Domain Specific Named Entity Recognition Referring to the Real World by Deep Neural Networks

Suzushi Tomori[†] **Takashi Ninomiya**[††] **Shinsuke Mori**[†††]
[†]Graduate School of Informatics, Kyoto University*
[††]Graduate School of Science and Engineering, Ehime University
[†††]Academic Center for Computing and Media Studies, Kyoto University
[†]tomori.suzushi.72e@st.kyoto-u.ac.jp
[††]ninomiya@cs.ehime-u.ac.jp
[†††]forest@i.kyoto-u.ac.jp

Abstract

In this paper, we propose a method for referring to the real world to improve named entity recognition (NER) specialized for a domain. Our method adds a stacked auto-encoder to a text-based deep neural network for NER. We first train the stacked auto-encoder only from the real world information, then the entire deep neural network from sentences annotated with NEs and accompanied by real world information. In our experiments, we took Japanese chess as the example. The dataset consists of pairs of a game state and commentary sentences about it annotated with game-specific NE tags. We conducted NER experiments and showed that referring to the real world improves the NER accuracy.

1 Introduction

In recent years there has been a surge of interest in relating natural language to the real world. And more and more language resources accompanied by nonlinguistic data are becoming available. Typical examples are image descriptions (Yang et al., 2011; Ushiku et al., 2011) and video (Hashimoto et al., 2014). Ferraro et al. (2015) summarized many other image and video datasets. These datasets allow us to attempt the task of connecting language expressions to the real world, which is called *symbol grounding* (Harnad, 1990). Bruni et al. (2014) proposed methods for acquiring multimodal representations by applying SVD to distributional semantics and bag-of-visual-words (BoVW). Ngiam et al. (2011) proposed unsupervised multimodal learning based on deep restricted boltzmann machines (RBMs). In the field of natural language processing (NLP) research, Kiela et al. (2015) proposed to acquire bilingual lexicon based on visual similarity. Ramisa et al. (2015) describe a method for predicting a preposition referring to positions in the image.

In this paper, we propose a method for enhancing a named entity (NE) recognizer referring to the real world. Because of the lack of datasets consisting of sentences annotated with the general NE tags such as names of people, organizations, and times (Sang and Meulder, 2003), with accompanying real world data, we take game states as the counterpart of the language and the NE tag set specialized for game commentaries such as defense formations and opening names (Mori et al., 2016). Similar to bio-medical NEs (Settles, 2004; Tateisi et al., 2002), these NEs are useful for applications in the game domain. Our method could be used to improve automatic game commentary systems (Kameko et al., 2015b; Chen et al., 2010) or to build a state search method that uses natural language queries instead of state notations (Ganguly et al., 2014). In addition to these interesting applications, game states have another advantage for NLP research. They are much easier to recognize than images and video, which allows us to concentrate on the NLP problem.

In order to incorporate the real world, *i.e.* game states, into NE recognition (NER), we propose to use deep neural networks (DNNs), which have been reported to be successful in various NLP tasks such as word embedding (Bengio et al., 2003; Mikolov et al., 2013b; Pennington et al., 2014; Mikolov et al., 2013a), part-of-speech tagging (Tsuboi, 2014), parsing (Socher et al., 2010; Socher et al., 2012; Socher et al., 2013a), parsing (Socher et al., 2013a), NER (Hammerton, 2003) , sentiment analysis (Socher et al., 2013b) and machine translation (Neubig et al., 2015). First we build a normal NE recognizer by referring only to the text information based on DNN. Each unit of its output layer corresponds to a BIO tag for

*This work was done when the first author was at Ehime University.

Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, pages 236–242,
Berlin, Germany, August 7-12, 2016.

the word (see Section 3). We use post processing based on the Viterbi algorithm to choose the best tag sequence by discarding inconsistent ones. This design allows us to train the model from partially annotated sentences, in which only some words are annotated with NE tags (Sasada et al., 2015). Next we extend the text-based DNN with a module that refers to game states. This module is a stacked-auto-encoder (SAE) (Bengio et al., 2007) and we first train it only from game states. The pre-training allows the model to learn game state embedding which abstracts game state information. Then we fine-tune the entire DNN for NER, consisting of both text-based DNN and SAE. As we show in later section of this paper, we end up with an NE recognizer that refers to real world information in addition to text information, which increases its accuracy.

2 Related Work

There are several lines of multimodal learning in the fields of pattern recognition and NLP. Most learn multimodal representations by solving unsupervised learning tasks or pseudo-supervised learning tasks, but there were only a few studies that directly learned multimodal representations for target tasks in NLP. Our method incorporates multimodal information in DNNs for NER.

Bruni et al. (2014) proposed methods for acquiring multimodal representations by applying SVD to distributional semantics and BoVW. Lopopolo and van Miltenburg (2015) proposed a similar method for acquiring sound-based distributional semantics. Textual vectors are acquired by using latent semantic analysis (LSA) and auditory vectors are acquired by the bag-of-audio-words (BoAW) method. The multimodal representations are acquired by applying SVD. Ngiam et al. (2011) and Srivastava and Salakhutdinov (2012) proposed unsupervised learning methods based on deep RBMs for learning multimodal representations in hidden layers. Providing paired information such as text-image pairs or audio-video pairs to RBMs, shared representations are learned in their hidden layers. Ngiam et al. (2011) also used deep auto-encoders for learning RBMs. After acquiring multimodal representations, they can be used as inputs for other supervised learning tasks, such as speech recognition and image retrieval, where standard linear classifiers are used for solving the tasks. Silberer and Lapata (2014) proposed a deep learning method for learning multimodal representations by solving pseudo-supervised tasks to predict the input's object label, such as 'boat,' given textual and visual attribute-based representations for the object. Their objective function is the weighted sum of the auto-encoding error and the classification error. Though their model is for supervised learning, Multimodal representations are learned In their experiments, the acquired multimodal representations were used for evaluating the word similarity task and word clustering task.

Lazaridou et al. (2015) extend word2vec (Mikolov et al., 2013a; Mikolov et al., 2013b) to incorporate visual information for acquiring multimodal representations. Word embedding methods including word2vec are often used for various NLP tasks instead of one hot representations, and were shown to improve the performance of NLP systems. Word embeddings are mappings from a word to a low-dimensional real vectors that represents word meanings and relations between words. Word2vec is a method for acquiring word embeddings from a neural network which solves a pseudo-supervised task to predict surrounding words. Kiela and Clark (2015) extend word2vec to incorporate bag-of-audio-words (BoAW). Gupta et al. (2015) have shown that word embeddings contain much information for predicting attributes. Herbelot and Vecchi (2015) proposed a method for predicting general quantifiers such as *some* for predicate-subject pairs.

Similar to this paper Kameko et al. (2015a) proposed a method for word segmentation using game states and DNNs. The main differences between their method and ours is that i) they use game states to build a term dictionary for word segmentation, but our method directly incorporates a game state to improve NER, and ii) they used manually developed features to extract game states while we automatically acquire game states by using pre-training.

3 Game Commentary Corpus

The game we chose for the experiments is Japanese chess, called *shogi* in Japanese. It is a two-player board game with professional players. The board has 9×9 squares and games are played with 40 pieces of 14 different types. Unlike chess, players can reuse captured pieces. In computer science terms, it is a deterministic perfect informa-

Tag	Meaning
Hu	Human
Tu	Turn
Po	Position
Pi	Piece
Ps	Piece specifier
Mc	Move compliment
Pa	Piece attribute
Pq	Piece quantity
Re	Region
Ph	Phase
St	Strategy
Ca	Castle
Me	Move eval.
Mn	Move name
Ee	Eval. element
Ev	Evaluation
Ti	Time
Ac	Player action
Ap	Piece action
Ao	Other action
Ot	Other notion

Table 1: The named entity tag set.

tion game, so we can completely specify a game state by the positions of the pieces on the board and the captured pieces held by on both sides.

Many matches between professional players have been recorded, and many game states have commentaries made for fans by other professional players.

A game commentary corpus[1] (Mori et al., 2016) defines 21 types of NEs, which are called *shogi*-NEs, as listed in Table 1. The words in the commentary sentences in the corpus are annotated with BIO-style tags. B, I, and O stand for beginning, intermediate, and others, respectively. B or I are used for representing the beginning or intermediate words of an NE as extension like Hu-B. And O is used for representing words that are not part of any NEs. Therefore there are $43 = 21 \times 2 + 1$ BIO tags.

The main idea of this paper is that the game state, *i.e.* the real world, provides information on the texts that describe it. In the next section, we propose a method for utilizing this information in the NER task.

[1] http://www.ar.media.kyoto-u.ac.jp/data/game/

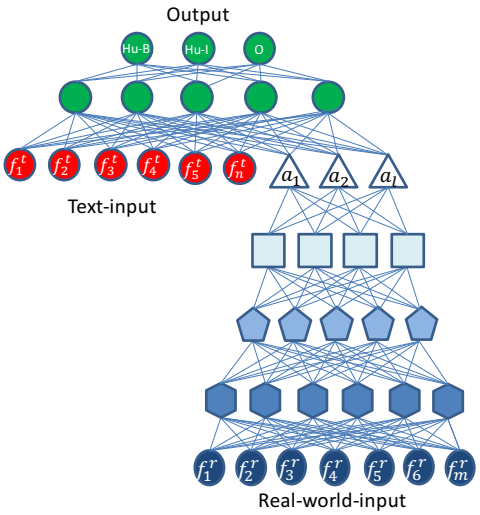

Figure 1: Deep neural networks for *shogi* NER.

Text features
w_{i-2}, w_{i-1}, w_i, w_{i+1}, w_{i+2}
$w_{i-2}w_{i-1}$, $w_{i-1}w_i$, w_iw_{i+1}, $w_{i+1}w_{i+2}$
$w_{i-2}w_{i-1}w_i$, $w_iw_{i+1}w_{i+2}$
$c(w_{i-2})$, $c(w_{i-1})$, $c(w_i)$, $c(w_{i+1})$, $c(w_{i+2})$
$pos(w_{i-2})$, $pos(w_{i-1})$, $pos(w_i)$, $pos(w_{i+1})$, $pos(w_{i+2})$

Table 2: Text features for DNN/CRF NER.

4 Utilizing Real World Information in a Named Entity Recognizer

Figure 1 shows the overall architecture of our DNN for NER. The left part is the DNN for text-based NER and the bottom right part is an additional DNN for referring to the real world.

4.1 Text-based NER

The text-based NER refers to the text only through the standard features for NER (Sang and Meulder, 2003) listed in Table 2. They consist of word n-grams in the window $\boldsymbol{w}_{i-2}^{i+2} = w_{i-2}w_{i-1}w_iw_{i+1}w_{i+2}$, where w_i is the word to be labelled, the part-of-speech tags $pos(w)$ and the character type $c(w)$[2] of a word w in the window $\boldsymbol{w}_{i-2}^{i+2}$. Each feature corresponds to a unit at the bottom left in Figure 1 ($f_1^t \ldots f_n^t$).

Each unit aligned at the top of Figure 1 corresponds to a BIO tag. Thus there are 43 units in the *shogi* NE case. The last layer is the softmax function and we choose the tag of the highest unit value

[2] In the target language in the experiments, Japanese, the types are *hiragana, katakana, kanji*, number, symbol, and combinations of them.

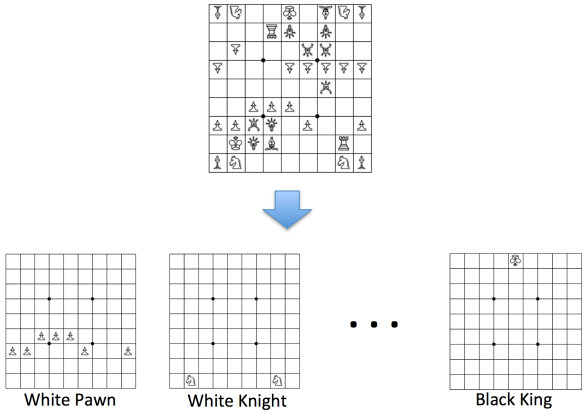

Figure 2: Game state features.

for the input word. As we mentioned in Section 1, this design makes it possible to use partially annotated data.[3] It can, however, generate inconsistent BIO tag sequences, *e.g.*, an NE starting with an I tag. We use a best path search module based on the Viterbi algorithm while limiting the search space into valid tag sequences (Sasada et al., 2015).

4.2 NER Referring to the Real World

To enable our NE recognizer to refer to the real world, we add a network to the DNN for text-based NER as shown in the bottom right in Figure 1. The input layer corresponds to the game state features depicted in Figure 2 ($f^r_1 \ldots f^r_m$). For *shogi* they are nine-by-nine binary features which represent the positions of pieces on the board for each piece type and each player. Thus we have $m = 2,268$ ($= 9 \times 9 \times 14 \times 2$) features for the pieces on the board and 14 ($= 7 \times 2$) integer features which represent the number of captured pieces for each type and each player.

To incorporate the game state features we propose using an SAE (Bengio et al., 2007) to abstract the game state information instead of directly adding the units for these features to the text-based NER. To build the SAE, we first prepare a three-layer neural network (with one hidden layer) as depicted on the left side of Figure 3 and train it providing the same game states to both input and output layers. With this process we can obtain the best reduced representations for the game states as the hidden layer that reconstructs the input game state features at the output layer.

[3]Tsuboi et al. (2008) extended conditional random fields to be trained from partially annotated data. One can extend sequence labeling DNN (RNN or LSTM) in a similar way. This is, however, clearly out of the scope of this paper.

Usage	#Sentences	#NEs	#Words	#Game states
Pretraining	-	-	-	213,195
Training	1,546	7,922	27,025	391
Test	492	2,365	7,161	156

Table 3: Game commentary corpus specifications.

Layer	0	1	2	3	4	5
Dimension	2,282	1,000	500	200	100	50

Table 4: Dimensions of the SAE layers.

Then we duplicate the hidden layer and put another hidden layer of smaller dimension between them (see the network in the middle of Figure 3) and train it in the same manner. This time the output layer is the duplicated former hidden layer and we train the new hidden layer by minimizing the difference between the duplicated former hidden layers. We repeat this process for a fixed number of times as shown on the right side of Figure 3. This process is called pre-training. Note that during pre-training only game states are used.

After the pre-training, we cut off the top layer to obtain a network with a trapezoid shape whose top layer abstracts game states ($a_1 \ldots a_l$ in Figure 1). Then we join it to the DNN for the text-based NER as shown in Figure 1. Finally, we fine-tune it from both game states and texts annotated with NE tags. Note that we also tune parameters in the pre-trained SAE.

5 Experimental Evaluation

In this section we describe the NER experiments we conducted to evaluate our method.

5.1 Experimental Settings

The corpus we used is the game commentary corpus (Mori et al., 2016) described in Section 3 briefly. Table 3 shows its specifications. Table 4 shows the number of dimensions in each layer for game state embeddings in pre-training. We set the number of layers in the SAE (Subsection 4.2) to four, with which we could maximize the accuracy on the development set held-out from the training data.

5.2 Models for Comparison

The baseline is text-based NER based on DNN as described in Subsection 4.1. In addition, we tested NER based on conditional random fields (CRFs) (Lafferty et al., 2001) with the same text features, because NER is a sequence labeling problem and

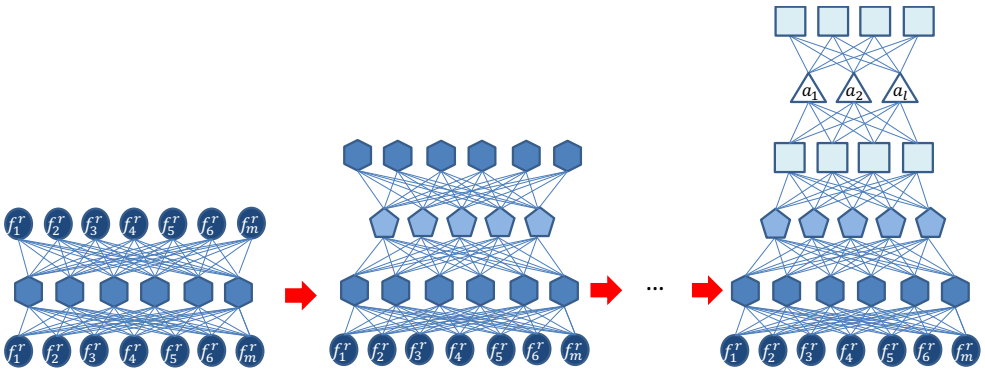

Figure 3: Building stacked-auto-encoder.

Method	BIO Accu.	Prec.	Recall	F-meas.
CRFs	89.76%	**90.58%**	76.87%	83.17
DNN	90.87%	89.27%	79.49%	84.10
DNN + R	**91.30%**	89.13%	**80.76%**	**84.74**

Table 5: NER results.

CRFs are the standard method used to solve it (McCallum and Li, 2003). We compared these baselines and our NER that refers to the real world (*DNN+R*) as described in Subsection 4.2. Its SAE was trained on 213,195 game states.

5.3 Results and Discussion

Table 5 shows the results. From the F-measures we see that *DNN* is better than *CRFs*. This is consistent with many works which apply DNN to NLP problems. A comparison between *DNN* and *DNN+R* tells us that we can achieve a further improvement by referring to real world information. The difference in BIO accuracies between them is statistically significant (McNemar's test, $p < 0.01$). Therefore we can say that our method successfully integrates real world information into text information to build a better solution to the NER problem.

When we take a close look at the precision and recall, *DNN+R* and *DNN* balance them better than *CRFs*. *CRFs* recognized *shogi*-NEs with high precision but with low recall. The NER results tell that *CRFs* tended to output O tags when they were not confident to classify correct *shogi*-NE tags. *DNN+R* and *DNN* can classify BIO tags more accurately than *CRFs* as can be seen in BIO accuracies in Table 5. As a consequence *DNN+R* and *DNN* confidently recognize more *shogi*-NEs, which makes their recall higher than that of *CRFs*.

From Table 5 we see that *DNN+R* is better than *DNN*. Followings are examples of *shogi*-NEs which *DNN+R* successfully recognized but *DNN* failed: Ot tag for "*tataki*," which means dropping a pawn in front of a piece of the opponent, and Mn tag for "*tsumero*" (threatmate). By referring to the game state, *DNN+R* was better at understanding the game situation and resulted better performance than *DNN*, the text-based NER.

6 Conclusion

In this paper, we proposed a method for referring to the real world to improve NER in a specialized domain. Our method adds an SAE to a text-based DNN for NER. We first pre-train the SAE using only real world information, and then we train the entire DNN from sentences annotated with NEs and accompanied by real world information.

In our experiments, we used *shogi* (Japanese chess) as the example. The dataset consists of pairs of a game state and commentary sentences on it annotated with 21 *shogi* NE tags. We conducted NER experiments and showed that referring to the real world improves NER accuracy.

Our method has the potential to be applied to various NER problems, such as general NER with pictures and financial NER with stock charts, by changing the SAE features. An interesting area of future work is preparing datasets in these domains and testing our method on them.

Acknowledgments

This work was supported by JSPS Grants-in-Aid for Scientific Research Grant Numbers 26540190 and 25280084. We are also grateful to Professor Yoshimasa Tsuruoka and Mr. Hirotaka Kameko for their valuable comments.

References

Yoshua Bengio, Rejean Ducharme, Pascal Vincent, and Christian Jauvin. 2003. A neural probabilistic language model. *Journal of Machine Learning Research*, 3:1137–1155.

Yoshua Bengio, Pascal Lamblin, Dan Popovici, and Hugo Larochelle. 2007. Greedy layer-wise training of deep networks. In B. Schölkopf, J. C. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*, pages 153–160. MIT Press.

E. Bruni, N. K. Tran, and M. Baroni. 2014. Multimodal distributional semantics. *Journal of Artificial Intelligence Research*, 49:1–47.

D. L. Chen, J. Kim, and R. J. Mooney. 2010. Training a multilingual sportscaster: Using perceptual context to learn language. *Journal of Artificial Intelligence Research*, 37:397–435.

Francis Ferraro, Nasrin Mostafazadeh, Ting-Hao Huang, Lucy Vanderwende, Jacob Devlin, Michel Galley, and Margaret Mitchell. 2015. A survey of current datasets for vision and language research. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 207–213.

Debasis Ganguly, Johannes Leveling, and Gareth J.F. Jones. 2014. Retrieval of similar chess positions. In *Proceedings of the 37th annual international ACM SIGIR conference*, pages 687–696. ACM.

Abhijeet Gupta, Gemma Boleda, Marco Baroni, and Sebastian Padó. 2015. Distributional vectors encode referential attributes. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 12–21, Lisbon, Portugal, September. Association for Computational Linguistics.

James Hammerton. 2003. Named entity recognition with long short-term memory. In Walter Daelemans and Miles Osborne, editors, *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 172–175.

Stevan Harnad. 1990. The symbol grounding problem. *Physica D*, 42:335–346.

Atsushi Hashimoto, Tetsuro Sasada, Yoko Yamakata, Shinsuke Mori, and Michihiko Minoh. 2014. Kusk dataset: Toward a direct understanding of recipe text and human cooking activity. In *Proceedings of the SixthInternational Workshop on Cooking and Eating Activities*.

Aurélie Herbelot and Eva Maria Vecchi. 2015. Building a shared world: mapping distributional to model-theoretic semantic spaces. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 22–32, Lisbon, Portugal, September. Association for Computational Linguistics.

Hirotaka Kameko, Shinsuke Mori, and Yoshimasa Tsuruoka. 2015a. Can symbol grounding improve low-level NLP? word segmentation as a case study. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2298–2303, Lisbon, Portugal, September. Association for Computational Linguistics.

Hirotaka Kameko, Shinsuke Mori, and Yoshimasa Tsuruoka. 2015b. Learning a game commentary generator with grounded move expressions. In *Proceedings of the 2015 IEEE Conference on Computational Intelligence and Games*.

Douwe Kiela and Stephen Clark. 2015. Multi- and cross-modal semantics beyond vision: Grounding in auditory perception. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2461–2470, Lisbon, Portugal, September. Association for Computational Linguistics.

Douwe Kiela, Ivan Vulić, and Stephen Clark. 2015. Visual bilingual lexicon induction with transferred convnet features. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 148–158.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, pages 282–289, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Angeliki Lazaridou, Nghia The Pham, and Marco Baroni. 2015. Combining language and vision with a multimodal skip-gram model. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 153–163, Denver, Colorado, May–June. Association for Computational Linguistics.

Alessandro Lopopolo and Emiel van Miltenburg. 2015. Sound-based distributional models. In *Proceedings of the 11th International Conference on Computational Semantics*, pages 70–75, London, UK, April. Association for Computational Linguistics.

Andrew McCallum and Wei Li. 2003. Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. In *Proceedings of the Seventh Conference on Computational Natural Language Learning*.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. In *Proceedings of workshop at the International Conference on Learning Representations (ICLR 2013)*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeffrey Dean. 2013b. Distributed representations of words and phrases and their composi-

tionality. *Advances in Neural Information Processing Systems*, 26:3111–3119.

Shinsuke Mori, John Richardson, Atsushi Ushiku, Tetsuro Sasada, Hirotaka Kameko, and Yoshimasa Tsuruoka. 2016. A Japanese chess commentary corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation*.

Graham Neubig, Makoto Morishita, and Satoshi Nakamura. 2015. Neural reranking improves subjective quality of machine translation: NAIST at WAT2015. In *Proceedings of the 2nd Workshop on Asian Translation (WAT2015)*, Kyoto, Japan, October.

J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Ng. 2011. Multimodal deep learning. In *Proceedings of the 28th International Conference on Machine Learning (ICML 2011)*, pages 689–696.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, October. Association for Computational Linguistics.

Arnau Ramisa, Josiah Wang, Ying Lu, Emmanuel Dellandrea, Francesc Moreno-Noguer, and Robert Gaizauskas. 2015. Combining geometric, textual and visual features for predicting prepositions in image descriptions. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 214–220.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the conll-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Computational Natural Language Learning*, pages 142–147.

Tetsuro Sasada, Shinsuke Mori, Tatsuya Kawahara, and Yoko Yamakata. 2015. Named entity recognizer trainable from partially annotated data. In *Proceedings of the Eleventh International Conference Pacific Association for Computational Linguistics*.

Burr Settles. 2004. Biomedical named entity recognition using conditional random fields and rich feature sets. In *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications*, pages 33–38.

Carina Silberer and Mirella Lapata. 2014. Learning grounded meaning representations with autoencoders. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 721–732, Baltimore, Maryland, June. Association for Computational Linguistics.

Richard Socher, Christopher D. Manning, and Andrew Y. Ng. 2010. Learning continuous phrase representations and syntactic parsing with recursive neural networks. In *Deep Learning and Unsupervised Feature Learning Workshop - NIPS 2010*.

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*, pages 1201–1211, Jeju Island, Korea, July. Association for Computational Linguistics.

Richard Socher, John Bauer, Christopher D. Manning, and Ng Andrew Y. 2013a. Parsing with compositional vector grammars. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 455–465, Sofia, Bulgaria, August. Association for Computational Linguistics.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013b. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA, October. Association for Computational Linguistics.

Nitish Srivastava and Ruslan R Salakhutdinov. 2012. Multimodal learning with deep boltzmann machines. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 2222–2230. Curran Associates, Inc.

Yuka Tateisi, Jin-Dong Kim, and Tomoko Ohta. 2002. The genia corpus: an annotated research abstract corpus in molecular biology domain. In *Proceedings of the HLT*, pages 73–77.

Yuta Tsuboi, Hisashi Kashima, Shinsuke Mori, Hiroki Oda, and Yuji Matsumoto. 2008. Training conditional random fields using incomplete annotations. In *Proceedings of the 22nd International Conference on Computational Linguistics*, pages 897–904.

Yuta Tsuboi. 2014. Neural networks leverage corpus-wide information for part-of-speech tagging. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 938–950, Doha, Qatar, October. Association for Computational Linguistics.

Yoshitaka Ushiku, Tatsuya Harada, and Yasuo Kuniyoshi. 2011. Automatic sentence generation from images. In *Proceedings of the 19th Annual ACM International Conference on Multimedia*, pages 1533–1536.

Yezhou Yang, Ching Lik Teo, Hal Daumé III, and Yiannis Aloimonos. 2011. Corpus-guided sentence generation of natural images. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*.