

Integrating Deep Linguistic Features in Factuality Prediction over Unified Datasets

Anonymous ACL submission

Abstract

Previous models for the assessment of commitment towards a predicate in a sentence (also known as factuality detection) were trained and tested against a specific annotated test set, subsequently limiting the generality of their results. In this work we propose a simple process which maps three previously annotated corpora onto a single factuality scale, thereby enabling models to be tested across these corpora. In addition, we devise a novel model for factuality detection which extends a previous hand-annotated predicate lexicon and applies it over an abstraction of dependency trees in both supervised and unsupervised settings. Finally, we show that the model performs well across our test sets.

1 Introduction

Factuality detection is the task of determining the level of commitment towards a predicate in a sentence according to a specific source, e.g., the author (Saurí and Pustejovsky, 2009). For instance, the author uses linguistic cues to mark the embedded proposition as factual in (1) (cue: *surprising*), as uncertain in (2) and (3) (cues: *risk*, *might*), and as counterfactual (cue: *did not manage*) or uncertain (cue: *will not manage*) in (4).

- (1) *It is not **surprising** that they work.*
- (2) *She takes the **risk** to find out the truth.*
- (3) *She might find out the truth.*
- (4) *He did/will **not manage** to be in time.*

Detecting factuality is hard as the linguistic means used to express it closely interact. For example, lexical cues, such as the proposition-embedding

predicates in (1) and (4) interact with negation (in (1), (4)) and tense (in (4)).

Detecting factuality has many potential applications. For instance, in knowledge base population, only propositions marked as factual should be admitted into the knowledge base, while hypothetical or negated ones should be left out. Similarly, for argumentation analysis and question answering, factuality can play a major role in backing a specific claim or supporting evidence for an answer to a question at hand.

Recent research efforts have approached the factuality task from two complementing directions: automatic prediction and large scale annotation. Previous attempts for automatic factuality detection either took a rule-based, deep syntactic approach (Lotan et al., 2013; Saurí and Pustejovsky, 2012) or a machine learning approach over more shallow features (Lee et al., 2015). In terms of annotation, each effort was largely carried out independently of the other efforts, picking up different factuality flavors and different annotation scales.

In correlation, the proposed algorithms have targeted a single annotated resource which they aim to recover. Subsequently, this separation between annotated corpora has prevented a comparison across testsets. Further, the models are non-portable, inhibiting advancements in one test set to carry over to any of the other annotations.

Our contribution in this work is twofold. First, we suggest that the task can benefit from a unified representation. We exemplify this by mapping the representation of two recent test sets (FactBank (Saurí and Pustejovsky, 2009) and MEAN-TIME (Minard et al., 2016)) onto the $[-3, +3]$ scale, as annotated by (Lee et al., 2015). This unification allows us to test the generality of models which were previously applicable on a single test set. Second, we design a new model for fac-

tuality detection that extends TruthTeller (Lotan et al., 2013), which employed implicative signatures (MacCartney and Manning, 2009; Karttunen, 2012) over dependency trees using a large predicate lexicon. We first extend TruthTeller’s lexicon by about 40% through a semi automatic process (following Eckle-Kohler (2016)). We then apply TruthTeller’s rules over an abstraction of dependency trees (Stanovsky et al., 2016), which represents predicate-argument structures more consistently, thereby allowing TruthTeller rules to apply on a wider range of syntactic constructions. Finally, we surpass previous methods by using the output from TruthTeller as deep linguistically-informed features in a supervised classifier, thus successfully integrating a rule-based approach in a machine learning framework.

Overall, we hope that our unified representation will allow training and testing on larger, more diverse test sets, and that the good performance of our new model (made publicly available upon publication) indicates its usability across different flavors of factuality detection.

2 Background

Factuality detection requires the identification of uncertainty, a concept which largely corresponds to the linguistic notion of *modality* (Hacquard, 2011). Modality expresses possibilities and necessities by means of negation, modal verbs (*may*, *might*, *can*), main verbs (*agree*, *refuse*), adjectives (*dishonest*), future tense (*will*, *won’t*), and more. Looking at the numerous and varied possibilities language offers to express all the different shades of modality, it is clear that factuality does not assume any fixed set of discrete values either. Instead, the underlying linguistic system forms a continuous spectrum ranging from factual to counterfactual (Saurí and Pustejovsky, 2009).

While linguistic theory assigns a spectrum of factuality values, recent years have seen many practical efforts to capture the notion of factuality in a consistent annotation (Saurí and Pustejovsky, 2009; Nissim et al., 2013; Lee et al., 2015; OGorman et al., 2016; Minard et al., 2016; Ghia et al., 2016). Each of these make certain decisions regarding the granularity of factuality that they aim to extract. In the course of this work we chose to set our focus on three of these annotations: FactBank (Saurí and Pustejovsky, 2009), MEANTIME (Minard et al., 2016) and the *UW*

corpus (Lee et al., 2015). We use these specific corpora as they represent recent efforts, display a range of different design choices (e.g., in their notion of factuality and method of annotation), and are made publicly available which ensures the ease of the reproducibility of our experiments. Table 1 sums the properties and variations of these corpora. For example, we can see that: (1) the *UW* corpus uses a continuous scale and is annotated by crowdsourcing, while MEANTIME and FactBank were annotated discretely by experts, (2) FactBank annotates factuality from different perspectives, and (3) MEANTIME is significantly smaller compared to the other corpora.

In parallel with the creation of these annotated resources, several efforts were made to predict factuality in an automatic manner. The methods for doing so can be largely divided into rule-based systems which examine deep linguistic features, and machine learning algorithms which generally extract more shallow features. The De Facto factuality profiler (Saurí and Pustejovsky, 2012) and TruthTeller algorithms (Lotan et al., 2013) take the rule-based approach and assign a discrete annotation of factuality (following the values assigned by FactBank) using a deterministic rule-based top-down approach on dependency trees, changing the factuality assessment when encountering factuality affecting predicates or modality and negation cues (following implicative signatures by Karttunen (2012)). In addition to a factuality assessment, TruthTeller assigns three values per predicate in the sentence: (1) implicative signature from a hand-coded lexicon indicating how this predicate changes the factuality of its embedded clause, in positive and negative contexts, (2) clause truth, marking the factuality assessment of the entire clause, and (3) negation and uncertainty, indicating whether this predicate is affected by negation or modality. Both of these algorithms rely on a hand-written lexicon of predicates, indicating how they modify the factuality status of their embedded predicates (e.g., **refuse** negates while **assure** asserts it). In this work we will make use of the more recent TruthTeller which uses a much larger lexicon of 1,700 predicates (verbs, adjectives and nouns) compared to De Facto’s lexicon, which contains 646 predicates.

In a separate attempt, Lee et al. (2015) have used SVM regression techniques to predict a *continuous* factuality value from lexical and syntactic

Corpus	#Tokens/Sentences	Factuality Values		Type	Annotators	Perspective
		Original	Our mapping			
FactBank	77231 / 3839	Factual (CT+/-)	+3.0 / -3.0	Discrete	Experts	Author's and discourse-internal sources
		Probable (PR+/-)	+2.0 / -2.0			
		Possible (PS+/-)	+1.0 / -1.0			
		Unknown (Uu/CTu)	0.0			
MEANTIME [†]	9743 / 631	Fact / Counterfact	+3.0 / -3.0	Discrete	Experts	Author's
		Possibility (uncertain)	+1.5 / -1.5			
		Possibility (future)	+0.5 / -0.5			
UW	106371 / 4234	[-3.0, 3.0]		Continuous	Crowdsource	Author's

Table 1: Factuality annotation statistics and mappings used in this paper - the number of tokens and sentences in each corpus, the original factuality value with the corresponding converted value to UW scale, the type of annotation (discrete or continuous), the annotators' proficiency, and the perspective to which the annotation refers. [†]This is an abstraction over the original MEANTIME annotation (suggested by the MEANTIME authors), which is composed of polarity, certainty and temporality.

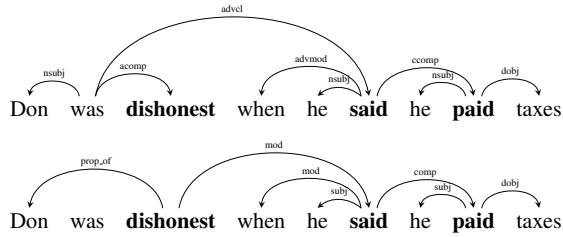


Figure 1: Dependency tree (top, obtained with spaCy) versus PropS representation (bottom, obtained via the online demo). Note that PropS posits *dishonest* as the head of *said*, while the dependency tree obstructs this relation.

features (lemma, part of speech, and dependency paths). Similarly to the TruthTeller approach, they also predict a single factuality value pertaining to the author's commitment towards the predicate.

3 Unified Representation and Model

First, we achieve a unified representation by mapping FactBank and MEANTIME onto the UW $[-3, +3]$ range in a simple automatic rule-based manner. Table 1 describes these rules (see column "Our mapping"), which were hand-written by consulting the annotation guidelines of each of the corpora.¹

Next, we devise a factuality detection model composed of three main components: (1) augmentation of the TruthTeller lexicon with about 800 adjectival, nominal and verbal predicates, (2) syntactic re-ordering with PropS (Stanovsky et al., 2016), (3) application of TruthTeller on top of PropS trees (Lotan et al., 2013). In the following we describe these components.

¹Full technical details are made available in the appendix.

Extending TruthTeller's lexicon We extended the TruthTeller lexicon of single-word predicates by integrating a large resource of modality markers. Following the approach of Eckle-Kohler (2016), we first induced the modality status of English adjectives and nouns from the subcategorization frames of their German counterparts listed in a large valency lexicon (using the "IMSLex German Lexicon" (Fitschen, 2004) and Google Translate for obtaining the translations²). We focused on four modality classes (the classes *wh-factual* and *wh/if-factual* indicating factuality, and the two classes *future-orientation* and *non-factual*, indicating uncertainty)³ and semi-automatically mapped them to the signatures used in TruthTeller. We performed the same kind of mapping for the modality classes of English verbs provided by Eckle-Kohler (2016). The result of this process extended TruthTeller's lexicon by roughly 40% (265 adjectives, 281 nouns, and 133 verbs).

Integrating PropS with TruthTeller PropS was recently presented as an abstraction over dependency trees. Most convenient in our case is its re-ordering of non-verbal predicates (adjectival, conditional, non-lexical, etc.) such that each predicate is the direct head of its respective arguments. For example, for adjectival predication, compare the different parses in Figure 1. PropS positions *dishonest* as the head of *said*, which is subsequently the head of *paid*. This chain allows the implicative signature encoded in TruthTeller to capture this complex relation. The dependency

²We used the translation function available as part of Google Sheets. <https://www.google.com/sheets> and removed all translation pairs with English multi-words.

³In Eckle-Kohler (2016), these are the classes containing the majority of the verb types.

syntax, in contrast, obstructs this relation by positing *dishonest* as a leaf node under *when*. Consistency of PropS annotation allows the top-down approach of TruthTeller to apply to predicates beyond the verbal case.

Finally, we take as features all four TruthTeller annotations (see Section 2) for: the target predicate, its PropS head and its children (padding or truncating to 4 children). For fair comparison with the UW algorithms, we use these features to train an SVM regression (Basak et al., 2007) model to predict the final factuality value.

4 Evaluation

In this section we describe the experiments we carried out on the three testsets (FactBank, MEANTIME, and UW). For fair comparison, we use the same train, development, test split of the test sets for all systems. We preprocess the data with the spaCy Python library⁴. In all our experiments we compute the metrics used in Lee et al. (2015): (1) *Mean Average Error*⁵ (MAE), which computes the absolute fit of the model and (2) *Pearson correlation coefficient* between automatic predictions and gold labels, useful in biased test sets as it assesses how well the model captures the variability in the gold data.

Baselines For state of the art rule-based approach, we create a variant of our extended TruthTeller, converting its factuality prediction to numerical values (similarly to how we converted FactBank’s annotation). The state-of-the art for supervised learning is represented by UW features. In addition, we tried applying UW features on the popular AMR formalism (Banarescu et al., 2013) using a state-of-the-art parser (Pust et al., 2015), hoping that its higher abstraction would improve performance. Finally, we compare against an **all-factual** baseline which assigns +3.0 to all predicates. Since the task is by nature heavily biased towards the factual label, it is interesting to compare against such a simple (yet strong) lower bound. See the supplemental material for a technical elaboration on the baselines implementation.

4.1 Results

Several observations can be made following the results on our test sets (Table 2).

⁴<https://spacy.io>

⁵Note that in our case this ranges between 0 (perfect performance) and 6 (worst performance).

Test Corpus	FactBank		UW		MEANTIME	
	MAE	<i>r</i>	MAE	<i>r</i>	MAE	<i>r</i>
All-factual	.80	0	.78	0	.31	0
UW[†]	.81	.66	.66	.61	.56	.33
AMR	.66	.66	.64	.58	.44	.30
Rule-based	.75	.62	.72	.63	.35	.23
Supervised	.59	.71	.43	.74	.34	.47

Table 2: Performance of the baselines (top) against our unsupervised and supervised approaches (bottom). [†]UW’s performance is different from that reported in their paper (.51 MAE / .71 *r*) as we used a different solver.

Unsupervised baseline is a good starting point

The unsupervised performance is well correlated with the gold predictions on FactBank and UW, showing its off-the-shelf usability.

Supervised setting improves performance

Adding our features provided a predictive signal for factuality assessment on all test sets. More significant improvement is observed in the larger FactBank and UW corpora.

UW achieves good correlation

UW gives a more diverse annotation thanks to its richer feature set (including lemma and dependency path). While this hurts MAE in some scenarios, it overall leads to good correlation with the gold data.

MEANTIME proves especially hard

None of the systems were able to surpass the all-factual baseline in terms of MAE on MEANTIME. This is due to its much smaller size and heavy factual bias (assigning +3.0 to 90% of the predicates).

AMR models achieve comparable performance

While AMR provides a more abstract representation, many aspects of factuality (interaction of verb tenses, modal verbs, negation) are not modeled. Noisy automatic parses also diminish the positive effect of richer feature representation.

5 Conclusions

We presented an intuitive method for mapping FactBank and MEANTIME onto the UW scale, and presented a novel factuality model which extends TruthTeller and applies it over PropS’ abstraction of dependency trees. We will make both the unified representation and the trained model publicly available, hoping that it will enable factuality research across larger, more diverse test sets.

References

- Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. Abstract meaning representation for sembanking .
- Debasish Basak, Srimanta Pal, and Dipak Chandra Patranabis. 2007. Support vector regression. *Neural Information Processing-Letters and Reviews* 11(10):203–224.
- Judith Eckle-Kohler. 2016. Verbs taking clausal and non-finite arguments as signals of modality revisiting the issue of meaning grounded in syntax. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL 2016)*. Association for Computational Linguistics, volume Volume 1: Long Papers, pages 811–822.
- Arne Fitschen. 2004. Ein Computerlinguistisches Lexikon als komplexes System. PhD Thesis, Universität Stuttgart, Germany.
- Elisa Ghia, Lennart Kloppenburg, Malvina Nissim, Paola Pietrandrea, and Valerio Cervoni. 2016. A construction-centered approach to the annotation of modality. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the 12th ISO Workshop on Interoperable Semantic Annotation*. European Language Resources Association (ELRA).
- Valentine Hacquard. 2011. Modality. In Claudia Maienborn, Klaus von Heusinger, and Paul Portner, editors, *Semantics: An International Handbook of Natural Language Meaning*. HSK 33.2, Berlin: Mouton de Gruyter, pages 1484–1515.
- Lauri Karttunen. 2012. Simple and Phrasal Implicatives. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics*. Montréal, Canada, pages 124–131.
- Kenton Lee, Yoav Artzi, Yejin Choi, and Luke Zettlemoyer. 2015. Event detection and factuality assessment with non-expert supervision. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Lisbon, Portugal.
- Amnon Lotan, Asher Stern, and Ido Dagan. 2013. Truthteller: Annotating predicate truth. In *HLT-NAACL*. pages 752–757.
- Bill MacCartney and Christopher D. Manning. 2009. An extended model of natural logic. In *Proceedings of the Eighth International Conference on Computational Semantics*. Tilburg, The Netherlands, IWCS-8 '09, pages 140–156.
- Anne-Lyse Minard, Manuela Speranza, Ruben Urizar, Begona Altuna, Marieke van Erp, Anneleen Schoen, and Chantal van Son. 2016. Meantime, the news-reader multilingual event and time corpus. *Proceedings of LREC2016* .
- Malvina Nissim, Paola Pietrandrea, Andrea Sanso, and Caterina Mauri. 2013. Cross-Linguistic Annotation of Modality: a Data-Driven Hierarchical Model. In *Proceedings of the 9th Joint ISO - ACL SIGSEM Workshop on Interoperable Semantic Annotation*. Potsdam, Germany, pages 7–14.
- Tim OGorman, Kristin Wright-Bettner, and Martha Palmer. 2016. Richer event description: Integrating event coreference with temporal, causal and bridging annotation. *Computing News Storylines* page 47.
- Michael Pust, Ulf Hermjakob, Kevin Knight, Daniel Marcu, and Jonathan May. 2015. Parsing english into abstract meaning representation using syntax-based machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Lisbon, Portugal, pages 1143–1154.
- Roser Saurí and James Pustejovsky. 2009. Factbank: a corpus annotated with event factuality. *Language resources and evaluation* 43(3):227.
- Roser Saurí and James Pustejovsky. 2012. Are You Sure That This Happened? Assessing the Factuality Degree of Events in Text. *Computational Linguistics* 38(2):261–299.
- Gabriel Stanovsky, Jessica Ficler, Ido Dagan, and Yoav Goldberg. 2016. Getting more out of syntax with props. *arXiv preprint* .