word tokens to **frames** they evoke, and for each frame, finds and labels its **argument** phrases with frame-specific **roles**. An example appears in figure 1.

In this paper, we address this **argument identification** subtask, a form of semantic role labeling (SRL), a task introduced by Gildea and Jurafsky (2002) using an earlier version of FrameNet. Our contribution addresses the paucity of annotated data for training using standard domain adaptation techniques. We exploit three annotation sources:

- the frame-to-frame relations in FrameNet, by using hierarchical features to share statistical strength among related roles (§3.2),
- FrameNet's corpus of partially-annotated **exemplar** sentences, by using "frustratingly easy" domain adaptation (§3.3), and


**Figure 1:** Part of a sentence from FrameNet full-text annotation. 3 frames and their arguments are shown: DESIRING is evoked by *want*, ACTIVITY_FINISH by *finish*, and HOLDING_OFF_ON by *hold off*. Thin horizontal lines representing argument spans are labeled with role names. (Not shown: *July* and *August* evoke CALENDRIC_UNIT and fill its **Unit** role.)

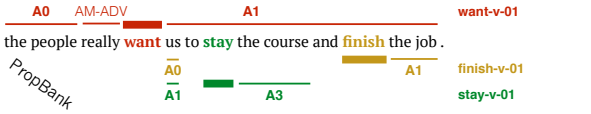



**Figure 2:** A PropBank-annotated sentence from OntoNotes (Hovy et al., 2006). The PB lexicon defines rolesets (verb sense–specific frames) and their core roles: e.g., `finish-v-01` 'cause to stop', `A0` 'intentional agent', `A1` 'thing finishing', and `A2` 'explicit instrument, thing finished with'. (`finish-v-03`, by contrast, means 'apply a finish, as to wood'.) Clear similarities to the FrameNet annotations in figure 1 are evident, though PB uses lexical frames rather than deep frames and makes some different decisions about roles (e.g., `want-v-01` has no analogue to **Focal_participant**).

- a PropBank-style SRL system, by using guide features (§3.4).[2]

These expansions of the *training corpus* and the *feature set* for supervised argument identification are integrated into SEMAFOR (Das et al., 2014), the leading open-source frame-semantic parser for English. We observe a 4% $F_1$ improvement in argument identification on the FrameNet test set, leading to a 1% $F_1$ improvement on the full frame-semantic parsing task. Our code and models are available at `http://www.ark.cs.cmu.edu/SEMAFOR/`.

## 2 FrameNet

FrameNet represents events, scenarios, and relationships with an inventory of **frames** (such as

‡ Corresponding author: `mkshirsa@cs.cmu.edu`

[1] `http://framenet.icsi.berkeley.edu`

[2] Preliminary experiments training on PropBank annotations mapped to FrameNet via SemLink 1.2.2c (Bonial et al., 2013) hurt performance, likely due to errors and coverage gaps in the mappings.



*Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Short Papers)*, pages 218–224, Beijing, China, July 26-31, 2015. 

SHOPPING and SCARCITY). Each frame is associated with a set of **roles** (or **frame elements**) called to mind in order to understand the scenario, and lexical **predicates** (verbs, nouns, adjectives, and adverbs) capable of evoking the scenario. For example, the BODY_MOVEMENT frame has **Agent** and **Body_part** as its core roles, and lexical entries including verbs such as bend, blink, crane, and curtsy, plus the noun use of curtsy. In FrameNet 1.5, there are over 1,000 frames and 12,000 lexical predicates.

### 2.1 Hierarchy

The FrameNet lexicon is organized as a network, with several kinds of **frame-to-frame relations** linking pairs of frames and (subsets of) their arguments (Ruppenhofer et al., 2010). In this work, we consider two kinds of frame-to-frame relations:

**Inheritance:** E.g., ROBBERY inherits from COMMITTING_CRIME, which inherits from MISDEED. Crucially, roles in inheriting frames are mapped to corresponding roles in inherited frames: ROBBERY.**Perpetrator** links to COMMITTING_CRIME.**Perpetrator**, which links to MISDEED.**Wrongdoer**, and so forth.

**Subframe:** This indicates a subevent within a complex event. E.g., the CRIMINAL_PROCESS frame groups together subframes ARREST, ARRAIGNMENT and TRIAL. CRIMINAL_PROCESS.**Defendant**, for instance, is mapped to ARREST.**Suspect**, TRIAL.**Defendant**, and SENTENCING.**Convict**.

We say that a *parent* of a role is one that has either the **Inheritance** or **Subframe** relation to it. There are 4,138 **Inheritance** and 589 **Subframe** links among role types in FrameNet 1.5.

Prior work has considered various ways of grouping role labels together in order to share statistical strength. Matsubayashi et al. (2009) observed small gains from using the **Inheritance** relationships and also from grouping by the role name (SEMAFOR already incorporates such features). Johansson (2012) reports improvements in SRL for Swedish, by exploiting relationships between both frames and roles. Baldewein et al. (2004) learn latent clusters of roles and role-fillers, reporting mixed results. Our approach is described in §3.2.

### 2.2 Annotations

Statistics for the annotations appear in table 1.

**Full-text (FT):** This portion of the FrameNet corpus consists of documents and has about 5,000 sentences for which annotators assigned frames and arguments to as many words as possible. Beginning with the SemEval-2007 shared task on FrameNet analysis, frame-semantic parsers have been trained and evaluated on the full-text data (Baker et al., 2007; Das et al., 2014).[3] The full-text documents represent a mix of genres, prominently including travel guides and bureaucratic reports about weapons stockpiles.

| | **Full-Text** | | **Exemplars** | |
|---|---|---|---|---|
| | *train* | *test* | *train* | *test* |
| Sentences | 2,780 | 2,420 | 137,515 | 4,132 |
| Frames | 15,019 | 4,458 | 137,515 | 4,132 |
| Overt arguments | 25,918 | 7,210 | 278,985 | 8,417 |
| TYPES | | | | |
| Frames | 642 | 470 | 862 | 562 |
| Roles | 2,644 | 1,420 | 4,821 | 1,224 |
| Unseen frames *vs. train:* | | 46 | | 0 |
| Roles in unseen frames *vs. train:* | | 178 | | 0 |
| Unseen roles *vs. train:* | | 289 | | 38 |
| Unseen roles *vs. combined train:* | | 103 | | 32 |

**Table 1:** Characteristics of the training and test data. (These statistics exclude the development set, which contains 4,463 frames over 746 sentences.)

**Exemplars:** To document a given predicate, lexicographers manually select corpus examples and annotate them *only with respect to the predicate in question*. These singly-annotated sentences from FrameNet are called lexicographic **exemplars**. There are over 140,000 sentences containing argument annotations and relative to the FT dataset, these contain an order of magnitude more frame annotations and over two orders of magnitude more sentences. As these were manually selected, the rate of overt arguments per frame is noticeably higher than in the FT data. The exemplars formed the basis of early studies of frame-semantic role labeling (e.g., Gildea and Jurafsky, 2002; Thompson et al., 2003; Fleischman et al., 2003; Litkowski, 2004; Kwon et al., 2004). Exemplars have not yet been exploited successfully to improve role labeling performance on the more realistic FT task.[4]

### 2.3 PropBank

PropBank (PB; Palmer et al., 2005) is a lexicon and corpus of predicate–argument structures that takes a shallower approach than FrameNet. FrameNet frames cluster lexical predicates that evoke sim-

[3] Though these were *annotated* at the document level, and train/development/test splits are by document, the frame-semantic parsing is currently restricted to the sentence level.

[4] Das and Smith (2011, 2012) investigated semi-supervised techniques using the exemplars and WordNet for frame identification. Hermann et al. (2014) also improve frame identification by mapping frames and predicates into the same continuous vector space, allowing statistical sharing.

ilar kinds of scenarios In comparison, PropBank frames are purely lexical and there are no formal relations between different predicates or their roles. PropBank's sense distinctions are generally coarser-grained than FrameNet's. Moreover, FrameNet lexical entries cover many different parts of speech, while PropBank focuses on verbs and (as of recently) eventive noun and adjective predicates. An example with PB annotations is shown in figure 2.

## 3 Model

We use the model from SEMAFOR (Das et al., 2014), detailed in §3.1, as a starting point. We experiment with techniques that augment the model's training data (§3.3) and feature set (§3.2, §3.4).

### 3.1 Baseline

In SEMAFOR, the argument identification task is treated as a structured prediction problem. Let the classification input be a dependency-parsed sentence $\mathbf{x}$, the token(s) $p$ constituting the predicate in question, and the frame $f$ evoked by $p$ (as determined by frame identification). We use the heuristic procedure described by (Das et al., 2014) for extracting candidate argument spans for the predicate; call this $spans(\mathbf{x}, p, f)$. *spans* always includes a special span denoting an empty or non-overt role, denoted $\varnothing$. For each candidate argument $a \in spans(\mathbf{x}, p, f)$ and each role $r$, a binary feature vector $\boldsymbol{\phi}(a, \mathbf{x}, p, f, r)$ is extracted. We use the feature extractors from (Das et al., 2014) as a baseline, adding additional ones in our experiments (§3.2–§3.4). Each $a$ is given a real-valued score by a linear model:

$$score_{\mathbf{w}}(a \mid \mathbf{x}, p, f, r) = \mathbf{w}^{\top} \boldsymbol{\phi}(a, \mathbf{x}, p, f, r) \qquad (1)$$

The model parameters $\mathbf{w}$ are learned from data (§4).

Prediction requires choosing a joint assignment of all arguments of a frame, respecting the constraints that a role may be assigned to at most one span, and spans of overt arguments must not overlap. Beam search, with a beam size of 100, is used to find this arg max.[5]

### 3.2 Hierarchy Features

We experiment with features shared between related roles of related frames in order to capture statistical generalizations about the kinds of arguments seen in those roles. Our hypothesis is that this will be beneficial given the small number of training examples for individual roles.

All roles that have a common parent based on the **Inheritance** and **Subframe** relations will share a set of features in common. Specifically, for each base feature $\phi$ which is conjoined with the role $r$ in the baseline model ($\phi \wedge$ "role=$r$"), and for each parent $r'$ of $r$, we add a new copy of the feature that is the base feature conjoined with the parent role, ($\phi \wedge$ "parent_role=$r'$"). We experimented with using more than one level of the hierarchy (e.g., grandparents), but the additional levels did not improve performance.

### 3.3 Domain Adaptation and Exemplars

Daumé (2007) proposed a feature augmentation approach that is now widely used in supervised domain adaptation scenarios. We use a variant of this approach. Let $\mathcal{D}_{\text{ex}}$ denote the exemplars training data, and $\mathcal{D}_{\text{ft}}$ denote the full text training data. For every feature $\phi(a, \mathbf{x}, p, f, r)$ in the base model, we add a new feature $\phi_{\text{ft}}(\cdot)$ that fires only if $\phi(\cdot)$ fires and $\mathbf{x} \in \mathcal{D}_{\text{ft}}$. The intuition is that each base feature contributes both a "general" weight and a "domain-specific" weight to the model; thus, it can exhibit a general preference for specific roles, but this general preference can be fine-tuned for the domain. Regularization encourages the model to use the general version over the domain-specific, if possible.

### 3.4 Guide Features

Another approach to domain adaptation is to train a supervised model on a source domain, make predictions using that model on the target domain, then use those predictions as additional features while training a new model on the target domain. The source domain model is effectively a form of preprocessing, and the features from its output are known as **guide features** (Johansson, 2013; Kong et al., 2014).[6]

In our case, the full text data is our target domain, and PropBank and the exemplars data are our source domains, respectively. For PropBank, we run the SRL system of Illinois Curator 1.1.4 (Pun-

[5] Recent work has improved upon global decoding techniques (Das et al., 2012; Täckström et al., 2015). We expect such improvements to be complementary to the gains due to the added features and data reported here.

[6] This is related to the technique of model stacking, where successively richer models are trained by cross-validation on the same dataset (e.g., Cohen and Carvalho, 2005; Nivre and McDonald, 2008; Martins et al., 2008).

yakanok et al., 2008)[7] on verbs in the full-text data. For the exemplars, we train baseline SEMAFOR on the exemplars and run it on the full-text data.

We use two types of guide features: one encodes the role label predicted by the source model, and the other indicates that a span $a$ was assigned *some* role. For the exemplars, we use an additional feature to indicate that the predicted role matches the role being filled.

## 4 Learning

Following SEMAFOR, we train using a **local** objective, treating each role and span pair as an independent training instance. We have made two modifications to training which had negligible impact on full-text accuracy, but decreased training time significantly:[8]

- We use the online optimization method AdaDelta (Zeiler, 2012) with minibatches, instead of the batch method L-BFGS (Liu and Nocedal, 1989). We use minibatches of size 4,000 on the full text data, and 40,000 on the exemplar data.
- We minimize squared structured hinge loss instead of a log-linear loss. Let $((\mathbf{x}, p, f, r), a)$ be the $i$th training example. Then the squared hinge loss is given by $L_{\mathbf{w}}(i) =$

$$\left(\max_{a'} \left\{ \begin{array}{c} \mathbf{w}^\top \boldsymbol{\phi}(a', \mathbf{x}, p, f, r) \\ +\mathbf{1}\{a' \neq a\} \end{array} \right\} - \mathbf{w}^\top \boldsymbol{\phi}(a, \mathbf{x}, p, f, r)\right)^2$$

We learn $\mathbf{w}$ by minimizing the $\ell_2$-regularized average loss on the dataset:

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} \frac{1}{N} \sum_{i=1}^{N} L_{\mathbf{w}}(i) + \frac{1}{2}\lambda \|\mathbf{w}\|_2^2 \qquad (2)$$

## 5 Experimental Setup

We use the same FrameNet 1.5 data and train/test splits as Das et al. (2014). Automatic syntactic dependency parses from MSTParserStacked (Martins et al., 2008) are used, as in Das et al. (2014).

**Preprocessing.** Out of 145,838 exemplar sentences, we removed 4,191 sentences which had no role annotations. We removed sentences that appeared in the full-text data. We also merged spans which were adjacent and had the same role label.

[7] http://cogcomp.cs.illinois.edu/page/software_view/SRL

[8] With SEMAFOR's original features and training data, the result of the above changes is that full-text $F_1$ decreases from 59.3% to 59.1%, while training time (running optimization to convergence) decreases from 729 minutes to 82 minutes.

| Training Configuration (Features) | Model Size | P (%) | R (%) | $F_1$ (%) |
|---|---|---|---|---|
| FT (Baseline) | 1.1 | 65.6 | 53.8 | 59.1 |
| FT (Hierarchy) | 1.9 | 67.2 | 54.8 | 60.4 |
| Exemplars $\xrightarrow{\text{guide}}$ FT | 1.2 | 65.2 | 55.9 | 60.2 |
| FT+Exemplars (Basic) | 5.0 | 66.0 | 58.2 | 61.9 |
| FT+Exemplars (DA) | 5.8 | 65.7 | 59.0 | 62.2 |
| PB-SRL $\xrightarrow{\text{guide}}$ FT | 1.2 | 65.0 | 54.8 | 59.5 |
| *Combining the best methods* | | | | |
| PB-SRL $\xrightarrow{\text{guide}}$ FT+Exemplars | 5.5 | 67.4 | 58.8 | 62.8 |
| FT+Exemplars (Hierarchy) | 9.3 | 66.0 | 60.4 | **63.1** |

**Table 2:** Argument identification results on the full-text test set. Model size is in millions of features.

**Hyperparameter tuning.** We determined the stopping criterion and the $\ell_2$ regularization parameter $\lambda$ by tuning on the FT development set, searching over the following values for $\lambda$: $10^{-5}$, $10^{-7}$, $10^{-9}$, $10^{-12}$.

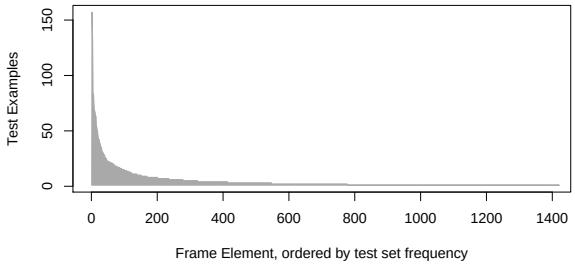
**Evaluation.** A complete frame-semantic parsing system involves frame identification and argument identification. We perform two evaluations: one assuming gold-standard frames are given, to evaluate argument identification alone; and one using the output of the system described by Hermann et al. (2014), the current state-of-the-art in frame identification, to demonstrate that our improvements are retained when incorporated into a full system.

## 6 Results

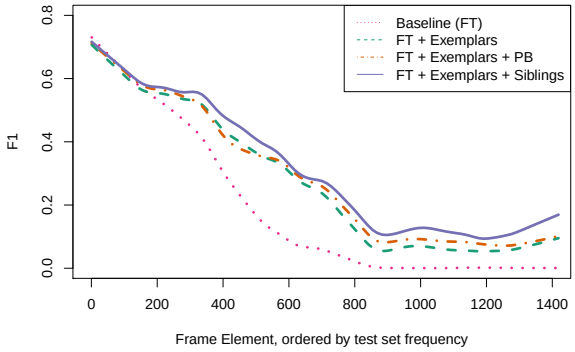
**Argument Identification.** We present precision, recall, and $F_1$-measure microaveraged across the test instances in table 2, for all approaches. The evaluation used in Das et al. (2014) assesses both frames and arguments; since our focus is on SRL, we only report performance for arguments, rendering our scores more interpretable. Under our argument-only evaluation, the system of Das et al. (2014) gets 59.3% $F_1$.

The first block shows baseline performance. The next block shows the benefit of FrameNet hierarchy features (+1.2% $F_1$). The third block shows that using exemplars as training data, especially with domain adaptation, is preferable to using them as guide features (2.8% $F_1$ vs. 0.9% $F_1$). PropBank SRL as guide features offers a small (0.4% $F_1$) gain.

The last two rows of table 2 show the performance upon combining the best approaches. Both use full-text and exemplars for training; the first uses PropBank SRL as guide features, and the second adds hierarchy features. The best result is the




**(a)** Frequency of each role appearing in the test set.




**(b)** $F_1$ of the best methods compared with the baseline.

**Figure 3:** $F_1$ for each role appearing in the test set, ranked by frequency. $F_1$ values have been smoothed with `loess`, with a smoothing parameter of 0.2. "Siblings" refers to hierarchy features.

latter, gaining 3.95% $F_1$ over the baseline.

**Role-level evaluation.** Figure 3(b) shows $F_1$ per frame element, for the baseline and the three best models. Each $x$-axis value is one role, sorted by decreasing frequency (the distribution of role frequencies is shown in figure 3(a)). For frequent roles, performance is similar; our models achieve gains on rarer roles.

**Full system.** When using the frame output of Hermann et al. (2014), $F_1$ improves by 1.1%, from 66.8% for the baseline, to 67.9% for our combined model (from the last row in table 2).

## 7 Conclusion

We have empirically shown that auxiliary semantic resources can benefit the challenging task of frame-semantic role labeling. The significant gains come from the FrameNet exemplars and the FrameNet hierarchy, with some signs that the PropBank scheme can be leveraged as well.

We are optimistic that future improvements to lexical semantic resources, such as crowdsourced lexical expansion of FrameNet (Pavlick et al., 2015) as well as ongoing/planned changes for PropBank (Bonial et al., 2014) and SemLink (Bonial et al., 2013), will lead to further gains in this task. Moreover, the techniques discussed here could be further explored using semi-automatic mappings between lexical resources (such as UBY; Gurevych et al., 2012), and correspondingly, this task could be used to extrinsically validate those mappings.

Ours is not the only study to show benefit from heterogeneous annotations for semantic analysis tasks. Feizabadi and Padó (2015), for example, successfully applied similar techniques for SRL of *implicit* arguments.[9] Ultimately, given the diversity of semantic resources, we expect that learning from heterogeneous annotations in different corpora will be necessary to build automatic semantic analyzers that are both accurate and robust.

## Acknowledgments

The authors are grateful to Dipanjan Das for his assistance, and to anonymous reviewers for their helpful feedback. This research has been supported by the Richard King Mellon Foundation and DARPA grant FA8750-12-2-0342 funded under the DEFT program.

## References

Collin Baker, Michael Ellsworth, and Katrin Erk. 2007. SemEval-2007 Task 19: frame semantic structure extraction. In *Proc. of SemEval*, pages 99–104. Prague, Czech Republic.

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet project. In *Proc. of COLING-ACL*, pages 86–90. Montreal, Quebec, Canada. URL `http://framenet.icsi.berkeley.edu`.

Ulrike Baldewein, Katrin Erk, Sebastian Padó, and Detlef Prescher. 2004. Semantic role labelling with similarity-based generalization using EM-based clustering. In Rada Mihalcea and Phil Edmonds, editors, *Proc. of SENSEVAL-3, the Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, pages 64–68. Barcelona, Spain.

Claire Bonial, Julia Bonn, Kathryn Conger, Jena D. Hwang, and Martha Palmer. 2014. PropBank: semantics of new predicate types. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proc. of LREC*, pages 3013–3019. Reykjavík, Iceland.

Claire Bonial, Kevin Stowe, and Martha Palmer. 2013. Renewing and revising SemLink. In *Proc. of the*

[9] They applied frustratingly easy domain adaptation to learn from FrameNet along with a PropBank-like dataset of nominal frames.

*2nd Workshop on Linked Data in Linguistics (LDL-2013): Representing and linking lexicons, terminologies and other language data*, pages 9–17. Pisa, Italy.

William W. Cohen and Vitor R. Carvalho. 2005. Stacked sequential learning. In *Proc. of IJCAI*, pages 671–676. Edinburgh, Scotland, UK.

Dipanjan Das, Desai Chen, André F. T. Martins, Nathan Schneider, and Noah A. Smith. 2014. Frame-semantic parsing. *Computational Linguistics*, 40(1):9–56. URL `http://www.ark.cs.cmu.edu/SEMAFOR`.

Dipanjan Das, André F. T. Martins, and Noah A. Smith. 2012. An exact dual decomposition algorithm for shallow semantic parsing with constraints. In *Proc. of *SEM*, pages 209–217. Montréal, Canada.

Dipanjan Das and Noah A. Smith. 2011. Semi-supervised frame-semantic parsing for unknown predicates. In *Proc. of ACL-HLT*, pages 1435–1444. Portland, Oregon, USA.

Dipanjan Das and Noah A. Smith. 2012. Graph-based lexicon expansion with sparsity-inducing penalties. In *Proc. of NAACL-HLT*, pages 677–687. Montréal, Canada.

Hal Daumé, III. 2007. Frustratingly easy domain adaptation. In *Proc. of ACL*, pages 256–263. Prague, Czech Republic.

Parvin Sadat Feizabadi and Sebastian Padó. 2015. Combining seemingly incompatible corpora for implicit semantic role labeling. In *Proc. of *SEM*, pages 40–50. Denver, Colorado, USA.

Charles J. Fillmore and Collin Baker. 2009. A frames approach to semantic analysis. In Bernd Heine and Heiko Narrog, editors, *The Oxford Handbook of Linguistic Analysis*, pages 791–816. Oxford University Press, Oxford, UK.

Michael Fleischman, Namhee Kwon, and Eduard Hovy. 2003. Maximum entropy models for FrameNet classification. In Michael Collins and Mark Steedman, editors, *Proc. of EMNLP*, pages 49–56.

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

Iryna Gurevych, Judith Eckle-Kohler, Silvana Hartmann, Michael Matuschek, Christian M. Meyer, and Christian Wirth. 2012. UBY - a large-scale unified lexical-semantic resource based on LMF. In *Proc. of EACL*, pages 580–590. Avignon, France.

Karl Moritz Hermann, Dipanjan Das, Jason Weston, and Kuzman Ganchev. 2014. Semantic frame identification with distributed word representations. In *Proc. of ACL*, pages 1448–1458. Baltimore, Maryland, USA.

Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2006. OntoNotes: the 90% solution. In *Proc. of HLT-NAACL*, pages 57–60. New York City, USA.

Richard Johansson. 2012. Non-atomic classification to improve a semantic role labeler for a low-resource language. In *Proc. of *SEM*, pages 95–99. Montréal, Canada.

Richard Johansson. 2013. Training parsers on incompatible treebanks. In *Proc. of NAACL-HLT*, pages 127–137. Atlanta, Georgia, USA.

Lingpeng Kong, Nathan Schneider, Swabha Swayamdipta, Archna Bhatia, Chris Dyer, and Noah A. Smith. 2014. A dependency parser for tweets. In *Proc. of EMNLP*, pages 1001–1012. Doha, Qatar.

Namhee Kwon, Michael Fleischman, and Eduard Hovy. 2004. FrameNet-based semantic parsing using maximum entropy models. In *Proc. of Coling*, pages 1233–1239. Geneva, Switzerland.

Ken Litkowski. 2004. SENSEVAL-3 task: Automatic labeling of semantic roles. In Rada Mihalcea and Phil Edmonds, editors, *Proc. of SENSEVAL-3, the Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, pages 9–12. Barcelona, Spain.

Dong C. Liu and Jorge Nocedal. 1989. On the Limited Memory BFGS Method for Large Scale Optimization. *Math. Program.*, 45(3):503–528.

André F. T. Martins, Dipanjan Das, Noah A. Smith, and Eric P. Xing. 2008. Stacking dependency parsers. In *Proc. of EMNLP*, pages 157–166. Honolulu, Hawaii.

Yuichiroh Matsubayashi, Naoaki Okazaki, and Jun'ichi Tsujii. 2009. A comparative study on generalization of semantic roles in FrameNet. In *Proc. of ACL-IJCNLP*, pages 19–27. Suntec, Singapore.

Joakim Nivre and Ryan McDonald. 2008. Integrating graph-based and transition-based dependency parsers. In *Proc. of ACL-HLT*, pages 950–958. Columbus, Ohio, USA.

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

Ellie Pavlick, Travis Wolfe, Pushpendre Rastogi, Chris Callison-Burch, Mark Drezde, and Benjamin Van Durme. 2015. FrameNet+: Fast paraphrastic tripling of FrameNet. In *Proc. of ACL-IJCNLP*. Beijing, China.

Vasin Punyakanok, Dan Roth, and Wen-tau Yih. 2008. The importance of syntactic parsing and inference in semantic role labeling. *Computational Linguistics*, 34(2):257–287. URL `http://cogcomp.cs.illinois.edu/page/software_view/SRL`.

Josef Ruppenhofer, Michael Ellsworth, Miriam R. L. Petruck, Christopher R. Johnson, and Jan Scheffczyk. 2010. FrameNet II: extended theory and practice. URL `https://framenet2.icsi.berkeley.edu/docs/r1.5/book.pdf`.

Oscar Täckström, Kuzman Ganchev, and Dipanjan Das. 2015. Efficient inference and structured learning for semantic role labeling. *Transactions of the Association for Computational Linguistics*, 3:29–41.

Cynthia A. Thompson, Roger Levy, and Christopher D. Manning. 2003. A generative model for semantic role labeling. In *Machine Learning: ECML 2003*, pages 397–408.

Matthew D. Zeiler. 2012. ADADELTA: An adaptive learning rate method. *arXiv:1212.5701 [cs]*. URL `http://arxiv.org/abs/1212.5701`, arXiv: 1212.5701.