

KG Cleaner: Identifying and Correcting Errors Produced by Information Extraction Systems

Ankur Padia, Francis Ferraro, Tim Finin
University of Maryland, Baltimore County
{pankurl, ferraro, finin}@umbc.edu

Abstract

KG Cleaner is a framework to *identify* and *correct* errors in data produced and delivered by an information extraction system. These tasks have been understudied and KG Cleaner is the first to address both. We introduce a multi-task model that jointly learns to predict if an extracted relation is credible and repair it if not. We evaluate our approach and other models as instance of our framework on two collections: a Wikidata corpus of nearly 700K facts and 5M fact-relevant sentences and a collection of 30K facts from the 2015 TAC Knowledge Base Population task. For credibility classification, we find that *parameter efficient*, simple shallow neural networks can achieve an absolute performance gain of 30 F_1 points on Wikidata and comparable performance on TAC. For the repair task, significant performance (at more than twice) gain can be obtained depending on the nature of the dataset and the models.

1 Introduction

Information Extraction (IE) systems extract entities, events and relations from text documents to create or update a knowledge graph. However, current IE systems are prone to various kinds of mistakes that result in errors in the knowledge graph data they produce. For example, in the Cold Start Knowledge Base Population task of the 2017 Text Analysis Conference (TAC) (Dang, 2017), none of the systems achieved an F1 score higher than 0.3. Moreover, the average precision of the best systems was even lower.

Consumers of data from IE systems could benefit if they had an independent way of evaluating the knowledge graph fragments to identify those likely to be incorrect. Even better would be a system that could, in some cases, repair a faulty fragment. To be truly independent, or **standalone**, such a system would not have access to the inner workings

Fact extracted by the IE system:

per:cause_of_death (Nelson Mandela, accident)

Provenance fetched by the IE system:

The bodies to be exhumed are those of three of Nelson Mandela's children: Mandla Mandela's father, Makgatho Mandela, who died in 2005; his first daughter, also named Makaziwe, who died as an infant in 1948; and another son, Madiba Thembekile Mandela, who died in a traffic accident in 1969.

Output from our system, KG Cleaner:

Is fact credible or incredible:	Incredible
Possible repair :	None

Figure 1: On this TAC 2016 example, KG Cleaner used the provenance sentence shown to judge the fact as *not* credible and find that the sentence expresses no known relation in TAC's schema.

of the IE systems producing the fragments or the detailed analytics or structures they might use in making their decisions. At best, the IE system might be assumed to provide simple *provenance* data, such as the document or a sentence that supports the fragment.

Figure 1 shows a fact and its provenance extracted by one of the systems participating in TAC 2015 to answer the query about Nelson Mandela's cause of death. KG Cleaner correctly judges the relation as not credible given the provenance, but is unable to offer a repaired or alternate fact based on provenance with respect to fixed subject.

Current research work focuses on determining the **credibility** of the extracted facts, i.e., determine whether facts are correct or incorrect given a collection of provenance information and *without* using human or outside knowledge. While human input or external knowledge is frequently not used, these credibility assessments often assume access to the inner-workings of IE systems (Yu et al., 2014); these are not *standalone*.

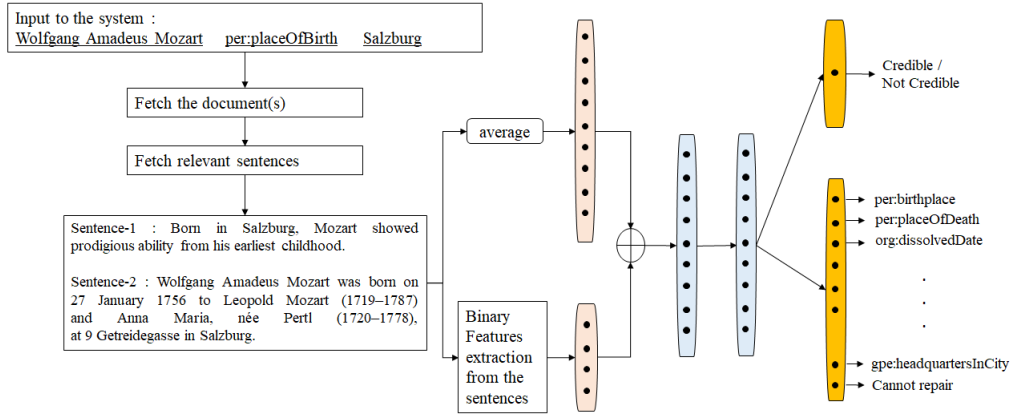


Figure 2: Overview of KG Cleaner’s joint credibility and repair model.

Though standalone approaches do exist, such as OpenEval (Samadi et al., 2013) and Yin et al. (2008), these rely on information retrieved from searching the Web. However, the retrieved results and facts are mostly biased towards popular entities and hence these systems do not perform well on emerging or long tail entities. Moreover, depending on the underlying data used for searching, the opaque personalization of the search engine, and other biases also effect the search result and there by the performance of the system like OpenEval. Hence making a controlled scientific comparison with OpenEval becomes very difficult.

We argue that the two tasks of provenance credibility and KG repair can leverage large-scale, automatic **semantic parses**; though we want to be standalone, we do not want to be eschew general knowledge that these semantic parses contain. We demonstrate the utility of semantic frame parsing, which can act as a malleable schema of sorts.

While our framework has the flexibility to use a range of models, from sophisticated neural networks (e.g., CNN, RNN) to simple linear regression, we opted for MLPs due to their simplicity (e.g., fewer parameters) and their success for verifying news in the 2017 Fake News Challenge (Pomerleau and Rao, 2017), where two of the top three teams used them. Moreover, we do not make strong assumptions on the availability of the provenance information, as the provenance to an extracted fact could be optional. Therefore, we consider two cases—when provenance is available, and when provenance is not available. Wikidata facts have document-level, but not sentence-level, provenance. We leverage automatic semantic parses of document sentences to find relevant

ones. TAC facts have sentence-level provenance, which we use directly.

We evaluate all approaches on two collections: a Wikidata corpus of 663,164 facts and a total of 4.76 million relevant sentences, and data from the Knowledge Base Population task of the 2015 Text Analysis Conference. Significant improvement can be achieved depending on the nature of the dataset and model. Demonstrate that system like KG Cleaner can perform significantly well when trained and test on same dataset using embeddings and semantic resources like FrameNet (Baker et al., 1998). We plan to make our relations mappings, datasets and implementation available upon publication.

2 Cataloging IE system errors

We analyzed the types of errors in facts sampled from the output of the 70 system that participated in the TAC 2015 Knowledge Base population task (Dang, 2015). For each of the 65 possible relations we considered ten facts with a given subject. This analysis is enlightening, as each system used different approaches and types of resources to extract potential facts. Each of the facts is assessed at two levels: (1) if the object value is correct, and (2) does the relation hold between the subject and the object. We use a sample of 10 facts across 65 relations with attempt to avoid repeating subject-predicate pairs, resulting in 643 facts with provenance.

We hand-analyzed this sample using LDC guidelines to understand different kinds of errors made by the IE systems and defined five error categories (examples can be seen in Table 1):

1. **Correct:** filler is correct and provenance sup-

Category	Extracted Fact and provenance text
Correct	<i>Kodak org:stateorprovince_of_headquarters New York</i> The following three memos were each sent via company e-mail to about 1,000 people at an Eastman Kodak Co. division at the company’s headquarters in Rochester, New York .
Subject missing	<i>Eleanor Catton gpe:subsidiaries Bain</i> Buying into Canada Goose is the latest Canadian investment for Bain .
Object missing	<i>Kermit Gosnell per:cities_of_residence America</i> Historic crowdfunding for movie about abortionist Kermit Gosnell - YouTube
Incorrect relation	<i>Harry Reid per:charges assault</i> Nevada’s Harry Reid switches longtime stance to support assault weapon ban
Misc	<i>Reginald Wayne Miller per:charges felony</i> Various news outlets have reported that federal agents have probable cause to charge Reginald Wayne Miller with forced labor, a felony that can carry up to a twenty-year prison sentence per charge.

Table 1: Example for each of the error category.

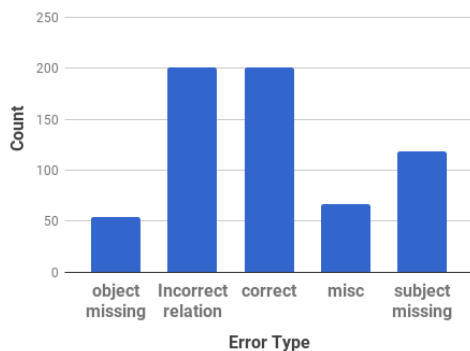


Figure 3: Frequency count vs. error types. Error analysis performed on the extracted facts

- ports fact
2. **Object missing:** object entity not mentioned in provenance
3. **Subject missing:** subject entity not mentioned in provenance
4. **Incorrect relation:** subject and object are present but relation is not entailed or triggered
5. **Misc:** fact does not match LDC guidelines e.g. “per:charge” should be alleged or convicted actual charge like “robbery” and not “5 year prison”.

Figure 3 shows that the most frequent error type is an incorrect relation, followed by missing subject, missing object and miscellaneous errors. In total roughly covering 2/3 of the sampled fact-provenance pair. We noted that regular expression based extraction for email addresses and website URLs were very accurate. We use this analysis to define features for our system and establish the initial motivation to attempt to repair relations.

3 Related work

We describe some prior research on the general problem of assessing the quality of facts extracted from text and note where and how their requirements or assumptions differ from ours.

Nakashole and Mitchell (2014) evaluates the credibility of facts extracted from a document using linguistic features that predict how *objective* or *subjective* the document is. Ojha and Talukdar (2017) estimates the quality of facts in a knowledge graph using a few sample seed annotations obtained by crowd sourcing and exploiting the knowledge graph’s schema to propagate the score across entire graph.

Ensemble-based approaches have been designed to determine the credibility of extracted facts. Yu et al. (2014) propose an unsupervised method using linguistic features to filter credible from incredible facts, but requires access to multiple IE system with different configuration settings that extract information from the same text corpus. Viswanathan et al. (2015) proposes a supervised approach to build a classifier from the confidence scores produced by multiple IE systems for the same triple. Such systems are not *standalone*, as they assume availability of multiple IE systems.

The assumption of having multiple IE systems can be relaxed by using iterative approach with linguistic features or by considering external schema information. Samadi et al. (2013) evaluates a fact’s correctness by using a web search engine to find sentences containing the subject and object and applying a bag-of-words classifier on the text. Such an approach considers the Web as a text corpus and hence are likely to handle popular entities. Samadi et al. (2016) the makes the

system more robust by considering conflicting information and resolving the conflict using probabilistic soft logic (Brocheler et al., 2012).

Pujara et al. (2013) tries to identify knowledge graph from a noisy knowledge graph by modeling schema as rules and reasoning using PSL to determine the optimal link combination in the knowledge graph that maximizes to satisfy the schema. However the work assumes the access of schema information and is limited to relations which are mathematically well defined like *inverse*, *disjoint*, *typeOf*, *domain* and *range*. As a result it cannot handle strings based relation like *bornIn*, *placeOfDeath*.

There has been recent work to verify statements within extended prose. Ferreira and Vlachos (2016) proposes a system for journalism while Patwari et al. (2017) tries to understand political debates to help humans to focus on check-worthy statements. Vlachos and Riedel (2015) tries to verify numerical statements like population and inflation rate made in text snippets.

FEVER (Thorne et al., 2018) is a system that is similar to ours, but performs provenance-based classification without attempting to repair errors. The “facts” being verified are text sentences that can correspond to a set of KG triples (e.g., “Washington was a soldier born in 1732.”). However, no schema is available to ground the semantics and the data would require conversion to a triple format for KG Cleaner to be applied.

Our approach differs from previous ones in several important ways. First, we jointly model fact credibility and repair within a knowledge graph. Second, our framework is designed as a standalone system that does not require access to the original IE system nor its enhanced, detailed output Yu et al. (2014); Viswanathan et al. (2015). Third, we do not assume output from an ensemble of IE systems on the same text collection. Finally, unlike (Lehmann et al., 2012; Li et al., 2011), our approach uses simple features which do not require tailoring based on the application being supported.

4 Approach

In this section we describe our model and the training procedure. The input to the system is a triple $\langle s, p, o \rangle$ where s is the subject, r the relation, and o the object and optional document (and/or entity/relation offset) as provenance. The system has two outputs: a classification as credible or in-

credible and a suggested correct fact if judged incredible. For each fact as input to the system, we fetch relevant sentences from the document, convert sentences into features, pass it to the model for multi-task classification.

4.1 Finding relevant sentences and features

For the document mentioned as the provenance id we process all sentences and consider a sentence to be relevant if it satisfies any of the following criteria: the subject or object or one of their aliases is mentioned in the sentence; a paraphrase of the object is mentioned in the sentence; the predicate or an alias is mentioned in the sentence; or the sentence triggers (i.e annotated by) a frame from FrameNet (Ferraro et al., 2014) for the predicate.

For each of the facts from Wikidata (Vrandečić and Krötzsch, 2014) we use exact string match on Wikipedia for the subject to retrieve information about the subject. We additionally created a database of subject, object, and predicate with their corresponding aliases and use it to fetch more documents when there is an alias match. For objects we also considered aliases when the object was paraphrased (Ganitkevitch et al., 2013). We manually mapping relations from Wikidata to frames from FrameNet. It took only an afternoon to create the mapping manually. We also experimented with an automatically created bag-of-words mapping and evaluated its effectiveness (Figure 6b). We considered the sentence to be relevant when the sentence triggered one of the mapped frame for the predicate. We consider only those facts for which we could fetch at least one sentence.

We used pre-trained word embedding and manual features to characterize the sentences. For each fact we compute the average of the embeddings of the words in its provenance sentences, using a zero-vector for out-of-vocabulary words. We derived binary features from the sentences corresponding to our selection criteria mentioned earlier. These features align with the error types described in Section 2) and help identify in which provenance does not support the object and/or entail the relationship between the subject and object; such signals have been used to determine fact credibility (Yu et al., 2014). However, our framework allows more sophisticated methods to make connections between provenance information and facts.

4.2 Multi-task neural network architecture

A feedforward multilayer perceptron was employed to jointly learn distribution of the credibility and repair tasks with shared parameters. As shown in Equation 2, each layer learns an abstract representation from the previous one.

$$\mathbf{h}^{(0)} = \mathbf{x}_f \quad (1)$$

$$\mathbf{h}^{(i)} = g(\mathbf{W}^{(i)}\mathbf{h}^{(i-1)} + \mathbf{b}^{(i)}) \quad (2)$$

$$\hat{y}_c = \text{sigmoid}(\mathbf{W}_{cred}^{(last)}\mathbf{h}^{(last)} + \mathbf{b}^{(last)}) \quad (3)$$

$$\hat{y}_r = \text{softmax}(\mathbf{W}_{repair}^{(last)}\mathbf{h}^{(last)} + \mathbf{b}^{(last)}) \quad (4)$$

Here i is the number of layers and variables with the *last* superscript indicating parameters and values from the previous hidden layer. We set g to tanh to introduce non-linearity. Each \mathbf{x}_f is the input to the system and has dimension $e+n$ where e is the word-embedding dimension and n is the number of features. Dimension of $\mathbf{W}^{(i)}$ except for last layer is $(e+n) \times (e+n)$. The dimension of $\mathbf{W}_{cred}^{(i)}$ and $\mathbf{W}_{repair}^{(i)}$ is $1 \times (e+n)$ for binary classification and $r \times (e+n)$ to obtain a probability distribution over all relations. $\mathbf{b}^{(i)}$, and $\mathbf{b}^{(last)}$ are the biases vectors. We use a sigmoid function for binary credibility classifier and softmax for relation classification. Regularized binary cross entropy was used for the credibility loss function and categorical cross entropy for relation repair.

4.3 Training and Negative Sampling

We trained KG Cleaner using stochastic gradient descent with a momentum of 0.9 (Sutskever et al., 2013) and a decay rate of 10^{-6} . We initialized the model with Xavier Uniform initialization (Glorot and Bengio, 2010) and used back propagation to learn the parameters.

Since our Wikidata facts provide only positive examples, we generated *faux facts* for training from them by first fixing the subject and then replacing the relation and object with randomly selected ones and associating a randomly selected provenance text segment from one of the positive examples. During our negative sampling we do not prefer any particular relation or object over other, but follow a uniform distribution to randomly pick relation, object, and provenance. We label all negative instances as not credible and a special ‘‘cannot repair’’ relation. During training, we used a balanced batch with an equal number of positive and negative examples.

5 Experiments

We present analysis of KG Cleaner and other models as instances of our general framework. We trained KG Cleaner and other models on our Wikidata training set and used pre-trained word embedding of dimension 400 produced using continuous bag-of-words and window size of five on a recent Wikipedia text corpus. For evaluation we used two separate test sets, one from Wikidata and another from TAC 2015 data.

Our evaluation showed that systems like,KG Cleaner, can outperform other bag-of-word models with a significant performance gain. We found using that using a simple network with two layers and with hand crafted features helped capture subtle nuances that improved the performance of both credibility classification and relation repair. Getting more relevant sentences also helped improve the performance of both tasks.

5.1 Dataset preparation

We manually created mappings from Wikidata relations (Table 2) to TAC KBP relations and also to FrameNet 1.7 frames. We also computed a mapping using lexical overlap using Wikidata descriptions/aliases and, for frames, the lexical unit that triggers the frame. For each of the Wikidata relations we picked as many as 100K instances. For each of these, we found relevant ‘‘provenance’’ sentences as explained in Section 4.1 to form positive examples for training. We created negative examples for training using negative sampling as described in Section 4.3

5.2 Dataset and hyper-parameters

Dataset : We divide 663,164 facts into 3 parts *train* (463,164), *test* (100K), and *Dev* (100K). We create negative instances of equal size for each part using negative sampling described above. We use TAC as provided by LDC with 9,215 positive examples and 21,019 as negative examples.

We tuned our hyperparameters using coordinate descent (Bengio, 2012), in which we change a hyperparameter and update its values if it improves the performance. We ran each model for five epoch and fixed batch size of 64. We tested with multiple learning rate $\{10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}\}$, Lasso regularization from $\{10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}\}$ and Dropout from $\{0.2, 0.3, 0.4, 0.5\}$ and chose the value of the hyperparameter that gave best per-

affiliation	child	educated at	member of political party	parent club
alternate names	convicted of	end time	mother	parent organization
birthday	country of origin	ethnic group	noble title	place of birth
birth name	crew member	father	number of seats	place of death
board member	date of birth	founded by	occupant	political alignment
business division	date of burial or cremation	inception	official name	posthumous name
capital	date of death	manner of death	official residence	reference URL
cause of death	discontinued date	member count	official website	religion
charge	dissolved, abolished or demolished	member of	owned by	residence
sibling	spouse	subsidiary		

Table 2: Wikidata relationships considered for credibility and repair.

formance. We hyper-tuned the parameter for the model and set to two hidden layers. We found using both Lasso (Tibshirani, 1996) and Dropout (Srivastava et al., 2014) to be helpful with performance gain of two F1 points.

5.3 Baselines

As this work is the first to introduce joint credibility and repair tasks, we consider Logistic Regression (LR) based models and approaches like ours, KG Cleaner, as instances of general credibility and repair framework.

We trained Logistic Regression (LR) using SAGA solver (Defazio et al., 2014) due to the large dataset. We fine-tuned the LR classifier with Lasso regularization using penalty values $\{0.01, 0.1, 1, 10, 100\}$ and chose the one that performed best on the validation dataset.

- **Bag-of-words + LR + count vector:** We calculated the frequency of each word.
- **Bag-of-words + LR + binary vector:** As above, but replacing non-zero values with 1.
- **LR + sum word2vec:** We summed the bag of word embeddings of all words.
- **LR + average word2vec:** As above, except taking the average rather than the sum.
- **LR + TFIDF:** We used a TF-IDF vector of the words of the sentences.
- **KG Cleaner:** An MLP with two hidden layers

5.4 Results

Tables 3 and 4 compare the performance of multiple logistic regression and MLP models for our framework. Overall, we find the jointly trained MLPs perform better when trained and tested on larger datasets (Wikidata); in contrast, logistic regression can perform quite well when tested on news and discussion forum articles (TAC). We note that not only is the Wikidata training and

test set larger than TAC, but we are able to leverage larger-scale semantic parses with Wikidata (Wikipedia). This suggests the benefit that even noisy semantic annotations can help.

For the credibility task, among the bag-of-word based models, binary vectors perform better than other LR based models. However, its F1 score on TAC for credibility is less than 0.5 like other methods. Training LR with word2vec-based embeddings provides comparable performance. Significant improvement of absolute 30% is achieved when the word embeddings are used with a two-layer MLP model.

For the repair task similar behavior is seen, with the MLPs performing well on Wikidata. The MLPs outperform LR by roughly 30 F1 points (both macro and micro), though the wide gulf between the macro and micro scores for all models indicates the difficulty in repairing the long tail of relations. Similarly, we consider our framework as a ranking repair system; looking at mean reciprocal rank (MRR), we see that while the models do well on Wikidata, the MLPs do exceptionally well, with an MRR of 0.875. Together with the micro F1 performance, this indicates that a majority of common relations can be correctly repaired.

On TAC, however, we find the performance of the models flipped: simple logistic regression models outperform both the independently and jointly trained MLPs. We note a number of possible explanations for this behavior. First, the TAC evaluations involved a type of out-of-domain testing: while all models were trained on encyclopedic text, TAC evaluation happened in news and discussion forum domains. Second, for Wikidata we were able to leverage automatic frame annotations and both find and utilize *multiple* possible sources of information; in contrast for the TAC evaluations, we considered only a *single* provenance sentence that was provided by another, noisy IE system.

	Wikidata Test (WD)			TAC		
	precision	recall	f1	precision	recall	f1
LR + count vector	0.501	0.536	0.518	0.312	0.478	0.378
LR + binary vector	0.502	0.628	0.558	0.294	0.590	0.392
LR + sum word2vec	0.500	0.516	0.508	0.290	0.452	0.353
LR + average word2vec	0.502	0.506	0.504	0.286	0.526	0.371
LR + TFIDF	0.501	0.536	0.518	0.312	0.480	0.378
MLP, independent training	0.995	0.820	0.891	0.299	0.461	0.355
KG Cleaner with 2 layers	0.975	0.819	0.890	0.289	0.462	0.355

Table 3: Credibility Performance

	Wikidata Test (WD)			TAC		
	macro	micro	mrr	macro	micro	mrr
LR + Count Vector	0.197	0.477	0.678	0.027	0.341	0.464
LR + Binary Vector	0.194	0.454	0.678	0.042	0.648	0.706
LR + Sum Word2Vec	0.089	0.465	0.657	0.022	0.261	0.381
LR + Average Word2Vec	0.124	0.483	0.687	0.035	0.607	0.676
LR + TFIDF	0.209	0.409	0.642	0.033	0.426	0.561
MLP, independent training	0.357	0.771	0.863	0.027	0.046	0.402
KG Cleaner with 2 layers	0.383	0.791	0.875	0.028	0.360	0.403

Table 4: Repair Performance. Macro and Micro are F1 scores.

We used a multi-task setting due to the co-relating nature of the cleaning and repair tasks. However, as we have a large dataset; our observation on MTL’s effectiveness is similar to recent large-scale MTL (Kaiser et al., 2017) where improvement is marginal with large datasets. Ablation experiments (Figure 6) showed improvement with an increasing amount of data. The 0.001-0.01 difference in F1 (Tables 3 & 4) is within allowed observed variance (Figure 6). We are motivated by a workflow where existing IE systems are black-box (e.g., with proprietary data or leveraging external systems output) and we dont know what extraction techniques were used.

5.5 Ablation studies

In this section, we study some of the individual components and design decisions in our framework, including the number of sentences to use as potential provenance information for Wikidata, the impact that larger MLPs can have on performance, and the utility of the automatic frame annotations themselves. Overall, we found that using more relevant sentences during training, using both frames and a higher quality relation to frame mapping improve performance.

5.5.1 Changing the number of sentences

In our approach we fetch relevant sentences, but not all of them can serve as provenance. For example, we do not consider a sentence that contains an object mention but does not trigger the mentioned relation. Practically, some articles are short with few sentences to use for provenance. We study the behavior of different models, both MLP and LR, as the number of sentences varies: for each fact, we randomly choose 1, 2, 3, 5, 7, 10 or all sentences from the set of *relevant* sentences and trained our model. When there are fewer sentence than the target number, we used all of the sentences. For an unbiased comparison, we use the same test set for Wikidata and TAC as above; we kept other parameters the same.

Figure 4 summarizes the performance for LR + binary and our approach. The MLPs perform better on the Wikidata test with fewer number of likely provenance sentences per fact. On the other hand LR + binary hand performs poor on Wikidata Test but does well TAC. The reason KG Cleaner is performing poor on TAC is due different distribution on which it our approach is trained.

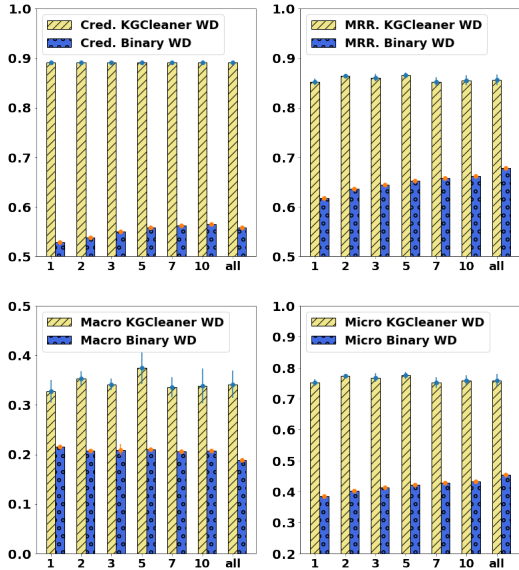


Figure 4: Effect of changing the number of sentences (horizontal axis) on task performance (vertical axis), averaged over three independent runs. For space, we only provide Wikidata results.

5.5.2 Effect of deeper networks

Does learning a deeper, potentially more abstract representation help the credibility and repair tasks? We explore this by considering our fine tuned best model. We make our network deep by adding more layers with the intuition that the later layers learns more abstract representation.

We keep all network configuration same beside activation function and number of layers. For two layer and less we used tanh as the activation function and for deeper network we used ReLU (Nair and Hinton, 2010) to reduce likelihood of vanishing gradient. Figure 5 show the behavior of adding more layers to the network, averaged over three different random initialization of the model. Overall we see that MLPs with two layers have better overall performance.

5.5.3 Effect of frame annotations

Figure 6(a) examines how effective our use of frame annotations was. First, we experiment with using sentences obtained from *only frames* (no subject or object filters), and also the sentences which were obtained without the use of any frame annotations (*no frames*). The horizontal lines in the chart are the performances of the full 2-layer MLP presented earlier. Using frames results in clear improvement trends across the board. On the Wikidata Test, none of the variants achieve better per-

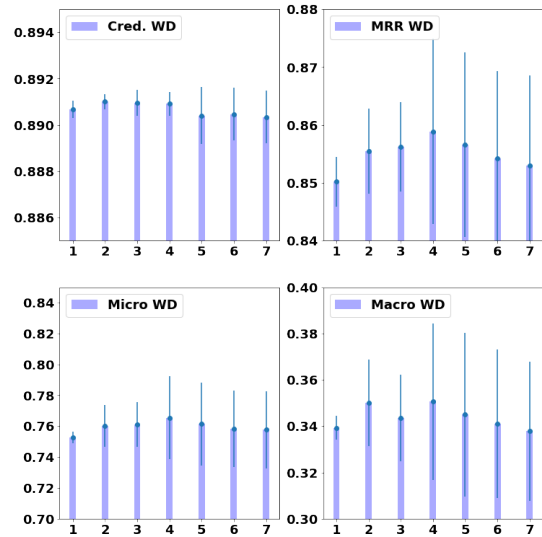


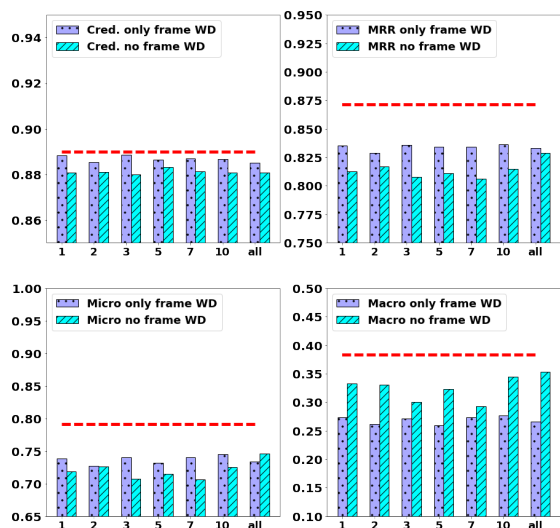
Figure 5: Effect of adding more layers to the MLP (horizontal axis) on the models' performance (vertical axis). In general deep networks have more variance in task performance compared to shallow networks. For space, we provide Wikidata results, though the same trends are evident in TAC.

formance than the frame-based MLP.

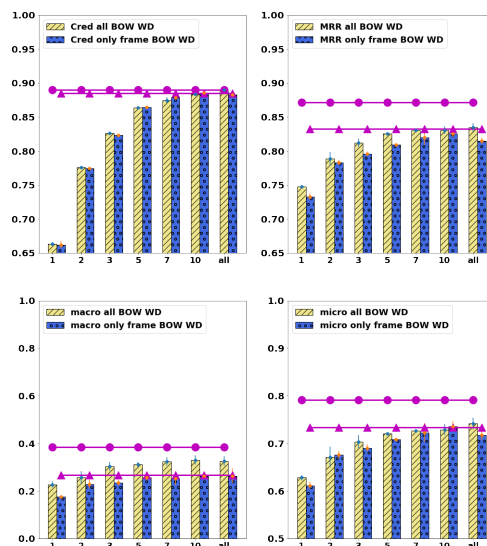
We also examine how we mapped Wikidata relations to FrameNet frames; manually constructing this mapping is time consuming. We explore constructing a mapping using a basic bag-of-words lexical overlapping among the description and aliases of Wikidata properties with lexical units of frames. In Figure 6(b) note none of the variants is able to match with expert mapping (filled circles and triangles) on Wikidata, demonstrating quality mappings are important, but competitive performance can be achieved automatically. Moreover, considering all sentences is better compared to choosing only frame based sentences. For TAC (not shown for space), choosing sentences based on frames give better support to ranking. The credibility task for TAC is improved when frame-based sentences are considered.

6 Conclusions

We described the KG Cleaner framework that can analyze facts produced by an IE system to perform two useful tasks: (1) identify facts that are likely to be incorrect and (2) suggest corrections for those thought to be wrong. It takes a standalone approach in which it only operates on the knowledge graph fragments and associated provenance text and has no knowledge of the IE system



(a) Performance when changing number of sentences are used and no FrameNet sentences are used. Change in performance when FrameNet is not used.



(b) Effect of changing the FrameNet mapping from an expert to a bag-of-word lexical overlap match. Target lines with *filler circle* and *triangle* indicate best performance using expert mapping for all provenance sentences and only frame net based provenance sentences, respectively.

Figure 6: The effect of using frame annotations at all (a), and how we map Wikidata and frames (b).

that produced the triples.

We evaluated our framework with our system and other instances on two large datasets: a collection of facts and false faux facts from Wikidata and a collection of facts produced by participants of the 2015 TAC Knowledge Base Population task. We plan to make our relations mappings, evaluation datasets and implementation available upon publication.

Acknowledgments

This research was partially supported by a gifts from the IBM AI Horizons Network and Northrop Grumman and by an NSF grant for UMBC’s high performance computing environment.

References

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The berkeley framenet project. In *36th Annual Meeting of the Association for Computational Linguistics*, pages 86–90. Association for Computational Linguistics.

Yoshua Bengio. 2012. Practical recommendations for gradient-based training of deep architectures. In *Neural networks: Tricks of the trade*, pages 437–478. Springer.

Matthias Brocheler, Lilyana Mihalkova, and Lise

Getoor. 2012. Probabilistic similarity logic. *arXiv preprint arXiv:1203.3469*.

Hoa Trang Dang, editor. 2015. *Proceedings of the Eighth Text Analysis Conference*. NIST.

Hoa Trang Dang, editor. 2017. *Proceedings of the 10th Text Analysis Conference*. NIST.

Aaron Defazio, Francis Bach, and Simon Lacoste-Julien. 2014. Saga: A fast incremental gradient method with support for non-strongly convex composite objectives. In *NIPS*.

Francis Ferraro, Max Thomas, Matthew R Gormley, Travis Wolfe, Craig Harman, and Benjamin Van Durme. 2014. Concretely annotated corpora. In *AKBC Workshop at NIPS*.

William Ferreira and Andreas Vlachos. 2016. Emergent: a novel data-set for stance classification. In *ACL*, pages 1163–1168.

Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. 2013. Ppdb: The paraphrase database. In *ACL*.

Xavier Glorot and Yoshua Bengio. 2010. Understanding the difficulty of training deep feedforward neural networks. In *AISTATS*.

Lukasz Kaiser, Aidan N Gomez, Noam Shazeer, Ashish Vaswani, Niki Parmar, Llion Jones, and Jakob Uszkoreit. 2017. One model to learn them all. *arXiv preprint arXiv:1706.05137*.

- Jens Lehmann, Daniel Gerber, Mohamed Morsey, and Axel-Cyrille Ngonga Ngomo. 2012. Defacto-deep fact validation. In *International Semantic Web Conference*. Springer.
- Xian Li, Weiyi Meng, and Clement Yu. 2011. T-verifier: Verifying truthfulness of fact statements. In *Data Engineering (ICDE)*. IEEE.
- Vinod Nair and Geoffrey E Hinton. 2010. Rectified linear units improve restricted boltzmann machines. In *27th international conference on machine learning*.
- Ndapandula Nakashole and Tom M Mitchell. 2014. Language-aware truth assessment of fact candidates. In *ACL*.
- Prakhar Ojha and Partha Talukdar. 2017. KGEval: Accuracy estimation of automatically constructed knowledge graphs. In *EMNLP*.
- Ayush Patwari, Dan Goldwasser, and Saurabh Bagchi. 2017. Tathya: A multi-classifier system for detecting check-worthy statements in political debates. In *Conference on Information and Knowledge Management*. ACM.
- Dean Pomerleau and Delip Rao. 2017. Fake news challenge. <http://fakenewschallenge.org/>.
- Jay Pujara, Hui Miao, Lise Getoor, and William Cohen. 2013. Knowledge graph identification. In *International Semantic Web Conference*, pages 542–557. Springer.
- Mehdi Samadi, Partha Pratim Talukdar, Manuela M Veloso, and Manuel Blum. 2016. Claimeval: Integrated and flexible framework for claim evaluation using credibility of sources. In *AAAI*, pages 222–228.
- Mehdi Samadi, Manuela M Veloso, and Manuel Blum. 2013. Openeval: Web information query evaluation. In *AAAI*.
- Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A simple way to prevent neural networks from overfitting. *JMLR*, pages 1929–1958.
- Ilya Sutskever, James Martens, George Dahl, and Geoffrey Hinton. 2013. On the importance of initialization and momentum in deep learning. In *International Conference on Machine Learning*.
- James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and verification. In *NAACL-HLT*.
- Robert Tibshirani. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288.
- Vidhoon Viswanathan, Nazneen Fatema Rajani, Yinon Bentor, and Raymond Mooney. 2015. Stacked ensembles of information extractors for knowledge-base population. In *ACL*.
- Andreas Vlachos and Sebastian Riedel. 2015. Identification and verification of simple claims about statistical properties. In *EMNLP*. Association for Computational Linguistics.
- Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*.
- Xiaoxin Yin, Jiawei Han, and S Yu Philip. 2008. Truth discovery with multiple conflicting information providers on the web. *IEEE TKDE*, 20(6).
- Dian Yu, Hongzhao Huang, Taylor Cassidy, Heng Ji, Chi Wang, Shi Zhi, Jiawei Han, Clare Voss, and Malik Magdon-Ismail. 2014. The wisdom of minority: Unsupervised slot filling validation based on multi-dimensional truth-finding. In *COLING*.