# A Unified Architecture for Natural Language Processing: Deep Neural Networks with Multitask Learning

Ronan Collobert and Jason Weston

International Conference on Machine Learning (ICML), 2008

## Problems Addressed

(a) Design a single CNN architecture that, given a sentence, outputs a host of language processing predictions : POS tags, chunks, NER, SRL, semantically similar words and the likelihood that the sentence makes sense (grammatically and semantically) using a language model.

(b) Correct the following failings in architectures required for deeper semantic tasks

(i) they are shallow in the sense that the classifier is often linear

(ii) for good performance with a linear classifier they must incorporate many hand-engineered features specific to the task

(iii) they cascade features learnt separately from other tasks, thus propagating errors.

## Proposed Solution

(a) All the tasks are basically assigning labels to words. This is done using a deep neural network, trained in an end to end fashion. The input sentence is processed by several layers for feature extraction, and the deep layers are automatically trained by backpropagation.

(b) The first layer (Lookup Table Layer) extracts features for each word. Since the NN is dealing with raw words, the first layer maps words into real valued vectors.

(1) Words are considered as indices in a finite dictionary D. Each word i in D is embedded into a d-dimensional space using a lookup table LTw() i.e. LTw(i) = Wi where W is a dX|D| matrix of parameters to be learnt, Wi is a column of the matrix (vector of size d), and d is the word vector size (wsz) chosen by the user.

(2) Therefore, a sentence is transformed into a series of vectors. The parameters W are automatically trained using backpropagation.

(c) The second layer (TDNN layer) extracts features from the sentence, treating it as a sequence with local and global structure.

(1) The first layer maps input to a sequence x(.) of n identically sized vectors of size d.

(2) Since n depends on sentence, TDNNs are used to perform convolution on x(.), outputting sequence o(.)

(d) The third layer (Max layer) captures most relevant features by taking max. over time (sentence) for each of nhu output features. This layer's output has fixed dimension, allowing subsequent layers to be normal NN layers.

(e) The word to be labeled is indicated with an additional lookup table.

(f) The Neural Networks are trained on relevant tasks, and the deep layers of the Neural Network are shared to improve features produced by these deep layers, thus improving generalization performance. The last layer is task specific.

(g) Training is done in a stochastic manner by selecting the next task, selecting a random training example for it, and updating the NN for that task.

## Claims

(a) SRL : Sections 2-21 if Propbank version 1 (about 1 million words) were used for training and Section 23 was used for testing.

(b) POS and Chunking : Used the same data split via the Penn TreeBank.

(c) NER labeled data : Obtained by running Stanford Named Entity Recognizer on the same dataset.

(d) Language Models : Trained on Wikipedia. Numeric numbers changed to "NUMBER" and accentuations removed from characters.

(e) Semantically Related Words : WordNet used.

(f) All tasks used the same dictionary of 30,000 most common words from Wikipedia, converted to lowercase. Other words were considered as unknown and mapped to special words.

## Results

*Table 2.* A Deep Architecture for SRL improves by learning auxiliary tasks that share the first layer that represents words as $wsz$-dimensional vectors. We give word error rates for $wsz$=15, 50 and 100 and various shared tasks.

| | $wsz$=15 | $wsz$=50 | $wsz$=100 |
|---|---|---|---|
| SRL | 16.54 | 17.33 | 18.40 |
| SRL + POS | 15.99 | 16.57 | 16.53 |
| SRL + Chunking | 16.42 | 16.39 | 16.48 |
| SRL + NER | 16.67 | 17.29 | 17.21 |
| SRL + Synonyms | 15.46 | 15.17 | 15.17 |
| SRL + Language model | 14.42 | 14.30 | 14.46 |
| SRL + POS + Chunking | 16.46 | 15.95 | 16.41 |
| SRL + POS + NER | 16.45 | 16.89 | 16.29 |
| SRL + POS + Chunking + NER | 16.33 | 16.36 | 16.27 |
| SRL + POS + Chunking + NER + Synonyms | 15.71 | 14.76 | 15.48 |
| SRL + POS + Chunking + NER + Language model | 14.63 | 14.44 | 14.50 |

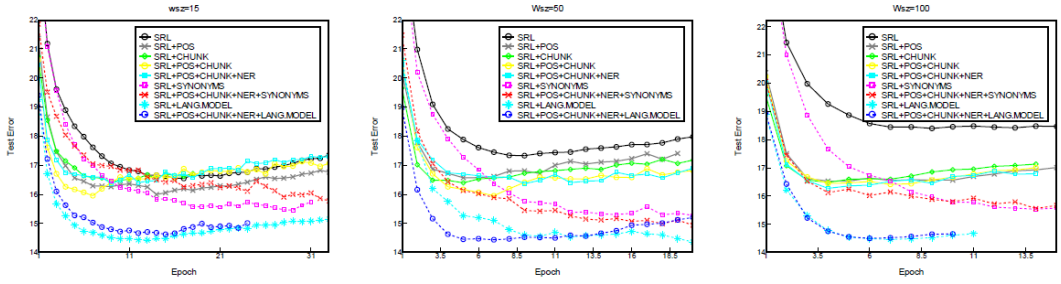


*Figure 3.* Test error versus number of training epochs over PropBank, for the SRL task alone and SRL jointly trained with various other NLP tasks, using deep NNs.

## Bibliography


1. Ando, R., & Zhang, T. (2005). A Framework for Learning Predictive Structures from Multiple Tasks and Unlabeled Data. JMLR, 6, 1817–1853.
2. Bengio, Y., & Ducharme, R. (2001). A neural probabilistic language model. NIPS 13.
3. Bridle, J. (1990). Probabilistic interpretation of feedforward classification network outputs, with relationships to statistical pattern recognition. In F. F. Souli´e and J. H´erault (Eds.), Neurocomputing : Algorithms, architectures and applications, 227–236. NATO ASI Series.
4. Caruana, R. (1997). Multitask Learning. Machine Learning, 28, 41–75.
5. Chapelle, O., Schlkopf, B., & Zien, A. (2006). Semisupervised learning. Adaptive computation and machine learning. Cambridge, Mass., USA : MIT Press.
6. Collobert, R., & Weston, J. (2007). Fast semantic extraction using a novel neural network architecture. Proceedings of the 45th Annual Meeting of the ACL (pp. 560–567).
7. Gildea, D., & Palmer, M. (2001). The necessity of parsing or predicate argument recognition. Proceedings of the 0th Annual Meeting of the ACL, 239–246.
8. Joachims, T. (1999). Transductive inference for text classification sing support vector machines. ICML.
9. LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-Based Learning Applied to Document Recognition. Proceedings of the IEEE, 86.
10. McClosky, D., Charniak, E., & Johnson, M. (2006). Effective self-training for parsing. Proceedings of HLTNAACL 2006.
11. Miller, S., Fox, H., Ramshaw, L., & Weischedel, R. (2000). A novel use of statistical parsing to extract information from text. 6th Applied Natural Language Processing Conference.
12. Musillo, G., & Merlo, P. (2006). Robust Parsing of the Proposition Bank. ROMAND 2006 : Robust Methods in Analysis of Natural language Data.
13. Okanohara, D., & Tsujii, J. (2007). A discriminative language model with pseudo-negative samples. Proceedings of the 45th Annual Meeting of the ACL, 73–80.
14. Palmer, M., Gildea, D., & Kingsbury, P. (2005). The proposition bank : An annotated corpus of semantic roles. Comput. Linguist., 31, 71–106.
15. Pradhan, S., Ward, W., Hacioglu, K., Martin, J., & Jurafsky, D. (2004). Shallow semantic parsing using support vector machines. Proceedings of HLT/NAACL-2004.
16. Rosenfeld, B., & Feldman, R. (2007). Using Corpus Statistics on Entities to Improve Semi-supervised Relation Extraction from the Web. Proceedings of the 45th Annual Meeting of the ACL, 600–607.
17. Schwenk, H., & Gauvain, J. (2002). Connectionist language modeling for large vocabulary continuous speech recognition. IEEE International Conference on Acoustics, Speech, and Signal Processing (pp. 765–768).
18. Sutton, C., & McCallum, A. (2005a). Composition of conditional random fields for transfer learning. Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing, 748–754.
19. Sutton, C., & McCallum, A. (2005b). Joint parsing and semantic role labeling. Proceedings of CoNLL-2005 (pp.225–228).
20. Sutton, C., McCallum, A., & Rohanimanesh, K. (2007). Dynamic Conditional Random Fields : Factorized Probabilistic Models for Labeling and Segmenting Sequence Data. JMLR, 8, 693–723.

21. Ueffing, N., Haffari, G., & Sarkar, A. (2007). Transductive learning for statistical machine translation. Proceedings of the 45th Annual Meeting of the ACL, 25–32.

22. Waibel, A., abd G. Hinton, T. H., Shikano, K., & Lang, K. (1989). Phoneme recognition using time-delay neural networks. IEEE Transactions on Acoustics, Speech, and Signal Processing, 37, 328–339.