Investigating LSTMs for Joint Extraction of Opinion Entities and Relations

Arzoo Katiyar and **Claire Cardie**
Department of Computer Science
Cornell University
Ithaca, NY, 14853, USA
arzoo, cardie@cs.cornell.edu

Abstract

We investigate the use of deep bi-directional LSTMs for joint extraction of opinion entities and the IS-FROM and IS-ABOUT relations that connect them — the first such attempt using a deep learning approach. Perhaps surprisingly, we find that standard LSTMs are not competitive with a state-of-the-art CRF+ILP joint inference approach (Yang and Cardie, 2013) to opinion entities extraction, performing below even the standalone sequence-tagging CRF. Incorporating sentence-level and a novel relation-level optimization, however, allows the LSTM to identify opinion relations and to perform within 1–3% of the state-of-the-art joint model for opinion entities and the IS-FROM relation; and to perform as well as the state-of-the-art for the IS-ABOUT relation — all without access to opinion lexicons, parsers and other preprocessing components required for the feature-rich CRF+ILP approach.

1 Introduction

There has been much research in recent years in the area of fine-grained opinion analysis where the goal is to identify subjective expressions in text along with their associated sources and targets. More specifically, fine-grained opinion analysis aims to identify three types of *opinion entities*:

- **opinion expressions**, O, which are direct subjective expressions (i.e., explicit mentions of otherwise private states or speech events expressing private states (Wiebe and Cardie, 2005));
- **opinion targets**, T, which are the entities or topics that the opinion is about; and
- **opinion holders**, H, which are the entities expressing the opinion.

In addition, the task involves identifying the IS-FROM and IS-ABOUT relations between an opinion expression and its holder and target, respectively. In the sample sentences, numerical subscripts indicate an IS-FROM or IS-ABOUT relation.

S1 [The sale]$_{T_1}$ [infuriated]$_{O_1}$ [Beijing]$_{H_{1,2}}$ which [regards]$_{O_2}$ [Taiwan]$_{T_2}$ an integral part of its territory awaiting reunification, by force if necessary.

S2 "[Our agency]$_{T_1,H_2}$ [seriously needs]$_{O_2}$ [equipment for detecting drugs]$_{T_2}$," [he]$_{H_1}$ [said]$_{O_1}$.

In S1, for example, "infuriated" indicates that there is an (negative) opinion from "Beijing" regarding "the sale."[1]

Traditionally, the task of extracting opinion entities and opinion relations was handled in a pipelined manner, i.e., extracting the opinion expressions first and then extracting opinion targets and opinion holders based on their syntactic and semantic associations with the opinion expressions (Kim and Hovy, 2006; Kobayashi et al., 2007). More recently, methods that *jointly* infer the opinion entity and relation extraction tasks (e.g., using Integer Linear Programming (ILP)) have been introduced (Choi et al., 2006; Yang and Cardie, 2013) and show that the existence of opinion relations provides clues for the identification of opinion entities and vice-versa, and thus results in better performance than a pipelined approach. However, the success of these methods depends critically on the availability of opinion lexicons, dependency parsers, named-entity taggers, etc.

[1] This paper does not attempt to determine the sentiment, i.e., the positive or negative polarity, of an opinion.

Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, pages 919–929,
Berlin, Germany, August 7-12, 2016.

Alternatively, neural network-based methods have been employed. In these approaches, the required latent features are automatically learned as dense vectors of the hidden layers. Liu et al. (2015), for example, compare several variations of recurrent neural network methods and find that long short-term memory networks (LSTMs) perform the best in identifying opinion expressions and opinion targets for the specific case of product/service reviews.

Motivated by the recent success of LSTMs on this and other problems in NLP, we investigate here the use of deep bi-directional LSTMs for joint extraction of opinion expressions, holders, targets and the relations that connect them. This is the first attempt to handle the full opinion entity and relation extraction task using a deep learning approach.

In experiments on the MPQA dataset for opinion entities (Wiebe and Cardie, 2005; Wilson, 2008), we find that standard LSTMs are not competitive with the state-of-the-art CRF+ILP joint inference approach of Yang and Cardie (2013), performing below even the standalone sequence-tagging CRF. Inspired by Huang et al. (2015), we show that incorporating sentence-level, and our newly proposed relation-level optimization, allows the LSTM to perform within 1–3% of the ILP joint model for all three opinion entity types and to do so without access to opinion lexicons, parsers or other preprocessing components.

For the primary task of identifying opinion entities together with their IS-FROM and IS-ABOUT relations, we show that the LSTM with sentence- and relation-level optimizations outperforms an LSTM baseline that does not employ joint inference. When compared to the CRF+ILP-based joint inference approach, the optimized LSTM performs slightly better for the IS-ABOUT[2] relation and within 3% for the IS-FROM relation.

In the sections that follow, we describe: related work (Section 2) and the multi-layer bi-directional LSTM (Section 3); the LSTM extensions (Section 4); the experiments on the MPQA corpus (Sections 5 and 6) and error analysis (Section 7).

[2]Target and IS-ABOUT relation identification is one important aspect of opinion analysis that hasn't been much addressed in previous work and has proven to be difficult for existing methods.

2 Related Work

LSTM-RNNs (Hochreiter and Schmidhuber, 1997) have recently been applied to many sequential modeling and prediction tasks, such as machine translation (Bahdanau et al., 2014; Sutskever et al., 2014), speech recognition (Graves et al., 2013), NER (Hammerton, 2003). The bi-directional variant of RNNs has been found to perform better as it incorporates information from both the past and the future (Schuster and Paliwal, 1997; Graves et al., 2013). Deep RNNs (stacked RNNs) (Schmidhuber, 1992; Hihi and Bengio, 1996) capture more abstract and higher-level representation in different layers and benefit sequence modeling tasks (İrsoy and Cardie, 2014). Collobert et al. (2011) found that adding dependencies between the tags in the output layer improves the performance of Semantic Role Labeling task. Later, Huang et al. (2015) also found that adding a CRF layer on top of bi-directional LSTMs to capture these dependencies can produce state-of-the-art performance on part-of-speech (POS), chunking and NER.

For fine-grained opinion extraction, earlier work (Wilson et al., 2005; Breck et al., 2007; Yang and Cardie, 2012) focused on extracting subjective phrases using a CRF-based approach from open-domain text such as news articles. Choi et al. (2005) extended the task to jointly extract opinion holders and these subjective expressions. Yang and Cardie (2013) proposed a ILP-based joint-inference model to jointly extract the opinion entities and opinion relations, which performed better than the pipelined based approaches (Kim and Hovy, 2006).

In the neural network domain, İrsoy and Cardie (2014) proposed a deep bi-directional recurrent neural network for identifying subjective expressions, outperforming the previous CRF-based models. Irsoy and Cardie (2013) additionally proposed a bi-directional recursive neural network over a binary parse tree to jointly identify opinion entities, but performed significantly worse than the feature-rich CRF+ILP approach of Yang and Cardie (2013). Liu et al. (2015) used several variants of recurrent neural networks for joint opinion expression and aspect/target identification on customer reviews for restaurants and laptops, outperforming the feature-rich CRF based baseline. In the product reviews domain, however, the opinion holder is generally the reviewer and the task

does not involve identification of relations between opinion entities. Hence, standard LSTMs are applicable in this domain. None of the above neural network based models can jointly model opinion entities and opinion relations.

In the relation extraction domain, several neural networks have been proposed for relation classification, such as RNN-based models (Socher et al., 2012) and LSTM-based models (Xu et al., 2015). These models depend on constituent or dependency tree structures for relation classification, and also do not model entities jointly. Recently, Miwa and Bansal (2016) proposed a model to jointly represent both entities and relations with shared parameters, but it is not a joint-inference framework.

3 Methodology

For our task, we propose the use of multi-layer bi-directional LSTMs, a type of recurrent neural network. Recurrent neural networks have recently been used for modeling sequential tasks. They are capable of modeling sequences of arbitrary length by repetitive application of a recurrent unit along the tokens in the sequence. However, recurrent neural networks are known to have several disadvantages like the problem of vanishing and exploding gradients. Because of these problems, it has been found that recurrent neural networks are not sufficient for modeling long term dependencies. Hochreiter and Schmidhuber (1997), thus proposed long short term memory (LSTMs), a variant of recurrent neural networks.

3.1 Long Short Term Memory (LSTM)

Long short term memory networks are capable of learning long-term dependencies. The recurrent unit is replaced by a memory block. The memory block contains two cell states – memory cell C_t and hidden state h_t; and three multiplicative gates – input gate i_t, forget gate f_t and output gate o_t. These gates regulate the addition or removal of information to the cell state thus overcoming vanishing and exploding gradients.

$$f_t = \sigma(W_f x_t + U_f h_{t-1} + b_f)$$
$$i_t = \sigma(W_i x_t + U_i h_{t-1} + b_i)$$

The forget gate f_t and input gate i_t above decides what part of the information we are going to throw away from the cell state and what new information we are going to store in the cell state. The sigmoid outputs a number between 0 and 1 where 0 implies that the information is completely lost and 1 means that the information is completely retained.

$$\widetilde{C}_t = \tanh(W_c x_t + U_c h_{t-1} + b_c)$$
$$C_t = i_t * \widetilde{C}_t + f_t * C_{t-1}$$

Thus, the intermediate cell state $\widetilde{C}_t$ and previous cell state C_{t-1} are used to update the new cell state C_t.

$$o_t = \sigma(W_o x_t + U_o h_{t-1} + V_o C_t + b_o)$$
$$h_t = o_t * \tanh(C_t)$$

Next, we update the hidden state h_t based on the output gate o_t and the cell state C_t. We pass both the cell state C_t and the hidden state h_t to the next time step.

3.2 Multi-layer Bi-directional LSTM

In sequence tagging problems, it has been found that only using past information for computing the hidden state h_t may not be sufficient. Hence, previous works (Graves et al., 2013; İrsoy and Cardie, 2014) proposed the use of bi-directional recurrent neural networks for speech and NLP tasks, respectively. The idea is to also process the sequence in the backward direction. Hence, we can compute the hidden state $\overrightarrow{h_t}$ in the forward direction and $\overleftarrow{h_t}$ in the backward direction for every token.

Also, in more traditional feed-forward networks, deep networks have been found to learn abstract and hierarchical representations of the input in different layers (Bengio, 2009). The multi-layer LSTMs have been proposed (Hermans and Schrauwen, 2013) to capture long-term dependencies of the input sequences in different layers.

For the first hidden layer, the computation proceeds similar to that described in Section 3.1. However, for higher hidden layers i the input to the memory block is the hidden state and memory cell from the previous layer $i-1$ instead of the input vector representation.

For this paper, we only use the hidden state from the last layer L to compute the output state y_t.

$$z_t = \overrightarrow{V}\overrightarrow{h_t}^{(L)} + \overleftarrow{V}\overleftarrow{h_t}^{(L)} + c$$
$$y_t = g(z_t)$$

4 Network Training

For our problem, we wish to predict a label y from a discrete set of classes Y for every word in a sentence. As is the norm, we train the network by

maximizing the log-likelihood

$$\sum_{(x,y)\in \mathrm{T}} \log p(y|x,\theta)$$

over the training data T, with respect to the parameters θ, where x is the input sentence and y is the corresponding tag sequence. We propose three alternatives for the log-likelihood computation.

4.1 Word-Level Log-Likelihood (WLL)

We first formulate a word-level log-likelihood (WLL) (adapted from Collobert et al. (2011)) that considers all words in a sentence independently. We interpret the score z_t corresponding to the i^{th} tag $[z_t]_i$ as a conditional tag probability $\log p(i|x,\theta)$ by applying a softmax operation.

$$\begin{aligned} p(i|x,\theta) &= softmax(z_t^i) \\ &= \frac{e^{z_t^i}}{\sum_j e^{z_t^j}} \end{aligned}$$

For the tag sequence y given the input sentence x the log-likelihood is :

$$\log p(y|x,\theta) = z^y - \operatorname*{logadd}_j z^j$$

4.2 Sentence-Level Log-Likelihood (SLL)

In the word-level approach above, we discard the dependencies between the tags in a tag sequence. In our sentence-level log-likelihood (SLL) formulation (also adapted from Collobert et al. (2011)) we incorporate these dependencies: we introduce a transition score $[A]_{i,j}$ for jumping from tag i to tag j of adjacent words in the tag sequence to the set of parameters $\widetilde{\theta}$. These transition scores are going to be trained.

We use both the transition scores $[A]$ and the output scores z to compute the sentence score $s(x|_{t=1}^T, y|_{t=1}^T, \widetilde{\theta})$.

$$s(x,y,\widetilde{\theta}) = \sum_{t=1}^{T} \left([A]_{y_{t-1},y_t} + z_t^{y_t} \right)$$

We normalize this sentence score over all possible paths of tag sequences $\widetilde{y}$ to get the log conditional probability as below :

$$\log p_{sent}(y|x,\widetilde{\theta}) = s(x,y,\widetilde{\theta}) - \operatorname*{logadd}_{\widetilde{y}} s(x,\widetilde{y},\widetilde{\theta})$$

Even though the number of tag sequences grows exponentially with the length of the sentence, we can compute the normalization factor in linear time (Collobert et al., 2011).

At inference time, we find the best tag sequence

$$\operatorname*{argmax}_{\widetilde{y}} s(x,\widetilde{y},\widetilde{\theta})$$

for an input sentence x using Viterbi decoding. In this case, we basically maximize the same likelihood as in a CRF except that a CRF is a linear model.

The above sentence-level log-likelihood is useful for sequential tagging, but it cannot be directly used for modeling relations between non-adjacent words in the sentence. In the next subsection, we extend the above idea to also model relations between non-adjacent words.

4.3 Relation-Level Log-Likelihood (RLL)

For every word x_t in the sentence x, we output the tag y_t and a distance d_t. If a word at position t is related to a word at position k and $k < t$, then $d_t = (t-k)$. If word t is not related to any other word to its left, then $d_t = 0$. Let D_{Left} be the maximum distance we model for such *left*-relations [3].

$$z_t = \overrightarrow{V_r}\overrightarrow{h_t}^{(L)} + \overleftarrow{V_r}\overleftarrow{h_t}^{(L)} + c_r$$

We let $\overrightarrow{V_r} \in \mathbb{R}^{(D_{Left}+1)\times Y \times d_h}$ (where d_h is the dimensionality of hidden units) such that the output state $z_t \in \mathbb{R}^{(D_{Left}+1)\times Y}$ as compared to $z_t \in \mathbb{R}^{(1)\times Y}$ in case of sentence-level log-likelihood.

In order to add dependencies between tags and relations, we introduce a transition score $[A]_{i,j,d',d"}$ for jumping from tag i and relation distance $d^{'}$ to tag j and relation distance $d^{"}$ of adjacent words in the tag sequence, to the set of parameters $\theta^{'}$. These transition scores are also going to be trained similar to the transition scores in sentence-level log-likelihood.

The sentence score $s(x|_{t=1}^T, y|_{t=1}^T, d|_{t=1}^T, \theta^{'})$ is:

$$s(x,y,d,\theta^{'}) = \sum_{t=1}^{T} \left([A]_{y_{t-1},y_t,d_{t-1},d_t} + z_t^{y_t,d_t} \right)$$

We normalize this sentence score over all possible paths of tag $\widetilde{y}$ and relation sequences $\widetilde{d}$ to get the log conditional probability as below :

$$\begin{aligned} \log p_{rel,Left}(y,d|x,\widetilde{\theta}) =& s(x,y,d,\theta^{'}) \\ & - \operatorname*{logadd}_{\widetilde{y},\widetilde{d}} s(x,\widetilde{y},\widetilde{d},\theta^{'}) \end{aligned}$$

[3] Later in this section, we will also add a similar likelihood in the objective function for *right*-relations, i.e., for each word the related words are in its right context.

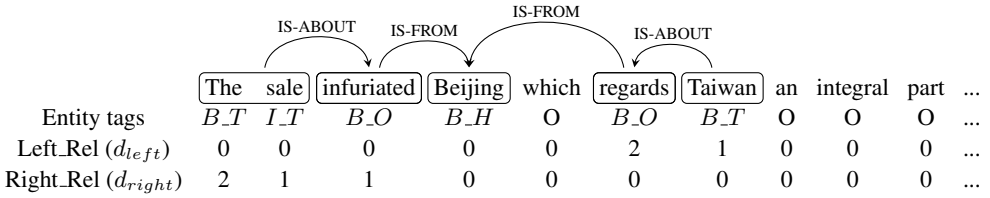

Figure 1: Gold standard annotation for an example sentence from MPQA dataset. O represents the 'Other' tag in the BIO scheme.

We can still compute the normalization factor in linear time similar to sentence-level log-likelihood.

At inference time, we jointly find the best tag and relation sequence

$$\underset{\widetilde{y},\widetilde{d}}{\operatorname{argmax}}\, s(x, \widetilde{y}, \widetilde{d}, \theta^{'})$$

for an input sentence x using Viterbi decoding.

For our task of joint extraction of opinion entities and relations, we train our model to predict tag y and relation distance d for every word in the sentence by maximizing the log-likelihood (**SLL+RLL**) below using Adadelta (Zeiler, 2012).

$$\sum_{(x,y)\in \mathrm{T}} \log p_{sent}(y|x,\theta^{'}) + \log p_{rel,Left}(y,d|x,\theta^{'}) + \log p_{rel,Right}(y,d|x,\theta^{'})$$

5 Experiments

5.1 Data

We use the MPQA 2.0 corpus (Wiebe and Cardie, 2005; Wilson, 2008). It contains news articles and editorials from a wide variety of news sources. There are a total of 482 documents in our dataset containing 9471 sentences with phrase-level annotations. We set aside 132 documents as a development set and use the remaining 350 documents as the evaluation set. We report the results using 10-fold cross validation at the document level to mimic the methodology of Yang and Cardie (2013).

The dataset contains gold-standard annotations for opinion entities — expressions, targets, holders. We use only the direct subjective/opinion expressions. There are also annotations for opinion relations – IS-FROM between opinion holders and opinion expressions; and IS-ABOUT between opinion targets and opinion expressions. These relations can overlap but we discard all relations that contain sub-relations similar to Yang and Cardie (2013). We also leave identification of overlapping relations for future work.

Figure 1 gives an example of an annotated sentence from the dataset: boxes denote opinion entities and opinion relations are shown by arcs. We interpret these relations arcs as directed — from an opinion expression towards an opinion holder, and from an opinion target towards an opinion expression.

In order to use the RLL formulation as defined in Section 4.3, we pre-process these relation arcs to obtain the left-relation distances (d_{left}) and right-relation distances (d_{right}) as shown in Figure 1. For each word in an entity, we find its distance to the nearest word in the related entity. These distances become our relation tags. The entity tags are interpreted using the BIO scheme, also shown in the figure. Our RLL model jointly models the entity tags and relation tags. At inference time, these entity tags and relation tags are used together to determine IS-FROM and IS-ABOUT relations. We use a simple majority vote to determine the final entity tag from SLL+RLL model.

5.2 Evaluation Metrics

We use precision, recall and F-measure (as in Yang and Cardie (2013)) as evaluation metrics. Since the identification of exact boundaries for opinion entities is hard even for humans (Wiebe and Cardie, 2005), soft evaluation methods such as Binary Overlap and Proportional Overlap are reported. Binary Overlap counts every overlapping predicted and gold entity as correct, while Proportional Overlap assigns a partial score proportional to the ratio of overlap span and the correct span (Recall) or the ratio of overlap span and the predicted span (Precision).

For the case of opinion relations, we report precision, recall and F-measure according to the Binary Overlap. It considers a relation correct if there is an overlap between the predicted opin-

	Opinion Expression			Opinion Target			Opinion Holder		
Method	P	R	F1	P	R	F1	P	R	F1
CRF	**84.42**$^{\mathbf{3.24}}$	61.61$^{3.20}$	*71.17*$^{2.66}$	**80.38**$^{\mathbf{2.72}}$	46.80$^{4.41}$	*59.10*$^{4.06}$	**73.37**$^{\mathbf{4.09}}$	49.71$^{3.46}$	59.21$^{3.49}$
CRF+ILP	73.53$^{3.90}$	**74.89**$^{\mathbf{2.51}}$	**74.11**$^{\mathbf{2.49}}$	*77.27*$^{3.49}$	56.94$^{3.94}$	**65.40**$^{\mathbf{3.07}}$	*67.00*$^{3.17}$	**67.22**$^{\mathbf{3.50}}$	**67.22**$^{\mathbf{2.54}}$
LSTM+WLL	67.88$^{4.49}$	66.13$^{3.20}$	66.87$^{2.66}$	58.71$^{4.87}$	54.92$^{3.23}$	56.50$^{1.51}$	60.33$^{4.54}$	*63.34*$^{2.33}$	61.65$^{2.37}$
LSTM+SLL	70.45$^{5.12}$	66.65$^{3.46}$	68.37$^{3.14}$	63.02$^{4.61}$	56.77$^{3.98}$	*59.65*$^{3.61}$	61.85$^{3.82}$	*63.12*$^{3.59}$	*62.35*$^{2.46}$
LSTM+SLL+RLL	71.73$^{5.35}$	*70.92*$^{3.96}$	*71.11*$^{2.71}$	64.52$^{5.52}$	**65.94**$^{\mathbf{4.74}}$	*64.84*$^{1.44}$	62.75$^{3.75}$	*67.17*$^{4.37}$	*64.71*$^{2.23}$
CRF	**80.78**$^{\mathbf{3.27}}$	57.62$^{3.24}$	*67.19*$^{2.63}$	**71.81**$^{\mathbf{3.22}}$	42.36$^{3.78}$	*53.23*$^{3.69}$	**71.56**$^{\mathbf{3.54}}$	48.61$^{3.51}$	57.86$^{3.43}$
CRF+ILP	71.03$^{4.03}$	**69.72**$^{\mathbf{2.37}}$	**70.22**$^{\mathbf{2.44}}$	**71.94**$^{\mathbf{3.25}}$	49.83$^{3.24}$	**58.72**$^{\mathbf{2.80}}$	*65.70*$^{3.07}$	**65.91**$^{\mathbf{3.63}}$	**65.68**$^{\mathbf{2.61}}$
LSTM+WLL	64.47$^{4.79}$	59.45$^{3.52}$	61.67$^{2.26}$	52.72$^{5.01}$	44.21$^{2.54}$	47.85$^{1.41}$	58.41$^{4.72}$	59.72$^{2.52}$	52.45$^{2.23}$
LSTM+SLL	65.97$^{5.46}$	61.76$^{3.69}$	63.60$^{3.05}$	54.46$^{4.49}$	50.16$^{4.38}$	52.01$^{3.05}$	59.80$^{3.29}$	*61.27*$^{3.75}$	60.40$^{2.26}$
LSTM+SLL+RLL	65.48$^{4.92}$	*65.54*$^{3.65}$	*65.56*$^{2.71}$	52.75$^{6.81}$	**60.54**$^{\mathbf{4.78}}$	*55.81*$^{1.96}$	59.44$^{3.56}$	*65.51*$^{4.22}$	*62.18*$^{2.50}$

Table 1: Performance on opinion entity extraction. Top table shows Binary Overlap performance; bottom table shows Proportional Overlap performance. Superscripts designate one standard deviation.

ion expression and the gold opinion expression as well as an overlap between the predicted entity (holder/target) and the gold entity (holder/target).

5.3 Baselines

CRF+ILP. We use the ILP-based joint inference model (Yang and Cardie, 2013) as baseline for both the entity and relation extraction tasks. It represents the state-of-the-art for fine-grained opinion extraction. Their method first identifies opinion entities using **CRF**s (an additional baseline) with a variety of features such as words, POS tags, and lexicon features (the subjectivity strength of the word in the Subjectivity Lexicon). They also train a relation classifier (logistic regression) by over-generating candidates from the CRFs (50-best paths) using local features such as word, POS tags, subjectivity lexicons as well as semantic and syntactic features such as semantic frames, dependency paths, WordNet hypernyms, etc. Finally, they use ILP for joint-inference to find the optimal prediction for both opinion entity and opinion relation extraction.

LSTM+SLL+Softmax. As an additional baseline for relation extraction, we train a softmax classifier on top of our SLL framework. We jointly learn the relation classifier and SLL model. For every entity pair $[x]_i^j$, $[x]_k^l$, we first sum the start and end word output representation $[z_t]$ and then concatenate them to learn softmax weight W' where $W' \in \mathbb{R}^{3 \times 2d_h}$.

$$y_{rel} = softmax(W' \begin{bmatrix} [z_t]_i + [z_t]_j \\ [z_t]_k + [z_t]_l \end{bmatrix})$$

The inference is pipelined in this case. At the time of inference, we first predict the entity spans and then use these spans for relation classification.

5.4 Hyperparameter and Training Details

We use multi-layer bi-directional LSTMs for all the experiments such that the number of hidden layers is 3 and the dimensionality of hidden units (d_h) is 50. We use Adadelta for training. We initialize our word representation using publicly available word2vec (Mikolov et al., 2013) trained on Google News dataset and keep them fixed during training. For RLL, we keep D_{Left} and D_{Right} as 15. All the weights in the network are initialized from small random uniform noise. We train all our models for 200 epochs. We do not pretrain our network. We regularize our network using dropout (Srivastava et al., 2014) with the dropout rate tuned using the development set. We select the final model based on development-set performance (average of Proportional Overlap for entities and Binary Overlap for relations).

6 Results

6.1 Opinion Entities

Table 1 shows the performance of opinion entity identification using the Binary Overlap and Proportional Overlap evaluation metrics. We discuss specific results in the paragraphs below.

WLL vs. SLL. SLL performs better than WLL on all entity types, particularly with respect to Proportional Overlap on opinion holder and target entities. A similar trend can be seen for the example sentences in Table 3. In S1, SLL extracts "has been in doubt" as the opinion expression whereas WLL only identifies "has". Similarly in S2, WLL annotates "Saudi Arabia's request on a case-by-case" as the target while SLL correctly includes "basis" in its annotation. Thus, we find that modeling the transitions between adjacent tags enables

	IS-ABOUT			IS-FROM		
Method	P	R	F1	P	R	F1
CRF+ILP	*61.57*$^{4.56}$	*47.65*$^{3.12}$	*54.39*$^{2.49}$	*64.04*$^{3.08}$	**58.79**$^{\mathbf{4.42}}$	**61.17**$^{\mathbf{3.02}}$
LSTM+SLL+Softmax	36.23$^{5.10}$	36.12$^{7.75}$	35.40$^{3.35}$	36.44$^{5.26}$	40.19$^{6.13}$	37.60$^{3.42}$
LSTM+SLL+RLL	**62.48**$^{\mathbf{3.87}}$	**49.80**$^{\mathbf{2.84}}$	**54.98**$^{\mathbf{2.54}}$	**64.19**$^{\mathbf{3.81}}$	*53.75*$^{6.00}$	*58.22*$^{3.01}$

Table 2: Performance on opinion relation extraction using Binary Overlap on the opinion entities. Superscripts designate one standard deviation.

SLL to find entire opinion entity phrases better than WLL, leading to better Proportional Overlap scores.

SLL vs. SLL+RLL. From Table 1, we see that the joint-extraction model (SLL+RLL) performs better than SLL as expected. More specifically, SLL+RLL model has better recall for all opinion entity types. The example sentences from Table 3 corroborate these results. In S1, SLL+RLL identifies "announced" as an opinion expression, which was missing in both WLL and SLL. In S3, neither the WLL nor the SLL model can annotate opinion holder (H_1) or the target (T_1), but SLL+RLL correctly identifies the opinion entities because of modeling the relations between the opinion expression "will decide" and the holder/target entities.

CRF vs. LSTM-based Models. From the analysis of the performance in Table 1, we find that our WLL and SLL models perform worse while our best SLL+RLL model can only match the performance of the CRF baseline on opinion expressions. Even though the recall of all our LSTM-based models is higher than the recall of the CRF-baseline for opinion expressions, we cannot match the precision of CRF baseline. We suspect that the reason for such high precision on the part of the CRF is its access to carefully prepared subjectivity-lexicons[4]. Our LSTM-based models do not rely on such features except via the word-vectors. With respect to holders and targets, we find that our SLL model performs similar to the CRF baseline. However, the SLL+RLL model outperforms CRF baseline.

CRF+ILP vs. SLL+RLL. Even though we find that our LSTM-based joint-model (SLL+RLL) outperforms our LSTM-based only-entity extraction model (SLL), the performance is still below the ILP-based joint-model (CRF+ILP). However, we perform comparably with respect to target entities (Binary Overlap). Also, our recall on targets is much better than all other models whereas the recall on holders is very similar to CRF+ILP. Our SLL+RLL model can identify targets such as "Australia's involvement in Kyoto" which the ILP-based model cannot, as observed for S1 in Table 3. In S3, the ILP-based model also erroneously divides the target "consider Saudi Arabia's request on a case-by-case basis" into a holder "Saudi Arabia's" and opinion expression "request", while SLL+RLL model can correctly identify it. We will compare the two models in detail in Section 7.

[4]http://mpqa.cs.pitt.edu/lexicons/ subj_lexicon/

6.2 Opinion Relations

The extraction of opinion relations is our primary task. Table 2[5] shows the performance on opinion relation extraction task using Binary Overlap.

SLL+Softmax vs. SLL+RLL. The opinion entities and relations are jointly modeled in both the models, but we see a significant improvement in performance by adding relation level dependencies to the model vs. learning a classifier on top of sentence-level dependencies to learn the relation between entities. LSTM+SLL+RLL performs much better in terms of both precision and recall on both IS-FROM and IS-ABOUT relations.

CRF+ILP vs. SLL+RLL. We find that our SLL+RLL model performs comparably and even slightly better on IS-ABOUT relations. Such performance is encouraging because our LSTM-based model does not rely on features such as dependency paths, semantic frames or subjectivity lexicons for our model. Our sequential LSTM model is able to learn these relations thus validating that LSTMs can model long-term dependencies. However, for IS-FROM relations, we find that our recall is lower than the ILP-based joint model.

[5]Yang and Cardie (2013) omitted a subset of targets and IS-ABOUT relations. We fixed this and re-ran their models on the updated dataset, obtaining the lower F-score 54.39 for IS-ABOUT relations.

S1 :	[Australia's involvement in Kyoto]$_{\mathbf{T_1}}$ [has been in doubt]$_{\mathbf{O_1}}$ ever since [the US President, George Bush]$_{\mathbf{H_2}}$, [announced]$_{\mathbf{O_2}}$ last year that [ratifying the protocol]$_{\mathbf{T_2}}$ would hurt the US economy.
CRF+ILP	Australia's involvement in Kyoto [has been in doubt]$_{\mathbf{O_1}}$ ever since the US President, George Bush, announced last year that [ratifying the protocol]$_{\mathbf{T_1}}$ would hurt the US economy.
WLL	[Australia's involvement in Kyoto]$_{\mathbf{T}}$ [has]$_{\mathbf{O}}$ been in doubt ever since the US [President]$_{\mathbf{H}}$, [George Bush]$_{\mathbf{H}}$, announced last year that ratifying the protocol would hurt the US economy.
SLL	[Australia's involvement in Kyoto]$_{\mathbf{T}}$ [has been in doubt]$_{\mathbf{O}}$ ever since the US President, George Bush, announced last year that ratifying the protocol would hurt the US economy.
SLL+RLL	[Australia's involvement in Kyoto]$_{\mathbf{T}}$ [has been in doubt]$_{\mathbf{O}}$ ever since the US President, [George Bush]$_{\mathbf{H_2}}$, [announced]$_{\mathbf{O_2}}$ last year that [ratifying the protocol]$_{\mathbf{T_2}}$ would hurt the US economy.
S2 :	Bush said last week [he]$_{\mathbf{H_{1,2}}}$ [was willing]$_{\mathbf{O_1}}$ [to consider]$_{\mathbf{O_2}}$ [Saudi Arabia's request on a case-by-case basis]$_{\mathbf{T_2}}$ but [U.S. officials]$_{\mathbf{H_3}}$ [doubted]$_{\mathbf{O_3}}$ [it would happen any time soon]$_{\mathbf{T_3}}$.
CRF+ILP	[Bush]$_{\mathbf{H_1}}$ [said]$_{\mathbf{O_1}}$ last week [he]$_{\mathbf{H_2}}$ [was willing to consider]$_{\mathbf{O_2}}$ [Saudi Arabia's]$_{\mathbf{H_3}}$ [request]$_{\mathbf{O_3}}$ on a case-by-case basis but [U.S. officials]$_{\mathbf{H_4}}$ [doubted]$_{\mathbf{O_4}}$ [it]$_{\mathbf{T_4}}$ would happen any time soon.
WLL	Bush said last week [he]$_{\mathbf{H}}$ [was willing]$_{\mathbf{O}}$ to [consider]$_{\mathbf{O}}$ [Saudi Arabia's request on a case-by-case]$_{\mathbf{T}}$ basis but [U.S. officials]$_{\mathbf{H}}$ [doubted]$_{\mathbf{O}}$ [it]$_{\mathbf{T}}$ would [happen any time soon]$_{\mathbf{T}}$.
SLL	Bush said last week [he]$_{\mathbf{H}}$ [was willing]$_{\mathbf{O}}$ to [consider Saudi Arabia's request on a case-by-case basis]$_{\mathbf{T}}$ but [U.S. officials]$_{\mathbf{H}}$ [doubted]$_{\mathbf{O}}$ [it]$_{\mathbf{T}}$ would happen any time soon.
SLL+RLL	Bush said last week [he]$_{\mathbf{H_1}}$ [was willing to consider]$_{\mathbf{O_1}}$ [Saudi Arabia's request on a case-by-case basis]$_{\mathbf{T_1}}$ but [U.S. officials]$_{\mathbf{H_2}}$ [doubted]$_{\mathbf{O_2}}$ [it would happen any time soon]$_{\mathbf{T_2}}$.
S3 :	Hence, [the Organization of Petroleum Exporting Countries (OPEC)]$_{\mathbf{H_1}}$, [will decide]$_{\mathbf{O_1}}$ at its meeting on Wednesday [whether or not to cut its worldwide crude production in an effort to shore up energy prices]$_{\mathbf{T_1}}$.
CRF+ILP	Hence, the Organization of Petroleum Exporting Countries (OPEC), [will decide]$_{\mathbf{O_1}}$ at its meeting on Wednesday whether [or not to cut its worldwide crude production in an effort to shore up energy prices]$_{\mathbf{T_1}}$.
WLL	Hence, the Organization of Petroleum Exporting Countries (OPEC), will [decide]$_{\mathbf{O}}$ at its meeting on Wednesday whether or not to cut its worldwide crude production in an effort to shore up energy prices.
SLL	Hence, the Organization of Petroleum Exporting Countries (OPEC), [will decide]$_{\mathbf{O}}$ at its meeting on Wednesday whether or not to cut its worldwide crude production in an effort to shore up energy prices.
SLL+RLL	Hence, [the Organization of Petroleum Exporting Countries (OPEC)]$_{\mathbf{H_1}}$, [will decide]$_{\mathbf{O_1}}$ at its meeting on Wednesday whether [or not to cut its worldwide crude production in an effort to shore up energy prices]$_{\mathbf{T_1}}$.

Table 3: Output from different models. The first row for each example is the gold standard.

7 Discussion

In this section, we discuss the various advantages and disadvantages of the LSTM-based SLL+RLL model as compared to the joint-inference (CRF+ILP) model. We provide examples from the dataset in Table 4.

From Table 2, we find that SLL+RLL model performs worse with respect to the opinion expression entities and opinion holder entities. On careful analysis of the output, we found cases such as S1 in Table 4. For such sentences SLL+RLL model prefers to annotate the opinion target (T_3) "US requests for more oil exports", whereas the ILP model annotates the embedded opinion holder (H_4) "US" and opinion expression (T_4) "requests". Both models are valid with respect to the gold-standard. In order to simplify our problem, we discard these embedded relations during training similar to Yang and Cardie (2013). However, for future work we would like to model these overlapping relations which could potentially improve our performance on opinion holders and opinion expressions.

We also found several cases such as S2, where the SLL+RLL model fails to annotate "said" as an opinion expression. The gold standard opinion expressions include speech events like "said" or "a statement", but not all occurrences of these speech events are opinion expressions, some are merely objective events. In S2, "was martyred" is an indication of an opinion being expressed, so "said" is annotated as an opinion expression. From our observation, the ILP model is more relaxed in annotating most of these speech events as opinion expressions and thus likely to identify corresponding

S1 :	However, [Chavez]$_{\mathbf{T_1}}$ who [is known for]$_{\mathbf{O_1}}$ [his]$_{\mathbf{H_2}}$ [ala Fidel Castro left-leaning anti-American philosophy]$_{O_2}$ had on a number of occasions [rebuffed]$_{O_3}$ [[US]$_{\mathbf{H_4}}$ [requests]$_{\mathbf{O_4}}$ for [more oil exports]$_{\mathbf{T_4}}$]$_{\mathbf{T_3}}$.
CRF+ILP	However, [Chavez]$_{\mathbf{H_1}}$ who [is known]$_{\mathbf{O}}$ for [his ala Fidel Castro]$_{\mathbf{H_2}}$ [left-leaning anti-American philosophy]$_{\mathbf{O_2}}$ had on a number of occasions [rebuffed]$_{\mathbf{O_1}}$ [US]$_{\mathbf{H_3}}$ [requests]$_{\mathbf{O_3}}$ for more oil exports.
SLL+RLL	However, Chavez who [is known]$_{\mathbf{O}}$ for his ala Fidel Castro left-leaning anti-American [philosophy]$_{\mathbf{O}}$ had on a number of occasions [rebuffed]$_{\mathbf{O_1}}$ [US requests for more oil exports]$_{\mathbf{T_1}}$.
S2 :	A short while ago, [our correspondent in Bethlehem]$_{\mathbf{H_1}}$ [said]$_{\mathbf{O_1}}$ that [Ra'fat al-Bajjali]$_{\mathbf{T_1}}$ **was martyred** of wounds sustained in the explosion.
CRF+ILP	A short while ago, [our correspondent]$_{\mathbf{H_1}}$ in Bethlehem [said]$_{\mathbf{O_1}}$ that [Ra'fat al-Bajjali]$_{T_1}$ was martyred of wounds sustained in the explosion.
SLL+RLL	A short while ago, our correspondent in Bethlehem said that Ra'fat al-Bajjali was martyred of wounds sustained in the explosion.
S3 :	This is no criticism, and is widely known and appreciated.
CRF+ILP	This is no criticism, and is widely known and appreciated.
SLL+RLL	[This]$_{\mathbf{T_1}}$ [is no criticism]$_{\mathbf{O_1}}$, and is widely [known and appreciated]$_{\mathbf{O}}$.
S4 :	From the fact that mothers care for their young, we can not deduce that they ought to do so, Hume argued.
CRF+ILP	From the fact that [mothers]$_{\mathbf{H_1}}$ [care]$_{\mathbf{O_1}}$ for their young, we can not deduce that they ought to do so, [Hume]$_{\mathbf{H_2}}$ [argued]$_{\mathbf{O_2}}$.
SLL+RLL	From the fact that mothers care for their young, [we]$_{\mathbf{H_1}}$ [can not deduce]$_{\mathbf{O_1}}$ that [they]$_{\mathbf{T_1}}$ ought to do so, [Hume]$_{\mathbf{H_2}}$ [argued]$_{\mathbf{O_2}}$.

Table 4: Examples from the dataset with label annotations from CRF+ILP and SLL+RLL models for comparison. The first row for each example is the gold standard.

opinion holders and opinion targets as compared to SLL+RLL model.

There were also instances such as S3 and S4 in Table 4 for which the gold standard does not have an annotation but the SLL+RLL output looks reasonable with respect to our task. In S3, SLL+RLL identifies "is no criticism" as an opinion expression for the target "This". However, it fails to identify the relation-link between "known and appreciated" and the target "This". Similarly, SLL+RLL also identifies reasonable opinion entities in S4, whereas the ILP model erroneously annotates "mothers" as the opinion holder and "care" as the opinion expression.

We handle the task of joint-extraction of opinion entities and opinion relations as a sequence labeling task in this paper and report the performance of the 1-best path at the time of Viterbi inference. However, there are approaches such as discriminative reranking (Collins and Koo, 2005) to rerank the output of an existing system that offer a means for further improving the performance of our SLL+RLL model. In particular, the oracle performance using the top-10 Viterbi paths from our SLL+RLL model has an F-score of 82.11 for opinion expressions, 76.77 for targets and 78.10 for holders. Similarly, IS-ABOUT relations have an F-score of 65.99 and IS-FROM relations, an F-score of 70.80. These scores are on average 10 points better than the performance of the current SLL+RLL model, indicating that substantial gains might be attained via reranking.

8 Conclusion

In this paper, we explored LSTM-based models for the joint extraction of opinion entities and relations. Experimentally, we found that adding sentence-level and relation-level dependencies on the output layer improves the performance on opinion entity extraction, obtaining results within 1-3% of the ILP-based joint model on opinion entities, within 3% for IS-FROM relation and comparable for IS-ABOUT relation.

In future work, we plan to explore the effects of pre-training (Bengio et al., 2009) and scheduled sampling (Bengio et al., 2015) for training our LSTM network. We would also like to explore re-ranking methods for our problem. With respect to the fine-grained opinion mining task, a potential future direction to be able to model overlapping and embedded entities and relations and also to extend this model to handle cross-sentential relations.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pages 41–48, New York, NY, USA. ACM.

Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. 2015. Scheduled sampling for sequence prediction with recurrent neural networks. *CoRR*, abs/1506.03099.

Yoshua Bengio. 2009. Learning deep architectures for ai. *Found. Trends Mach. Learn.*, 2(1):1–127, January.

Eric Breck, Yejin Choi, and Claire Cardie. 2007. Identifying expressions of opinion in context. In *Proceedings of the 20th International Joint Conference on Artifical Intelligence*, IJCAI'07, pages 2683–2688, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Yejin Choi, Claire Cardie, Ellen Riloff, and Siddharth Patwardhan. 2005. Identifying sources of opinions with conditional random fields and extraction patterns. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT '05, pages 355–362, Stroudsburg, PA, USA. Association for Computational Linguistics.

Yejin Choi, Eric Breck, and Claire Cardie. 2006. Joint extraction of entities and relations for opinion recognition. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, EMNLP '06, pages 431–439, Stroudsburg, PA, USA. Association for Computational Linguistics.

Michael Collins and Terry Koo. 2005. Discriminative reranking for natural language parsing. *Comput. Linguist.*, 31(1):25–70, March.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *J. Mach. Learn. Res.*, 12:2493–2537, November.

Alex Graves, Navdeep Jaitly, and Abdel-rahman Mohamed. 2013. Hybrid speech recognition with deep bidirectional LSTM. In *2013 IEEE Workshop on Automatic Speech Recognition and Understanding, Olomouc, Czech Republic, December 8-12, 2013*, pages 273–278.

James Hammerton. 2003. Named entity recognition with long short-term memory. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003 - Volume 4*, CONLL '03, pages 172–175, Stroudsburg, PA, USA. Association for Computational Linguistics.

Michiel Hermans and Benjamin Schrauwen. 2013. Training and analysing deep recurrent neural networks. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States.*, pages 190–198.

Salah El Hihi and Yoshua Bengio. 1996. Hierarchical recurrent neural networks for long-term dependencies.

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

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional LSTM-CRF models for sequence tagging. *CoRR*, abs/1508.01991.

Ozan Irsoy and Claire Cardie. 2013. Bidirectional recursive neural networks for token-level labeling with structure. *arXiv preprint arXiv:1312.0493*.

Ozan İrsoy and Claire Cardie. 2014. Opinion mining with deep recurrent neural networks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 720–728.

Soo-Min Kim and Eduard Hovy. 2006. Extracting opinions, opinion holders, and topics expressed in online news media text. In *Proceedings of the Workshop on Sentiment and Subjectivity in Text*, SST '06, pages 1–8, Stroudsburg, PA, USA. Association for Computational Linguistics.

Nozomi Kobayashi, Kentaro Inui, and Yuji Matsumoto. 2007. Extracting aspect-evaluation and aspect-of relations in opinion mining. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL.*

Pengfei Liu, Shafiq Joty, and Helen Meng. 2015. Fine-grained opinion mining with recurrent neural networks and word embeddings. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1433–1443, Lisbon, Portugal, September. Association for Computational Linguistics.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In C.J.C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates, Inc.

Makoto Miwa and Mohit Bansal. 2016. End-to-end relation extraction using lstms on sequences and tree structures. *CoRR*, abs/1601.00770.

Jürgen Schmidhuber. 1992. Learning complex, extended sequences using the principle of history compression. *Neural Comput.*, 4(2):234–242, March.

M. Schuster and K.K. Paliwal. 1997. Bidirectional recurrent neural networks. *Trans. Sig. Proc.*, 45(11):2673–2681, November.

Richard Socher, Brody Huval, Christopher D. Manning, and Andrew Y. Ng. 2012. Semantic compositionality through recursive matrix-vector spaces. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, EMNLP-CoNLL '12, pages 1201–1211, Stroudsburg, PA, USA. Association for Computational Linguistics.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15:1929–1958.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 3104–3112. Curran Associates, Inc.

Janyce Wiebe and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. language resources and evaluation. In *Language Resources and Evaluation (formerly Computers and the Humanities*, page 2005.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT '05, pages 347–354, Stroudsburg, PA, USA. Association for Computational Linguistics.

Theresa Ann Wilson. 2008. *Fine-grained Subjectivity and Sentiment Analysis: Recognizing the intensity, polarity, and attitudes of private states*. Ph.D. thesis, The University of Pittsburgh, June.

Yan Xu, Lili Mou, Ge Li, Yunchuan Chen, Hao Peng, and Zhi Jin. 2015. Classifying relations via long short term memory networks along shortest dependency paths. In *In Proceedings of Conference on Empirical Methods in Natural Language Processing*.

Bishan Yang and Claire Cardie. 2012. Extracting opinion expressions with semi-markov conditional random fields. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, EMNLP-CoNLL '12, pages 1335–1345, Stroudsburg, PA, USA. Association for Computational Linguistics.

Bishan Yang and Claire Cardie. 2013. Joint inference for fine-grained opinion extraction. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1640–1649, Sofia, Bulgaria, August. Association for Computational Linguistics.

Matthew D. Zeiler. 2012. ADADELTA: an adaptive learning rate method. *CoRR*, abs/1212.5701.