The New Propbank: Aligning Propbank with AMR through POS Unification

Tim O'Gorman[1], Sameer Pradhan[23], Julia Bonn[1], Katie Conger[1], James Gung[1,4], Martha Palmer[1]
[1]University of Colorado [2]cemantix.org [3]Vassar College [4]IPSoft
ogormant@colorado.edu

Abstract

We present a corpus which converts the sense labels of existing Propbank resources to a new unified format which is more compatible with AMR and more robust to sparsity. This adopts an innovation of the Abstract Meaning Representation project (Banarescu et al., 2013) in which one abstracts away from different, related parts of speech, so that related forms such as "insert" and "insertion" could be represented by the same roleset and use the same semantic roles. We note that this conversion also serves to make the different English Propbank corpora released over the years consistent with each other, so that one might train and evaluate systems upon that larger combined data. We present analysis of some appealing characteristics of this final dataset, and present preliminary results of training and evaluating SRL systems on this combined set, to spur usage of this challenging new dataset.

Keywords: Propbank,SRL, Semantic Roles, Corpora

1. Introduction

We introduce the conversion of all existing Propbank data — constituting more than half a million predicate instances in English — into a format in which etymologically related senses from different parts are speech are merged, making that data compatible with the predicate senses used for Abstract Meaning Representation (Banarescu et al., 2013). This constitutes a large set of data made consistent to use the same frames and conventions, both increasing the amount of training data available for Propbank SRL and also providing a large corpus of semantic role labeling whose rolesets and numbered arguments match those of the Abstract Meaning Representation data (while containing over twice as many predicate instances).

We describe the combination of automatic and manual methods used in converting these corpora, provide some analysis to characterize the resulting corpora, and present preliminary SRL results against the test sets presented here, as a baseline for future evaluation of SRL. We suggest that evaluating against the combined test sets (OntoNotes, English Web Treebank and BOLT) can provide a challenging test set that could encourage the community to build SRL systems with a greater coverage over nominal, adjectival and light verb data, and with robustness to a range of difficult genres.

1.1. Motivations for Unifying Parts of Speech

Propbank (Kingsbury and Palmer, 2002) is a paradigm for the development of semantic role labeling corpora, designed for large-scale annotation. It focuses upon annotation of coarse-grained senses ("rolesets") which provide predicate-specific definitions of numbered arguments (ARG0, ARG1, etc.) to represent semantic roles. By using coarse-grained sense labels and these predicate-specific arguments (following the "individual thematic roles" of Dowty (1991)), Propbank approaches can achieve a high inter-annotator agreement rate, and the methodology has been adapted to Chinese (Palmer et al., 2005), Korean (Palmer et al., 2006), Hindi/Urdu (Bhatt et al., 2009), Finnish (Haverinen et al., 2013), Turkish (Sahin, 2016) and Brazilian Portuguese (Duran and Aluísio, 2011).

Such Propbank semantic role labels are generally annotated by labeling individual phrases within a constituency parse (which can then be converted into surface forms (Carreras and Màrquez, 2005; Pradhan et al., 2011) or dependency parses (Surdeanu et al., 2008)), but AMR annotation is done by directly building a semantic graph for a sentence – utilizing Propbank senses and numbered arguments – without explicit linking of that graph to phrases within a sentence.

While Propbank 1.0 (Kingsbury and Palmer, 2002) annotated only verbal predicates, it was later expanded to nouns (Hwang et al., 2010) and predicative adjectives (Bonial et al., 2017), creating new Propbank senses (called "rolesets") for those nouns and adjectives. Parallel work in the Abstract Meaning Representation project also handled nouns and adjectives, but did so by representing them with etymologically related verbal rolesets – so that a noun such as "insertion" would not have its own rolesets, but would instead be labeled with a verbal sense of "insert". This approach has a range of useful properties in reducing the number of senses with small amounts of training data, and also better conforms to the approach of FrameNet (Baker et al., 1998).

In order to merge these different rolesets into the new "unified" form, Propbank rolesets were given any number of "alias" entries, each of which expresses a different surface form and part of speech. These alias fields are analogous to the Lexical Units of FrameNet (Baker et al., 1998), although this does not result in a Framenet-like lexicon; Propbank rolesets rarely contain more than three aliases, and maintain the same coarse-grained sense distinctions developed in earlier Propbank works. For example, one might look at the different usages of "appeal" in that regard: there is a legal sense of the verb and noun "appeal", a begging sense for the same terms, or an attractiveness sense for which one might use verbal "appeal", nominal "appeal", the adjective "appealing" or the light verb "have appeal". In the pre-unification methodology, each sense of each lemma would receive its own roleset, resulting in seven different rolesets. Instead, these are merged into three rolesets, each with 2–3 aliases.

This approach results in a reduction in sparsity, and reduces the number of redundant senses which must be added as a lexicon increases. Section 4 will attempt to quantify such gains. Conversion to unified forms was also an appropriate context to update a range of prior Propbank annotations in order to make them more compatible – resulting in a larger collective landscape of SRL resources for English.

2. Methodology of Corpus Conversion

2.1. Conversion of Lexicon

It is intuitive to a casual speaker of English that "insert" and "insertion", or "appeal" and "appealing", are related lemmas. We focus upon that simple, coarse-grained level of etymological relatedness, focusing upon clusters of lemmas that would be verbalized into the same verbal lemma. Thus, "appealing" is clustered with "appeal" we do not dive into distantly related terms with the same root, such as "appellation" or "compel." After determining which lemmas were to be treated as related, the challenge was (a) to get an alignment between those rolesets — discerning which verbal senses corresponded with which nominal senses — and (b) to get an alignment between the numbered arguments in each aligned pair of rolesets.

The nominal and adjectival rolesets were developed with an awareness of both the related verbal senses and related senses in Nombank (Meyers et al., 2004). When nominal or adjectival forms had the same meaning as a verbal form, care was therefore taken to maintain consistency in between their sense descriptions and numbered argument descriptions. These allowed many direct, automatic mappings to be done.

However, senses and numbered arguments did not always have such identical descriptions. To align senses without clear matches, we started with automatic pre-annotations using existing information, as rolesets were sometimes mapped to corresponding FrameNet classes (Palmer, 2009), but then overviewed all such alignments manually, using experienced Propbank frame builders. For aligning numbered arguments between those aligned senses, a range of resources exist that help one generalize beyond predicate-specific numbered arguments of Propbank, such as the Propbank function tags (labels such as LOC, TMP, PAG (more proto-agentive core role) and PPT (more proto-patientive core role); cf Bonial et al. (2014)), the number of the numbered argument itself, which are designed to capture general tendencies (Kingsbury and Palmer, 2002), and mappings to Verbnet or FrameNet semantic role types (Palmer, 2009). As with roleset alignments, these were then manually checked by expert framers to confirm each mapping.

This resulted not simply in a new lexicon, but in mappings from the older lexicon to the new one, with manually crafted labels regarding which mappings were deterministic, and which mappings might require manual checking. This resulted in the merging of 13,460 rolesets delineated by part of speech into 10,183 unified rolesets, 57 of which were determined by framers to require manual instance-by-instance sense disambiguation. The alignment of numbered arguments converted 34,469 different arguments down to a set of 25,452 unified numbered arguments, only 388 of which required manual instance-by-instance retrofitting.

2.2. Conversion of Annotated Data

The conversion of the Propbank lexicon resulted in a set of direct mappings between senses and numbered arguments, a small percentage of which required manual disambiguation. All senses which were labeled as ambiguous were double-annotated to revise them. In addition, whenever a numbered argument was labeled as ambiguous between numbered arguments in the new frames, every instance with that numbered argument was also selected for retrofitting. The manual retrofitting covered 12,000 instances in Propbank across all corpora (roughly 2% of all predicates), each of which was double-annotated and adjudicated. Subsequent data generated after that conversion was annotated directly using these unified frames.

2.3. Revision of Formatting

In the process of this conversion, other inconsistencies between different releases of Propbank have also been resolved in order to make the data more consistent in format and behavior. In addition to minor consistency decisions (such as replacement of a "ARGM-PNC" role with "ARGM-PRP"), we made the treatment of control and relative clause chains (appearing as LINK-PRO and LINK-SLC) consistent, using postprocessing tools in ClearNLP (Choi, 2012). A set of scripts used convert to that data to the bracketing over surface forms used in evaluations (Pradhan et al., 2011) has been revised and corrected, to improve the consistency of discontinuous material ("R-" and "C-" prefixed arguments in the CoNLL-2012 data).

2.4. Relationship of this Unified Data to AMR

The Abstract Meaning Representation project annotates predicate argument structures using the Propbank lexicon, supplemented with a small set of AMR-specific rolesets, ending with *-91*, for specific semantic functions and the reification of semantic roles, such as INCLUDE-91 for set operations or HAVE-ORG-ROLE-91 for organizational membership. Other than those AMR rolesets ending in *-91*, the rolesets used within AMR are a subset of the rolesets used in Propbank, adopting nearly every verbal form in Propbank and most nominal and adjectival senses. The most common reason why a verbal roleset is not in AMR is because AMR deletes "semantically light" predicates, such as copular "be" or auxiliary "have". As the use of existing SRL systems has been shown to help AMR parsing (Wang et al., 2015), we expect that the introduction of a larger SRL dataset with closer alignment to the AMR lexicon should increase that utility. The AMR SUBSET line in Table 1 illustrates the size of these corpora when limited to only the rolesets used in AMR; while there is a large drop in verbal senses (notably due to the omission of semantically light predicates), it still remains a very large corpus. Work is ongoing adding more of the Propbank nominal and adjectival rolesets to AMR.

2.5. Expansion of multi-word predicate coverage

More recent work has expanded coverage of the Propbank lexicon to encompass multi-word predicates as well, such as *take with a grain of salt*, *cut slack*, or *jump on bandwagon*. Propbank has long annotated certain classes of multi-word predicates such as verb-particle constructions and light verbs (Bonial et al., 2014), and is now expanded to arbitrarily structured semi-fixed expressions. We added coverage to many of the most high-frequency multi-word predicates in the corpus, and created lexical entries for each multi-word predicate. The important contribution is not simply the detection of these MWP elements (which can also be found in larger resources such as PARSEME (Savary et al., 2017)), but the annotation of semantic roles for each MWP. For example, something like "jump on the bandwagon" would be as follows:

- **jump-on-bandwagon.09**: *join an activity or group because of its popularity*
  - Arg0: person jumping on the bandwagon
  - Arg1: popular thing joined
  - Arg2: action done which gets one on the bandwagon

This is a step forward in not simply detecting these MWPs, but being able to represent them in structured semantic representations such as AMR.

3. Larger Landscape of Propbank Resources

The OntoNotes corpus (Hovy et al., 2006), most recently released as part of the Conll-2012 (Pradhan et al., 2011), was developed during the DARPA-GALE annotation project covering a wide range of domains, and has been the largest resource for training and evaluating semantic role labeling systems. However, the range of other corpora that have been annotated with Propbank roles since OntoNotes – most notably, the English Web Treebank and the BOLT corpora – collectively constitute an amount of additional predicate annotations the same size as OntoNotes itself. The English Web Treebank encompasses a range of genres of the web, such as reviews and emails (Bies et al., 2012), and the BOLT datasets encompass informal corpora of English discussion forum data, SMS text, and translations of conversational data (Garland et al., 2012; Song et al., 2014). All of these resources are either currently released or in the process of being released, with stand-off SRL annotations available at propbank.github.io .

We suggest that the combination of the test sets of the three major corpora provides a more interesting and challenging dataset against which to evaluate a semantic role labeling system. This is due to both the challenging informal domains (such as SMS messages and discussion forum posts) as well as to the increase in the coverage of nouns and adjectives in the data. Table 1 illustrates the size of these corpora, broken down by the parts of speech seen. Due to annotations done at the end of the OntoNotes data collection phase, additional rolesets were also added to the OntoNotes corpus release.

Propbank data also exists for related domains. The SHARP and THYME clinical data (Albright et al., 2013) encompass nearly a million words of clinical text; there are also annotations in the same frames for the LORELEI English core data (Strassel and Tracey, 2016), image captions of the Flickr 8k corpus (Hodosh et al., 2013), MASC data (Ide et al., 2008), and Earth Science data affiliated with the ClearEarth project (Duerr et al., 2016). These additional corpora – as noted in Table 2 below – illustrate the range of domain-specific annotations of Propbank data which might be utilized for semantic role labeling in specific domains.

	verbs	nouns (light v.)	adjectives
OntoNotes (ON)	349,352	40,163 (2,215)	750
EWT	44,736	9,453 (732)	3,305
BOLT	132,642	18,839 (1973)	10,957
ON+EWT+BOLT	526,730	68,455 (4920)	15,012
in AMR Subset	349,783	63,585 (4714)	10,121
Conll-2012	319,239	20,305	0

Table 1: Core Corpora Annotated with Propbank rolesets for general English. Light verbs are annotated using nominal frames (Hwang et al. 2010) and therefore included in those counts

	verbs	nouns (light v.)	adjectives
LORELEI	18,871	4,089 (196)	780
MASC	14,150	70 (3)	0
flickr-5k	5,897	551 (91)	51
earth science	10,070	5713 (8)	468
clinical – SHARP	27,667	15,807 (22)	0
clinical – THYME	49,649	17,906 (89)	756
All current	653,034	112,591 (5,329)	17,067

Table 2: Additional corpora annotated with Propbank rolesets

4. Data Analysis

4.1. Effect of Unification on Sense Sparsity

The largest anticipated advantage to unifying across parts of speech is to reduce the number of very low frequency rolesets, which are therefore both hard to detect and whose numbered arguments are difficult to learn. Using the running example of the "appeal" senses: while three verbal senses of "appeal" each have 20+ examples in training data, the adjectival sense of "appealing" and three nouns of "appeal" all have fewer than six examples, making all of them problematic for learning. One way of quantifying this is to look at a new English dataset (we use the LORELEI corpus of disaster-related newswire texts) and to measure how many predicate instances have a sufficient number of examples in prior Propbank corpora. Figure 1 illustrates this over a range of different thresholds for sufficiency. One may see that there is a reduction of these out-of-vocabulary and low-frequency senses due to the added nominal rolesets now in OntoNotes, further reductions of sparsity when one looks at the combined set of OntoNotes, English Web Treebank, and BOLT, and further reductions of sparsity with the unification of rolesets across parts of speech.

	ON (CoNLL-2012)		ON (Unified)		BOLT		Google	
	Valid	Test	Valid	Test	Valid	Test	Valid	Test
Our System	78.7	79.0	78.2	78.3	65.2	66.0.	72.0	72.4
Comparable Past Results								
Täckström et al. (2015)	79.1	79.4	-	-	-	-	-	-
FitzGerald et al. (2015) (Single)	79.2	79.6	-	-	-	-	-	-
FitzGerald et al. (2015) (POE)	79.7	80.1	-	-	-	-	-	-
Zhou and Xu (2015)	81.1	81.3	-	-	-	-	-	-

Table 3: Preliminary results on the part-of-speech *unified* corpora along with results reported on existing partition by other researchers.

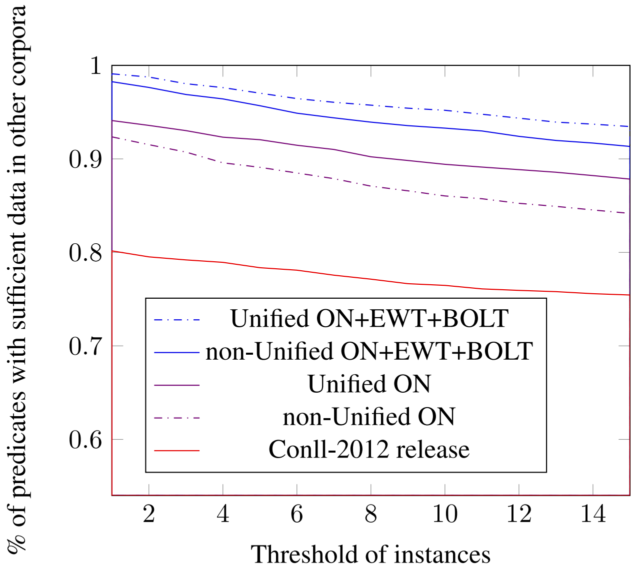

Figure 1: Percent of predicate instances in LORELEI with N or more training instances

4.2. Nature of the Expansion of Nominal Coverage

The coverage over nominals is dramatically expanded in this release of Propbank, having 3934 rolesets with nominal *aliases*, focused upon nouns that correspond with predicates or events, and roughly as many as NomBank corpus(Meyers et al., 2004). This differs from two other core resources for semantic roles of nominals, FrameNet (Baker et al., 1998) and NomBank (Meyers et al., 2004), which often annotated non-eventive nominals. Because of this, while Propbank has coverage over only 20% of nominal lexical units in FrameNet, and 30% of those of NomBank (Meyers et al., 2004), that coverage encompasses most of the nouns which assert an event or dynamic situation. Table 4 illustrates how the coverage differs between each resource, by illustrating the kinds of nominal predicates unique to each annotation project. Propbank rolesets which overlap with FrameNet and NomBank (the left column) illustrate prototypical nominalizations. In contrast, nouns not represented in Propbank but captured in FrameNet or NomBank show their coverage over more traditional entities, objects and relational nouns.

In all resources	FN only	NomBank only
permission	business card	pirate
assistance	knife	mound
prohibition	rotunda	normalcy
invasion	pattern	ire
kidnapping	raincoat	glamour

Table 4: Random samples of Propbank nominal rolesets overlapping with Framenet and NomBank (left), and those unique to FrameNet (center), or to NomBank (right), illustrating that most eventive nouns are in that intersection

5. Preliminary Results

5.1. Semantic Role Labeler

Our SRL system (Gung and Pradhan, 2018) uses a deep neural network model which does not include explicit syntactic information. We closely follow Zhou and Xu (2015), treating SRL as an IOB tagging problem and using deep bidirectional LSTMs with a linear chain conditional random field (Lafferty et al., 2001) loss function.

Long-short term memory networks (LSTMs) are a form of recurrent neural network (RNN) that has been successfully applied to many NLP tasks. Sequential inputs are often processed using pairs of RNNs, with one RNN processing from the first to last element and the other RNN processing from the last element to the first, concatenating outputs for each element (Graves et al., 2013). In our approach, and that of Zhou and Xu (2015), the result of the forward pass is used as input to the backward pass, enabling repeated stacking of these layers to form a deep topology. Instead of scoring each label locally, the addition of a CRF loss function allows for globally normalized scoring of all possible sequences of labels, maximizing the sequence-level log-likelihood (Collobert et al., 2011). This approach has been shown to improve performance on a variety of tasks (Huang et al., 2015; Ma and Hovy, 2016; Lample et al., 2016).

Following standard practices for applying neural architectures to NLP tasks, we initialize our network with word embeddings trained on orders of magnitude more data than is available for our task (SRL). Specifically, we use publicly available GloVe 100-dimensional vectors trained on 6 billion words from Wikipedia and Gigaword (Pennington et al., 2014). These embeddings are updated during training as network parameters along with a single out-of-vocabulary (OOV) vector, which is randomly initialized. We simplify the features used in the original model of (Zhou and Xu, 2015), using only the vector associated with the current word as well as the distance from the current word to the predicate. The distance feature uses a trainable lookup table to map each discrete distance to a low-dimensional representation.

To improve the handling of OOV words, such as names and numbers not found in the original word vectors, we also use character-level vector representations produced using a convolutional neural network. Recent work has demonstrated the effectiveness of using neural networks to extract character-level (morphological) features (Chiu and Nichols, 2015; Lample et al., 2016; dos Santos et al., 2015; dos Santos and Zadrozny, 2014; Kim et al., 2015; Ma and Hovy, 2016). In these approaches, characters are assigned fixed-dimensional embeddings, which are composed into a single fixed-length vector using a neural network-based reduction function. We follow the approach described in Ma and Hovy (2016), using a convolutional neural network with max-over-time pooling. The resulting character-based representations are concatenated with each word and distance vector to form the input to the deep bidirectional LSTM. We train the full neural network end-to-end using Adam (Kingma and Ba, 2014).

5.2. Preliminary Results on the English Unified Set

In this section we will report preliminary results on this updated corpus using the aforementioned semantic role labeler model. For these experiments we used the OntoNotes training partition to train the model.

Table 3 summarizes the performance of our system on the validation and test partitions of the following three subcorpora— the CoNLL-2012 subset of OntoNotes, the revised OntoNotes with additional predicates and unification across the part-of-speeches, as well as the BOLT and Google corpora which will be released in their unified format. Further details of the partitions, and further analysis and genre-wise breakdowns of results, will be provided in the extended form of the paper.

6. Discussion

We've outlined the methods for converting Propbank to a unified form, and the advantages provided by that unified form and by the larger size of the Propbank corpora now available. We suggest that testing against the combination of OntoNotes, English Web Treebank and BOLT corpora can provide a more challenging SRL evaluation, requiring systems to better handle challenging web domains and non-verbal predicates.

7. Acknowledgements

This work is supported by Defense Advanced Research Projects Agency (DARPA) under DARPA HR001117S0026 - AIDA, and BOLT- HR0011-11-C-0145, DEFT-FA-8750-13-2-0045 and DARPA Lorelei via subcontracts from LDC. We thank Google in supporting the annotation of the English Web Treebank, and thank the LDC for cooperation in distributing the source material.

8. Bibliographical References

Albright, D., Lanfranchi, A., Fredriksen, A., Styler IV, W. F., Warner, C., Hwang, J. D., Choi, J. D., Dligach, D., Nielsen, R. D., Martin, J., et al. (2013). Towards comprehensive syntactic and semantic annotations of the clinical narrative. *Journal of the American Medical Informatics Association*, 20(5):922–930.

Banarescu, L., Bonial, C., Cai, S., Georgescu, M., Griffitt, K., Hermjakob, U., Knight, K., Koehn, P., Palmer, M., and Schneider, N. (2013). Abstract meaning representation for sembanking.

Bonial, C., Bonn, J., Conger, K., Hwang, J., and Palmer, M. (2014). PropBank: Semantics of New Predicate Types. pages 3013–3019.

Bonial, C., Conger, K., Hwang, J. D., Mansouri, A., Aseri, Y., Bonn, J., OGorman, T., and Palmer, M. (2017). Current directions in english and arabic propbank. In *Handbook of Linguistic Annotation*, pages 737–769. Springer.

Carreras, X. and Màrquez, L. (2005). Introduction to the conll-2005 shared task: Semantic role labeling. In *Proceedings of the Ninth Conference on Computational Natural Language Learning*, pages 152–164. Association for Computational Linguistics.

Chiu, J. P. and Nichols, E. (2015). Named entity recognition with bidirectional lstm-cnns. *arXiv preprint arXiv:1511.08308*.

Choi, J. D. (2012). *Optimization of natural language processing components for robustness and scalability*. Ph.D. thesis, University of Colorado at Boulder.

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., and Kuksa, P. (2011). Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12(Aug):2493–2537.

dos Santos, C. N. and Zadrozny, B. (2014). Learning character-level representations for part-of-speech tagging. In *Proceedings of ICML-14*.

dos Santos, C., Guimaraes, V., Niterói, R., and de Janeiro, R. (2015). Boosting named entity recognition with neural character embeddings. In *Proceedings of NEWS 2015 The Fifth Named Entities Workshop*, page 25.

Duerr, R., Thessen, A., Jenkins, C., Palmer, M., Myers, S., and Ramdeen, S. (2016). The clearearth project: Preliminary findings from experiments in applying the cleartk nlp pipeline and annotation tools developed for biomedicine to the earth sciences. In *AGU Fall Meeting Abstracts*.

FitzGerald, N., Täckström, O., Ganchev, K., and Das, D. (2015). Semantic role labeling with neural network factors. In *Proceedings of EMNLP*, pages 960–970.

Graves, A., Jaitly, N., and Mohamed, A.-r. (2013). Hybrid speech recognition with deep bidirectional lstm. In *Automatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop on*, pages 273–278. IEEE.

Gung, J. and Pradhan, S. (2018). How redundant is syntax for deep semantic role labeling? Ms.

Huang, Z., Xu, W., and Yu, K. (2015). Bidirectional lstm-crf models for sequence tagging. *arXiv preprint arXiv:1508.01991*.

Hwang, J. D., Bhatia, A., Bonial, C., Mansouri, A., Vaidya, A., Xue, N., and Palmer, M. (2010). Propbank annotation of multilingual light verb constructions. In *Proceedings of the Fourth Linguistic Annotation Workshop*, pages 82–90. Association for Computational Linguistics.

Kim, Y., Jernite, Y., Sontag, D., and Rush, A. M. (2015). Character-aware neural language models. *arXiv preprint arXiv:1508.06615*.

Kingma, D. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Lafferty, J., McCallum, A., and Pereira, F. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the eighteenth international conference on machine learning, ICML*, volume 1, pages 282–289.

Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., and Dyer, C. (2016). Neural architectures for named entity recognition. *arXiv preprint arXiv:1603.01360*.

Ma, X. and Hovy, E. (2016). End-to-end sequence labeling via bi-directional lstm-cnns-crf. *arXiv preprint arXiv:1603.01354*.

Palmer, M. (2009). Semlink: Linking propbank, verbnet and framenet. In *Proceedings of the generative lexicon conference*, pages 9–15. Pisa Italy.

Pennington, J., Socher, R., and Manning, C. (2014). Glove: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Pradhan, S., Ramshaw, L., Marcus, M., Palmer, M., Weischedel, R., and Xue, N. (2011). Conll-2011 shared task: Modeling unrestricted coreference in ontonotes. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*, pages 1–27.

Savary, A., Ramisch, C., Cordeiro, S., Sangati, F., Vincze, V., QasemiZadeh, B., Candito, M., Cap, F., Giouli, V., Stoyanova, I., et al. (2017). The parseme shared task on automatic identification of verbal multiword expressions. In *Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017)*, pages 31–47.

Surdeanu, M., Johansson, R., Meyers, A., Màrquez, L., and Nivre, J. (2008). The conll-2008 shared task on joint parsing of syntactic and semantic dependencies. In *Proceedings of the Twelfth Conference on Computational Natural Language Learning*, pages 159–177. Association for Computational Linguistics.

Täckström, O., Ganchev, K., and Das, D. (2015). Efficient inference and structured learning for semantic role labeling. *Transactions of the Association for Computational Linguistics*, 3:29–41.

Wang, C., Xue, N., and Pradhan, S. (2015). Boosting transition-based amr parsing with refined actions and auxiliary analyzers. *Volume 2: Short Papers*, page 857.

Zhou, J. and Xu, W. (2015). End-to-end learning of semantic role labeling using recurrent neural networks. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

9. Language Resource References

Baker, C. F., Fillmore, C. J., and Lowe, J. B. (1998). The Berkeley FrameNet project. In *Proceedings of the 17th International Conference on Computational Linguistics (COLING/ACL-98)*, pages 86–90, Montreal.

Bhatt, R., Narasimhan, B., Palmer, M., Rambow, O., Sharma, D. M., and Xia, F. (2009). A multi-representational and multi-layered treebank for hindi/urdu. In *Proceedings of the Third Linguistic Annotation Workshop*, pages 186–189. Association for Computational Linguistics.

Bies, A., Mott, J., Warner, C., and Kulick, S. (2012). English web treebank. *Linguistic Data Consortium, Philadelphia, PA*.

Duran, M. S. and Aluísio, S. M. (2011). Propbank-br: a brazilian portuguese corpus annotated with semantic role labels. In *Proceedings of the 8th Symposium in Information and Human Language Technology, Cuiabá/MT, Brazil*.

Garland, J., Strassel, S., Ismael, S., Song, Z., and Lee, H. (2012). Linguistic resources for genre-independent language technologies: user-generated content in bolt. In *Workshop Programme*, page 34.

Haverinen, K., Laippala, V., Kohonen, S., Missilä, A., Nyblom, J., Ojala, S., Viljanen, T., Salakoski, T., and Ginter, F. (2013). Towards a dependency-based propbank of general finnish. In *Proceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA 2013); May 22-24; 2013; Oslo University; Norway. NEALT Proceedings Series 16*, number 085, pages 41–57. Linköping University Electronic Press.

Hodosh, M., Young, P., and Hockenmaier, J. (2013). Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research*, 47:853–899.

Hovy, E., Marcus, M., Palmer, M., Ramshaw, L., and Weischedel, R. (2006). OntoNotes: the 90% solution. In *Proceedings of the human language technology conference of the NAACL, Companion Volume: Short Papers*, pages 57–60. Association for Computational Linguistics.

Ide, N., Baker, C., Fellbaum, C., and Fillmore, C. (2008). Masc: The manually annotated sub-corpus of american english. In *In Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC*. Citeseer.

Kingsbury, P. and Palmer, M. (2002). From treebank to propbank. In *LREC*, pages 1989–1993.

Meyers, A., Reeves, R., Macleod, C., Szekely, R., Zielinska, V., Young, B., and Grishman, R. (2004). The nombank project: An interim report. In *HLT-NAACL 2004 workshop: Frontiers in corpus annotation*, volume 24, page 31.

Palmer, M., Xue, N., Babko-Malaya, O., Chen, J., and Snyder, B. (2005). A parallel proposition bank ii for chinese and english. In *Proceedings of the Workshop on Frontiers in Corpus Annotations II: Pie in the Sky*, pages 61–67. Association for Computational Linguistics.

Palmer, M., Ryu, S., Choi, J., Yoon, S., and Jeon, Y. (2006).

Korean propbank. *LDC Catalog No.: LDC2006T03 ISBN*, pages 1–58563.

Pradhan, S., Ramshaw, L., Marcus, M., Palmer, M., Weischedel, R., and Xue, N. (2011). Conll-2011 shared task: Modeling unrestricted coreference in ontonotes. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*, pages 1–27.

Sahin, G. (2016). Verb sense annotation for turkish propbank via crowdsourcing. In *Proceedings of 17th international conference on intelligent text processing and computational linguistics. CICLING*.

Song, Z., Strassel, S., Lee, H., Walker, K., Wright, J., Garland, J., Fore, D., Gainor, B., Cabe, P., Thomas, T., et al. (2014). Collecting natural sms and chat conversations in multiple languages: The bolt phase 2 corpus. In *LREC*, pages 1699–1704.

Strassel, S. and Tracey, J. (2016). Lorelei language packs: Data, tools, and resources for technology development in low resource languages. In *LREC*.