Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 2460–2467
Marseille, 20-25 June 2022

BeSt: The Belief and Sentiment Corpus

Jennifer Tracey[*], Owen Rambow[⋈], Michael Arrigo[*], Claire Cardie[♢], Adam Dalton[♡]
Hoa Dang[♠], Mona Diab[★], Bonnie Dorr[¶], Louise Guthrie[℧]
Magdalena Markowska[⋈], Smaranda Muresan[◎], Vinodkumar Prabhakaran[♁]
Samira Shaikh[◁], Tomek Strzalkowski[♮], Janyce Wiebe[†]
[*]Linguistic Data Consortium [⋈]Stony Brook University [♢]Cornell University [♡]IHMC [♠]NIST
[★]George Washington University [¶]University of Florida [℧]University of Texas at El Paso [◎]Columbia University
[♁]Stanford University [◁]University of North Carolina at Charlotte [♮]RPI [†]Deceased
[*]Corresponding author: garjen@ldc.upenn.edu

Abstract

We present the BeSt corpus, which records cognitive State: who believes what (i.e., factuality), and who has what sentiment towards what. This corpus is inspired by similar source-and-target corpora, specifically MPQA and FactBank. The corpus comprises two genres, newswire and discussion forums, in three languages, Chinese (Mandarin), English, and Spanish. The corpus is distributed through the LDC.

1. Introduction

Much of natural language processing (NLP) has been oriented towards extracting propositional content: who did what to whom. Examples include semantic annotations (e.g., Propbank (Kingsbury et al., 2002) and the related OntoNotes (Hovy et al., 2006) and Abstract Meaning Representation (O'Gorman et al., 2018), and FrameNet (Baker et al., 1998)) and entity-relation-event annotations (such as (Song et al., 2015)). Over the last ten years, interest has grown in not only understanding what content (propositions) is mentioned in text, but propositional attitudes of different agents towards this content (e.g., (Mather et al., 2021)). A *propositional attitude* is a cognitive attitude, including belief and sentiment, towards a proposition.

The set of propositional attitudes of an agent is her cognitive state (also called "private state"). The author can use language to report her own propositional attitudes, but she can also use language to report on other agents' propositional attitudes. One of the goals of NLP is, given a set of propositions mentioned in a text, to also understand who, according to the author, believes in which of these propositions with what level of commitment, and who has what sentiment towards the propositions and the entities mentioned in them.

The first corpus to address such concerns is the MPQA (Multi-Purpose Question-Answer) corpus of (Wiebe et al., 2005), which grew out of Wiebe's interest in modeling and detecting private state, i.e., cognitive state, which includes sentiment and belief/factuality. FactBank (Saurí and Pustejovsky, 2009) simplifies MPQA and concentrates on belief/factuality. The BeSt corpus which we present in this paper follows FactBank closely but extends it to sentiment and adds multiple genres and more languages.

This paper is structured as follows: Section 2 discusses related work in more detail. In Section 3 and Section 4 we present the goals and basic conceptual structure of the corpus. Section 5 provides details about the corpus, including annotation mechanism and size, and Section 6 discusses some example annotations. We summarize some initial work using the corpus in Section 7 and conclude.

2. Related Corpora

Many corpora explore the notion of belief/factuality and/or sentiment. Those include: LU (Diab et al., 2009), FactBank (FB) (Saurí and Pustejovsky, 2009), MPQA 3.0 (Deng and Wiebe, 2015), UW (Lee et al., 2015), LDCCB (Prabhakaran et al., 2015), MEANTIME (MT) (Minard et al., 2016), MegaVeridicality (MV) (White et al., 2018), UDS-IH2 (UD2) (Rudinger et al., 2018), CommitmentBank (CB) (de Marneffe et al., 2019), and RP (Ross and Pavlick, 2019). All of those corpora tackle similar issues, such as identifying whether an event is considered a fact or what attitude does the source have toward the event, in order to better understand cognitive state. Below we highlight several dimensions that differentiate the corpora.

The most salient differences are what type of data are used to build the corpus, and whether or not the data are manipulated. For example, MV utilizes the MegaAttitude dataset and selects only 6 syntactic frames and lexically "bleaches" the sentences. Lexical bleaching replaces noun phrases for persons in subject positions by *someone* (or in subjects of passive constructions, *a particular person*), noun phrases for things by *a particular thing*, entire finite embedded clause by *a particular thing happened*, and entire embedded clauses under a control verb by *to do a particular thing*, all inflected appropriately. The goal is to concentrate annotator effort on the matrix verb and its effect on annotation. In contrast, FactBank tags all events introduced in a corpus of complete, naturally occurring texts.

The next difference is the definition of annotatable event. In MV only past events are taken into consideration (the corpus does not contain other events); in UD2, both past and present events (the annotators are

instructed to ignore future events, though all three exist in the corpus); UW, LU, FB, and LDCCB consider also future events.

Another dimension concerns the annotators themselves: FB, LDCCB, LU, and MPQA use trained annotators, while CB, RP, UDS-IH2, and UW argue for the value of crowd-sourced judgements collected from naive annotators. This correlates with whether or not an annotator should use her world knowledge to indicate her judgments. For example, the RP and LU annotators were explicitly directed not to use any information unavailable in the data presented, while the UDS-IH2 instructions were ambiguous about the matter. Unclear instructions may affect the quality of the judgements.

The events can be annotated with belief and/or sentiment values on various scales. Annotations can be continuous numerical values (often derived from averaging naive annotators' judgments), typically [-3,3] (CB, UW or UDS-IH2), or categorial labels (FB, LU, and LDCCB). Those corpora distinguish between committed belief (CT in FactBank) and non-committed belief (NCB). In LU and LDCCB, the cases where an event is not presented as factual (but as a wish or hypothetical) are marked with NA. FactBank uses a more fine-grained scale, where NCB is subdivided into possible (PS), and probable (PR). These labels are augmented with polarity marks, e.g. PS-. There is also a label for types of non-factuals such as reported beliefs, wishes, and hypotheticals; for FactBank, this is UU, while LDCCB distinguishes between reported beliefs (ROB) and wishes and Hypotheticals (NA).

MPQA 3.0 differs from the other corpora. In addition to distinguishing factual from non-factual presentation of events, it focuses on other properties of events, e.g., a distinction is drawn between direct subjective and objective speech events. In addition, sentiment and polarity of the source(s) toward the event are also annotated.

The annotation instructions may or may not be specific with regards to the perspective of the event. CB, UW, and MT annotate the text from the author's perspective. On the other hand, LDCCB and LU instruct annotators to report what they believe about the factuality. In RP and MV, the perspective of annotation is unclear. FB follows MPQA (Wiebe et al., 2005; Deng and Wiebe, 2015) in annotating factuality judgments of the author and other agents mentioned in the text. Both FactBank and MPQA use a nested source reference, so that they are attributed to the agent "according to the author" (as we have no independent evidence of that agent's factuality assessment), leading to attribution chains.

While nested sources are useful for identifying possibly different attitudes towards an event, they complicate the annotation procedures. In an attempt to unify the representation and to allow comparison across datasets, Stanovsky et al. (2017) remove the FB non-author perspective annotation and map the discrete annotations of factuality in FB and MT onto the continuous scale used in UW.

One important contribution of both CB and MPQA, which distinguishes these from other corpora, is the consideration of the context of annotatable sentences. In CB, the clause containing the event at issue is preceded by up to 2 prior sentences creating discourse segments. In MPQA, the trained annotators are specifically directed to analyze sentences "with respect to the context in which they appear". It has been suggested that context (Tonhauser et al., 2018), as well as other grammatical factors, such as tense or number and person (de Marneffe et al., 2019) matter in making inferences about events.

3. Corpus Annotation Goals

The goal of BeSt annotation is to determine from text what the writer's and other agents' cognitive states are. Specifically, we are interested in beliefs and sentiments. Note that this is the technical notion of *belief* as used in cognitive science, AI, and philosophy. It should not be confused with a common usage of *belief* which contrasts with *knowledge* and which connotes absence of certainty, or even absence of truth. The notion of "factuality" is closely related to belief. The term *factuality* refers to the presentation by an agent (the author) of a proposition as true. Belief differs from factuality when the author is lying: if the author is lying, a proposition presented by her as factual does not actually correspond to her belief. However, BeSt does not assess whether the author is lying, and the same is true of all corpora discussed in Section 2. We therefore consider belief and factuality to be the same concept for the purpose of this paper. Furthermore, the BeSt corpus is not interested in whether a proposition is actually true in the world, only in cognitive states, i.e., whether agents use language to express that they believe something to be true. The same is true of all corpora discussed in Section 2. For a fuller discussion of these concepts, and a direct comparison with FactBank, please see (Prabhakaran et al., 2015, Section 2).

Like FactBank, BeSt is annotated on top of an existing corpus annotated with propositional content. In the case of BeSt, this is an annotation of entities, relations, and events (ERE), which is described fully in the work of Song et al. (2015). One important consequence is that the private states are only annotated with respect to what the ERE annotation provides. Only certain types of entities are annotated (and these entities are the sources or bearers of private states), and only certain types of events and relations are annotated.

With respect to the dimensions used to discuss the previous literature in Section 2, our annotation goals are as follows:

- BeSt annotates naturally occurring texts. It is based on entire news stories, and parts of web forum discussions (written conversations). The data is not changed.
- The annotatables are given by the prior ERE annotation (see above).

- Annotation is carried out by trained annotators.
- Annotation is from the point of view of the author: what does her language use tell us about her cognitive state, and her model of others' cognitive state?
- Annotators use the whole text, but not world knowledge, to deduce this cognitive state.

We refer to Section 5 for the details on how the annotation happened.

4. Conceptual Corpus Structure

The conceptual basis of the BeSt annotation are private state tuples (PSTs), which are 7-tuples of the following form:

> (type, source-entity, target-object, value, polarity, sarcasm, anchor)

The 7-tuples express the belief or sentiment of the source-entity (which is always an entity) towards the target-object (which can be an entity, a relation, or an event). The meaning of the remaining components of the PST is as follows.

1. The **type** is either "belief" or "sentiment".
2. The **value** is used only for belief and is empty for sentiment. It can take the following values:
   (a) CB = committed belief, meaning that the source is convinced the target is true. Note that this does not mean it "happened" in the past, a source can hold a committed belief about an event in the past, present, or future.
   (b) NCB = non-committed belief, meaning that the source thinks it is possible or probable that the target is true, but is not certain.
   (c) ROB = reported belief. Sometimes, a writer reports on a different source's belief, without letting the reader know what her own belief state is.
   (d) NA = not applicable. This is not a belief at all, but a wish, desire, or hypothetical.
3. The **polarity** is a binary flag indicating the following:
   (a) For sentiment, whether the sentiment is positive (a like) or negative (a dislike).
   (b) For belief, whether the proposition that the belief value is about is affirmed or negated. Note that negation can happen in the expression of the proposition (*I think John will not come*) or in the context (*I doubt John will come*) – in both examples the annotation for the coming event would carry the negative polarity annotation.
4. **Sarcasm** is a flag indicating whether the identified belief or sentiment is expressed sarcastically by the author, i.e., the reader is intended to interpret the utterance as meaning the opposite of its surface interpretation. Thus, even when sarcasm is detected, polarity is annotated based only on the surface characteristics of the language.
5. The **anchor** is a pointer to the text passage which supports the identified claim about belief or sentiment. Specifically, the anchor is the target mention ID, along with the file name.

Note that we do not annotate a "trigger", i.e., a linguistic item (such as a word or syntactic construction or morphological feature) that signals the private state. The reason for this is that we want the corpus to allow researchers to determine the triggers (presumably using machine learning) rather than the annotators.
All the private states expressed in a document collection can be expressed as a collection of PSTs. The same (source-entity, target-object) pair can occur several times with different values. There are two reasons for this:

1. A source can have several different private states with respect to the same target. For example, the writer can have positive sentiment towards the election of Clinton, and also have a non-committed belief towards it. A source can even have conflicting private states, for example both positive and negative sentiment. This happens when someone changes their mind, or when they react to different aspects of the target.
2. Because the provided ERE files only record in-document co-reference, it is possible that what is in fact the same source and target and the same private state get recorded multiple times (if they are expressed in multiple documents).

We discuss some examples in Section 6.

5. Corpus Details

Three languages are annotated in the BeSt corpus: English, Mandarin, and Spanish. The documents annotated in the BeSt corpus come from two genres: news and discussion forums. The source documents were manually selected using a topic-driven approach to ensure appropriate coverage of certain features in documents slated for annotation (including target language, coverage of annotatable entity, relation, and event types, etc.). News documents always include the full text of the news article, while discussion forum documents may include partial threads due to the extreme length of many forum threads. For annotation, discussion forum threads are truncated to consist of a continuous run of posts approximately 800 words in length (excluding metadata and text within `<quote>` elements). When taken from a short thread, a document

Data Set	Documents					
	English		Mandarin		Spanish	
	news	forum	news	forum	news	forum
Training	37	209	0	200	0	95
2016 Test	81	84	79	82	84	84
2017 Test	83	84	84	83	83	83

Table 1: Corpus details by language and genre

may comprise the entire thread. However, when taken from longer threads, a document is a truncated version of its source.

The source data consists of news documents and discussion forum threads from a variety of sources. The source texts were drawn from existing LDC collections as well as new collection, and were selected to serve as a shared source data collection for use in multiple annotation tasks within the Defense Advanced Research Projects Agency (DARPA) Deep Exploration and Filtering of Text (DEFT) program and the National Institute of Standards and Technology (NIST) Text Analysis Conference Knowledge Base Population (TAC KBP) evaluations. As such, the source texts were not specifically selected for sentiment or belief features, but rather as a broad sample of formal and informal text suitable for several different annotation and evaluation tasks. There is no specific topic focus in the source texts, and the documents come from a variety of news sources and discussion forum websites. Many of the factors considered in selection of documents, beyond language and genre distribution, were oriented toward the entity, relation, and event annotations that comprise the targets of the belief and sentiment annotated in BeSt (e.g., presence of event types, multiple mentions of the same entities and events across documents and languages, etc.).

All annotation was fully manual and performed by trained annotators with native fluency in the language, using a custom web-based annotation interface that presents annotators with the source document and the existing ERE annotations. The annotators who performed the BeSt annotation were not the same annotators who performed the ERE annotation that served as input to the BeSt annotation task.

Average inter-annotator agreement on a first-pass annotation is 70% (the range is 66%-76% for belief, and 55%-78% for sentiment). To address the low agreement by first-pass annotators (especially on sentiment), we conducted additional training with one senior annotator per language, who performed a sentiment-specific second pass on all documents.

The BeSt corpus consists of three data sets: a training data set, a 2016 test set, and a 2017 test set. All three sets are include annotated texts in English, Mandarin, and Spanish Genres include news and discussion forums, with an emphasis on discussion forums in the training set (in fact, only English contains any news documents in the training set). The distribution by language and genre is summarized in Table 1.

The Belief and Sentiment annotations have as the target of belief or sentiment the entities, relations, and events annotated as part of the DEFT ERE task. Each event and relation annotated in ERE is labeled for belief and sentiment, and each entity is labeled for sentiment. Belief annotation marks the belief-holder's commitment to a belief in the occurrence of an event (event-target), the participation of an entity in an annotated event (entity-target), and/or the existence of a relation (relation-target), while sentiment annotation takes entities (independent of their role in an event or relation), relations, and events as targets. Thus, the number of possible belief and sentiment targets is determined by the presence of the ERE annotations. Table 2 lists the number of annotated entity, relation, and event mentions available as targets of belief and sentiment annotation in each data set.

Table 3 shows the breakdown of belief annotations in each data set (CB or committed, NCB or non-committed, ROB or reported, and NA or not a belief). There are several observations we can make about expressed beliefs.

- The distribution of belief and sentiment types is different in the two genres represented in the BeSt corpus, newswire (NW) and discussion forums (DF). Note that these are not evenly distributed among the training and two test corpora, nor across the languages (see Table 1).
- We see that non-committed beliefs (NCB) are very rare, as is already found in the LU, LDCCB, and FactBank corpora. This holds for both genres and all three languages.
- News has more reported belief (ROB) than discussion forums. This holds across languages, with the exception of English in the Training corpus. Reported news stories (but not necessarily opinion pieces) typically attribute much of the conveyed information to specific sources. We also note that the percentages of reported belief in discussion forums are somewhat higher in Mandarin than in English, and higher in Spanish than in Mandarin. This may be due to the selection of forums in these languages.
- Concerning non-belief statements (wishes, hypotheticals, etc., marked NA), we see that discussion

Data Set	Belief/Sentiment Targets		
	English	Mandarin	Spanish
Training	29,279 entities 4,216 relations 4,729 events	21,397 entities 3,531 relations 2,959 events	10,335 entities 2,255 relations 1,277 events
2016 Test	16,674 entities 4,045 relations 4,099 events	13,296 entities 3,056 relations 2,426 events	11,625 entities 2,528 relations 2,369 events
2017 Test	13,860 entities 3,505 relations 4,375 events	15,726 entities 3913 relations 3,884 events	12,297 entities 3,115 relations 3,428 events

Table 2: Number of annotated entity, relation, and event mentions available as targets of belief and sentiment annotation in each data set

Sentiment		English				Mandarin				Spanish			
		CB	NCB	ROB	NA	CB	NCB	ROB	NA	CB	NCB	ROB	NA
Training	NW	4268	36	317	394	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
		85%	1%	6%	8%	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
	DF	9355	256	1180	3109	8522	167	1822	2681	4264	214	1956	2972
		67%	2%	8%	22%	65%	1%	14%	20%	45%	2%	21%	32%
2016 Test	NW	9164	80	4782	1054	5572	24	1688	320	4400	59	2587	787
		61%	1%	32%	7%	73%	0%	22%	4%	56%	1%	33%	10%
	DF	3856	71	410	1949	3304	47	396	832	3034	83	770	891
		61%	1%	7%	31%	72%	1%	9%	18%	63%	2%	16%	19%
2017 Test	NW	8024	61	3395	950	8174	43	2825	644	5862	50	2449	618
		65%	0%	27%	8%	70%	0%	24%	6%	65%	1%	27%	7%
	DF	4704	226	693	1977	5263	49	793	1063	4100	186	1060	1203
		62%	3%	9%	26%	73%	1%	11%	15%	63%	3%	16%	18%

Table 3: Distribution of belief labels by language and subcorpus

forums have a higher percentage than do news stories. This observation holds across languages. We attribute this to the goal of newswire (informing about what happened) as opposed to discussion forums (frequently, writing about the wishes or ideas of authors without regard to what has happened).

- We note an interesting difference between the languages. Mandarin has more committed belief (CB) in both newswire and discussion forums than English and Spanish, and less non-belief (NA) than the other two languages, again across both genres. Future research will need to investigate why this is the case.

We now turn to sentiment. Table 4 shows the sentiment annotations (pos, neg, none) in each data set.

- First, we see that no sentiment is expressed towards most targets (entities, relations, events): label "none" is by far the most common, with percentages over 90% for most subcorpora. This contrasts with belief, where the non-belief label NA is far less common. This is because that any simple declarative sentence (such as *Paolo is coming to dinner*) will be interpreted as expressing a belief (in our example, a committed belief or CB towards the event of Paolo coming), but not necessarily a sentiment (in our example, there is no sentiment expressed).

- Furthermore, we see that, contrary to expectations, discussion forums do not actually have substantially more sentiment than newswire texts, with the exception of the English Training corpus (and to a lesser degree, the Mandarin corpora). This is presumably because of reported sentiment in newswire, rather than sentiment expressed by the journalists.

- We see that across languages and genres, there is more negative sentiment expressed than positive sentiment. Furthermore, for all three languages there is more negative sentiment in the forums compared to news. Presumably, people with positive sentiment are less likely to use a discussion forum to express such positive sentiments.

- Comparing the languages to each other, Mandarin

Sentiment		English			Mandarin			Spanish		
		positve	negative	none	positve	negative	none	positve	negative	none
Training	NW	110	296	5653	n/a	n/a	n/a	n/a	n/a	n/a
		2%	5%	93%	n/a	n/a	n/a	n/a	n/a	n/a
	DF	1953	4875	25777	114	440	27428	367	1236	12696
		6%	15%	79%	0%	2%	98%	3%	9%	89%
2016 Test	NW	620	663	12434	168	85	8077	450	533	7208
		5%	5%	91%	2%	1%	97%	5%	7%	88%
	DF	526	555	10681	126	561	9963	273	768	8172
		4%	5%	91%	1%	5%	94%	3%	8%	89%
2017 Test	NW	250	780	9938	82	227	9958	165	673	8516
		2%	7%	91%	1%	2%	97%	2%	7%	91%
	DF	363	668	10371	227	499	12768	219	895	9154
		3%	6%	91%	2%	4%	95%	2%	9%	89%

Table 4: Distribution of sentiment labels by language and subcorpus

has far less negative sentiment than English and Spanish. Future work with the BeSt corpus will probe whether these differences are artifacts of the particular discussion forums that got annotated, or whether they perhaps reflect different communication styles or cultural values.

6. Example Annotations

We discuss two examples in detail in this section. The examples are taken from Englilsh discussion forums, and the bold-faced names are the names of the poster.

(1) **Fat.The.Gangster**: I have tickets to this Friday's Brewers/Cubs game but the memory of my three hour nightmare getting to Miller Park last time is making me nauseous.

In the underlying ERE annotation, the identified entities are: the person **Fat.The.Gangster**, who is coreferenced with *I*, *my*, and *me*; the two teams *Brewers* and *Cubs*; the location *Miller Park*.
There is one identified relation: the location relation between **Fat.The.Gangster** and Miller Park.
There is one event, the transportation event which results in **Fat.The.Gangster** being at Miller Park.
For the BeSt annotation, we find the location relation and the transportation event annotated as committed belief of the author (**Fat.The.Gangster**). Interestingly, both are evoked in complicated syntax (*the memory of my [...] nightmare getting to Miller Park*) – the complement of a noun phrase which is itself the complement of a noun phrase which is the subject of a main clause. Here, it is only the specifics of lexical items that confirm the CB annotation for the *getting* event and the resulting location relation; compare:

(2) The thought of a three hour nightmare getting to Miller Park next time is making me nauseous.

In (2), we change *memory* to *thought* and make *nightmare* indefinite, and the event becomes a non-belief (NA) since we are not told whether the author (**Fat.The.Gangster**) believes it had or would happen, just that he or she is entertaining the thought of it.
In terms of sentiment, the sentence makes clear that **Fat.The.Gangster** has negative sentiment towards the getting event, and the larger context makes clear that he or she has positive sentiment towards the location relation (he or she really did want to be at Miller Park).
We now turn to the second example.

(3) **NeoNerd**: It's just appeared on the AP wires and the BBC live coverage. People who spilled over from the "protest" attacked their car as it went down Regent Street. No harm done to them, apparently.

In this example, we will concentrate on the harm event evoked in the last sentence. **NeoNerd** is relaying what the AP and BBC are reporting (as underlined by the use of *apparently*), so from his or her perspective, the harm event is labeled as a reported belief (ROB). However, from the perspective of the AP and the BBC, it is labeled as committed belief (CB), since they are doing the reporting. Thus, the harm event is annotated three times, with three different sources (**NeoNerd**, the AP, the BBC). In all annotations, the polarity is negative (it did not actually happen). Note that for FactBank and MPQA, the belief attribution of committed belief to the AP and BBC would be couched as being "according to" **NeoNerd**; the BeSt corpus omits this nested attribution, because it captures it indirectly through the additional belief annotation for **NeoNerd** as ROB.

7. Initial Experiments

The BeSt corpus has been used in the DARPA DEFT program as a basis for evaluations in 2016 and 2017, resulting in the two test sets labeled with these years. For 2016, a summary of the evaluation can be found in (Rambow et al., 2016). Four system descriptions for the submitted systems are available at `https://tac.nist.gov/publications/`

2016/papers.html (note: not all papers on the page relate to the BeSt corpus). For the 2017 evaluation, six system descriptions for the submitted systems are available at https://tac.nist.gov/publications/2017/papers.html. Since the corpus has only been available for participants in the evaluation and during the evaluation, no subsequent publications have used the corpus. We expect that to change with its official publication in 2022.

8. Availability and Conclusion

The corpus will be distributed through the Linguistic Data Consortium (LDC). It will become part of the regular catalog in the near future; at the time of publication of this paper, the catalog number has not yet been determined. The LDC catalog can be searched at https://catalog.ldc.upenn.edu. The reader will be able to find the BeSt corpus by searching on "TAC KBP Belief and Sentiment".

Acknowledgments

The corpus presented in this paper was developed under the Defense Advanced Research Projects Agency (DARPA) DEFT program between 2012 and 2017. Authors were supported under contracts with Columbia University (Diab, Muresan, Prabhakaran, Rambow, Contract No. FA87501220347), Cornell University (Cardie, Contract No. FA87501320015), IHMC (Dalton, Guthrie, Contract No. FA87501220348), the Linguistic Data Consortium (Tracey, Arrigo, Contract No. FA87501320045), NIST (Dang, Contract Nos. HR001131727 and HR001149781), the University at Albany (Shaikh, Strzalkowski, Contract No. FA87501220348 to IHMC), the University of Pittsburgh (Wiebe, Contract No. FA87501320015 to Cornell). This paper was written with additional support from DARPA under Contracts No. HR001121C0186 (Dorr, Rambow), No. HR001120C0037 (Dorr), and No. HR0011154158 (Dorr, Markowska, Rambow). We would like to thank Lisa Ferro (MITRE) for very helpful suggestions about how to improve the presentation of this paper, as well as three anonymous reviewers for their comments. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the US Government.

9. Bibliographical References

Baker, C. F., Fillmore, J., and Lowe, J. B. (1998). The Berkeley FrameNet project. In *36th Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics (COLING-ACL'98)*, pages 86–90, Montréal.

de Marneffe, M.-C., Simons, M., and Tonhauser, J. (2019). The commitmentbank: Investigating projection in naturally occurring discourse. *Proceedings of Sinn und Bedeutung*, 23(2):107–124, Jul.

Deng, L. and Wiebe, J. (2015). MPQA 3.0: An entity/event-level sentiment corpus. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1323–1328, Denver, Colorado, May–June. Association for Computational Linguistics.

Diab, M., Levin, L., Mitamura, T., Rambow, O., Prabhakaran, V., and Guo, W. (2009). Committed belief annotation and tagging. In *Proceedings of the Third Linguistic Annotation Workshop (LAW III)*, pages 68–73, Suntec, Singapore, August. Association for Computational Linguistics.

Hovy, E., Marcus, M., Palmer, M., Ramshaw, L., and Weischedel, R. (2006). OntoNotes: The 90% solution. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, pages 57–60, New York City, USA, June. Association for Computational Linguistics.

Kingsbury, P., Palmer, M., and Marcus, M. (2002). Adding semantic annotation to the Penn TreeBank. In *Proceedings of the Human Language Technology Conference*, San Diego, CA.

Lee, K., Artzi, Y., Choi, Y., and Zettlemoyer, L. (2015). Event detection and factuality assessment with non-expert supervision. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1643–1648, Lisbon, Portugal, September. Association for Computational Linguistics.

Mather, B., Dorr, B. J., Rambow, O., and Strzalkowski, T. (2021). A general framework for domain-specialization of stance detection A covid-19 response use case. In Eric Bell et al., editors, *Proceedings of the Thirty-Fourth International Florida Artificial Intelligence Research Society Conference, North Miami Beach, Florida, USA, May 17-19, 2021*.

Minard, A.-L., Speranza, M., Urizar, R., Altuna, B., van Erp, M., Schoen, A., and van Son, C. (2016). MEANTIME, the NewsReader multilingual event and time corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4417–4422, Portorož, Slovenia, May. European Language Resources Association (ELRA).

O'Gorman, T., Regan, M., Griffitt, K., Hermjakob, U., Knight, K., and Palmer, M. (2018). AMR beyond the sentence: the multi-sentence AMR corpus. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3693–3702, Santa Fe, New Mexico, USA, August. Association for Computational Linguistics.

Prabhakaran, V., By, T., Hirschberg, J., Rambow, O., Shaikh, S., Strzalkowski, T., Tracey, J., Arrigo, M., Basu, R., Clark, M., Dalton, A., Diab, M., Guthrie, L., Prokofieva, A., Strassel, S., Werner, G., Wilks, Y., and Wiebe, J. (2015). A new dataset and evaluation for belief/factuality. In *Proceedings of the Fourth Joint Conference on Lexical and Computational Semantics*, pages 82–91, Denver, Colorado, June. Association for Computational Linguistics.

Rambow, O., Alagesan, M., Arrigo, M., Bauer, D., Cardie, C., Dalton, A., Diab, M., Dubbin, G., Katsios, G., Radeva, A., Strzalkowski, T., and Tracey, J. (2016). The 2016 TAC KBP BeSt evaluation. In *Proceedings of the 2016 TAC KBP Conference*. NIST.

Ross, A. and Pavlick, E. (2019). How well do NLI models capture verb veridicality? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2230–2240, Hong Kong, China, November. Association for Computational Linguistics.

Rudinger, R., White, A. S., and Van Durme, B. (2018). Neural models of factuality. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 731–744, New Orleans, Louisiana, June. Association for Computational Linguistics.

Saurí, R. and Pustejovsky, J. (2009). FactBank: a corpus annotated with event factuality. *Language Resources and Evaluation*, 43:227–268. 10.1007/s10579-009-9089-9.

Song, Z., Bies, A., Strassel, S., Riese, T., Mott, J., Ellis, J., Wright, J., Kulick, S., Ryant, N., and Ma, X. (2015). From light to rich ERE: Annotation of entities, relations, and events. In *Proceedings of the The 3rd Workshop on EVENTS: Definition, Detection, Coreference, and Representation*, pages 89–98, Denver, Colorado, June. Association for Computational Linguistics.

Stanovsky, G., Eckle-Kohler, J., Puzikov, Y., Dagan, I., and Gurevych, I. (2017). Integrating deep linguistic features in factuality prediction over unified datasets. In *ACL*.

Tonhauser, J., Beaver, D. I., and Degen, J. (2018). How Projective is Projective Content? Gradience in Projectivity and At-issueness. *Journal of Semantics*, 35(3):495–542, 06.

White, A. S., Rudinger, R., Rawlins, K., and Van Durme, B. (2018). Lexicosyntactic inference in neural models. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4717–4724, Brussels, Belgium, October-November. Association for Computational Linguistics.

Wiebe, J., Wilson, T., and Cardie, C. (2005). Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, 39(2/3):164–210.