Assessing The Factual Accuracy of Generated Text

Ben Goodrich
Vinay Rao
Peter J. Liu
Mohammad Saleh
bgoodrich@google.com
vinaysrao@google.com
peterjliu@google.com
msaleh@google.com
Google Brain

ABSTRACT

We propose a model-based metric to estimate the factual accuracy of generated text that is complementary to typical scoring schemes like ROUGE (Recall-Oriented Understudy for Gisting Evaluation) and BLEU (Bilingual Evaluation Understudy). We introduce and release a new large-scale dataset based on Wikipedia and Wikidata to train relation classifiers and end-to-end fact extraction models. The end-to-end models are shown to be able to extract complete sets of facts from datasets with full pages of text. We then analyse multiple models that estimate factual accuracy on a Wikipedia text summarization task, and show their efficacy compared to ROUGE and other model-free variants by conducting a human evaluation study.

KEYWORDS
datasets, neural networks, fact extraction, deep learning, metric, end-to-end

ACM Reference Format:
Ben Goodrich, Vinay Rao, Peter J. Liu, and Mohammad Saleh. 2019. Assessing The Factual Accuracy of Generated Text. In The 25th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '19), August 4–8, 2019, Anchorage, AK, USA. ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3292500.3330955

1 INTRODUCTION

Recently, there has been wide empirical success in text summarization [15, 21, 27], machine translation [1, 36, 39], dialogue response generation [12, 28, 29], and other text generation tasks. For evaluation, these models generally rely on metrics like ROUGE (Recall-Oriented Understudy for Gisting Evaluation) [13], BLEU (Bilingual Evaluation Understudy) [23] and perplexity [3] that measure locally oriented text generation. For instance, these models generally rely on metric scores. For example, ROUGE-1 (1-gram overlap) measures 0.83. A real example is highlighted in Table 1 where the summarization model commits such a mistake.

A fact \( f \) is defined to be a relation tuple \( (subject, relation, object) \), where \( subject \) has a binary relation to \( object \) and can be assumed to have been inferred from text or a knowledge base, e.g. \( Barack Hussein Obama II (born August 4, 1961) is an American politician who served as the 44th President of the United States from January 20, 2009 to January 20, 2017 implies a set of facts such as \( (Barack Obama, president of, United States), (Barack Obama, born on, August 4 1961) \).

In this paper, we limit our scope to the task of evaluating text summarization. To evaluate a text summarization model, we compare the ground-truth summary text \( T \) and the generated summary \( G \). Let \( f_t, f_g \in F \) and \( F_T, F_G \subseteq F \) where \( F \) is a set of relation tuples.

\[
F_T = \{ f_t | f_t \text{ is inferred from ground-truth } T \} \\
F_G = \{ f_g | f_g \text{ is inferred from generated-text } G \}
\]

The models used in the metric we propose do not make use of world knowledge (e.g. knowledge base) during inference, and to account for that we filter \( F_T \) and \( F_G \) by only considering claims made in \( G \) that can either be verified or refuted by statements in \( T \). Concretely, if \( f_t = (subject_t, relation_t, object_t) \in F_T \) and \( f_g = (subject_g, relation_g, object_g) \in F_G \):

\[
F_T' = \{ f_t | \exists f_g \text{ and } subject_t = subject_g, relation_t = relation_g \} \\
F_G' = \{ f_g | \exists f_t \text{ and } subject_g = subject_t, relation_g = relation_t \}
\]

We can then define factual accuracy \( fact_{acc} \) as the precision between \( F_T' \) and \( F_G' \).

\[
\text{fact}_{acc} = \frac{|F_T' \cap F_G'|}{|F_G'|} \tag{1}
\]

For example, consider ground-truth summary \( T: Brad Pitt was born in 1963 and generated summary \( G: Brad Pitt was born in 1961. \) Then, \( F_T = \{(Brad Pitt, born-in, 1963)\}, F_G = \{(Brad Pitt, born-in, 1961)\}. \) The metric \( fact_{acc} = 0 \) indicates there is no factual consistency between the two summaries, whereas another metric like ROUGE-1 (1-gram overlap) measures 0.83. A real example is highlighted in Table 1 where the summarization model commits such a mistake. It is important to be able to measure these mistakes accurately to aid in training factually accurate summarization models.

Extracting fact tuples from text has been previously studied in methods like OpenIE (Open Information Extraction) [2]. OpenIE extracts triplets with an unspecified schema, and the relation is usually the text linking the two entities. However, it does not leverage information from a knowledge base and leads to outputs that are hard to compare. For example, \( Person \ was \ born \ in \ that \ town \Rightarrow \)
Table 1: Example of factual inaccuracy noted in a summarization model [15]. In this example, the summarization model uses the subject (Peter Duryea’s) father, Dan Duryea’s birthday.

| Target | Peter Duryea (July 14, 1939 – March 24, 2013) was an American actor. He is best known for appearing in a pilot episode of Star Trek: The Original Series, “The Cage” (1964), most of which was reused in “The Menagerie” (1966), as Lieutenant Tyler. His father, Dan Duryea (1907 – 1968), was also an actor. |
| Output | Peter Duryea (April 23, 1907 – March 24, 2013) was an American actor. He is best known for his role as Lt. Jose Tyler in the original Star Trek pilot, “The Cage” |

(related to the content of the image)
We also make use of our proposed dataset that is bigger, more diverse and has more relation types. Our dataset also has article-level information that can be used to train models like in Section 4.2. Since using two-step processes may be affected by compounding of errors across the models, some end-to-end approaches [18, 19] have been proposed, where the models extract entities and relations in one pass through the model. However, the method used in Miwa and Sasaki [19] required designing hand-crafted features and task-specific algorithms. Miwa and Bansal [18] has a two-phase model that first extracts entity candidates and then predicts relations based on the parsed tree-structure of the sentence. We instead propose a sequence-to-sequence model that is able to output fact tuples directly, and does not require any feature engineering.

We found that the abstractive summarization models such as those described in Liu et al. [15] may generate sentences with factual inaccuracies (e.g. incorrect month in date of birth, wrong city in the state, etc.). Cao et al. [4] found that 30% of summaries generated by a state-of-the-art abstractive summarization model contained factual inaccuracies. We found by running a large-scale experiment as described in Section 8.1, that the summarization model had factual inaccuracy rate of approximately 17%. We believe that this is because such mistakes are not heavily penalized by cross-entropy or n-gram based model loss metrics and losses.

As further motivation, we synthesized factually inaccurate samples by making simple corruptions to Wikipedia lead sections. We replaced mentions of dates (day and month only), locations or people with other entities of the same type in the text. For example, Barack was born on August 4, 1961 in Honolulu. He married Michelle on October 3, 1992 in Chicago. He married Michelle on August 4, 1992 in Honolulu. Table 2 shows that model-free metrics such as ROUGE and OpenIE-based tuple comparison do not reflect the decline in factual accuracy due to such corruption as much as the model-based metrics do.

### 4 MODEL-BASED METRICS

In this section we describe models that can extract fact tuples from text and how we use them to define the factual accuracy metric as defined in Eq 1. Given some input text $X$, we then extract claims made in $X$ as fact tuples.

#### 4.1 Named Entity Recognition (NER) + Relation Classifier

This approach consists of two steps, where we first recognize all the named entities $e_i$ from $X$ and then classify relations between entity pairs $(e_i, e_j)$.

##### 4.1.1 Named Entity Recognition

Entities are real-world objects like people, locations, organizations etc that can be identified by a proper name$^3$. Entities can be identified with name-entity recognition (NER) systems like Chiu and Nichols [5], Finkel et al. [7], Lample et al. [9] etc. NER is followed by co-reference resolution$^4$ [6, 11, 24, 25]. Publicly available NER and co-reference systems include Stanford’s CoreNLP$^5$ and NLTK$^6$.

##### 4.1.2 Relation Classifier

For each pair $(e_i, e_j) \neq (e_j, e_i)$ we consider all sentences $S_j$ in $X$ that contain both entities. The input to the classifier is then each of these sentences $S_j$. Because a sentence may contain multiple entities, we also add a prefix $SUBj$ for $e_i$ and OBJ to $e_j$ as a hint. For example, $X = \text{Person1 was born in City1 becomes Person2 was born in City2}$. Unlike Sorokin and Gurevych [32], our classifier does not require additional context. Let $s_i$ be a token in the input sentence $S_j$ after NER, and $r_k$ denote the $k$th relation. Our classifier takes in input tokens $s_i$ that are first embedded onto a latent space, and then a stack of Transformer encoder-only layers process the whole sequence. A subsequent max-pooling layer selects one of these outputs that is then converted to a probability estimate of relations by a sigmoid operation.

The exact series of operations can be viewed as:

\[
\begin{align*}
    w_{1:n} &= \text{embed}(s_{1:n}) \\
    h_{1:n} &= \text{transformer}_\text{encoder}(w_{1:n}) \\
    h_i &= \text{max} (h_{1:n}); h_i \in \mathbb{R}^k \\
    p(r_k) &= \frac{1}{1 + e^{-h_i}} = \text{sigmoid}(h_i)
\end{align*}
\]

$^1$https://en.wikipedia.org/wiki/Named_entity

$^2$While we use an NER and co-reference resolution system that is not available to the public, the dataset we release (Section 3) has the positions of all the recognized and resolved entities that we use for training our classifier.

$^3$https://stanfordnlp.github.io/CoreNLP/coref.html

$^4$https://www.nltk.org/
We propose an end-to-end fact extraction model to avoid compounding of errors across components in multi-stage approaches like Section 4.1 [16]. This model also does not require any feature engineering or context. The input to the model is text \( X \) of any length (sentence/paragraph/article) and the subject entity \( e_i \) prefixed to \( X \). All the inputs in \( \{e_i; X\} \) are first embedded onto a latent space. A Transformer model consisting of a stack of encoder layers followed by decoder layers produces an output sequence of fact tuples. For example, if \( \langle Person1; born in; Country1; Painter \rangle \), we are able to extract a set of tuples of the form \( \langle e_i, rel, e_j \rangle \) from both \( T \) and \( G \). To use eq 1 to define the factual accuracy, we filter the set by considering only entity pairs \( \{e_i, e_j\} \) that are found in both \( T \) and \( G \) to then compare the predicted label \( rel \) between them.

5 MODEL-FREE METRICS

We describe model-free automatic metrics in this section. Unlike model-based metrics, they are not susceptible to changes in training data, and might be considered easier to interpret or understand.

5.1 ROUGE

ROUGE [13] has been used as an automatic metric to judge the quality of generated text, and has shown to correlate well with human judgment of overall linguistic quality of the text.

5.2 OpenIE

OpenIE [2] is a tool that can extract relation tuples from text, without a specified schema. We use it to extract sets of relation tuples from \( T \) and \( G \), and then compute the precision like in eq 1.

6 MODEL EXPERIMENTS

In this section, we describe the methods we used to train and evaluate our relation extraction models. All of our proposed classifiers and end-to-end models have 6 Transformer layers and 1 embedding layer, with number of neurons (hidden layer size) set to 512. In the Transformer-based models, we use 8 attention heads. Our models are trained using the AdaFactor [30] optimizer. We use the publicly available TensorFlow2Tensor [35] framework for our experiments and will be releasing our code extensions as part of that framework. On our proposed dataset, the classifiers are trained for 50,000 iterations with batch-size of 1024 and the end-to-end models are trained for 50,000 iterations with batch-size of 256. We evaluate classifiers and end-to-end models on our dataset. These results are presented in Table 3. The end-to-end model is learning to recognize entities, resolving entity co-references, and reason
about their relation in one pass through the model. To the best of our knowledge, we are not aware of other end-to-end structured relation extraction models and therefore do not include a comparison against other approaches. Some examples of extracting facts on our dataset are shown in A.2, where we include a comparison to OpenIE’s triplet extraction.

We calculate precision and recall in the above experiments by matching ground-truth fact tuples exactly. This implies that the end-to-end model is not only learning to identify entities and resolve co-references, but also predict structured output, and its outputs can be used for reasoning. Their performance is competitive against relation classifiers while having a simple training and inference routine.

For each model, we sort and select the ten most frequent relation types that appear in our test sets. The $F_1$ measure on these relations for classifiers are shown in Table 4, and end-to-end models are shown in Table 5.

| Relation                      | P     | R     | F1    |
|-------------------------------|-------|-------|-------|
| No relation                   | 0.9830| 0.9817| 0.9824|
| Country of citizenship        | 0.6446| 0.9394| 0.7646|
| Date of birth                 | 0.9330| 0.9850| 0.9582|
| Country                       | 0.6049| 0.9484| 0.7386|
| Located in territory          | 0.6260| 0.8118| 0.7069|
| Instance of                   | 0.5097| 0.7015| 0.6094|
| Place of birth                | 0.6430| 0.7436| 0.6897|
| Member of sports team         | 0.5179| 0.9248| 0.6640|
| Occupation                    | 0.6934| 0.7770| 0.7354|
| Date of death                 | 0.9163| 0.9875| 0.9506|

Table 4: Precision (P), Recall (R) and $F_1$ measure of the relation classifier (Section 4.1) on our test sets on ten most frequent relations.

| Relation                              | P     | R     | F1    |
|---------------------------------------|-------|-------|-------|
| Country of citizenship                | 0.8247| 0.8359| 0.8302|
| Instance of                           | 0.7122| 0.6676| 0.6934|
| Date of birth                         | 0.9342| 0.9798| 0.9564|
| Country                               | 0.8387| 0.8267| 0.8327|
| Cast member                           | 0.5889| 0.4910| 0.5355|
| Place of birth                        | 0.7012| 0.7348| 0.7176|
| Located in the administrative territorial entity | 0.7293| 0.7700| 0.7491|
| Member of sports team                 | 0.7405| 0.7027| 0.7036|
| Occupation                            | 0.5911| 0.5774| 0.5842|
| Educated at                           | 0.5432| 0.7278| 0.6221|

Table 5: Precision (P), Recall (R) and $F_1$ measure of our end-to-end model (Section 4.2) on our test sets on ten most frequent relations.
7 ERROR ANALYSIS OF MODEL PREDICTIONS

Distant supervision [17] is a way to create training data by using weak signals. In our dataset, we assign a relation label \( r_k \) for every entity pair \((e_i, e_j)\) in the input text \(X\) if the relation tuple \((e_i, r_k, e_j)\) exists in the Wikidata knowledge base \(W_{KB}\). However, the sentence \(S_l\) containing \((e_i, e_j)\) may not necessarily entail \(r_k\). This leads to inaccurate estimates of the true-positive rate for our fact extraction models. We evaluate the effect of this distant supervision by gathering the set of facts extracted from our models that are marked false-positive by the distant supervision scheme. We present a pair of input text (Wikipedia articles) and facts extracted by our models to human evaluators, and ask them to mark a fact to be True only if the relation tuple \((subject, relation, object)\) is implied by the input text. We asked two evaluators to score facts marked false-positive from a random set of 30 Wikipedia articles. We consider the fact to be true if both evaluators agree. We present the results in Table 6, where we can see the rate of false-positive facts that were marked true by the evaluators. This suggests that the end-to-end models could benefit from a better labeling scheme.

| Model                  | % True-positives |
|------------------------|------------------|
| End-to-end             | 77.8             |
| Relation Classifier    | 46.6             |

Table 6: Percentage of true facts that were inaccurately labelled wrong by the distant supervisor. The End-to-end model is the best model from Section 4.2 and Classifier is the best from 4.1 (Transformer-Sigmoid). The End-to-end model (in bold) predicts facts that are likelier to be true.

8 EVALUATION OF \( f_{acc} \) AS A METRIC

In this section, we show the effectiveness of our proposed metric on judging the factual accuracy of generated text. We use the text summarization model proposed in [15] to generate lead sections of Wikipedia articles using the dataset and model in that paper, and compare the generated summary against the real lead section. In the following section, we describe the methodology used to compare human judgment of factual accuracy and how we compare our metric against that baseline.

8.1 Human Evaluation

Every claim made in the generated text \(G\) can be considered to belong to one of three categories: supported by a sentence in ground-truth \(T\), refuted by \(T\) or cannot be verified by \(T\). The evaluators were asked to only consider claims that are either supported or refuted by \(T\). This ensures that no external knowledge is used in comparing \(T\) and \(G\), and ignores all claims that cannot be verified by \(T\). Four evaluators were asked to rate 30 examples of generated text \(G\) and then give it a score of 1-5 with 5 being highest factual accuracy. A special case is where the generated text has no verifiable claims. In this case, they were asked to give it a score of 1. Figure 2 shows the interface a human evaluator uses in our experiment.

We conduct the same experiment on two sets of data: first is a random sampling from summaries generated for Actors. We consider this an easier subset because we expect our fact extraction models to do well on this subset due to the summaries and Wikipedia lead sections generally containing relationships our models perform well on (see tables 4 and 5). We present these results in Table 7. We analyzed the inter-rater agreement on the scores given to each example, and found that Krippendorff’s alpha (allows for ordinal rankings) was 0.6897. The second is a random sampling from all categories in Wikipedia. The results are presented in Table 8. The inter-rater agreement on this sample was found to be 0.7530.

We see that our end-to-end model (Section 4.2) has the best correlation on both subsets, indicating that it generalizes better to generated text. This may also be because the classifier suffers from a compounding of errors, where it is unable to predict relations if the NER system fails to recognize entities.

| Metric                  | Correlation with human scores |
|-------------------------|------------------------------|
| ROUGE-1                 | 0.583                        |
| ROUGE-2                 | 0.639                        |
| ROUGE-L                 | 0.634                        |
| OpenIE                  | 0.258                        |
| \( f_{acc} \)-Binary Classifier | 0.596                     |
| \( f_{acc} \)-Relation Classifier | 0.523                   |
| \( f_{acc} \)-E2E       | 0.645                        |
| \( f_{acc} \)-E2E-Reduced | 0.668                     |

Table 7: Spearman correlation of different metrics with human evaluation of factual accuracy on the ‘Actors’ subset of summaries. ROUGE and OpenIE are described in Sec 5, and the model-based \( f_{acc} \) metrics are described in Sec 4. The best metric is shown in bold.

| Metric                  | Correlation with human scores |
|-------------------------|------------------------------|
| ROUGE-1                 | 0.384                        |
| ROUGE-2                 | 0.435                        |
| ROUGE-L                 | 0.339                        |
| OpenIE                  | 0.128                        |
| \( f_{acc} \)-Binary Classifier | 0.200                     |
| \( f_{acc} \)-Relation Classifier | 0.250                   |
| \( f_{acc} \)-E2E       | 0.314                        |
| \( f_{acc} \)-E2E-Reduced | 0.453                     |

Table 8: Spearman correlation of different metrics with human evaluation of factual accuracy on a random subset of summaries. ROUGE and OpenIE are described in Sec 5, and the model-based \( f_{acc} \) metrics are described in Sec 4. The best metric is shown in bold.

9 CONCLUSION

9.1 Limitations

The dataset we create only makes use of sentences found in Wikipedia, and facts found in WikiData. This means that our models are biased to sentences structured to the neutral tone set in Wikipedia, and towards popular types of facts expressed in WikiData such as date of birth, profession, etc. Other sources of text may have more complex structures and styles of writing that may make it hard for our
models to adapt to easily. An simple example of this is negating a binary relationship with ‘not’, and different ways of expressing the same idea such as ‘wife/husband’ instead of ‘spouse’. WikiData is an incomplete knowledge base, and this also leads to many sentences that in reality imply a fact to be marked containing no facts. This is a very typical problem faced by any work using distant supervision, and is combated with methods like active learning [31]. It should be noted that ROUGE and to the best of our knowledge, most other automatic metrics, are also susceptible to changes in linguistic style and structure. However, elaborate labeling and bigger datasets will allow for our models to learn to overcome these challenges.

9.2 Discussion and future work
We have shown that our proposed metric is able to indicate the factual accuracy of generated text, and agrees with human judgment on our datasets. By leveraging a new dataset for both relation classification and end-to-end fact extraction, we also showed that classifiers and end-to-end models with straightforward architectures are able to perform competitive fact extraction. Our end-to-end model avoids compounding of errors over sub-components typically used in other fact-extraction pipelines. We will release the code and datasets used to train this model, so that the proposed metric can be used to standardize comparison. We are in the process of building a bigger dataset that will contain multiple text domains, stronger human supervision and a larger collection of relation tuples that will help overcome many of the limitations discussed in the previous section (9.1). We encourage further development and use of this metric for automating the assessment of factual accuracy of generated text, and the development of better end-to-end models with structured outputs for fact extraction.

REFERENCES
[1] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural Machine Translation by Jointly Learning to Align and Translate. In International Conference on Learning Representations.
[2] Michele Banko, Michael J. Cafarella, Stephen Soderland, Matt Broadhead, and Oren Etzioni. 2007. Open Information Extraction from the Web. In Proceedings of the 20th International Joint Conference on Artificial Intelligence. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2670–2676.
[3] Peter F. Brown, Vincent J. Della Pietra, Robert L. Mercer, Stephen A. Della Pietra, and Jennifer C. Lai. 1992. An Estimate of an Upper Bound for the Entropy of English. Computational Linguistics 18, 1 (March 1992), 31–40.
[4] Ziqiang Cao, Furu Wei, Wenjie Li, and Sujian Li. 2017. Faithful to the Original Fact Aware Neural Abstractive Summarization. CoRR abs/1711.04434 (2017). arXiv:1711.04434 http://arxiv.org/abs/1711.04434
[5] Jason Chiu and Eric Nichols. 2016. Named Entity Recognition with Bidirectional LSTM-CNNs. Transactions of the Association for Computational Linguistics 4 (2016), 357–370.
[6] Kevin Clark and Christopher D. Manning. 2016. Improving Coreference Resolution by Learning Entity-Level Distributed Representations. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, 643–653. https://doi.org/10.18653/v1/P16-1061
[7] Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. In Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL 2005) 363–370.
[8] Eduard Hovy, Chin-Yew Lin, Liang Zhou, and Junichi Fukumoto. 2006. Automated Summarization Evaluation with Basic Elements. In Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC’06).
[9] Guillaume Lample, Miguel Ballesteros, Sander Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural Architectures for Named Entity Recognition. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, 260–270. https://doi.org/10.18653/v1/N16-1030
[10] Alon Lavie and Abhaya Agarwal. 2007. Meteor: An Automatic Metric for MT Evaluation with High Levels of Correlation with Human Judgments. In Proceedings of the Second Workshop on Statistical Machine Translation. Association for Computational Linguistics, Stroudsburg, PA, USA, 228–231.
[11] Heeyoung Lee, Yves Perseman, Angel Chang, Nathanael Chambers, Mihai Surdeanu, and Dan Jurafsky. 2011. Stanford’s Multi-Pass SaCcore Coreference Resolution System at the CoNLL-2011 Shared Task. In In Proceedings of the CoNLL-2011 Shared Task.
[12] Jiwei Li, Will Monroe, Tianlin Shi, Sébastien Jean, Alan Ritter, and Dan Jurafsky. 2017. Adversarial Learning for Neural Dialogue Generation. In Conference on Empirical Methods in Natural Language Processing, 2157–2169.
[13] Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In Text Summarization Branches Out: Proceedings of the ACL-04 Workshop. Stan Szpakowicz Marie-Francine Moens (Ed.). Association for Computational Linguistics, Barcelona, Spain, 74–81.
[14] 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, Stroudsburg, PA, USA, 228–231.
[15] Peter J. Liu, Mohammad Saleh, Etienne Pot, Ben Goodrich, Ryan Sepassi, Lukasz Kaiser, and Noam Shazeer. 2018. Generating Wikipedia by Summarizing Long Sequences. In Proceedings of the 2018 International Conference on Learning Representations.
[16] Andrew Mccallum and David Jensen. 2003. A Note on the Unification of Information Extraction and Data Mining using Conditional-Probability, Relational Models. In In Proceedings of the IJCAI-2003 Workshop on Learning Statistical Models from Relational Data.
A APPENDIX

A.1 Reproducibility

We release code to train our fact extraction models as part of the Tensor2Tensor framework along with trained model weights at https://github.com/tensorflow/tensor2tensor/tree/master/tensor2tensor/data_generators/wikifact. A large fact extraction dataset (Sec 3) based on Wikidata and Wikipedia is made available. To train our end-to-end and classifier models for fact extraction, we use the hyper-parameter set “transformer_base” defined in the Tensor2Tensor framework. We further release code to use our end-to-end models as a fact extractor and calculate the factual accuracy metric at https://github.com/tensorflow/tensor2tensor/tree/master/tensor2tensor/data_generators/wikifact.

A.2 Fact extraction example

We include an example of facts extracted from text using our models where we compare it against OpenIE’s [2] triplet extraction in Table 9. This example illustrates the advantage of using structured approaches to fact extraction. OpenIE yields many triplets that mostly cannot be used for reasoning.

---

11https://github.com/tensorflow/tensor2tensor
12https://github.com/tensorflow/tensor2tensor/tree/master/tensor2tensor/data_generators/wikifact
13https://github.com/tensorflow/tensor2tensor/blob/master/tensor2tensor/models/transformer.py
Christopher Simon (born 5 June 1963) is an Australian actor and producer. Born in Sydney, Australia. He produced the film Miss You Already directed by Catherine Hardwicke. Simon is also a producer of such films as The Sweeney (2012 film) directed by Nick Love, Pusher, I, Anna, Still Life, Me and Me Dad, Boogie Woogie, The Proposition, Beyond the Ocean, The Trouble with Men and Women. He also produced short films by Joe Wright such as The End and Nick Love’s Love Story. Simon’s various television acting roles include Eddie in The Long Firm, Pedro in Gimme Gimme Gimme, Michael Hassan in The Bill, Lee Andersen in Casualty, Abdel in Lovejoy Samir in Ultimate Force, Da Souza in Lynda La Plante’s Supply and Demand, Nathan Morgan in Wire In The Blood and he appeared in Lenny Henry in Pieces. Film acting roles include Room To Rent, The Delivery and O Jerusalem. Simon has acted in such plays as 12 Angry Men and Taking Sides both directed by Harold Pinter in London’s west end, The Kitchen directed by Stephen Daldry at the Royal Court, the Amnesty award winning one man show When The Bulbull Stopped Singing for which he was nominated for the Acting Excellence Award (Best Actor) at the Edinburgh Festival Fringe, which premiered at the Traverse theatre and toured to Iran, New York and Jordan. Other theatre roles include Welcome to Ramallah, which toured York and London, at the Arcola and the Theatre Royal York, The Present at the Royal Court and the Bush, and Poor Superman at the Hampstead and the Traverse.

| Input | Christopher Simon (born 5 June 1963) is an Australian actor and producer. Born in Sydney, Australia. He produced the film Miss You Already directed by Catherine Hardwicke. Simon is also a producer of such films as The Sweeney (2012 film) directed by Nick Love, Pusher, I, Anna, Still Life, Me and Me Dad, Boogie Woogie, The Proposition, Beyond the Ocean, The Trouble with Men and Women. He also produced short films by Joe Wright such as The End and Nick Love’s Love Story. Simon’s various television acting roles include Eddie in The Long Firm, Pedro in Gimme Gimme Gimme, Michael Hassan in The Bill, Lee Andersen in Casualty, Abdel in Lovejoy Samir in Ultimate Force, Da Souza in Lynda La Plante’s Supply and Demand, Nathan Morgan in Wire In The Blood and he appeared in Lenny Henry in Pieces. Film acting roles include Room To Rent, The Delivery and O Jerusalem. Simon has acted in such plays as 12 Angry Men and Taking Sides both directed by Harold Pinter in London’s west end, The Kitchen directed by Stephen Daldry at the Royal Court, the Amnesty award winning one man show When The Bulbull Stopped Singing for which he was nominated for the Acting Excellence Award (Best Actor) at the Edinburgh Festival Fringe, which premiered at the Traverse theatre and toured to Iran, New York and Jordan. Other theatre roles include Welcome to Ramallah, which toured York and London, at the Arcola and the Theatre Royal York, The Present at the Royal Court and the Bush, and Poor Superman at the Hampstead and the Traverse. |
| Targets | (Christopher Simon, date of birth, June 5 1963), (Christopher Simon, country of citizenship, Australian), (Christopher Simon, place of birth, Sydney) |
| OpenIE | (Abdel, is in, Ultimate Force), (Casualty, Abdel in, Ultimate Force), (Nathan Morgan, is in, Blood), (Lee Andersen, is in, Casualty), (Da Souza, is in, Lynda La Plante’s Supply), (Simon’s various television acting roles, include, Eddie), (Simon, is producer of, films as Sweeney directed by Nick Love), (Simon, is also producer of, such films as Sweeney), (Simon’s television roles, include, Eddie in Firm), (Simon, is producer of, such films), (Simon’s various television roles, include, Eddie), (Eddie, is in, Long Firm), (Simon, is producer of, such films as Sweeney), (Simon, is producer of, such films as Sweeney directed by Nick Love), (Michael Hassan, is in, Bill), (Bill, Andersen in, Casualty), (You, Already directed by, Catherine Hardwicke), (Simon, is producer of, films), (Simon, has, various television acting roles), (Simon’s television acting roles, include, Eddie), (Abdel, is in, Lovejoy Samir), (Simon’s television roles, include, Eddie in Long Firm), (Simon’s television acting roles, include, Eddie in Firm), (Simon, is producer of, films as Sweeney directed), (Simon’s various television acting roles, include, Eddie in Firm), (Simon, is also producer of, films as Sweeney), (Simon’s various television acting roles, include, Eddie in Long Firm), (Simon, is, producer), (Rent, To Room is, Delivery), (Simon’s television roles, include, Eddie), (Simon, is also producer of, films as Sweeney directed), (Lynda La Plante, in, Supply), (Pedro, is in, Gim), ... |
| Seq2Seq | (Christopher Simon, date of birth, June 5 1963), (Christopher Simon, country of citizenship, Australian), (Christopher Simon, place of birth, Sydney), (Christopher Simon, occupation, Actor) |
| Classifier | (Christopher Simon, date of birth, June 5 1963), (Christopher Simon, country of citizenship, Australian) |

Table 9: Comparison of fact tuples extracted from this example, using OpenIE, our end-to-end model (Section 4.2), and our classifier (Section 4.1). A triplet consists of (subject, relation, object).