# STransE: a novel embedding model of entities and relationships in knowledge bases

**Dat Quoc Nguyen**[1], **Kairit Sirts**[1], **Lizhen Qu**[2] and **Mark Johnson**[1]

[1] Department of Computing, Macquarie University, Sydney, Australia
dat.nguyen@students.mq.edu.au, {kairit.sirts, mark.johnson}@mq.edu.au
[2] NICTA, ACT 2601, Australia
lizhen.qu@nicta.com.au

## Abstract

Knowledge bases of real-world facts about entities and their relationships are useful resources for a variety of natural language processing tasks. However, because knowledge bases are typically incomplete, it is useful to be able to perform *link prediction*, i.e., predict whether a relationship not in the knowledge base is likely to be true. This paper combines insights from several previous link prediction models into a new embedding model *STransE* that represents each entity as a low-dimensional vector, and each relation by two matrices and a translation vector. STransE is a simple combination of the SE and TransE models, but it obtains better link prediction performance on two benchmark datasets than previous embedding models. Thus, STransE can serve as a new baseline for the more complex models in the link prediction task.

## 1 Introduction

Knowledge bases (KBs), such as WordNet (Fellbaum, 1998), YAGO (Suchanek et al., 2007), Freebase (Bollacker et al., 2008) and DBpedia (Lehmann et al., 2015), represent relationships between entities as triples $(\mathrm{head\ entity}, \mathrm{relation}, \mathrm{tail\ entity})$. Even very large knowledge bases are still far from complete (Socher et al., 2013; West et al., 2014). *Link prediction* or *knowledge base completion* systems (Nickel et al., 2015) predict which triples not in a knowledge base are likely to be true (Taskar et al., 2004; Bordes et al., 2011). A variety of different kinds of information is potentially useful here, including information extracted from external corpora (Riedel et al., 2013; Wang et al., 2014a) and the other relationships that hold between the entities (Angeli and Manning, 2013; Zhao et al., 2015). For example, Toutanova et al. (2015) used information from the external ClueWeb-12 corpus to significantly enhance performance.

While integrating a wide variety of information sources can produce excellent results, there are several reasons for studying simpler models that directly optimize a score function for the triples in a knowledge base, such as the one presented here. First, additional information sources might not be available, e.g., for knowledge bases for specialized domains. Second, models that don't exploit external resources are simpler and thus typically much faster to train than the more complex models using additional information. Third, the more complex models that exploit external information are typically extensions of these simpler models, and are often initialized with parameters estimated by such simpler models, so improvements to the simpler models should yield corresponding improvements to the more complex models as well.

*Embedding models* for KB completion associate entities and/or relations with dense feature vectors or matrices. Such models obtain state-of-the-art performance (Nickel et al., 2011; Bordes et al., 2011; Bordes et al., 2012; Bordes et al., 2013; Socher et al., 2013; Wang et al., 2014b; Guu et al., 2015) and generalize to large KBs (Krompa et al., 2015). Table 1 summarizes a number of prominent embedding models for KB completion.

Let $(h, r, t)$ represent a triple. In all of the models

*Proceedings of NAACL-HLT 2016*, pages 460–466,
San Diego, California, June 12-17, 2016. ©2016 Association for Computational Linguistics

| **Model** | Score function $f_r(h,t)$ | **Opt.** |
|---|---|---|
| SE | $\|\mathbf{W}_{r,1}\mathbf{h} - \mathbf{W}_{r,2}\mathbf{t}\|_{\ell_{1/2}}$ ; $\mathbf{W}_{r,1}, \mathbf{W}_{r,2} \in \mathbb{R}^{k \times k}$ | SGD |
| Unstructured | $\|\mathbf{h} - \mathbf{t}\|_{\ell_{1/2}}$ | SGD |
| TransE | $\|\mathbf{h} + \mathbf{r} - \mathbf{t}\|_{\ell_{1/2}}$ ; $\mathbf{r} \in \mathbb{R}^k$ | SGD |
| DISTMULT | $\mathbf{h}^\top \mathbf{W}_r \mathbf{t}$ ; $\mathbf{W}_r$ is a diagonal matrix $\in \mathbb{R}^{k \times k}$ | AdaGrad |
| NTN | $\mathbf{u}_r^\top \, tanh(\mathbf{h}^\top \mathbf{M}_r \mathbf{t} + \mathbf{W}_{r,1}\mathbf{h} + \mathbf{W}_{r,2}\mathbf{t} + \mathbf{b}_r)$ ; $\mathbf{u}_r, \mathbf{b}_r \in \mathbb{R}^d$; $\mathbf{M}_r \in \mathbb{R}^{k \times k \times d}$; $\mathbf{W}_{r,1}, \mathbf{W}_{r,2} \in \mathbb{R}^{d \times k}$ | L-BFGS |
| TransH | $\|(\mathbf{I} - \mathbf{r}_p\mathbf{r}_p^\top)\mathbf{h} + \mathbf{r} - (\mathbf{I} - \mathbf{r}_p\mathbf{r}_p^\top)\mathbf{t}\|_{\ell_{1/2}}$ ; $\mathbf{r}_p, \mathbf{r} \in \mathbb{R}^k$ ; $\mathbf{I}$: Identity matrix size $k \times k$ | SGD |
| TransD | $\|(\mathbf{I} + \mathbf{r}_p\mathbf{h}_p^\top)\mathbf{h} + \mathbf{r} - (\mathbf{I} + \mathbf{r}_p\mathbf{t}_p^\top)\mathbf{t}\|_{\ell_{1/2}}$ ; $\mathbf{r}_p, \mathbf{r} \in \mathbb{R}^d$ ; $\mathbf{h}_p, \mathbf{t}_p \in \mathbb{R}^k$ ; $\mathbf{I}$: Identity matrix size $d \times k$ | AdaDelta |
| TransR | $\|\mathbf{W}_r\mathbf{h} + \mathbf{r} - \mathbf{W}_r\mathbf{t}\|_{\ell_{1/2}}$ ; $\mathbf{W}_r \in \mathbb{R}^{d \times k}$ ; $\mathbf{r} \in \mathbb{R}^d$ | SGD |
| Our STransE | $\|\mathbf{W}_{r,1}\mathbf{h} + \mathbf{r} - \mathbf{W}_{r,2}\mathbf{t}\|_{\ell_{1/2}}$ ; $\mathbf{W}_{r,1}, \mathbf{W}_{r,2} \in \mathbb{R}^{k \times k}$; $\mathbf{r} \in \mathbb{R}^k$ | SGD |

**Table 1:** The score functions $f_r(h,t)$ and the optimization methods (Opt.) of several prominent embedding models for KB completion. In all of these the entities $h$ and $t$ are represented by vectors $\mathbf{h}$ and $\mathbf{t} \in \mathbb{R}^k$ respectively.

discussed here, the head entity $h$ and the tail entity $t$ are represented by vectors $\mathbf{h}$ and $\mathbf{t} \in \mathbb{R}^k$ respectively. The *Unstructured* model (Bordes et al., 2012) assumes that $\mathbf{h} \approx \mathbf{t}$. As the Unstructured model does not take the relationship $r$ into account, it cannot distinguish different relation types. The *Structured Embedding* (SE) model (Bordes et al., 2011) extends the unstructured model by assuming that $h$ and $t$ are similar only in a relation-dependent subspace. It represents each relation $r$ with two matrices $\mathbf{W}_{r,1}$ and $\mathbf{W}_{r,2} \in \mathbb{R}^{k \times k}$, which are chosen so that $\mathbf{W}_{r,1}\mathbf{h} \approx \mathbf{W}_{r,2}\mathbf{t}$. The *TransE* model (Bordes et al., 2013) is inspired by models such as Word2Vec (Mikolov et al., 2013) where relationships between words often correspond to translations in latent feature space. The TransE model represents each relation $r$ by a translation vector $\mathbf{r} \in \mathbb{R}^k$, which is chosen so that $\mathbf{h} + \mathbf{r} \approx \mathbf{t}$.

The primary contribution of this paper is that two very simple relation-prediction models, SE and TransE, can be combined into a single model, which we call *STransE*. Specifically, we use relation-specific matrices $\mathbf{W}_{r,1}$ and $\mathbf{W}_{r,2}$ as in the SE model to identify the relation-dependent aspects of both $h$ and $t$, and use a vector $\mathbf{r}$ as in the TransE model to describe the relationship between $h$ and $t$ in this subspace. Specifically, our new KB completion model STransE chooses $\mathbf{W}_{r,1}$, $\mathbf{W}_{r,2}$ and $\mathbf{r}$ so that $\mathbf{W}_{r,1}\mathbf{h} + \mathbf{r} \approx \mathbf{W}_{r,2}\mathbf{t}$. That is, a TransE-style relationship holds in some relation-dependent subspace, and crucially, this subspace may involve very different projections of the head $h$ and tail $t$. So $\mathbf{W}_{r,1}$ and $\mathbf{W}_{r,2}$ can highlight, suppress, or even change the sign of, relation-specific attributes of $h$ and $t$. For example, for the "purchases" relationship, certain attributes of individuals $h$ (e.g., age, gender, marital status) are presumably strongly correlated with very different attributes of objects $t$ (e.g., sports car, washing machine and the like).

As we show below, STransE performs better than the SE and TransE models and other state-of-the-art link prediction models on two standard link prediction datasets WN18 and FB15k, so it can serve as a new baseline for KB completion. We expect that the STransE will also be able to serve as the basis for extended models that exploit a wider variety of information sources, just as TransE does.

## 2 Our approach

Let $\mathcal{E}$ denote the set of entities and $\mathcal{R}$ the set of relation types. For each triple $(h, r, t)$, where $h, t \in \mathcal{E}$ and $r \in \mathcal{R}$, the STransE model defines a *score function* $f_r(h, t)$ of its implausibility. Our goal is to choose $f$ such that the score $f_r(h, t)$ of a plausible triple $(h, r, t)$ is smaller than the score $f_{r'}(h', t')$ of an implausible triple $(h', r', t')$. We define the STransE score function $f$ as follows:

$$f_r(h,t) \quad = \quad \|\mathbf{W}_{r,1}\mathbf{h} + \mathbf{r} - \mathbf{W}_{r,2}\mathbf{t}\|_{\ell_{1/2}}$$

using either the $\ell_1$ or the $\ell_2$-norm (the choice is made using validation data; in our experiments we found that the $\ell_1$ norm gave slightly better results). To learn the vectors and matrices we minimize the following margin-based objective function:

$$\mathcal{L} \;=\; \sum_{\substack{(h,r,t)\in\mathcal{G} \\ (h',r,t')\in\mathcal{G}'_{(h,r,t)}}} [\gamma + f_r(h,t) - f_r(h',t')]_+$$

where $[x]_+ = \max(0, x)$, $\gamma$ is the margin hyperparameter, $\mathcal{G}$ is the training set consisting of correct triples, and $\mathcal{G}'_{(h,r,t)} = \{(h', r, t) \mid h' \in \mathcal{E}, (h', r, t) \notin \mathcal{G}\} \cup \{(h, r, t') \mid t' \in \mathcal{E}, (h, r, t') \notin \mathcal{G}\}$ is the set of incorrect triples generated by corrupting a correct triple $(h, r, t) \in \mathcal{G}$.

We use Stochastic Gradient Descent (SGD) to minimize $\mathcal{L}$, and impose the following constraints during training: $\|\mathbf{h}\|_2 \leqslant 1$, $\|\mathbf{r}\|_2 \leqslant 1$, $\|\mathbf{t}\|_2 \leqslant 1$, $\|\mathbf{W}_{r,1}\mathbf{h}\|_2 \leqslant 1$ and $\|\mathbf{W}_{r,2}\mathbf{t}\|_2 \leqslant 1$.

## 3 Related work

Table 1 summarizes related embedding models for link prediction and KB completion. The models differ in the score functions $f_r(h, t)$ and the algorithms used to optimize the margin-based objective function, e.g., SGD, AdaGrad (Duchi et al., 2011), AdaDelta (Zeiler, 2012) and L-BFGS (Liu and Nocedal, 1989).

DISTMULT (Yang et al., 2015) is based on a Bilinear model (Nickel et al., 2011; Bordes et al., 2012; Jenatton et al., 2012) where each relation is represented by a diagonal rather than a full matrix. The neural tensor network (NTN) model (Socher et al., 2013) uses a bilinear tensor operator to represent each relation. Similar quadratic forms are used to model entities and relations in KG2E (He et al., 2015) and TATEC (Garcia-Duran et al., 2015b).

The TransH model (Wang et al., 2014b) associates each relation with a relation-specific hyperplane and uses a projection vector to project entity vectors onto that hyperplane. TransD (Ji et al., 2015) and TransR/CTransR (Lin et al., 2015b) extend the TransH model using two projection vectors and a matrix to project entity vectors into a relation-specific space, respectively. TransD learns a relation-role specific mapping just as STransE, but represents this mapping by projection vectors rather than full matrices, as in STransE. Thus STransE can be viewed as an extension of the TransR model, where head and tail entities are associated with their own project matrices, rather than using the same matrix for both, as in TransR and CTransR.

| Dataset | #E | #R | #Train | #Valid | #Test |
|---|---|---|---|---|---|
| WN18 | 40,943 | 18 | 141,442 | 5,000 | 5,000 |
| FB15k | 14,951 | 1,345 | 483,142 | 50,000 | 59,071 |

**Table 2:** Statistics of the experimental datasets used in this study (and previous works). #E is the number of entities, #R is the number of relation types, and #Train, #Valid and #Test are the numbers of triples in the training, validation and test sets, respectively.

Recently, Lao et al. (2011), Neelakantan et al. (2015), Gardner and Mitchell (2015), Luo et al. (2015), Lin et al. (2015a), Garcia-Duran et al. (2015a) and Guu et al. (2015) showed that relation paths between entities in KBs provide richer information and improve the relationship prediction. Nickel et al. (2015) reviews other approaches for learning from KBs and multi-relational data.

## 4 Experiments

For link prediction evaluation, we conduct experiments and compare the performance of our STransE model with published results on the benchmark WN18 and FB15k datasets (Bordes et al., 2013). Information about these datasets is given in Table 2.

### 4.1 Task and evaluation protocol

The link prediction task (Bordes et al., 2011; Bordes et al., 2012; Bordes et al., 2013) predicts the head or tail entity given the relation type and the other entity, i.e. predicting $h$ given $(?, r, t)$ or predicting $t$ given $(h, r, ?)$ where ? denotes the missing element. The results are evaluated using the ranking induced by the score function $f_r(h, t)$ on test triples.

For each test triple $(h, r, t)$, we corrupted it by replacing either $h$ or $t$ by each of the possible entities in turn, and then rank these candidates in ascending order of their implausibility value computed by the score function. Following the protocol described in Bordes et al. (2013), we remove any corrupted triples that appear in the knowledge base, to avoid cases where a correct corrupted triple might be ranked higher than the test triple. We report the mean rank and the Hits@10 (i.e., the proportion of test triples in which the target entity was ranked in the top 10 predictions) for each model. Lower mean rank or higher Hits@10 indicates better link prediction performance.

Following TransR/CTransR (Lin et al., 2015b), TransD (Ji et al., 2015), TATEC (Garcia-Duran et al., 2015b), RTransE (Garcia-Duran et al., 2015a) and PTransE (Lin et al., 2015a), we used the entity and relation vectors produced by TransE (Bordes et al., 2013) to initialize the entity and relation vectors in STransE, and we initialized the relation matrices with identity matrices. Following Wang et al. (2014b), Lin et al. (2015b), He et al. (2015), Ji et al. (2015) and Lin et al. (2015a), we applied the "*Bernoulli*" trick for generating head or tail entities when sampling incorrect triples. We ran SGD for 2,000 epochs to estimate the model parameters. Following Bordes et al. (2013) we used a grid search on validation set to choose either the $l_1$ or $l_2$ norm in the score function $f$, as well as to set the SGD learning rate $\lambda \in \{0.0001, 0.0005, 0.001, 0.005, 0.01\}$, the margin hyper-parameter $\gamma \in \{1, 3, 5\}$ and the number of vector dimensions $k \in \{50, 100\}$. The lowest mean rank on the validation set was obtained when using the $l_1$ norm in $f$ on both WN18 and FB15k, and when $\lambda = 0.0005, \gamma = 5$, and $k = 50$ for WN18, and $\lambda = 0.0001, \gamma = 1$, and $k = 100$ for FB15k.

## 4.2 Main results

Table 3 compares the link prediction results of our STransE model with results reported in prior work, using the same experimental setup. The first twelve rows report the performance of models that do not exploit information about alternative paths between head and tail entities. The next two rows report results of the RTransE and PTransE models, which are extensions of the TransE model that exploit information about relation paths. The last row presents results for the log-linear model Node+LinkFeat (Toutanova and Chen, 2015) which makes use of textual mentions derived from the large external ClueWeb-12 corpus.

It is clear that Node+LinkFeat with the additional external corpus information obtained best results. In future work we plan to extend the STransE model to incorporate such additional information. Table 3 also shows that models RTransE and PTransE employing path information achieve better results than models that do not use such information. In terms of models not exploiting path information or external information, the STransE model scores better than the other models on WN18 and produces the highest Hits@10 score on FB15k. Compared to the closely related models SE, TransE, TransR, CTransR and TransD, STransE does better than these models on both WN18 and FB15k.

| Method | WN18 | | FB15k | |
|---|---|---|---|---|
| | MR | H10 | MR | H10 |
| SE (Bordes et al., 2011) | 985 | 80.5 | 162 | 39.8 |
| Unstructured (Bordes et al., 2012) | 304 | 38.2 | 979 | 6.3 |
| TransE (Bordes et al., 2013) | 251 | 89.2 | 125 | 47.1 |
| TransH (Wang et al., 2014b) | 303 | 86.7 | 87 | 64.4 |
| TransR (Lin et al., 2015b) | 225 | 92.0 | 77 | 68.7 |
| CTransR (Lin et al., 2015b) | 218 | 92.3 | 75 | 70.2 |
| KG2E (He et al., 2015) | 348 | 93.2 | 59 | 74.0 |
| TransD (Ji et al., 2015) | 212 | 92.2 | 91 | 77.3 |
| TATEC (Garcia-Duran et al., 2015b) | - | - | **58** | 76.7 |
| NTN (Socher et al., 2013) | - | $66.1^+$ | - | $41.4^+$ |
| DISTMULT (Yang et al., 2015) | - | $94.2^+$ | - | $57.7^+$ |
| Our STransE model | **206** | **93.4** | 69 | **79.7** |
| RTransE (Garcia-Duran et al., 2015a) | - | - | **50** | 76.2 |
| PTransE (Lin et al., 2015a) | - | - | 58 | 84.6 |
| NLFeat (Toutanova and Chen, 2015) | - | **94.3** | - | **87.0** |

**Table 3:** Link prediction results. MR and H10 denote evaluation metrics of mean rank and Hits@10 (in %), respectively. "NLFeat" abbreviates Node+LinkFeat. The results for NTN (Socher et al., 2013) listed in this table are taken from Yang et al. (2015) since NTN was originally evaluated on different datasets. The results marked with $^+$ are obtained using the optimal hyper-parameters chosen to optimize Hits@10 on the validation set; trained in this manner, STransE obtains a mean rank of 244 and Hits@10 of **94.7**% on WN18, while producing the same results on FB15k.

Following Bordes et al. (2013), Table 4 analyzes Hits@10 results on FB15k with respect to the relation categories defined as follows: for each relation type $r$, we computed the averaged number $a_h$ of heads $h$ for a pair $(r, t)$ and the averaged number $a_t$ of tails $t$ for a pair $(h, r)$. If $a_h < 1.5$ and $a_t < 1.5$, then $r$ is labeled **1-1**. If $a_h \geq 1.5$ and $a_t < 1.5$, then $r$ is labeled **M-1**. If $a_h < 1.5$ and $a_t \geq 1.5$, then $r$ is labeled as **1-M**. If $a_h \geq 1.5$ and $a_t \geq 1.5$, then $r$ is labeled as **M-M**. 1.4%, 8.9%, 14.6% and 75.1% of the test triples belong to a relation type classified as **1-1**, **1-M**, **M-1** and **M-M**, respectively.

Table 4 shows that in comparison to prior models not using path information, STransE obtains highest Hits@10 result for **M-M** relation category at $(80.1\% + 83.1\%)/2 = 81.6\%$. In addition, STransE

| Method | Predicting head $h$ | | | | Predicting tail $t$ | | | |
|---|---|---|---|---|---|---|---|---|
| | 1-1 | 1-M | M-1 | M-M | 1-1 | 1-M | M-1 | M-M |
| SE | 35.6 | 62.6 | 17.2 | 37.5 | 34.9 | 14.6 | 68.3 | 41.3 |
| Unstr. | 34.5 | 2.5 | 6.1 | 6.6 | 34.3 | 4.2 | 1.9 | 6.6 |
| TransE | 43.7 | 65.7 | 18.2 | 47.2 | 43.7 | 19.7 | 66.7 | 50.0 |
| TransH | 66.8 | 87.6 | 28.7 | 64.5 | 65.5 | 39.8 | 83.3 | 67.2 |
| TransR | 78.8 | 89.2 | 34.1 | 69.2 | 79.2 | 37.4 | 90.4 | 72.1 |
| CTransR | 81.5 | 89.0 | 34.7 | 71.2 | 80.8 | 38.6 | 90.1 | 73.8 |
| KG2E | **92.3** | 94.6 | **66.0** | 69.6 | **92.6** | **67.9** | **94.4** | 73.4 |
| TransD | 86.1 | **95.5** | 39.8 | 78.5 | 85.4 | 50.6 | **94.4** | 81.2 |
| TATEC | 79.3 | 93.2 | 42.3 | 77.2 | 78.5 | 51.5 | 92.7 | 80.7 |
| STransE | 82.8 | 94.2 | 50.4 | **80.1** | 82.4 | 56.9 | 93.4 | **83.1** |

**Table 4:** Hits@10 (in %) by the relation category on FB15k. “Unstr.” abbreviates Unstructured.

also performs better than TransD for **1-M** and **M-1** relation categories. We believe the improved performance of the STransE model is due to its use of full matrices, rather than just projection vectors as in TransD. This permits STransE to model diverse and complex relation categories (such as **1-M**, **M-1** and especially **M-M**) better than TransD and other similiar models. However, STransE is not as good as TransD for the **1-1** relations. Perhaps the extra parameters in STransE hurt performance in this case (note that 1-1 relations are relatively rare, so STransE does better overall).

## 5 Conclusion and future work

This paper presented a new embedding model for link prediction and KB completion. Our STransE combines insights from several simpler embedding models, specifically the Structured Embedding model (Bordes et al., 2011) and the TransE model (Bordes et al., 2013), by using a low-dimensional vector and two projection matrices to represent each relation. STransE, while being conceptually simple, produces highly competitive results on standard link prediction evaluations, and scores better than the embedding-based models it builds on. Thus it is a suitable candidate for serving as future baseline for more complex models in the link prediction task.

In future work we plan to extend STransE to exploit relation path information in knowledge bases, in a manner similar to Lin et al. (2015a), Garcia-Duran et al. (2015a) or Guu et al. (2015).

## Acknowledgments

This research was supported by a Google award through the Natural Language Understanding Focused Program, and under the Australian Research Council’s *Discovery Projects* funding scheme (project number DP160102156).

NICTA is funded by the Australian Government through the Department of Communications and the Australian Research Council through the ICT Centre of Excellence Program. The first author is supported by an International Postgraduate Research Scholarship and a NICTA NRPA Top-Up Scholarship.

## References

Gabor Angeli and Christopher Manning. 2013. Philosophers are Mortal: Inferring the Truth of Unseen Facts. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 133–142.

Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: A Collaboratively Created Graph Database for Structuring Human Knowledge. In *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data*, pages 1247–1250.

Antoine Bordes, Jason Weston, Ronan Collobert, and Yoshua Bengio. 2011. Learning Structured Embeddings of Knowledge Bases. In *Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence*, pages 301–306.

Antoine Bordes, Xavier Glorot, Jason Weston, and Yoshua Bengio. 2012. A Semantic Matching Energy Function for Learning with Multi-relational Data. *Machine Learning*, 94(2):233–259.

Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. 2013. Translating Embeddings for Modeling Multi-relational Data. In *Advances in Neural Information Processing Systems 26*, pages 2787–2795.

John Duchi, Elad Hazan, and Yoram Singer. 2011. Adaptive Subgradient Methods for Online Learning and Stochastic Optimization. *The Journal of Machine Learning Research*, 12:2121–2159.

Christiane D. Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. MIT Press.

Alberto Garcia-Duran, Antoine Bordes, and Nicolas Usunier. 2015a. Composing Relationships with Translations. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 286–290.

Alberto Garcia-Duran, Antoine Bordes, Nicolas Usunier, and Yves Grandvalet. 2015b. Combining Two And Three-Way Embeddings Models for Link Prediction in Knowledge Bases. *CoRR*, abs/1506.00999.

Matt Gardner and Tom Mitchell. 2015. Efficient and Expressive Knowledge Base Completion Using Subgraph Feature Extraction. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1488–1498.

Kelvin Guu, John Miller, and Percy Liang. 2015. Traversing Knowledge Graphs in Vector Space. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 318–327.

Shizhu He, Kang Liu, Guoliang Ji, and Jun Zhao. 2015. Learning to Represent Knowledge Graphs with Gaussian Embedding. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, pages 623–632.

Rodolphe Jenatton, Nicolas L. Roux, Antoine Bordes, and Guillaume R Obozinski. 2012. A latent factor model for highly multi-relational data. In *Advances in Neural Information Processing Systems 25*, pages 3167–3175.

Guoliang Ji, Shizhu He, Liheng Xu, Kang Liu, and Jun Zhao. 2015. Knowledge Graph Embedding via Dynamic Mapping Matrix. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 687–696.

Denis Krompa, Stephan Baier, and Volker Tresp. 2015. Type-Constrained Representation Learning in Knowledge Graphs. In *Proceedings of the 14th International Semantic Web Conference*, pages 640–655.

Ni Lao, Tom Mitchell, and William W. Cohen. 2011. Random Walk Inference and Learning in a Large Scale Knowledge Base. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 529–539.

Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N. Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick van Kleef, Sören Auer, and Christian Bizer. 2015. DBpedia - A Large-scale, Multilingual Knowledge Base Extracted from Wikipedia. *Semantic Web*, 6(2):167–195.

Yankai Lin, Zhiyuan Liu, Huanbo Luan, Maosong Sun, Siwei Rao, and Song Liu. 2015a. Modeling Relation Paths for Representation Learning of Knowledge Bases. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 705–714.

Yankai Lin, Zhiyuan Liu, Maosong Sun, Yang Liu, and Xuan Zhu. 2015b. Learning Entity and Relation Embeddings for Knowledge Graph Completion. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence Learning*, pages 2181–2187.

D. C. Liu and J. Nocedal. 1989. On the Limited Memory BFGS Method for Large Scale Optimization. *Mathematical Programming*, 45(3):503–528.

Yuanfei Luo, Quan Wang, Bin Wang, and Li Guo. 2015. Context-Dependent Knowledge Graph Embedding. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1656–1661.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013. Linguistic Regularities in Continuous Space Word Representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 746–751.

Arvind Neelakantan, Benjamin Roth, and Andrew McCallum. 2015. Compositional Vector Space Models for Knowledge Base Completion. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 156–166.

Maximilian Nickel, Volker Tresp, and Hans-Peter Kriegel. 2011. A Three-Way Model for Collective Learning on Multi-Relational Data. In *Proceedings of the 28th International Conference on Machine Learning*, pages 809–816.

Maximilian Nickel, Kevin Murphy, Volker Tresp, and Evgeniy Gabrilovich. 2015. A Review of Relational Machine Learning for Knowledge Graphs. *Proceedings of the IEEE, to appear*.

Sebastian Riedel, Limin Yao, Andrew McCallum, and Benjamin M. Marlin. 2013. Relation Extraction with Matrix Factorization and Universal Schemas. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 74–84.

Richard Socher, Danqi Chen, Christopher D Manning, and Andrew Ng. 2013. Reasoning With Neural Tensor Networks for Knowledge Base Completion. In *Advances in Neural Information Processing Systems 26*, pages 926–934.

Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum. 2007. YAGO: A Core of Semantic Knowledge. In *Proceedings of the 16th International Conference on World Wide Web*, pages 697–706.

Ben Taskar, Ming fai Wong, Pieter Abbeel, and Daphne Koller. 2004. Link Prediction in Relational Data. In

*Advances in Neural Information Processing Systems 16*, pages 659–666.

Kristina Toutanova and Danqi Chen. 2015. Observed Versus Latent Features for Knowledge Base and Text Inference. In *Proceedings of the 3rd Workshop on Continuous Vector Space Models and their Compositionality*, pages 57–66.

Kristina Toutanova, Danqi Chen, Patrick Pantel, Hoifung Poon, Pallavi Choudhury, and Michael Gamon. 2015. Representing Text for Joint Embedding of Text and Knowledge Bases. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1499–1509.

Zhen Wang, Jianwen Zhang, Jianlin Feng, and Zheng Chen. 2014a. Knowledge Graph and Text Jointly Embedding. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1591–1601.

Zhen Wang, Jianwen Zhang, Jianlin Feng, and Zheng Chen. 2014b. Knowledge Graph Embedding by Translating on Hyperplanes. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence*, pages 1112–1119.

Robert West, Evgeniy Gabrilovich, Kevin Murphy, Shaohua Sun, Rahul Gupta, and Dekang Lin. 2014. Knowledge Base Completion via Search-based Question Answering. In *Proceedings of the 23rd International Conference on World Wide Web*, pages 515–526.

Bishan Yang, Wen-tau Yih, Xiaodong He, Jianfeng Gao, and Li Deng. 2015. Embedding Entities and Relations for Learning and Inference in Knowledge Bases. In *Proceedings of the International Conference on Learning Representations*.

Matthew D. Zeiler. 2012. ADADELTA: An Adaptive Learning Rate Method. *CoRR*, abs/1212.5701.

Yu Zhao, Sheng Gao, Patrick Gallinari, and Jun Guo. 2015. Knowledge Base Completion by Learning Pairwise-Interaction Differentiated Embeddings. *Data Mining and Knowledge Discovery*, 29(5):1486–1504.

NAACL HLT 2016

# The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies

**Proceedings of the Conference**

June 12-17, 2016
San Diego, California, USA

Logo design by Mark Klingsporn

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)
209 N. Eighth Street
Stroudsburg, PA 18360
USA
Tel: +1-570-476-8006
Fax: +1-570-476-0860
acl@aclweb.org

ISBN 978-1-941643-91-4

# Message from the General Chair

Greetings,

Welcome to NAACL HLT 2016! This year's conference is held in San Diego, California, where we have assembled an exciting program of computational linguistics research.

The main program features a wide array of topics, and it includes excellent invited talks by Prof. Regina Barzilay and Prof. Ehud Reiter. In addition, we have six tutorials on the day before the main program, plus fifteen workshops on the following two days. Some of these workshops are back for their 10th or 11th incarnation, while others are brand-new. In parallel, we have a live demonstration track, and a Student Research Workshop that showcases work by the junior members of our research community.

This NAACL HLT meeting takes place only through the hard work of many people who deserve our gratitude.

Thanks to Priscilla Rasmussen for making local arrangements, handling registration, setting up social events, writing visa invitation letters, and solving a myriad of issues. Priscilla, your experience is a great asset to any conference!

The NAACL HLT organizing committee took all the steps to bring you a great conference. Many thanks to Ani Nenkova and Owen Rambow (Program Co-chairs), Mohit Bansal and Alexander M. Rush (Tutorial Co-chairs), Radu Soricut and Adrià de Gispert (Workshop Co-chairs), Jacob Andreas, Eunsol Choi, and Angeliki Lazaridou (Student Research Workshop Co-Chairs) and their faculty advisors Jacob Eisenstein and Nianwen Xue, Aliya Deri (Student Volunteer Coordinator), Julie Medero (Local Sponsorship Chair), Mark Finlayson, Sravana Reddy, and John DeNero (Demonstration Co-chairs), Adam Lopez and Margaret Mitchell (Publications Co-chairs), Jason Riesa (Website Chair), Wei Xu (Publicity Chair), and Jonathan May (Social Media Chair).

Thanks also to the NAACL Board for providing excellent advice, and thanks to previous chairs for their suggestions and timelines.

Sponsors of NAACL HLT 2016 include Baidu and Google (Platinum Sponsors), Amazon, Bloomberg, eBay, Microsoft Research, and UnitedHealth Group (Gold Sponsors), Huawei (Silver Sponsors), Civis Analytics, Facebook, @newsela, and Nuance (Bronze Sponsors), and the University of Washington (Supporter). Thanks for your extremely valuable contributions!

Finally, thanks to the scientists, engineers, authors, and attendees who come to share and learn at this leading venue for computational linguistics research!

Kevin Knight
Information Sciences Institute, University of Southern California
NAACL HLT 2016 General Chair

# Message from the Program Co-Chairs

Welcome to San Diego for the 15th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies!

The conference has grown remarkably in the past five years: we had 698 submissions this year, despite our deadline right after the end-of-the-year holidays. As we worked on organizing the conference program, we made many changes to reflect the growth of the NAACL community, the increasing diversity of topics covered by the field, and the acceleration of the pace of the publication cycle.

We had a record short time between paper submission and author notification—less than two months. We settled on such compressed timeline in order to avoid spreading the reviewing period over the winter holidays, to ensure that papers spend only a short time under submission, and to coordinate submission deadlines with ACL. Our incredible team of area chairs and reviewers ensured that the planned schedule went smoothly.

As the computational linguistics field has expanded, it has become increasingly difficult to recruit a sufficient number of knowledgeable reviewers. We decided to reach out to the largest possible pool of computational linguists and provide convenient ways for the area chairs to control which reviewers they end up working with: we invited all researchers actively working in the area of computational linguistics/language processing to review for the conference. We defined "active researchers" to be those who have published at least five papers in the last ten years in the ACL, NAACL, EMNLP, EACL or COLING conferences. In order to be inclusive of the amazing young researchers who became active in the field only more recently, we also included everyone who had published at least three papers in the same venues for the last five years. This yielded a list of over 1,400 researchers that we invited to serve as reviewers for the conference. Of these, 685 agreed and participated in the review process. This is another record for NAACL HLT 2016, no previous NAACL has had such a large program committee. Among these, the area chairs recognized 120 as best reviewers.

Working with the reviewers were the 42 area chairs. We asked the area chairs to work in pairs, so they can have a back-up in case other obligations need their attention during the review period and to ensure that all decisions about reviewer assignment and paper recommendation are discussed in detail. All area chairs and reviewers submitted a list of keywords that describe their area of expertise (the full list appears in the conference call for papers). The area chairs were paired based on the keyword overlap.

To match reviewers to area chairs, we used a bidding system. For bidding, each area received a list of the 140 reviewers with best matching keyword profiles. If the area chairs did not know the work of a potential reviewer on their bidding list, they looked him or her up on DBLP or Google Scholar before making their final bid. Areas were assigned only reviewers for which the area chairs bid positively. Area chairs were free as usual to recruit additional reviewers they wished to work with.

Submissions were assigned to areas by taking into account the match between the paper keywords and the area chair keywords. Areas were capped at 40 submissions maximum (long and short combined). As in the past, reviewers bid on papers they wanted to review. 69% of the reviews were written by reviewers who had bid indicating that they want to review the paper; 29% of the reviews were written by reviewers who had bid indicating they are ok with reviewing the paper. The remaining 2% of reviews were written by reviewers who did not bid on the paper but were asked by an area chair to review it. Three reviewers were assigned a paper that they did not want to review according to their bid. The average reviewer load was 3 papers, which included a mix of long and short submissions. Only 43 reviewers had more than four papers to review.

Area chairs wrote meta-reviews, for use only by us, justifying their accept/reject recommendation. In making difficult decisions, we drew on these meta-reviews, the reviews themselves, the discussion among the reviewers, and the author response to the initial reviews.

We are happy with our changes to the review process: area chairs had control over the reviewers they worked with, reviewers were assigned papers they wanted to review and the overall reviewing load was low. Needless to say, there is room for further improvements. The reviewing process is crucial to the quality of this conference; only if the community has confidence in the quality of the reviewing process will this conference continue to be a leading conference in our field. Our goal has been to make sure that every single submission receives a complete and fair review and decision, and to make sure that the authors of every single submission understand why their paper was accepted or declined for the conference. We would like to thank our 685 reviewers, and we would especially like to thank our 42 area chairs, who were patient in allowing us to pursue some of the innovative aspects of this year's reviewing cycle.

Eighteen of the 698 initial submissions were withdrawn by the authors or rejected without review because of formatting violations. A total of 396 long and 284 short papers underwent review; 100 long and 82 short papers were accepted, for an acceptance rate of 25% and 29% respectively. In addition, ten TACL papers will be presented at the conference.

This year we decided to have shorter slots for oral presentations, in order to have more of the accepted papers presented as talks. In the program, long papers are allotted 20-minute slots (15 min presentation + 5 min questions). Short papers are allotted 10-minute slots (6 min presentation + 4 min questions).

The best paper award committee consisted of NAACL general and program chairs from the last three years. Not all past chairs could participate in the selection. The final best paper committee included Joyce Chai, Katrin Kirchhoff, Rada Mihalcea, Kristina Toutanova, Lucy Vanderwende and Hua Wu. They selected two best long papers and one best short paper, along with two runner-ups in each category.

**Best Short Paper**
*Improving sentence compression by learning to predict gaze*
Sigrid Klerke, Yoav Goldberg and Anders Søgaard

**Short Paper, Runners Up**
*Patterns of Wisdom: Discourse-Level Style in Multi-Sentence Quotations*
Kyle Booten and Marti A. Hearst

*A Joint Model of Orthography and Morphological Segmentation*
Ryan Cotterell, Tim Vieira and Hinrich Schütze

**Best Long Papers**
*Feuding Families and Former Friends: Unsupervised Learning for Dynamic Fictional Relationships*
Mohit Iyyer, Anupam Guha, Snigdha Chaturvedi, Jordan Boyd-Graber and Hal Daumé III

*Learning to Compose Neural Networks for Question Answering*
Jacob Andreas, Marcus Rohrbach, Trevor Darrell and Dan Klein

**Long Paper, Runners Up**
*Multi-way, Multilingual Neural Machine Translation with a Shared Attention Mechanism*
Orhan Firat, Kyunghyun Cho and Yoshua Bengio

*Black Holes and White Rabbits: Metaphor Identification with Visual Features*
Ekaterina Shutova, Douwe Kiela and Jean Maillard

The conference program includes two inspiring invited talks by Regina Barzilay and Ehud Reiter. Both push the boundaries of the field, discussing the potential for real-world impact of language technologies.

Finally we would like to thank all other people who supported us in the past year in our work for NAACL HLT 2016. Last year's program chairs, Anoop Sarkar and Joyce Chai shared their valuable advice and promptly answered the many questions we had throughout the process. The NAACL board chair for 2015 (Hal Daumé III) and 2016 (Emily Bender) were our effective link with the NAACL board. The conference general chair, Kevin Knight, was always available to us when we needed to consult about decisions we were making. The conference business manager, Priscilla Rasmussen, gave us details about the venue and coordinated with us at the final stages of making the conference schedule. The ACL treasurer, Greame Hirst, answered questions about the venue. The conference webmaster, Jason Riesa, put content on the conference webpage as soon as we made it available to him. The publication chairs, Meg Mitchell and Adam Lopez, answered all lingering author questions about formatting for submission and final versions. Many talks to all of them!

We look forward to an exciting conference!

NAACL HLT 2016 Program Co-Chairs
Ani Nenkova, University of Pennsylvania
Owen Rambow, Columbia University