Integrating Lexical Information into Entity Neighbourhood Representations for Relation Prediction

Ian David Wood
Macquarie University
and CSIRO, Australia
ian.wood@mq.edu.au

Stephen Wan
CSIRO, Australia
Stephen.Wan@
data61.csiro.au

Mark Johnson
Oracle Digital Assistant
mark.mj.johnson@
oracle.com

Abstract

Relation prediction informed from a combination of text corpora and curated knowledge bases, combining knowledge graph completion with relation extraction, is a relatively little studied task. A system that can perform this task has the ability to extend an arbitrary set of relational database tables with information extracted from a document corpus. OpenKi (Zhang et al., 2019) addresses this task through extraction of named entities and predicates via OpenIE tools then learning relation embeddings from the resulting entity-relation graph for relation prediction, outperforming previous approaches. We present an extension of OpenKi that incorporates embeddings of text-based representations of the entities and the relations. We demonstrate that this results in a substantial performance increase over a system without this information.
https://github.com/drevicko/OpenKI

1 Introduction

Curated knowledge repositories such as knowledge bases and relational databases provide powerful tools for many practical knowledge related tasks. They require, however, substantial effort to create and maintain. Many applications deal with knowledge that is continuously changing, presenting prohibitive maintenance costs and limiting the utility of explicit knowledge representation technologies. The new knowledge is often available in text based formats such as reports, news items and memos. In this work, we use the term "proposition" to describe a triple (e_1, r, e_2) that indicates that a relation r holds between two entities e_1 and e_2.

Work in the field has largely focussed on either extracting propositions directly from text or inferring missing propositions by examining knowledge graphs. What we are interested in here combines the two in a single model, utilising information from the knowledge base and collections of text together to infer relations, both mentioned in the text and implied by the text in combination with existing knowledge.

Previous work following this approach draws on patterns in the curated knowledge graph in combination with the graph of entity mentions in texts, allowing prediction of new knowledge base relations (Riedel et al., 2013; Verga et al., 2015, 2017). Zhang et al. (2019) extend this work by incorporating text predicates connecting entity mentions extracted using OpenIE tools (Fader et al., 2011; Lockard et al., 2019) and introducing the concept of "entity neighbourhoods" consisting of the binary OpenIE predicates and knowledge base relations[1] that occur with a given entity as their subject or object. Drawing on the success of text based representations incorporated into entity recognition tasks (Gillick et al., 2019), we extend Zhang et.al.'s model by incorporating text based embeddings of entities and relations into the entity neighbourhood representations. Texts are drawn from knowledge base metadata and occurrences in source texts. We use fasttext (Mikolov et al., 2018) word embeddings and BERT (Devlin et al., 2018) to obtain text embeddings. The resulting models achieve state of the art results on two knowledge base extension data sets.

2 Related Work

Open information extraction (OpenIE) attempts to find relations expressed in collections of texts through identification of entity and relation spans (Fader et al., 2011; Stanovsky et al., 2018). Our work can be taken as an approach to incorporate this extracted information into an existing knowledge base.

Relation extraction, the identification of relations expressed in text between given entity mentions, has received much attention in recent years

[1] we refer to relations from a knowledge base as "relations" or "KB relations" and predicates extracted from text as "predicates" or "text predicates".

Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 3429–3436
June 6–11, 2021. ©2021 Association for Computational Linguistics

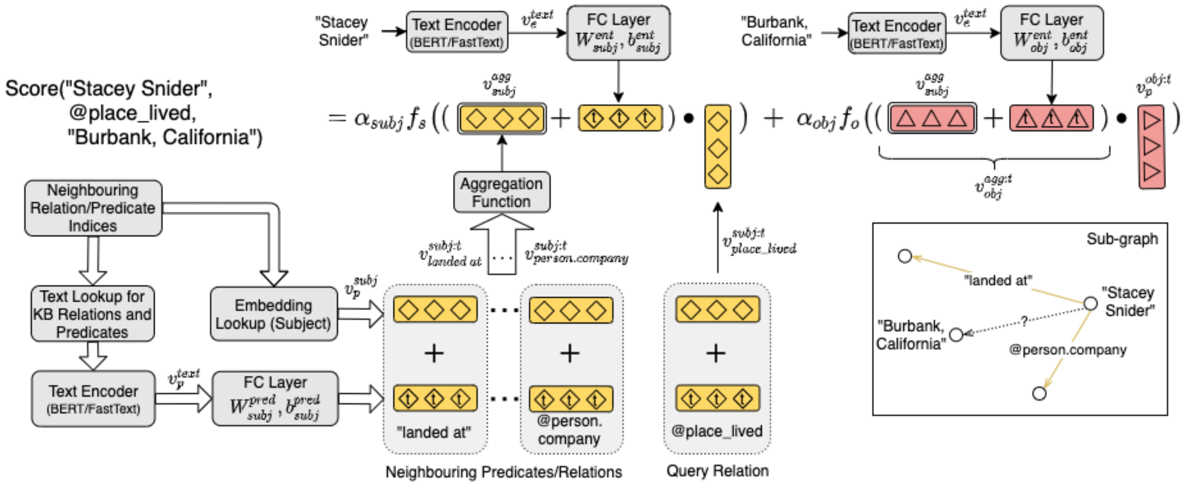
"Stacey Snider"
Text Encoder (BERT/FastText)
v_e^{text}
FC Layer $W_{subj}^{ent}, b_{subj}^{ent}$
"Burbank, California"
Text Encoder (BERT/FastText)
v_e^{text}
FC Layer $W_{obj}^{ent}; b_{obj}^{ent}$
Score("Stacey Snider", @place_lived, "Burbank, California")
$= \alpha_{subj} f_s((\; + \;) \bullet \;) + \alpha_{obj} f_o((\; + \;) \bullet \;)$
v_{subj}^{agg}
v_{subj}^{agg}
$v_{obj}^{agg:t}$
$v_p^{obj:t}$
Aggregation Function
Neighbouring Relation/Predicate Indices
Text Lookup for KB Relations and Predicates
Embedding Lookup (Subject)
v_p^{subj}
Text Encoder (BERT/FastText)
v_p^{text}
FC Layer $W_{subj}^{pred}, b_{subj}^{pred}$
$v_{landed\ at}^{subj:t}$
$v_{person.company}^{subj:t}$
$v_{place\_lived}^{subj:t}$
"landed at"
@person. company
@place_lived
Neighbouring Predicates/Relations
Query Relation
Sub-graph
"landed at"
"Stacey Snider"
"Burbank, California"
?
@person.company

Figure 1: Overview of text enhanced ENE model. The box bottom-right represents a relevant portion of the graph of predicate and KB propositions. In this example, we consider `@place_lived` as a candidate KB relation between entities Stacey Snider and Burbank, California. KB relation `@person.company` and predicate `"landed at"` are among those that have a triple in the data with Stacey Snider as subject. Their embeddings v_p^{subj} (yellow diamonds) contribute to her subject neighbour representation. We enhance this representation with encodings of text forms (diamonds with "t") of the respective predicates and KB relations (eqn. 1). The aggregate of the neighbour representations v_{subj}^{agg} (eqn. 2) is further enhanced with an encoding of a text representation (either the name itself or the name and a description from the KB) of the entity Stacey Snider (eqn. 3). Details for object entity representations (v_{obj}^{agg}, pink triangles) are similar to subject entity representations. The dot product of enhanced aggregate representations and enhanced representation of the query relation (eqn. 4) are pased through activation functions $f_{s/o}$ and summed with learnable weights $\alpha_{subj/obj}$. Here f_s and f_o are sigmoid functions with trainable temperature $a_{subj/obj}$ and threshold $b_{subj/obj}$ (eqn. 5)

(e.g.: (Cohen et al., 2020; Wang et al., 2019; Peters et al., 2019)[2]) including the creation of many annotated data sets (e.g.: (Zhang et al., 2017; Alt et al., 2020; Mesquita et al., 2019; Elsahar et al., 2019)). These tasks consider only the recognition of knowledge directly expressed in individual texts, whereas we seek to utilise the combined knowledge from both a collection of texts and a knowledge base, allowing implicit and automatic association between expressions in texts and knowledge base relations and inference of propositions not directly expressed in individual texts.

A number of works present a distant supervision approach that utilises entity pairs in texts as a signal for the presence of propositions that may be incorporated in a knowledge base. This signal is inherently noisy, and several approaches have been devised do deal with this (e.g.: (Hoffmann et al., 2011; Zeng et al., 2015; Lin et al., 2016)). Closer to what we propose, Han et al. (2018) propose a neural attention mechanism between a knowlege graph and supporting texts, outperforming previous approaches. These approaches do not utilise graph information in the form of connections between the texts and can only extract relations explicitly mentioned in the texts. We note that the OpenKI model (Zhang et al., 2019), which we use as a baseline, outperforms these models (see Table 3).

3 Enhanced Entity Neighbourhood Model

We build on the Entity Neighbourhood Encoding (ENE) model proposed by Zhang et al. (2019). We then combine our enhanced neighbourhood encodings with the more complex "dual attention" model coined as "OpenKI".

Input data consists of a knowledge base or "KB" (a curated collection of proposition triples) and a collection of texts with entities identified and linked to knowledge base entities (where possible). In addition, text predicates linking entity mentions in source texts may be extracted (for example with OpenIE tools such as Reverb (Fader et al., 2011) or Ceres (Lockard et al., 2019)). Alternatively, sentences can be used as proxies for text predicates. The task then is to decide whether a query proposition (e_1, r, e_2) with KB relation r is true and should be added to the KB.

A graph of the propositions drawn from both

[2] https://paperswithcode.com/sota/relation-extraction-on-tacred

the knowledge base and source texts is constructed. Here the entities are nodes and KB relations and text predicates are directed links from the subject entity to the object entity. "Neighbourhoods" of entities are then defined as the set of outward links (subject neighbourhoods) and inward links (object neighbourhoods) from/to an entity.

Each relation and predicate p is associated with two unique, trainable embeddings $v_p^{subj}, v_p^{obj} \in \mathbb{R}^T$. We combine these learned relation/predicate embeddings with embeddings v_p^{text} derived from associated texts to obtain enhanced representations $v^{\cdots:t}$ as follows.

$$
\begin{aligned}
v_p^{subj:t} &= v_p^{subj} + tanh(W_{subj}^{pred} v_p^{text} + b_{subj}^{pred}) \\
v_p^{obj:t} &= v_p^{obj} \ + tanh(W_{obj}^{pred} v_p^{text} + b_{obj}^{pred})
\end{aligned} \tag{1}
$$

where $W_{subj}^{pred}, W_{obj}^{pred} \in \mathbb{R}^{D,T}$ and $b_{subj}^{pred}, b_{obj}^{pred} \in \mathbb{R}^D$ are trainable weight matrices and bias vectors respectively. We use a $tanh$ activation function to allow the model to adapt the learned representations v_p^{subj} and v_p^{obj} in both a positive and negative direction. In this work, the text representations v_p^{text} are static and do not vary during training.

Given subject/object entities s and o in our query, we aggregate relation/predicate representations from their respective entity neighbourhoods $R(s,\cdot)$ and $R(\cdot,o)$, as follows. We use vector average as the aggregation function $Agg(\cdot)$. Note that entities have no associated learned embedding, and are represented only as these aggregate representations.

$$
\begin{aligned}
v_{subj}^{agg}(s) &= Agg_{p \in R(s,\cdot)}(v_p^{subj:t}) \\
v_{obj}^{agg}(o) &= Agg_{p \in R(\cdot,o)}(v_p^{obj:t})
\end{aligned} \tag{2}
$$

Zhang et.al. (Zhang et al., 2019) posit that the aggregated representations $v_{subj}^{agg}(e)$ and $v_{obj}^{agg}(e)$ provide ultra-fine grained type information about entities when playing the respective roles and observe that including entity type information into their models does not notably improve performance, suggesting that type information is already present. Taking inspiration from that, we propose combining these aggregated representations with text based entity embeddings $v_e^{text} \in \mathbb{R}^T$ derived from entity names and descriptions.

$$
\begin{aligned}
v_{subj}^{agg:t}(e) &= v_{subj}^{agg}(e) + tanh(W_{subj}^{ent} v_e^{text} + b_{subj}^{ent}) \\
v_{obj}^{agg:t}(e) &= v_{obj}^{agg}(e) \ + tanh(W_{obj}^{ent} \ v_e^{text} + b_{obj}^{ent} \ )
\end{aligned} \tag{3}
$$

where $W_{subj}^{ent}, W_{obj}^{ent} \in \mathbb{R}^{D,T}$ and $b_{subj}^{ent}, b_{obj}^{ent} \in \mathbb{R}^D$, are trainable weight matrices and bias vectors respectively.

We then obtain association scores for a candidate predicate p, candidate subject entity s and candidate object entity o via a vector similarity measure (dot product in our case).

$$
\begin{aligned}
S_{subj}^{ENE}(s,p) &= v_{subj:t}^{agg}(s) \cdot v_p^{subj:t} \\
S_{obj}^{ENE}(p,o) &= v_{obj:t}^{agg}(o) \cdot v_p^{obj:t}
\end{aligned} \tag{4}
$$

These scores are then passed through sigmoid functions with trainable temperatures $a_{subj}, a_{obj} \in \mathbb{R}$ and thresholds $b_{subj}, b_{obj} \in \mathbb{R}$, then summed with trainable mixing weights $\alpha_{subj}, \alpha_{obj} \in \mathbb{R}$. The mixing weights are passed through the $ReLU$ function to ensure that the raw scores can only contribute positively to the final score without canceling each other out.

$$
\begin{aligned}
&score(s,p,o) \\
&= ReLU(\alpha_{subj}) \cdot \sigma(a_{subj} S_{subj}^{ENE}(s,p) + b_{subj}) \\
&+ ReLU(\alpha_{obj} \ ) \cdot \sigma(a_{obj} \ S_{obj}^{ENE}(p,o) + b_{obj} \ )
\end{aligned} \tag{5}
$$

The resulting score, trained with a max-margin loss, allows us to rank propositions, with true propositions ranked higher.

The full OpenKi model incorporates a third scoring component that combines aggregated neighbour representations (Equation 2) with a "query attention mechanism" similar to (Verga et al., 2017) — see (Zhang et al., 2019) for details. For text enhanced models we replace the neighbour representations with Equation 3.

4 Data Sets

Following (Zhang et al., 2019) we test our models on two data sets: 1) English language extractions from the New Your Times (NYT) (Riedel et al., 2010) consisting of sentences with named entities identified and linked to FreeBase (FB) and 2) REVERB (Fader et al., 2011) (an OpenIE tool) extractions from ClueWeb (Lin et al., 2012) (English language web texts) as preprocessed by OpenKI authors[3] also with entities linked to FreeBase.

For the NYT data, we use sentences as proxies for text predicates and for predicate texts we use whole sentences (including the entity mentions). Texts for Freebase relations are derived

[3] https://github.com/zhangdongxu/relation-inference-naacl19

Table 1: Data Statistics

	OpenIE	NYT
Training data		
# entity pairs	40,878	377,013
# without KB relations	0	359,197
# KB relation types	250	57
# Predicate types	124,836	320,711
Test data		
# test triples	4,938	1,761

from their identifiers, which are paths in the freebase relation hierarchy. We convert these paths to texts consisting of the sequence of relation class names separated by full stops. For example, "`location.us_state.capital`" is converted to "Location. US state. Captial." See Appendix C for details of NYT data preprocessing.

The second data set consists of REVERB (Fader et al., 2011) (an OpenIE tool) extractions from ClueWeb with entities linked to FreeBase (Lin et al., 2012), as provided by OpenKI authors[4]. Text predicates in this data are provided in text form, which are used directly. We obtain texts for FreeBase relations in a similar way to the NYT data. Note that original sentences from ClueWeb are not readily available for this data.

To obtain entity texts for both data sets, we use the property `type.object.name` of the associated FreeBase entity, where present, or the entity span in the NYT source text in other cases[5]. Most FreeBase entities also include a longer description text (the `common.topic.description` property). We concatenate the entity names and their descriptions to obtain a second text representation, used in the "…+ Desc" columns in Table 2. Where the description text is missing or the entity was not found in FreeBase, we use only the shorter text for the "…+ desc" results.

5 Experiments

We follow the experiments presented in (Zhang et al., 2019) for effective comparison. In preliminary experiments, we additionally trained all model variants using only text representations (effectively fixing all learned representations $v_p^{subj/obj}$ to zero vectors), and found performance to be substantially degraded in all cases. Similarly, the SOTA knowledge base completion model Tucker (Balazevic et al., 2019) performed very poorly when applied to the combined text predicate + KB relation graph for both data sets. Source code for our experiments including data download links is available on GitHub[6].

We use text embeddings derived from fasttext word embeddings (Mikolov et al., 2018) and BERT-SMALL (Devlin et al., 2018). For fasttext, we average the embeddings for words in the text. For BERT we use two strategies: the average of token representations and the representation of the special "[CLS]" token appended to all texts during standard BERT pre-processing.

We use 100 dimensional learned embeddings, learning rate 0.005 with the RAdam optimiser (Liu et al., 2019) and batch size 128 for 150 epochs (ClueWeb data) and 70 epochs (NYT data). We train with max-margin loss with a margin of 1.0, using 16 negative examples for each positive example. Negative samples consist of the entity pair and a uniformly sampled (no-positive) relation or predicate. For evaluation, with the NYT data we use the area under the Precision-Recall graph (AUC-PR) for relation prediction over entity pairs. With the OpenIE data we use mean average precision (MAP) on the task of ranking entity pairs. Reported results are from the best of 5 runs for each configuration (as measured by development set performance).

6 Discussion:

In Table 2 we see that inclusion of text based information provides a substantial boost to performance across all model variants, with improvements up to 9% in MAP and 16% in AUC-PR.

We observe that including entity texts performs better than relation/predicate texts, even when entity texts are included as well (mostly ~3% improvement). This can probably be explained by the paucity of the predicate text representations for KB relations and that whole sentences contain extraneous information not relevant to the relationship between entities. Future work with, for example, contextual BERT representations of predicate spans and excluding KB relation texts may perform better.

[4] https://github.com/zhangdongxu/relation-inference-naacl19

[5] Two entities in the OpenIE data were not found in FreeBase, zero vectors were used for their entity text embeddings.

[6] https://github.com/drevicko/OpenKI

Table 2: Performance of OpenKi ("dual attention") and Entity Neighbourhood Encoding (ENE) models with and without text enhancements on REVERB ClueWeb Extractions (MAP scores) and NYT (AUC-PR). "Entity" uses only entity names, "Ent+Desc" concatenates entity descriptions to the names, "Pred/Rel" uses only predicate/relation texts, "Both" combines entity names and predicate/relation texts and "Both+Desc" utilises all text information.

	Entity Neighbourhood Encoding (ENE)						OpenKI	
Included Texts:	None	Entity	Ent+D	Pred/Rel	Both	Both+D	None	Ent+D
OpenIE + ClueWeb (MAP)								
+ FastText	0.516	0.576	**0.610**	0.549	0.549	0.551	0.512	**0.618**
+ BERT (cls)		0.553	0.552	0.532	0.538	0.543		0.573
+ BERT (avg)		0.559	0.591	0.561	0.516	0.560		0.567
NYT (AUC-PR)								
+ FastText	0.674	0.831	**0.838**	0.726	0.738	0.699	0.681	**0.826**
+ BERT (cls)		0.757	0.824	0.636	0.718	0.711		0.819
+ BERT (avg)		0.769	0.813	0.729	0.758	0.756		0.813

Including entity descriptions is either similar or better than not including them (up to ~4%), in particular for BERT. It is not surprising that BERT can leverage the long-form entity descriptions effectively. Average BERT token embeddings perform better than the CLS token embedding in most cases.

The most surprising result is the relative performance between BERT and FastText, with FastText outperforming BERT with entity only text enhancement and providing the best performing models. It is not clear to us why this is the case. One hypothesis is that the fully conected layers projecting text representations to the learned embedding dimension may do better with a different, lower learning rate, and that this effect may be more pronounced with the larger BERT representations. We plan to explore this in future work.

It is worth noting that using sentences as proxies for text predicates is a rather weak setup. The majority of sentences contain a single entity pair, meaning that the sentences (as a predicate proxy) only appears in one subject and one object neighbour list. This provides little graph information for the model to utilise. The small proportion that do overalap appear to provide benefit however.

OpenKI identifies a compatibility between relations and entities through their co-occurrences in a graph. Though a strong signal, our results indicate that this information is further enhancded by the detailed and nuanced information that can be found in both task source texts and entity and relation descriptions. Text based information alone, however, has not been found to provide sufficient information for good performance on these tasks, as seen in both our preliminary experiments without learned graph-based embeddings and previous work that relies on text based inference (Table 3).

Table 3: Other Baseline Models on NYT data.

model	AUC-PR
ENE + Entity Descriptions (ours)	0.838
OpenKI (Zhang et al., 2019)	0.461
JointD + KATT (Han et al., 2018)	0.369
PYCNN + Att. (Lin et al., 2016)	0.341

7 Conclusion

We investigated the task of integrating new information in the form of a collection of texts such as news articles into a knowledge base (KB), building on previous models that utilised information from the combined graph of knowledge base relations and predicates extracted from the texts using OpenIE tools. We propose a mechanism for incorporating text representations of entities, KB relations and text predicates into the state of the art OpenKI model, providing a substantial improvement in performance. From this we can conclude that source texts and entity and relation descriptions contain nuanced information useful to the task beyond that contained in graph structures in the knowledge base and extracted predicate propositions. Our models represent a new state of the art on two data sets for this task.

Acknowledgments

We thank anonymous reviewers for their insightful suggestions to improve this paper. This research was conducted under the Australian Research Councils Discovery Projects funding scheme (project number DP160102156).

References

Christoph Alt, Aleksandra Gabryszak, and Leonhard Hennig. 2020. Tacred revisited: A thorough evaluation of the tacred relation extraction task. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Ivana Balazevic, Carl Allen, and Timothy Hospedales. 2019. Tucker: Tensor factorization for knowledge graph completion. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5185–5194, Hong Kong, China. Association for Computational Linguistics.

Amir DN Cohen, Shachar Rosenman, and Yoav Goldberg. 2020. Relation extraction as two-way span-prediction. *arXiv:2010.04829 [cs]*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv:1810.04805 [cs]*.

Hady Elsahar, Pavlos Vougiouklis, Arslen Remaci, Christophe Gravier, Jonathon Hare, Elena Simperl, and Frederique Laforest. 2019. T-rex: A large scale alignment of natural language with knowledge base triples. In *LREC 2018 - 11th International Conference on Language Resources and Evaluation*, pages 3448–3452.

Anthony Fader, Stephen Soderland, and Oren Etzioni. 2011. Identifying relations for open information extraction. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1535–1545, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Daniel Gillick, Sayali Kulkarni, Larry Lansing, Alessandro Presta, Jason Baldridge, Eugene Ie, and Diego Garcia-Olano. 2019. Learning dense representations for entity retrieval. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 528–537, Hong Kong, China. Association for Computational Linguistics.

Xu Han, Zhiyuan Liu, and Maosong Sun. 2018. Neural knowledge acquisition via mutual attention between knowledge graph and text. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1).

Raphael Hoffmann, Congle Zhang, Xiao Ling, Luke Zettlemoyer, and Daniel S. Weld. 2011. Knowledge-based weak supervision for information extraction of overlapping relations. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 541–550, Portland, Oregon, USA. Association for Computational Linguistics.

Thomas Lin, Mausam, and Oren Etzioni. 2012. Entity linking at web scale. In *Proceedings of the Joint Workshop on Automatic Knowledge Base Construction and Web-scale Knowledge Extraction (AKBC-WEKEX)*, pages 84–88, Montréal, Canada. Association for Computational Linguistics.

Yankai Lin, Shiqi Shen, Zhiyuan Liu, Huanbo Luan, and Maosong Sun. 2016. Neural relation extraction with selective attention over instances. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2124–2133, Berlin, Germany. Association for Computational Linguistics.

Liyuan Liu, Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Jiawei Han. 2019. On the variance of the adaptive learning rate and beyond.

Colin Lockard, Prashant Shiralkar, and Xin Luna Dong. 2019. Openceres: When open information extraction meets the semi-structured web. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3047–3056, Minneapolis, Minnesota. Association for Computational Linguistics.

Filipe Mesquita, Matteo Cannaviccio, Jordan Schmidek, Paramita Mirza, and Denilson Barbosa. 2019. Knowledgenet: A benchmark dataset for knowledge base population. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 749–758, Hong Kong, China. Association for Computational Linguistics.

Tomas Mikolov, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin. 2018. Advances in pre-training distributed word representations. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

Matthew E. Peters, Mark Neumann, Robert Logan, Roy Schwartz, Vidur Joshi, Sameer Singh, and Noah A. Smith. 2019. Knowledge enhanced contextual word representations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 43–54, Hong Kong, China. Association for Computational Linguistics.

Sebastian Riedel, Limin Yao, and Andrew McCallum. 2010. Modeling relations and their mentions without labeled text. In *Machine Learning and Knowledge Discovery in Databases*, Lecture Notes in Computer Science, pages 148–163, Berlin, Heidelberg. Springer.

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, Atlanta, Georgia. Association for Computational Linguistics.

Gabriel Stanovsky, Julian Michael, Luke Zettlemoyer, and I. Dagan. 2018. Supervised open information extraction. In *NAACL-HLT*.

Patrick Verga, David Belanger, Emma Strubell, Benjamin Roth, and Andrew McCallum. 2015. Multilingual relation extraction using compositional universal schema.

Patrick Verga, Arvind Neelakantan, and Andrew McCallum. 2017. Generalizing to unseen entities and entity pairs with row-less universal schema. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 613–622, Valencia, Spain. Association for Computational Linguistics.

Xiaozhi Wang, Tianyu Gao, Zhaocheng Zhu, Zhiyuan Liu, Juanzi Li, and Jian Tang. 2019. Kepler: A unified model for knowledge embedding and pre-trained language representation. *arXiv:1911.06136 [cs]*.

Daojian Zeng, Kang Liu, Yubo Chen, and Jun Zhao. 2015. Distant supervision for relation extraction via piecewise convolutional neural networks. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1753–1762, Lisbon, Portugal. Association for Computational Linguistics.

Dongxu Zhang, Subhabrata Mukherjee, Colin Lockard, Luna Dong, and Andrew McCallum. 2019. Openki: Integrating open information extraction and knowledge bases with relation inference. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 762–772, Minneapolis, Minnesota. Association for Computational Linguistics.

Yuhao Zhang, Victor Zhong, Danqi Chen, Gabor Angeli, and Christopher D. Manning. 2017. Position-aware attention and supervised data improve slot filling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 35–45, Copenhagen, Denmark. Association for Computational Linguistics.

A Observations and Discussion

It is worth noting that the learned mixing weights between scoring components have the effect of learning the margin for the max-margin loss. We observed that these weights all increased during inference to values from 7 to 10, resulting in an effective margin of 0.10 to 0.14. Setting and fixing the margin (so the model learns relative weighting, but cannot scale the whole model) however did not perform as well, indicating that this learned scaling helped the model to navigate the parameter space.

The bert dual models with average entity text encoding with descriptions on reverb data performed better (by $\tilde{0}.5\%$) on the development split, but worse on the test split, indicating an element of overfitting. With CLS token encoding, dev (and test) results with and without descriptions were not notably different (<0.1%). FastText development set results were around 2.5% better with descriptions, which is similar to test results.

B Choice of Model Parameters:

Due to large run times, substantial parameter tuning was not practical. We chose 100 dimensional embeddings as a reasonable compromise between improved expressiveness and increased run times. We found that in most cases the parameters used in (Zhang et al., 2019) (batch size 128, lr 0.005 etc. . . — see Section 4) were able to produce reasonable results from model checkpoints with the best development set performance. We typically ran two models and chose the best dev set performance from the two. We found that larger batch sizes were more stable in training (ie: less variation in loss and dev set performance between epochs), however lacked the inherent exploration of the parameter space provided by those variations and hence resulted in lower (best epoch) development set performance.

C Details of NYT Data Preparation

The original NYT data has of two splits: training data from 2005-2006 and evaluation data from 2007. We further randomly split the training data into train and development subsets such that entity pairs in the development subset are not present in the training subset and further exclude test set entity pairs from training subset. We use sentences as proxies for text predicates in this data. Development and test data consist of all FreeBase relations connecting entity pairs in the development subset and 2007 data respectively. Note that entity pairs for which no FreeBase relation is indicated are not included in dev/test data, following previous literature (e.g.: (Han et al., 2018; Zhang et al., 2019)). Training data consists of remaining FreeBase relations and sentences found in the train split. Note that no development or test triples are present in the training data.

D Experimental Environment and Run Times

Experiments were performed on dual P100 GPUs with 10gb of memory in the CSIRO "Bracewell" high performance computing cluster.

Runtime per epoch with the given parameters varied by dataset and model complexity. For the REVERB extractions from ClueWeb, ENE without text and entity text enhanced models requiring approximately 10 minutes, ENE with predicate and combined entity-predicate text enhancement 15 minutes, OpenKi without text and entity text enhanced models requiring approximately 40 minutes, OpenKi with predicate and combined entity-predicate text enhancement 50 minutes.

For the NYT data, ENE without text and entity text enhanced models requiring approximately 50-60 minutes, ENE with predicate and combined entity-predicate text enhancement 60-70 minutes, OpenKi without text and entity text enhanced models requiring approximately 180 minutes, OpenKi with predicate and combined entity-predicate text enhancement 240 minutes.

E Low Dimensinal Models (from original OpenKI paper)

We ran 12 dimensional models on the NYT data to match the configuration used in the original OpenKI paper (Zhang et al., 2019). Entity neighbourhood encoding (ENE) and "dual attention" (OpenKI), without text enhancement (W/O Text) and with entity text enhancement including entity descriptions (Ent+D) — see Table 4.

These results improve on those presented in the original OpenKI paper (0.421 and 0.462 respectively) primarily due to our use of sentences as predicate proxies, which allow for a modest level of predicate co-occurrence between neighbour lists due to sentences with more than two entities. They use a window around the entity pair with the two entities masked, resulting in no co-occurrence between neighbour lists.

Table 4: 12D NYT Models

model	W/O Text	Ent+D
ENE	0.528	0.722
OpenKI	0.558	0.581