Polyglot Semantic Role Labeling

Phoebe Mulcaire[♡] **Swabha Swayamdipta**[♢] **Noah A. Smith**[♡]
[♡]Paul G. Allen School of Computer Science & Engineering, University of Washington
[♢]School of Computer Science, Carnegie Mellon University
{pmulc,nasmith}@cs.washington.edu, swabha@cs.cmu.edu

Abstract

Previous approaches to multilingual semantic dependency parsing treat languages independently, without exploiting the similarities between semantic structures across languages. We experiment with a new approach where we combine resources from a pair of languages in the CoNLL 2009 shared task (Hajič et al., 2009) to build a *polyglot* semantic role labeler. Notwithstanding the absence of parallel data, and the dissimilarity in annotations between languages, our approach results in an improvement in SRL performance on multiple languages over a monolingual baseline. Analysis of the polyglot model shows it to be advantageous in lower-resource settings.

1 Introduction

The standard approach to multilingual NLP is to design a single architecture, but tune and train a separate model for each language. While this method allows for customizing the model to the particulars of each language and the available data, it also presents a problem when little data is available: extensive language-specific annotation is required. The reality is that most languages have very little annotated data for most NLP tasks.

Ammar et al. (2016a) found that using training data from multiple languages annotated with Universal Dependencies (Nivre et al., 2016), and represented using multilingual word vectors, outperformed monolingual training. Inspired by this, we apply the idea of training one model on multiple languages—which we call polyglot training—to PropBank-style semantic role labeling (SRL). We train several parsers for each language in the CoNLL 2009 dataset (Hajič et al., 2009): a traditional monolingual version, and variants which additionally incorporate supervision from English portion of the dataset. To our knowledge, this is the first multilingual SRL approach to combine supervision from several languages.

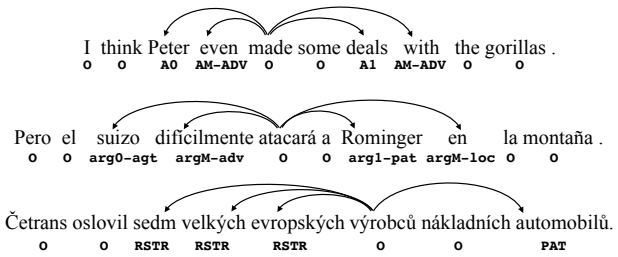

Figure 1: Example predicate-argument structures from English, Spanish, and Czech. Note that the argument labels are different in each language.

The CoNLL 2009 dataset includes seven different languages, allowing study of trends across the same. Unlike the Universal Dependencies dataset, however, the semantic label spaces are entirely language-specific, making our task more challenging. Nonetheless, the success of polyglot training in this setting demonstrates that sharing of statistical strength across languages does not depend on explicit alignment in annotation conventions, and can be done simply through parameter sharing. We show that polyglot training can result in better labeling accuracy than a monolingual parser, *especially* for low-resource languages. We find that even a simple combination of data is as effective as more complex kinds of polyglot training. We include a breakdown into label categories of the differences between the monolingual and polyglot models. Our findings indicate that polyglot training consistently improves label accuracy for common labels.

	# sentences	# sentences w/ 1+ predicates	# predicates
CAT	13200	12876	37444
CES	38727	38579	414133
DEU	36020	14282	17400
ENG	39279	37847	179014
JPN	4393	4344	25712
SPA	14329	13836	43828
ZHO	22277	21073	102827

Table 1: Train data statistics. Languages are indicated with ISO 639-3 codes.

2 Data

We evaluate our system on the semantic role labeling portion of the CoNLL-2009 shared task (Hajič et al., 2009), on all seven languages, namely Catalan, Chinese, Czech, English, German, Japanese and Spanish. For each language, certain tokens in each sentence in the dataset are marked as predicates. Each predicate takes as arguments, other words in the same sentence, their relationship marked by labeled dependency arcs. Sentences may contain no predicates.

Despite the consistency of this format, there are significant differences between the training sets across languages.[1] English uses PropBank role labels (Palmer et al., 2005). Catalan, Chinese, English, German, and Spanish include (but are not limited to) labels such as "arg_0-agt" (for "agent") or "A_0" that may correspond to some degree to each other and to the English roles. Catalan and Spanish share most labels (being drawn from the same source corpus, AnCora; Taulé et al., 2008), and English and German share some labels. Czech and Japanese each have their own distinct sets of argument labels, most of which do not have clear correspondences to English or to each other.

We also note that, due to semi-automatic projection of annotations to construct the German dataset, more than half of German sentences do *not* include labeled predicate and arguments. Thus while German has almost as many sentences as Czech, it has by far the fewest training examples (predicate-argument structures); see Table 1.

[1]This is expected, as the datasets were annotated independently under diverse formalisms and only later converted into CoNLL format (Hajič et al., 2009).

3 Model

Given a sentence with a marked predicate, the CoNLL 2009 shared task requires disambiguation of the sense of the predicate, and labeling all its dependent arguments. The shared task assumed predicates have already been identified, hence we do not handle the predicate identification task.

Our basic model adapts the span-based dependency SRL model of He et al. (2017). This adaptation treats the dependent arguments as argument spans of length 1. Additionally, BIO consistency constraints are removed from the original model—each token is tagged simply with the argument label or an empty tag. A similar approach has also been proposed by Marcheggiani et al. (2017).

The input to the model consists of a sequence of pretrained embeddings for the surface forms of the sentence tokens. Each token embedding is also concatenated with a vector indicating whether the word is a predicate or not. Since the part-of-speech tags in the CoNLL 2009 dataset are based on a different tagset for each language, we do not use these. Each training instance consists of the annotations for a single predicate. These representations are then passed through a deep, multilayer bidirectional LSTM (Graves, 2013; Hochreiter and Schmidhuber, 1997) with highway connections (Srivastava et al., 2015).

We use the hidden representations produced by the deep biLSTM for both argument labeling and predicate sense disambiguation in a multitask setup; this is a modification to the models of He et al. (2017), who did not handle predicate senses, and of Marcheggiani et al. (2017), who used a separate model. These two predictions are made independently, with separate softmaxes over different last-layer parameters; we then combine the losses for each task when training. For predicate sense disambiguation, since the predicate has been identified, we choose from a small set of valid predicate senses as the tag for that token. This set of possible senses is selected based on the training data: we map from lemmatized tokens to predicates and from predicates to the set of all senses of that predicate. Most predicates are only observed to have one or two corresponding senses, making the set of available senses at test time quite small (less than five senses/predicate on average across all languages). If a particular lemma was not observed in training, we heuristically predict it as the first sense of that predicate. For Czech and

Japanese, the predicate sense annotation is simply the lemmatized token of the predicate, giving a one-to-one predicate-"sense" mapping.

For argument labeling, every token in the sentence is assigned one of the argument labels, or NULL if the model predicts it is not an argument to the indicated predicate.

3.1 Monolingual Baseline

We use pretrained word embeddings as input to the model. For each of the shared task languages, we produced GloVe vectors (Pennington et al., 2014) from the news, web, and Wikipedia text of the Leipzig Corpora Collection (Goldhahn et al., 2012).[2] We trained 300-dimensional vectors, then reduced them to 100 dimensions with principal component analysis for efficiency.

3.2 Simple Polyglot Sharing

In the first polyglot variant, we consider multilingual sharing between each language and English by using pretrained *multilingual* embeddings. This polyglot model is trained on the union of annotations in the two languages. We use stratified sampling to give the two datasets equal effective weight in training, and we ensure that every training instance is seen at least once per epoch.

Pretrained multilingual embeddings. The basis of our polyglot training is the use of pretrained multilingual word vectors, which allow representing entirely distinct vocabularies (such as the tokens of different languages) in a shared representation space, allowing crosslingual learning (Klementiev et al., 2012). We produced multilingual embeddings from the monolingual embeddings using the method of Ammar et al. (2016b): for each non-English language, a small crosslingual dictionary and canonical correlation analysis was used to find a transformation of the non-English vectors into the English vector space (Faruqui and Dyer, 2014).

Unlike multilingual word representations, argument label sets are disjoint between language pairs, and correspondences are not clearly defined. Hence, we use separate label representations for each language's labels. Similarly, while (for example) ENG:look and SPA:mira may be semantically connected, the senses `look.01` and `mira.01` may not correspond. Hence, predicate sense representations are also language-specific.

3.3 Language Identification

In the second variant, we concatenate a language ID vector to each multilingual word embedding and predicate indicator feature in the input representation. This vector is randomly initialized and updated in training. These additional parameters provide a small degree of language-specificity in the model, while still sharing most parameters.

3.4 Language-Specific LSTMs

This third variant takes inspiration from the "frustratingly easy" architecture of Daume III (2007) for domain adaptation. In addition to processing every example with a shared biLSTM as in previous models, we add language-specific biLSTMs that are trained only on the examples belonging to one language. Each of these language-specific biLSTMs is two layers deep, and is combined with the shared biSLTM in the input to the third layer. This adds a greater degree of language-specific processing while still sharing representations across languages. It also uses the language identification vector and multilingual word vectors in the input.

4 Experiments

We present our results in Table 2. We observe that simple polyglot training improves over monolingual training, with the exception of Czech, where we observe no change in performance. The languages with the fewest training examples (German, Japanese, Catalan) show the most improvement, while large-dataset languages such as Czech or Chinese see little or no improvement (Figure 2).

The language ID model performs inconsistently; it is better than the simple polyglot model in some cases, including Czech, but not in all. The language-specific LSTMs model performs best on a few languages, such as Catalan and Chinese, but worst on others. While these results may reflect differences between languages in the optimal amount of crosslingual sharing, we focus on the simple polyglot results in our analysis, which sufficiently demonstrate that polyglot training can improve performance over monolingual training.

We also report performance of state-of-the-art systems in each of these languages, all of which make explicit use of syntactic features, Marcheg-

[2]For English we used the vectors provided on the GloVe website `nlp.stanford.edu/projects/glove/`.

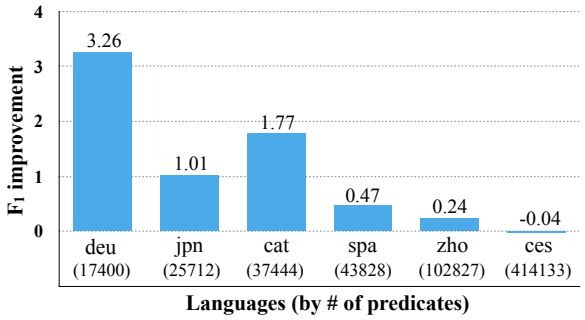

Figure 2: Improvement in absolute F_1 with polyglot training with addition of English. Languages are sorted in order of increasing number of predicates in the training set.

giani et al. (2017) excepted. While this results in better performance on many languages, our model has the advantage of not relying on a syntactic parser, and is hence more applicable to languages with lower resources. However, the results suggest that syntactic information is critical for strong performance on German, which has the fewest predicates and thus the least semantic annotation for a semantics-only model to learn from. Nevertheless, our baseline is on par with the best published scores for Chinese, and it shows strong performance on most languages.

Label-wise results. Table 3 gives the F_1 scores for individual label categories in the Catalan and Spanish datasets, as an illustration of the larger trend. In both languages, we find a small but consistent improvement in the most common label categories (e.g., arg_1 and arg_M). Less common label categories are sensitive to small changes in performance; they have the largest changes in F_1 in absolute value, but without a consistent direction. This could be attributed to the addition of English data, which improves learning of representations that are useful for the most common labels, but is essentially a random perturbation for the rarer ones. This pattern is seen across languages, and consistently results in overall gains from polyglot training.

One exception is in Czech, where polyglot training reduces accuracy on several common argument labels, e.g., PAT and LOC. While the effect sizes are small (consistent with other languages), the overall F_1 score on Czech decreases slightly in the polyglot condition. It may be that the Czech dataset is too large to make use of the comparatively small amount of English data, or that differences in the annotation schemes prevent effective crosslingual transfer.

Future work on language pairs that do not include English could provide further insights. Catalan and Spanish, for example, are closely related and use the same argument label set (both being drawn from the AnCora corpus) which would allow for sharing output representations as well as input tokens and parameters.

Polyglot English results. For each language pair, we also evaluated the simple polyglot model on the English test set from the CoNLL 2009 shared task (Table 4). English SRL consistently benefits from polyglot training, with an increase of 0.25–0.7 absolute F_1 points, depending on the language. Surprisingly, Czech provides the smallest improvement, despite the large amount of data added; the absence of crosslingual transfer in both directions for the English-Czech case, breaking the pattern seen in other languages, could therefore be due to differences in annotation rather than questions of dataset size.

Labeled vs. unlabeled F_1. Table 5 provides unlabeled F_1 scores for each language pair. As can be seen here, the unlabeled F_1 improvements are generally positive but small, indicating that polyglot training can help both in structure prediction and labeling of arguments. The pattern of seeing the largest improvements on the languages with the smallest datasets generally holds here: the largest F_1 gains are in German and Catalan, followed by Japanese, with minimal or no improvement elsewhere.

5 Related Work

Recent improvements in multilingual SRL can be attributed to neural architectures. Swayamdipta et al. (2016) present a transition-based stack LSTM model that predicts syntax and semantics jointly, as a remedy to the reliance on pipelined models. Guo et al. (2016) and Roth and Lapata (2016) use deep biLSTM architectures which use syntactic information to guide the composition. Marcheggiani et al. (2017) use a simple LSTM model over word tokens to tag semantic dependencies, like our model. Their model predicts a token's label based on the combination of the token vector and the predicate vector, and saw benefits from using POS tags, both improvements that could be added to our model. Marcheggiani and Titov (2017) apply the recently-developed graph

Model	CAT	CES	DEU	ENG	JPN	SPA	ZHO
Marcheggiani et al. (2017)	-	86.00	-	87.60	-	80.30	81.20
Best previously reported	*80.32*	86.00	*80.10*	*89.10*	*78.15*	*80.50*	*81.20*
Monolingual	77.31	84.87	66.71	86.54	74.99	75.98	81.26
+ ENG(simple polyglot)	79.08	84.82	69.97	–	76.00	76.45	81.50
+ ENG(language ID)	79.05	85.14	69.49	–	75.77	77.32	81.42
+ ENG(language-specific LSTMs)	79.45	84.78	68.30	–	75.88	76.86	81.89

Table 2: Semantic F_1 scores (including predicate sense disambiguation) on the CoNLL 2009 dataset. State of the art for Catalan and Japanese is from Zhao et al. (2009), for German and Spanish from Roth and Lapata (2016), for English and Chinese from Marcheggiani and Titov (2017). Italics indicate use of syntax.

	arg$_0$	**arg**$_1$	**arg**$_2$	**arg**$_3$	**arg**$_4$	**arg**$_L$	**arg**$_M$
Gold label count (CAT)	2117	4296	1713	61	71	49	2968
Monolingual CAT F_1	82.06	79.06	68.95	28.89	42.42	39.51	60.85
+ ENG improvement	+2.75	+2.58	+4.53	+18.17	+9.81	+1.35	+1.10
Gold label count (SPA)	2438	4295	1677	49	82	46	3237
Monolingual SPA F_1	82.44	77.93	70.24	28.89	41.15	22.50	58.89
+ ENG improvement	+0.37	+0.43	+1.35	-3.40	-3.48	+4.01	+1.26

Table 3: Per-label breakdown of F_1 scores for Catalan and Spanish. These numbers reflect labels for each argument; the combination is different from the overall semantic F_1, which includes predicate sense disambiguation.

ENG-only	+CAT	+CES	+DEU	+JPN	+SPA	+ZHO
86.54	86.79	87.07	87.07	87.11	87.24	87.10

Table 4: Semantic F_1 scores on the English test set for each language pair.

Model	CAT	CES	DEU	ENG	JPN	SPA	ZHO
Monolingual	93.92	91.92	87.95	92.87	85.55	93.61	87.93
+ ENG	94.09	91.97	89.01	–	86.17	93.65	87.90

Table 5: Unlabeled semantic F_1 scores on the CoNLL 2009 dataset.

convolutional networks to SRL, obtaining state of the art results on English and Chinese. All of these approaches are orthogonal to ours, and might benefit from polyglot training.

Other polyglot models have been proposed for semantics. Richardson et al. (2018) train on multiple (natural language)-(programming language) pairs to improve a model that translates API text into code signature representations. Duong et al. (2017) treat English and German semantic parsing as a multi-task learning problem and saw improvement over monolingual baselines, especially for small datasets. Most relevant to our work is Johannsen et al. (2015), which trains a polyglot model for *frame*-semantic parsing. In addition to sharing features with multilingual word vectors, they use them to find word translations of target language words for additional lexical features.

6 Conclusion

In this work, we have explored a straightforward method for polyglot training in SRL: use multilingual word vectors and combine training data across languages. This allows sharing without crosslingual alignments, shared annotation, or parallel data. We demonstrate that a polyglot model can outperform a monolingual one for semantic analysis, particularly for languages with less data.

Acknowledgments

We thank Luke Zettlemoyer, Luheng He, and the anonymous reviewers for helpful comments and feedback. This research was supported in part by the Defense Advanced Research Projects Agency (DARPA) Information Innovation Office (I2O) under the Low Resource Languages for Emergent Incidents (LORELEI) program issued by DARPA/I2O under contract HR001115C0113 to BBN. Views expressed are those of the authors alone.

References

Waleed Ammar, George Mulcaire, Miguel Ballesteros, Chris Dyer, and Noah Smith. 2016a. Many languages, one parser. *Transactions of the Association for Computational Linguistics*, 4:431–444.

Waleed Ammar, George Mulcaire, Yulia Tsvetkov, Guillaume Lample, Chris Dyer, and Noah A Smith. 2016b. Massively multilingual word embeddings. arXiv:1602.01925.

Hal Daume III. 2007. Frustratingly easy domain adaptation. In *Proceedings of ACL*.

Long Duong, Hadi Afshar, Dominique Estival, Glen Pink, Philip Cohen, and Mark Johnson. 2017. Multilingual semantic parsing and code-switching. In *Proceedings of CoNLL*.

Manaal Faruqui and Chris Dyer. 2014. Improving vector space word representations using multilingual correlation. In *Proceedings of EACL*.

Dirk Goldhahn, Thomas Eckart, and Uwe Quasthoff. 2012. Building large monolingual dictionaries at the Leipzig corpora collection: From 100 to 200 languages. In *Proceedings of LREC*.

Alex Graves. 2013. Generating sequences with recurrent neural networks. arXiv:1308.0850.

Jiang Guo, Wanxiang Che, Haifeng Wang, Ting Liu, and Jun Xu. 2016. A unified architecture for semantic role labeling and relation classification. In *Proceedings of COLING*.

Jan Hajič, Massimiliano Ciaramita, Richard Johansson, Daisuke Kawahara, Maria Antònia Martí, Lluís Màrquez, Adam Meyers, Joakim Nivre, Sebastian Padó, Jan Štěpánek, et al. 2009. The conll-2009 shared task: Syntactic and semantic dependencies in multiple languages. In *Proceedings of CoNLL*.

Luheng He, Kenton Lee, Mike Lewis, and Luke Zettlemoyer. 2017. Deep semantic role labeling: What works and what's next. In *Proceedings of ACL*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Anders Johannsen, Héctor Martínez Alonso, and Anders Søgaard. 2015. Any-language frame-semantic parsing. In *Proceedings of EMNLP*.

Alexandre Klementiev, Ivan Titov, and Binod Bhattarai. 2012. Inducing crosslingual distributed representations of words. *Proceedings of COLING*.

Diego Marcheggiani, Anton Frolov, and Ivan Titov. 2017. A simple and accurate syntax-agnostic neural model for dependency-based semantic role labeling. arXiv:1701.02593.

Diego Marcheggiani and Ivan Titov. 2017. Encoding sentences with graph convolutional networks for semantic role labeling. In *Proceedings of EMNLP*, Copenhagen, Denmark.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D. Manning, Ryan T. McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. Universal dependencies v1: A multilingual treebank collection. In *Proceedings of LREC*.

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

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of EMNLP*.

Kyle Richardson, Jonathan Berant, and Jonas Kuhn. 2018. Polyglot semantic parsing in APIs. In *Proceedings of NAACL*.

Michael Roth and Mirella Lapata. 2016. Neural semantic role labeling with dependency path embeddings. arXiv:1605.07515.

Rupesh Kumar Srivastava, Klaus Greff, and Jürgen Schmidhuber. 2015. Training very deep networks. In *NIPS*.

Swabha Swayamdipta, Miguel Ballesteros, Chris Dyer, and Noah A. Smith. 2016. Greedy, joint syntactic-semantic parsing with stack LSTMs. In *Proceedings of CoNLL*.

Mariona Taulé, M. Antònia Martí, and Marta Recasens. 2008. AnCora: Multilevel annotated corpora for Catalan and Spanish. In *Proceedings of LREC*.

Hai Zhao, Wenliang Chen, Jun'ichi Kazama, Kiyotaka Uchimoto, and Kentaro Torisawa. 2009. Multilingual dependency learning: Exploiting rich features for tagging syntactic and semantic dependencies. In *Proceedings of CoNLL*.