Entity-Based Semantic Adequacy for Data-to-Text Generation

Juliette Faille
Université de Lorraine
CNRS/LORIA
juliette.faille@loria.fr

Albert Gatt
Utrecht University
a.gatt@uu.nl

Claire Gardent
CNRS/LORIA
Université de Lorraine
claire.gardent@loria.fr

Abstract

While powerful pre-trained language models have improved the fluency of text generation models, it remains difficult to ensure that the generated texts are semantically faithful to the input. In this paper, we introduce a novel automatic evaluation metric, Entity-Based Semantic Adequacy, which can be used to assess to what extent generation models that verbalise RDF (Resource Description Framework) graphs produce text that contains mentions of the entities occurring in the RDF input. This is important as RDF subject and object entities make up 2/3 of the input. We use our metric to compare 25 models from the WebNLG Shared Tasks and we examine correlation with results from human evaluations of semantic adequacy. We show that while our metric correlates with human evaluation scores, this correlation varies with the specifics of the human evaluation setup. This suggests that in order to measure the entity-based adequacy of generated texts, an automatic metric such as the one proposed here might be more reliable, as less subjective and more focused on correct verbalisation of the input, than human evaluation measures.

1 Introduction

With the introduction of pretrained models, the fluency of text generation systems has improved. However, semantic adequacy (faithfulness to the input) remains an unsolved issue. It remains difficult to ensure that the generated text faithfully captures the input (Wiseman et al., 2017; Gehrmann et al., 2018).

In this paper, we focus on semantic adequacy for RDF-Verbalisers i.e., models such as those submitted to the WebNLG 2017 and 2020 shared tasks (Gardent et al., 2017; Castro Ferreira et al., 2020b) which map an RDF graph to a text verbalising the content of that graph. In this case, the input to Natural Language Generation (NLG) is a set of triples of the form $(e1, p, e2)$ where $e1, e2$ are RDF entities and p is a property. RDF triplestores are used in particular to model Semantic Web data and their verbalisation aims at making the information from these knowledge-bases easily accessible to users. As exemplified in Figure 1, one necessary condition for the generated text to be semantically adequate is that all entities present in the input should be mentioned at least once in the output. We refer to this requirement as entity-based semantic adequacy (ESA for short). ESA offers one way of formalising the requirement that the output of a generator should reflect the information in the input. Thus, its significance extends beyond the specific problem domain of RDF verbalisation, though the latter provides a useful testcase.

We make the following contributions.

We devise metrics which assess to what extent a text verbalising an RDF graph respects entity-based semantic adequacy. These metrics rely on an algorithm designed to automatically detect whether an entity present in the input graph has a corresponding mention in the output text. We evaluate this algorithm on a corpus of 25,173 (RDF, Text) pairs with manually annotated entity mentions from Castro Ferreira et al. (2018) and show that our algorithm has a recall of 0.74 and a precision of 0.75.

We apply these metrics to the output of 25 RDF verbalisers developed for the WebNLG 2017 and 2020 challenges and show that some of the systems which rank highest in terms of BLEU scores actually rank in the lower half with respect to entity-based semantic adequacy. This indicates that ESA is measuring a different quality from that measured by surface-based metrics such as BLEU.

We compute correlation between our metrics and both automatic and human evaluation scores collected by the WebNLG organisers for these 25 models. We find a stronger correlation with human metrics related to semantic adequacy than with automatic metrics. Among the automatic metrics,

Findings of the Association for Computational Linguistics: EMNLP 2021, pages 1530–1540
November 7–11, 2021. ©2021 Association for Computational Linguistics

1 (**short**)	**Output text** Liselotte Grschebina is a German national who was born in the German Empire and has a total area of 20769100000. 0.			
		Liselotte_Grschebina	nationality	Israel
		Israel	areaTotal	20769100000.0
	RDF Input	Israel	officialLanguage	Modern_Standard_Arabic
		Liselotte_Grschebina	birthPlace	German_Empire
		Liselotte_Grschebina	training	School_of_Applied_Arts_in_Stuttgart
2 (**hal**)	**Output text** Born in the Kingdom of England in 1726-01-01, and living in India, on the 18th of July, 1776, the country is the birth place of Joh Davutoglu.			
		Lady_Anne_Monson	birthPlace	Darlington
		Lady_Anne_Monson	birthDate	1726-01-01
	RDF Input	Lady_Anne_Monson	deathDate	1776-02-18
		Lady_Anne_Monson	birthPlace	Kingdom_of_England
		Lady_Anne_Monson	residence	India
3 (**deg**)	**Output text** The distributor of the distributor of the distributor of the distributor of the distributor of the distribution of the distribution of the distribution of the distribution of the dish, Roadside Attrón, is Tom Botta, who starred in the preparation of the tennis Katzman.			
		Super_Capers	editing	Stacy_Katzman
		Super_Capers	starring	Michael_Rooker
	RDF Input	Super_Capers	starring	Tom_Sizemore
		Super_Capers	language	English_language
		Super_Capers	distributor	Roadside_Attractions

Figure 1: Examples of outputs with low Entity-based Semantic Adequacy. RDF input entities that are missing in the text are underlined (short: the short output fails to mention all input entities, deg: degenerate output, hal: the text hallucinates entities not present in the input and omits to mention others)

correlations are highest with METEOR. Interestingly, we also find that the correlation with human scores varies with the specifics of the human evaluation setup. This suggests that our automatic metric might be a more reliable means of identifying models with low entity-based semantic adequacy than human evaluation. We are publicly releasing our source code. [1]

2 Related work

Various methods have been proposed to evaluate the semantic adequacy of generated texts.

Commonly-used metrics are surface-based (either word- or character-based) such as BLEU (Papineni et al., 2002), TER (Snover et al., 2006) or chrF (Popović, 2015). As these methods fail to account for paraphrases, alternative metrics have been proposed such as METEOR (Banerjee and Lavie, 2005), which measures n-gram overlap but integrates synonyms and BERTscore, a trained metric based on word-embeddings similarity (Zhang* et al., 2020). Semantic similarity has also been modeled in terms of propositional content. In computer vision for instance, SPICE transforms both generated and reference captions into a scene graph encoding the objects and relations present in these captions and computes an F-score over the semantic propositions in the scene graph (Anderson et al., 2016). Similarly, the MEANT metric applies Semantic Role Labelling to generated and reference texts and computes similarity by matching the resulting semantic frames. In data-to-text generation, (Dhingra et al., 2019) uses custom entailment models to determine whether an n-gram in the generated text is entailed by the input and computes an F-score based on these n-grams. In text summarisation, Goodrich et al. (2019) compare relation tuples extracted from a ground-truth summary and a generated one using either a Named entity Recognition and a Relation Classifier or an end-to-end Transformer model to extract these tuples.

Rather than abstract over the lexical content of the generated and reference text, other work has focused on developing metrics which model human judgement in particular, judgments of semantic similarity. Thus Sellam et al. (2020) introduced BLEURT, an automatic metric pre-trained on synthetic and automatically rated data and fine-tuned on human judgments.

Closest to our approach are metrics which evaluate the generated output, not with respect to the reference or to human judgments, but with respect to the input. Wiseman et al. (2017) define Relation Generation score as the precision of input relations found in the output texts (the relation ex-

[1] https://gitlab.nl4xai.eu/juliette.faille/entity-based-semantic-adequacy

traction is performed by a neural model). Reed et al. (2018) define information extraction patterns to measure the occurrence of the input attributes and their values in the outputs and compute semantic adequacy using the Slot Error Rate. Ribeiro et al. (2020); Dušek and Kasner (2020) use natural language inference (NLI) to detect two way entailment between the generated text and the input. Sulem et al. (2020) introduce SAMSA which assesses simplification quality by comparing the predicate/argument structures contained in the input with those contained in the output summary.

Similarly, we evaluate the semantic adequacy of a generated text by comparing it with the input. We focus on entities however and provide a detailed assessment of both the reliability of our metrics and its correlation with human and with automatic metrics.

3 Defining E-Based Semantic Adequacy

We assume a corpus of (R, T) instances where R is an RDF graph (a set of RDF triples) and T is a text verbalising that graph. RDF triples are of the form (s, p, o) where p is a binary relation holding between a subject (s) and an object (o)[2]. We use the term "entity" to refer to both RDF triple subjects and objects and we write E_R for the set of RDF entities occurring in RDF graph R.

Entity Mentions. Given a corpus instance (R, T), an *entity mention* m is a text segment in T which denotes an entity e present in the input graph ($e \in E_R$). We write M_T for the set of entity mentions occurring in T and $\llbracket\ m\ \rrbracket = e$ to indicate that the mention $m \in T$ denotes entity $e \in E_R$.

(Un)Detected Entities. A detected entity $e \in E_R$ is an entity which has a matching mention in M_T i.e., there is a mention $m \in M_T$ such that $\llbracket\ m\ \rrbracket = e$. Conversely, an undetected entity is an entity $e \in E_R$ which has no corresponding mention in M_T. We define $E_T \subseteq E_R$ as the set of RDF entities which have a corresponding mention in T.

Entity-Based Semantic Adequacy. Given an (R, T) pair, we define entity-based semantic adequacy (ESA) as the proportion of RDF entities in E_R which have a corresponding mention $m \in M_T$. In other words, ESA is the ratio between the number of entities for which a mention was found (E_T) and the total number of entities occurring in the input RDF (E_R).

$$\text{ESA}_I = \frac{\mid E_T \mid}{\mid E_R \mid}$$

Given a corpus of (R, T) pairs, we also compute the proportion of texts in that corpus with at least n undetected entities. We refer to this metric as corpus-level, entity-based semantic inadequacy at n ($\text{ESI}_C{}^n$ for short).

4 Computing E-Based Semantic Adequacy

The metrics introduced in the previous section rely on being able to determine which entities in the RDF input have a matching mention in the corresponding text. We present an algorithm for entity mention detection and we report on an evaluation of that algorithm using a dataset of 25,173 (RDF graph, Text) pairs where entity mentions have been manually annotated.

4.1 Detecting Entity Mentions

We define our entity mention detection algorithm using a combination of existing tools and heuristics.

Entity linker We use the state-of-the-art REL entity linker from van Hulst et al. (2020). When applied to a text, REL returns a list of entity mentions and their corresponding DBPedia entities. We filter out the mentions for which the related DBPedia entity does not match any of the input RDF entities.

Approximate string matching of text n-grams and RDF entities We match text n-grams to candidate RDF entities, using approximate string matching with a fixed maximum allowed edit distance (normalized Levenshtein distance). This value is experimentally fixed at 0.4 [3]. To improve results, we create a dictionary of RDF entity synonyms and compute the approximate match between text n-grams and all RDF entities, including their synonyms. The synonym dictionary was initially created using DBPedia aliases and rules (handcrafted to improve the detection of frequent entities in the WebNLG corpus such as places or quantities). During the evaluation of the entity detection (cf section 4.2), an expanded version of

[2] Subjects are Uniform Resource Identifiers (URI) and objects are either URIs or literals. Intuitively, RDF subjects and objects refer to things such as persons, locations, abstract entities, dates or phone numbers.

[3] The algorithm used by our string matching procedure is described in detail in Algorithm 1 in the Appendix.

the dictionary was developed by updating it with entities for which no mention was detected in an evaluated text; these were manually included in the dictionary. This dictionary provides a symbolic means to improve entity mention detection and more generally, to adapt the algorithm to a new domain. However, it should be noted that, on the WebNLG 2017 dataset, adding this dictionary only slightly improves entity mention detection and is not essential.

Pronominal entity mentions In order to detect which input entity a pronoun refers to, we use two methods. We first use `Stanza` (Qi et al., 2020) to compute co-reference chains in our texts, keeping only the pronominal mentions. For the pronouns that were not detected by the previous method, we used a simple heuristic. In the WebNLG corpus, RDF graphs are created with a single entity as "root". Other entities are meant to describe and provide information about this root. As the texts are quite short we assume that most pronominal anaphors refer to the "root" of the graph. We therefore associate all remaining pronouns to the root entity of the RDF graph.

Dates We use the python library `dateparser` to normalise dates both in the text and in the RDF. Results are further filtered using entity type information from the input RDF graph.

Putting it all together. Each method described above yields a list of mentions found in a text for each RDF entity in an input graph. In case different methods identify the same (or overlapping) mentions, we select those mentions with the lowest edit distance to their matched RDF entity (or one of its synonyms). In case of equality, we keep the longest mention.

4.2 Evaluating automatic entity mention detection

Castro Ferreira et al. (2018) manually annotated entity mentions in the WebNLG 2017 dataset (an example of annotation is shown in the Appendix). We use these manual annotations as gold standard to evaluate our entity mentions detection algorithm. Given M^{auto}, the set of mentions detected on this corpus by our mention detection algorithm and M^{human}, the set of manually annotated mentions, we compute Recall and Precision in the usual way: $Recall = \frac{|M^{auto} \cap M^{human}|}{|M^{human}|}$ and $Precision = \frac{|M^{auto} \cap M^{human}|}{|M^{auto}|}$. The intersection $M^{auto} \cap M^{human}$ is the number of exact string matches between the sets of mentions M^{auto} and M^{human}. We obtain a recall of 0.74 and a precision of 0.75.

If we consider approximate string matching in the computation of $M^{auto} \cap M^{human}$ with a maximum allowed normalized edit distance of 0.2, we obtain a recall of 0.82 and a precision of 0.83. This shows that although some of the automatically detected mentions do not match the gold standard annotations exactly, they are nonetheless close to them. In Section 5.2 below, we show that even though imperfect, our entity mention detection algorithm permits reliably identifying models which have low entity-based semantic adequacy.

5 Evaluating RDF-to-text Generation Models

25 models participated in the WebNLG 2017 and 2020 challenges. We apply our entity-based semantic adequacy metrics to the output of these models on the WebNLG 2017 and 2020 test data[4]. We group models with respect to BLEU and $\text{ESI}_C{}^1$ rank. As the text output by the models might differ from the crowdsourced texts we used for the evaluation presented in Section 4.2, we report on a manual verification of our entity mention detection algorithm on a sample from these model outputs. Finally, we show some example outputs illustrating different ways in which a generated text might have low entity-based semantic adequacy.

5.1 Entity-Based Semantic Adequacy in the WebNLG Shared Tasks

For each model in the WebNLG 2017 and 2020 Shared Tasks, we compute the $\text{ESI}_C{}^1$ score (proportion of texts with one or more RDF entities lacking a matching text mention) and the ESA_I score (proportion of RDF entities in the input with a matching entity mention in the output). The ESA_I scores are averaged over the corpus in three different ways, over all texts (ESA_C score), texts that have at least one undetected entity ($\text{ESA}_C\backslash 1$) and texts with at least two undetected entities ($\text{ESA}_C\backslash 2$). Table 1 shows the results together with the distribution of undetected entities.

[4] Examples of outputs of the entity mentions detection are given in the Appendix.

	Model= NUIG-DSI, BLEU=41.33, ESA$_I$=0.63
Text	The record label of Bootleg Series Volume 1: The Quine Tapes is Polydor Records and it was recorded in St Louis, Missouri, United States. The album was preceded by Squeeze (The Velvet Underground album).
RDF Input	Bootleg_Series_Volume_1:_The_Quine_Tapes | recordedIn | United_States Bootleg_Series_Volume_1:_The_Quine_Tapes | recordedIn | St._Louis,_Missouri Bootleg_Series_Volume_1:_The_Quine_Tapes | precededBy | Squeeze_(The_Velvet_Underground_album) Bootleg_Series_Volume_1:_The_Quine_Tapes | recordLabel | Polydor_Records Bootleg_Series_Volume_1:_The_Quine_Tapes | recordLabel | Universal_Music_Group Bootleg_Series_Volume_1:_The_Quine_Tapes | releaseDate | 2001-10-16 Bootleg_Series_Volume_1:_The_Quine_Tapes | runtime | 230.05
	Model= CycleGT, BLEU=44.59, ESA$_I$=0.75
Text	the 11th Mississippi Infantry Monument was established in 2000 and is located in Cumberland County, Pennsylvania.
RDF Input	11th_Mississippi_Infantry_Monument | established | 2000 11th_Mississippi_Infantry_Monument | location | Adams_County,_Pennsylvania Adams_County,_Pennsylvania | hasToItsNorth | Cumberland_County,_Pennsylvania
	Model= NUIG-DSI, BLEU=47.92, ESA$_I$=0.67
Text	The Acharya Institute of Technology is located in Soldevanahalli, Acharya Dr. Sarvapalli Radhakrishnan Road, Hessarghatta Main Road, Bangalore – 560090. Its director is Dr. G. P. Prabhukumar and it is located in Mumbai.
RDF Input	Acharya_Institute_of_Technology | campus | "In Soldevanahalli, Acharya Dr. Sarvapalli Radhakrishnan Road, Hessarghatta Main Road, Bangalore – 560090." All_India_Council_for_Technical_Education | location | Mumbai Acharya_Institute_of_Technology | director | "Dr. G. P. Prabhukumar" Acharya_Institute_of_Technology | city | Bangalore Acharya_Institute_of_Technology | wasGivenTheTechnicalCampusStatusBy | All_India_Council_for_Technical_Education

Figure 2: Examples of texts with high BLEU and low ESA$_I$. (Missing RDF input entities are underlined.)

2017 vs. 2020. We see a marked improvement between 2017 and 2020. While in 2017, the ratio of generated texts failing to mention at least one entity varies from 10 to 77% whereas in 2020 it ranges between 3% and 51%. The trend is similar for the various ESA$_C$ scores with e.g., an ESA$_C\backslash 2$ range of [0.17,0.64] in 2017 against [0.36,0.71] in 2020. This corroborates the impression that Natural Language Generation (NLG) models have strongly improved in recent years.

2020 NLG. Zooming in on the more state-of-the-art 2020 models, we find that out of a total of 1779 texts and 16 model outputs, the average ESI$_C{}^1$ score is 17% and the median 10%. In other words, on average, models fail to mention at least one entity 17% of the time.

There are strong differences between the models however. The rule-based models (RALI, Baseline-2017, DANGNT-SGU, Baseline-2020) have low ESI$_C{}^1$. This is unsurprising as such models can integrate lexicons mapping RDF entities to natural language mentions. Interestingly, among the other five models with an ESI$_C{}^1$ less than 11%, four are bilingual neural NLG models i.e., models which were trained to transform RDF data not only in English but also in Russian.

High BLEU does not guarantee Entity-Based Semantic Adequacy. Figure 3 clusters models with respect to both BLEU and ESI$_C{}^1$ ranks. Models that occur right of the vertical axis have high ESI$_C$ rank (they are in the first 8 group), models that occur above the horizontal axis have high BLEU rank. We see that from the 8 models with highest BLEU rank, only three are also among the 8 models with highest ESI$_C{}^1$ rank (cuni-ufal, FB-ConvAI and Amazon_AI). The five other models which rank among the first eight in terms of BLEU score (OSU, CycleGT, NUIG, TGen, bt5) have a BLEU score ranging between 0.45 and 0.54 yet their ESI$_C{}^1$ score ranges between 10 and 22%. This highlights the fact that a high BLEU score does not guarantee semantic adequacy: while their BLEU score is high, on average these models fail to mention at least one of the input entities 10 to 22% of the time. Figure 2 shows some examples of 2020 outputs with low ESA$_I$ and high BLEU score.

Figure 3 further shows that no model ranks high in term of both BLEU and entity-based semantic adequacy (no model in the top right corner).

5.2 Manual Verification of the ESI$_C$ results

Our mention detection algorithm does not detect all mentions, while texts generated by the WebNLG

Model	1	2	3	4	5-8	>1	↓ $\text{ESI}_C{}^1$	Type	BLEU	↑ESA_C	↑$\text{ESA}_C\backslash 1$	↑$\text{ESA}_C\backslash 2$
2020												
RALI	50	1			0	51	3%	Symb	11	0.99	0.8	0.6
Baseline-2020	55	1			0	56	3%	Symb	10	0.99	0.78	0.6
Huawei	62	4			0	66	4%	T5	12	0.99	0.79	0.65
DANGNT-SGU	98	2	1		0	101	6%	Symb	9	0.99	0.76	0.56
Baseline-2017	94	27	0	2	7	130	7%	Symb	15	0.97	0.65	0.36
FBConvAI	154	4			0	158	9%	BART	3	0.98	0.77	0.62
cuni-ufal	169	11	2		0	182	10%	mBART	7	0.98	0.78	0.65
Amazon_AI	175	10			0	185	10%	T5	1	0.98	0.78	0.69
OSU	171	13	1		0	185	10%	T5	2	0.98	0.78	0.65
CycleGT	240	25	1		0	266	15%	T5	8	0.97	0.81	0.71
NUIG-DSI	203	62	17	0	1	283	16%	T5	4	0.96	0.76	0.67
bt5	305	45	4		0	354	20%	T5	5	0.95	0.76	0.62
TGen	264	78	26	15	17	400	22%	T5	6	0.94	0.72	0.58
NILC	499	123	13	5	0	640	36%	BART	16	0.89	0.69	0.56
ORANGE	600	190	45	9	2	846	48%	BART	14	0.83	0.65	0.5
UPC-POE	589	230	63	19	7	908	51%	T5	13	0.84	0.7	0.59
2017												
Tilburg SMT	179	8				187	10%	SMT	2	0.97	0.7	0.55
UPF-FORGe	203	18				221	12%	Symb	4	0.97	0.73	0.56
Melbourne	371	74	11			456	24%	NMT	1	0.94	0.76	0.64
Tilburg NMT	555	171	20	3		749	40%	NMT	6	0.89	0.72	0.62
Tilburg Pipeline	304	233	122	72	49	780	42%	Symb	5	0.76	0.42	0.2
Adapt	482	295	130	52	15	974	52%	NMT	8	0.76	0.54	0.38
PKUWriter	529	282	135	106	60	1112	60%	NMT	3	0.71	0.52	0.36
UIT-DANGNT	47	138	238	317	630	1370	74%	Symb	9	0.28	0.02	0
Baseline	377	398	249	207	206	1437	77%	NMT	7	0.47	0.31	0.17

Table 1: Entity-Based Semantic Adequacy of the WebNLG Challenge 2020 and 2017 Participant Models. $\text{ESI}_C{}^1$: Proportion of texts with at least one undetected mention (lower is better). The second to sixth columns indicate the number of texts with n undetected entities. The last three columns give the corpus average of the text level ESA_I score, for all texts (ESA_C), for texts with at least one undetected entity ($\text{ESA}_C\backslash 1$) and for texts with at least two undetected entities ($\text{ESA}_C\backslash 1$). For ESA_I scores, higher is better. BLEU indicates the rank of the model in terms of BLEU in the WebNLG Shared Task and Type, the type of model (Symb: the model integrates a symbolic component, BART, mBART, T5: the pre-trained model used).

models might differ from the crowdsourced texts on which we evaluated our entity mention detection algorithm (cf. Section 4.2). Therefore, we manually verify the result of our mentions detection algorithm for different types of models. We focus on five models with contrasting BLEU and $\text{ESI}_C{}^1$ rank, two models with high $\text{ESI}_C{}^1$ rank but low BLEU rank; one model with high rank for both dimensions; one model with high BLEU rank and low $\text{ESI}_C{}^1$ rank; and one model with low rank in both dimensions. For each of these models, we check texts with different numbers of missing entities (one or two missing entities for the models with high $\text{ESI}_C{}^1$ and one, three and five missing entities for the models with low $\text{ESI}_C{}^1$) and computed the rate of false positives, i.e. entities which were labeled as undetected by our algorithm but which are in fact present in the generated text. While for the three models which rank high in terms of entity-based semantic adequacy, the rate of false positive is high (100% for RALI, 81% for Huawei and 52% for FBConvAI)[5], for models with low ESI_C rank, the number of false positives is much lower (49% for bt5, 13% for Orange). In other words, our entity mention detection algorithm is best at detecting models with low semantic adequacy.

5.3 Qualitative Analysis

Examining outputs with low ESA score, we found three main causes for low semantic adequacy: short output, hallucination and degenerate output. Figure 1 shows some examples. When the output text is much shorter than expected, many mentions are missing (Ex.1). When the model hallucinates entities not present in the input, it also often simultaneously fails to mention those that are (Ex. 2).

[5] The repetition of the same entities is different entries of the dataset has a strong impact here. For instance, for the RALI system, on the 1779 texts we checked, our algorithm finds 51 texts with undetected entities but only nine of these entities are distinct. That is, the algorithm fails to detect nine entities and as these are in multiple corpus instances, it has 100% of false positives on these corpus instances.

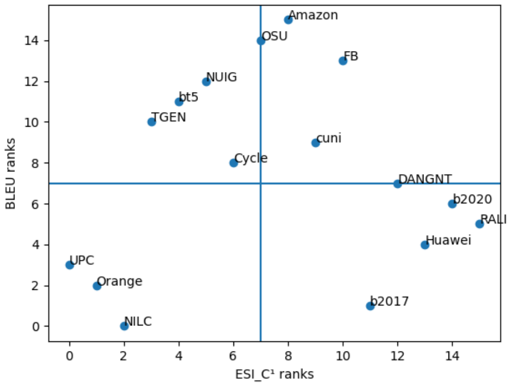

Figure 3: **BLEU vs ESI$^1_\mathrm{C}$ ranks for WebNLG2020 models** (higher is better i.e., position 8 in the graph indicates the highest ranked system). The top part of the figure shows the top 8 ranked models w.r.t. BLEU, the right most part the top 8 ranked models w.r.t. ESA. Of the 8 top ranked models w.r.t. BLEU (top part of the figure), only two (FB and OSU) are among the 8 top ranked w.r.t. ESA scores.

Finally, entity mentions may be missing because of degenerate output (Ex.3).

6 Correlation with Human and Automatic Metrics

We study the correlation of our ESA$_I$ metric with the human and automatic metrics used in the WebNLG challenges.

6.1 Evaluation Set-Up

During the WebNLG Challenges 2017 and 2020, 223 texts were sampled from the outputs of the participants models for human evaluation in 2017, and 178 in 2020 (Shimorina et al., 2018; Castro Ferreira et al., 2020a). For our correlation study, we therefore use the automatic and human evaluation scores collected by the WebNLG organisers for 2,007 texts (223 for each of the 9 models) in 2017 and 2,848 texts (178 for each of the 16 models) in 2020. We use the results and scripts from Shimorina et al. (2018) and Castro Ferreira et al. (2020a).

In 2017, the human evaluation metric which focuses on semantic adequacy is Semantics where the annotator is asked to assess semantic faithfulness of the generated output w.r.t. the input (1-low, 2-medium or 3-good). In 2020, the human evaluation metrics concerned with semantic adequacy are Data Coverage, Correctness, Relevance (between 0 and 100). For Data Coverage, evaluators were asked to check whether all input RDF properties were in the text; for Relevance, whether the text describes only such predicates which were present in the input; and for Correctness, whether the output text correctly describes the subject and object of those predicates which matched a property in the input graph. Note that while all these criteria bear on semantic adequacy, none of them specifically target entities.

6.2 Results

We compute the correlations with ESA$_I$ metric at text level in three different set-ups (for all texts, for texts with at least one undetected entity and for texts with at least two undetected entities) using three correlation metrics (Pearson correlation, the Spearman rank correlation and Kendall's Tau). Tables 2 and 3 only report Pearson correlations on texts with at least one undetected entity ($n = 822$ for 2017; $n = 470$ for 2020). Detailed correlation results for the three metrics and considering the three text setups are reported in the Appendix.

Human vs. Automatic Metrics. The correlation between ESA$_I$ and human metrics of semantic adequacy is strong in 2017 and moderate in 2020, indicating that ESA$_I$ correctly captures what humans judge to be semantically adequate. ESA also has very strong (2017) and moderate (2020) correlation with METEOR, which suggests that variants of entity mentions involve synonyms and stemming like modifications.

Varying Correlation Strengths and Scale. The strength of the correlations and their relative order vary with the shared tasks. For automatic metrics, this is likely due to greater variance in metric scores between systems. For human judgments, it might also result from the different criteria used in 2017 vs. 2020 and from their subjectivity. Indeed, as shown below, some of the collected human judgments are in fact incorrect. Not shown here, but reported in the appendix, correlations with Human, METEOR and BLEU scores also tend to be higher for texts with at least one or two undetected entities – that is, texts which are likely to have semantic adequacy problems – compared to correlations computed over all texts (compare Table 3 to correlations over all texts, in the Appendix).

Cases of strong disagreement between ESA$_I$ and human evaluation We manually checked some of the texts which received high human evaluation scores but had a low proportion of detected

Metrics	METEOR	TER	Fluency	Grammar	Semantics	ESA$_I$
BLEU	0.74	-0.57	0.39	0.43	0.53	0.59
METEOR	x	-0.54	0.57	0.63	0.72	**0.87**
TER	x	x	-0.42	-0.45	-0.4	-0.42
Fluency	x	x	x	0.89	0.51	0.49
Grammar	x	x	x	x	0.57	0.57
Semantics	x	x	x	x	x	**0.66**

Table 2: Pearson correlation coefficients for WebNLG 2017 metrics and ESA$_I$. Only for texts with at least one undetected entity (i.e. 822 texts). All the p-values are <0.01. Bold numbers indicate the highest correlations of ESA$_I$with surface-based (top block) and human evaluation (bottom block) metrics.

Metrics	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1 BLEU	0.97	0.71	0.82	-0.67	0.69	0.66	0.71	0.49	0.42	0.3	0.34	0.33	0.31	<u>0.41</u>
2 BLEU NLTK	x	0.77	0.87	-0.74	0.74	0.72	0.77	0.54	0.45	0.34	0.39	0.36	0.36	0.39
3 METEOR	x	x	0.9	-0.62	0.67	0.82	0.78	0.67	0.49	0.49	0.4	0.42	0.36	**0.45**
4 chrF++	x	x	x	-0.69	0.74	0.82	0.82	0.6	0.51	0.46	0.41	0.43	0.37	**0.45**
5 TER	x	x	x	x	-0.76	-0.67	-0.75	-0.61	-0.41	-0.31	-0.42	-0.39	-0.4	-0.24
6 BERT-score P	x	x	x	x	x	0.83	0.95	0.73	0.6	0.41	0.52	0.56	0.5	0.39
7 BERT-score R	x	x	x	x	x	x	0.95	0.75	0.57	0.52	0.49	0.49	0.45	<u>0.43</u>
8 BERT-score F1	x	x	x	x	x	x	x	0.77	0.61	0.49	0.53	0.55	0.5	**0.44**
9 BLEURT	x	x	x	x	x	x	x	x	0.62	0.54	0.52	0.59	0.5	<u>0.43</u>
10 Correctness	x	x	x	x	x	x	x	x	x	0.75	0.71	0.83	0.67	<u>0.56</u>
11 DataCoverage	x	x	x	x	x	x	x	x	x	x	0.62	0.76	0.57	**0.57**
12 Fluency	x	x	x	x	x	x	x	x	x	x	x	0.67	0.86	0.41
13 Relevance	x	x	x	x	x	x	x	x	x	x	x	x	0.65	0.53
14 TextStructure	x	x	x	x	x	x	x	x	x	x	x	x	x	0.36
15 ESA$_I$	x	x	x	x	x	x	x	x	x	x	x	x	x	1

Table 3: Pearson correlation coefficients for WebNLG 2020 metrics ESA$_I$. Only for texts with at least one undetected entity (i.e. 470 texts). All the p-values are <0.01. Bold (resp. underlined) numbers indicate the highest (resp. second highest) correlation scores between ESA$_I$and different categories of evaluation metrics, i.e. surface-based similarity metrics (top block), embedding-based similarity (middle block) and human evaluation metrics.

entities ESA$_I$ (resp. texts with low human evaluation scores but high ESA$_I$). [6]

For WebNLG 2017, there are 7 texts that have ESA$_I$ < 0.4 and Semantics≥ 2. For 6 of them we find that there are indeed missing entities, while the remaining one is a degenerate text. The relatively high scores given by human evaluators for Semantics (2 out of 3) suggest that such scoring tends to be subjective, perhaps especially so with a broad evaluation criterion such as 'Semantics'. Among the 41 texts which received the lowest possible rating for semantics (Semantics=1), but had ESA$_I$ > 0.9, we find that 25 texts do not have missing entities but are indeed semantically incorrect (usually because of mistakes or hallucinations of predicates); 3 texts have missing entities and 9 texts have hallucinated entities. The remaining 4 texts contain all input entities and have correct semantics.

The same kind of observations can be made for WebNLG 2020 models. There are 4 texts which received high human evaluation scores for semantic adequacy-related criteria (Data Coverage> 80 or Correctness> 80 or Relevance> 80) and low ESA$_I$ (ESA$_I$ < 0.4) and all of them have missing entities. In contrast, there are 20 texts which got low human evaluation scores (Data Coverage< 30 or Correctness< 30 or Relevance< 30) and high ESA$_I$ (ESA$_I$ > 0.9). Fifteen have wrong or hallucinated predicates, two have significant spelling or repetition problems. Three texts are correct.

From these observations we can draw two main conclusions. First, detecting input RDF entities in the output text is no sufficient condition to assess a model's semantic adequacy. It does not give information about hallucination of entities or about correct verbalization of RDF predicates, which are also necessary conditions for semantic adequacy. These observations also illustrate the subjectivity of human evaluation. Sometimes correct texts can be rated badly by human annotators or vice versa.

Detection of Hallucinations We can use our entity mention detection algorithm in reverse to detect hallucinations i.e., mentions that have no corresponding RDF entity in the input graph. We gather all (entity, mention) pairs found by the entity linker (4.1) for which the entity does not occur in the in-

[6] Examples are given in the appendix.

1 (**add**)	**Output text** Bananaman was created by Steve Bright and starred Bill Oddie. It was broadcast by the BBC, which is based in the Broadcasting House in London, and last aired on 15th April 1986.			
		BBC	city	Broadcasting_House
		Bananaman	starring	Bill_Oddie
	RDF Input	Bananaman	creator	Steve_Bright
		Bananaman	lastAired	"1986-04-15"
		Bananaman	broadcastedBy	BBC
2 (**repl**)	**Output text** Aaron Turner performs Trance music and played with the band Bobina.			
	RDF Input	Andrew_Rayel	associatedBand/associatedMusicalArtist	Bobina
		Andrew_Rayel	genre	Trance_music
3 (**inac**)	**Output text** Cyril frankel is the director of the film "it's Great to be New (1956"), which was written by "ted willis" and starred cecil Parker and John mills. Mr. millis died in 2005.			
		Its_Great_to_Be_Young_(1956_film)	starring	Cecil_Parker
		Its_Great_to_Be_Young_(1956_film)	writer	Ted_Willis
	RDF Input	Its_Great_to_Be_Young_(1956_film)	starring	John_Mills
		Its_Great_to_Be_Young_(1956_film)	director	Cyril_Frankel
		John_Mills	deathYear	2005

Figure 4: Examples of hallucinations (underlined in the texts). add: output contains additional information, repl: an input RDF entity is replaced by another with the same context (here the name of another musician), inac: the name of the input entities are inaccurate which makes them difficult to link with input entities

put RDF graph. Table 4 summarizes the results for each model of the WebNLG challenges. Figure 4 also shows examples of different types of hallucinations.

Model	>1	$>1_{\checkmark}$	Dist	↓ $\text{ESI}_C{}^1$
RALI	0	0	0	0%
B-2017	1	1	1	0.1%
B-2020	1	1	1	0.1%
NUIG	4	3	3	0.2%
UPC	4	4	3	0.2%
DANGNT	5	5	5	0.3%
TGen	8	7	2	0.5%
cuni-ufal	9	7	6	0.5%
Amazon	9	9	3	0.5%
FBConvAI	17	11	6	1%
CycleGT	19	18	10	1%
OSU	20	19	3	1%
bt5	36	17	3	2%
Huawei	48	47	28	3%
NILC	117	99	66	7%
ORANGE	288	288	60	16%
UIT	1	0	1	0.1%
Tilburg SMT	4	0	4	0.2%
Tilburg NMT	11	4	7	0.6%
UPF	12	8	4	0.6%
Tilburg Pl	14	11	6	0.8%
Melbourne	114	112	24	6%
Adapt	241	234	151	13%
PKUWriter	286	283	135	15%
Baseline	754	144*	147	40%

Table 4: Hallucinations (>1 and $>1_{\checkmark}$: number of texts with at least one hallucination before and after a manual check of automatically detected hallucinations. *Verification on 144 randomly chosen texts. Dist: number of distinct detected hallucinated entities.

7 Conclusion

RDF stores have become increasingly popular as a means to make knowledge available on the web (Assi et al., 2020). We propose an automatic metric for assessing the entity-based semantic adequacy of RDF verbalisers and show that it is effective in highlighting semantic inadequacy even for state-of-the-art models with high BLEU scores. We further show that models detected by this metric as having low entity-based semantic adequacy can still have high scores on surface-based metrics, and that while ESA correlates with human scores on semantic criteria, it may in fact be more reliable as a means of detecting low performing models than human-based evaluation protocols, which tend to be subjective.

Acknowledgments

Research reported in this paper has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 860621. This document reflects the views of the authors and does not necessarily reflect the views or policy of the European Com-mission. The REA cannot be held responsible for any use that may be made of the information this document contains.

References

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. Spice: Semantic propositional image caption evaluation. In *European conference on computer vision*, pages 382–398. Springer.

Ali Assi, Hamid Mcheick, and Wajdi Dhifli. 2020. Data linking over rdf knowledge graphs: A survey. *Concurrency and Computation: Practice and Experience*, 32(19):e5746.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Thiago Castro Ferreira, Claire Gardent, Nikolai Ilinykh, Chris van der Lee, Simon Mille, Diego Moussallem, and Anastasia Shimorina. 2020a. The 2020 bilingual, bi-directional WebNLG+ shared task: Overview and evaluation results (WebNLG+ 2020). In *Proceedings of the 3rd International Workshop on Natural Language Generation from the Semantic Web (WebNLG+)*, pages 55–76, Dublin, Ireland (Virtual). Association for Computational Linguistics.

Thiago Castro Ferreira, Claire Gardent, Nikolai Ilinykh, Chris van der Lee, Simon Mille, Diego Moussallem, and Anastasia Shimorina, editors. 2020b. *Proceedings of the 3rd International Workshop on Natural Language Generation from the Semantic Web (WebNLG+)*. Association for Computational Linguistics, Dublin, Ireland (Virtual).

Thiago Castro Ferreira, Diego Moussallem, Emiel Krahmer, and Sander Wubben. 2018. Enriching the WebNLG corpus. In *Proceedings of the 11th International Conference on Natural Language Generation*, pages 171–176, Tilburg University, The Netherlands. Association for Computational Linguistics.

Bhuwan Dhingra, Manaal Faruqui, Ankur Parikh, Ming-Wei Chang, Dipanjan Das, and William Cohen. 2019. Handling divergent reference texts when evaluating table-to-text generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4884–4895, Florence, Italy. Association for Computational Linguistics.

Ondřej Dušek and Zdeněk Kasner. 2020. Evaluating semantic accuracy of data-to-text generation with natural language inference. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 131–137, Dublin, Ireland. Association for Computational Linguistics.

Claire Gardent, Anastasia Shimorina, Shashi Narayan, and Laura Perez-Beltrachini. 2017. The WebNLG challenge: Generating text from RDF data. In *Proceedings of the 10th International Conference on Natural Language Generation*, pages 124–133, Santiago de Compostela, Spain. Association for Computational Linguistics.

Sebastian Gehrmann, Falcon Dai, Henry Elder, and Alexander Rush. 2018. End-to-end content and plan selection for data-to-text generation. In *Proceedings of the 11th International Conference on Natural Language Generation*, pages 46–56, Tilburg University, The Netherlands. Association for Computational Linguistics.

Ben Goodrich, Vinay Rao, Peter J. Liu, and Mohammad Saleh. 2019. Assessing the factual accuracy of generated text. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery Data Mining*, KDD '19, page 166–175, New York, NY, USA. Association for Computing Machinery.

Python library. Dateparser.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online. Association for Computational Linguistics.

Lena Reed, Shereen Oraby, and Marilyn Walker. 2018. Can neural generators for dialogue learn sentence planning and discourse structuring? In *Proceedings of the 11th International Conference on Natural Language Generation*, pages 284–295, Tilburg University, The Netherlands. Association for Computational Linguistics.

Leonardo F. R. Ribeiro, Yue Zhang, Claire Gardent, and Iryna Gurevych. 2020. Modeling global and local node contexts for text generation from knowledge graphs. *Transactions of the Association for Computational Linguistics*, 8:589–604.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Anastasia Shimorina, Claire Gardent, Shashi Narayan, and Laura Perez-Beltrachini. 2018. WebNLG Challenge: Human Evaluation Results. Technical report, Loria & Inria Grand Est.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *In Proceedings of Association for Machine Translation in the Americas*, pages 223–231.

Elior Sulem, Omri Abend, and Ari Rappoport. 2020. Semantic structural decomposition for neural machine translation. In *Proceedings of the Ninth Joint Conference on Lexical and Computational Semantics*, pages 50–57, Barcelona, Spain (Online). Association for Computational Linguistics.

Johannes M. van Hulst, Faegheh Hasibi, Koen Dercksen, Krisztian Balog, and Arjen P. de Vries. 2020. Rel: An entity linker standing on the shoulders of giants. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20. ACM.

Sam Wiseman, Stuart Shieber, and Alexander Rush. 2017. Challenges in data-to-document generation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2253–2263, Copenhagen, Denmark. Association for Computational Linguistics.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.