A Pragmatics-Centered Evaluation Framework for Natural Language Understanding

Damien Sileo, Tim van de Cruys, Camille Pradel, Philippe Muller

Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 2382–2394
Marseille, 20-25 June 2022

A Pragmatics-Centered Evaluation Framework for Natural Language Understanding

Damien Sileo[3,2], **Tim Van de Cruys**[2], **Camille Pradel**[1], **Philippe Muller**[3,4]
1:Synapse Développement, 2:KU Leuven, 3:IRIT, University of Toulouse
4: Artificial and Natural Intelligence Toulouse Institute (ANITI)
damien.sileo@kuleuven.be

Abstract

New models for natural language understanding have recently made an unparalleled amount of progress, which has led some researchers to suggest that the models induce *universal* text representations. However, current benchmarks are predominantly targeting semantic phenomena; we make the case that pragmatics needs to take center stage in the evaluation of natural language understanding. We introduce PragmEval, a new benchmark for the evaluation of natural language understanding, that unites 11 pragmatics-focused evaluation datasets for English. PragmEval can be used as supplementary training data in a multi-task learning setup, and is publicly available, alongside the code for gathering and preprocessing the datasets. Using our evaluation suite, we show that natural language inference, a widely used pretraining task, does not result in genuinely universal representations, which presents a new challenge for multi-task learning.

1. Introduction

Over the last few years, pretrained for natural language understanding (NLU) have made a remarkable amount of progress on a number of widely accepted evaluation benchmarks. The GLUE benchmark (Wang et al., 2018), for example, was designed to be a set of challenging NLU tasks, such as question answering, sentiment analysis, and textual entailment; yet, current state of the art systems surpass human performance estimates on the average score of its subtasks (Yang et al., 2019). Similarly, the NLU subtasks that are part of the SentEval (Conneau et al., 2017) framework, a widely used benchmark for the evaluation of sentence-to-vector encoders, are successfully dealt with by current neural models, with scores that exceed the 90% mark.[1]

The results on these benchmarks are impressive, but sometimes lead to excessive optimism regarding the ability of current NLU models. For example, based on the resulting performance on the above-mentioned benchmarks, a considerable number of researchers has even put forward the claim that their models induce *universal* representations (Cer et al., 2018; Kiros and Chan, 2018; Subramanian et al., 2018; Wieting et al., 2015; Liu et al., 2019). It is important to note, however, that benchmarks like SentEval and GLUE are primarily focusing on semantic aspects, i.e. the literal and uncontextualized content of text. While the semantics of language is without doubt an important aspect of language, we believe that a single focus on semantic aspects leads to an impoverished model of language. For a versatile model of language, other aspects of language, viz. pragmatic aspects, equally need to be taken into account. Pragmatics focuses on the larger context that surrounds a particular textual instance, and it is of vital importance for meaning representations that aspire to lay a claim to universality. Consider the following utterance :

(1) You're standing on my foot.

The utterance in (1) has a number of direct implications that are logically entailed, such as the implication that the hearer is standing on a body part of the speaker, or the implication that the speaker is touching the hearer. But there are also more indirect implications, that are not literally expressed, but need to be inferred from the context, such as the implication that the speaker wants the hearer to move away from them. The latter kind of implication, that is indirectly implied by the context of an utterance, is called *implicature*—a term coined by Grice (1975). In real world applications, recognizing the implicatures of a statement is arguably more important than recognizing its mere semantic content.

The implicatures that are conveyed by an utterance are highly dependent on its illocutionary force (Austin, 1975). In Austin's framework, the *lo-*

[1] `http://nlpprogress.com/english/`

cution is the literal meaning of an utterance, while the *illocution* is the goal that the utterance tries to achieve. When we restrict the meaning of (1) to its locution, the utterance is reduced to the mere statement that the hearer is standing on the speaker's foot. However, when we also take its illocution into account, it becomes clear that the speaker actually formulates the request that the speaker step away. The utterance's illocution is clearly an important part of the entire meaning of the utterance, that is complementary to the literal content (Green, 2000).[2]

The example above makes clear that pragmatics is a fundamental aspect of the meaning of an utterance. Semantics focuses on the literal content of utterances, but not on the kind of goal the speaker is trying to achieve. Pragmatics and discourse tasks focus on the actual use of language, so a pragmatics-centered evaluation could *by construction* be a better fit to evaluate how NLU models perform in practical use cases—and in any case it should at least be used as a complement to semantics-focused evaluation benchmarks. Ultimately, many use cases of NLP models are related to conversations with end users or analysis of structured documents. In such cases, discourse analysis (i.e. the ability to parse high-level textual structures that take into account the global context) is a prerequisite for human level performance. Moreover, standard benchmarks often strongly influence the evolution of NLU models, which means they should be as exhaustive as possible, and closely related to the models' end use cases.

In this work, we compile a list of eleven pragmatics-focused tasks for English that are meant to complement existing benchmarks. We propose: (i) a new evaluation benchmark, named *PragmEval*, which is publicly available; [3] (ii) derivations of human accuracy estimates for some of the tasks; (iii) evaluation on these tasks of a state of the art generalizable NLU model, viz. BERT (both with and without auxiliary finetunings); (iv) new comparisons of discourse-based and natural language inference based training signals, showing that the most widely used auxiliary finetuning dataset (MNLI) is not the best performing on PragmEval, which suggests a margin for improvement.

[2]In order to precisely determine their illocution, utterances have been categorized into classes called speech acts (Searle et al., 1980), such as ASSERTION, QUESTION or ORDER which have different kinds of effects on the world. For instance, constative speech acts (e.g. *the sky is blue*) describe a state of the world and are either true or false while performative speech acts (e.g. *I declare you husband and wife*) can change the world upon utterance (Austin, 1975).

[3]`https://github.com/synapse-developpement/PragmEval` and `https://huggingface.co/datasets/pragmeval`

2. Related Work

Evaluation methods of NLU have been the object of heated debates since the proposal of the Turing Test. Automatic evaluations relying on annotated datasets are arguably limited but they have become customary practice. A popular method of evaluation is to predict sentence similarity (Agirre et al., 2012), leveraging human annotated scores of similarity between sentence pairs. This task requires some representation of the sentences' semantic content beyond their surface form, and sentence similarity estimation tasks can potentially encompass many aspects. However, it is not clear how human annotators weigh semantic, stylistic, and discursive aspects while rating.

Using a set of more focused and clearly defined tasks has been another popular approach. Kiros et al. (2015) proposed a set of tasks and tools for sentence understanding evaluation. These thirteen tasks were compiled in the SentEval (Conneau et al., 2017) evaluation suite designed for automatic evaluation of pre-trained sentence embeddings. SentEval tasks are mostly based on sentiment analysis, semantic sentence similarity and natural language inference. Since SentEval evaluates sentence embeddings, the users have to provide a sentence encoder that is not fine-tuned during the evaluation.

GLUE (Wang et al., 2018) proposes to evaluate language understanding with less constraints than SentEval, allowing users not to rely on explicit sentence embedding based models. GLUE consists of nine classification or regression tasks that are carried out for sentences or sentence pairs. Three tasks focus on semantic similarity, and four tasks are based on NLI, which makes GLUE arguably semantics-based, even though it also includes sentiment classification (Socher et al., 2013) and grammaticality judgment (Warstadt et al., 2018).

NLI can be regarded as a universal framework for evaluation. In the *Recast* framework (Poliak et al., 2018), existing datasets (e.g. sentiment analysis) are formulated as NLI tasks. For instance,

based on the sentence *don't waste your money*, annotated as a negative review, they use handcrafted rules to generate the following example: (PREMISE: *When asked about the product, Liam said "don't waste your money"* , HYPOTHESIS: *Liam didn't like the product*, LABEL: entailment). However, the generated datasets do not allow to directly measure how well a model deals with the semantic phenomena present in the original dataset, since some sentences use artificially generated reported speech. Likewise, NLI data could be used to evaluate pragmatics and discourse analysis, but it is not clear how to generate examples that are not overly artificial. Moreover, it is unclear to what extent instances in existing NLI datasets need to deal with pragmatic aspects (Bowman, 2016).

SuperGLUE (Wang et al., 2018) updates GLUE with six novel tasks that are selected to be even more challenging. Two of those tasks deal with contextualized lexical semantics, another two tasks are a form of question answering, and the remaining two are NLI problems. Only one of these NLI tasks, viz. CommitmentBank (de Marneffe et al., 2019), is related to pragmatics.

Another effort towards evaluation of general purpose NLP systems is DecaNLP (McCann et al., 2018). The ten tasks of this benchmark are all framed as question answering. For example, a question answering task is derived from a sentiment analysis task using artificial questions such as *Is this sentence positive or negative?* Four of these tasks deal with semantic parsing, and other tasks include NLI and sentiment analysis. Pragmatic phenomena can be involved in some tasks (e.g. the summarization task) although it is hard to assess to what extent.

Discourse relation prediction has punctually been used for sentence representation learning evaluation, by Nie et al. (2019) and Sileo et al. (2019b), but they all used only one dataset (viz. PDTB; Prasad et al., 2008), which we included in our benchmark. Discourse has also been considered for evaluation in the field of machine translation. Läubli et al. (2018) showed that neural models achieve superhuman results on sentence-level translations but that current models yield underwhelming results when considering document-level translations, also making a case for discourse-aware evaluations. DiscoEval (Chen et al., 2019) proposed a more principled evaluation of discourse modeling in NLP models. However, they mirror SentEval in that they rely on sentence embeddings and fixed compositions, which has been shown to be restrictive (Sileo et al., 2019a), and not necessarily in line with state of the art systems. Moreover, they focus on rather shallow aspects of document structure such as the position of sentences within a document.

Other evaluations, such as linguistic probing or GLUE diagnostics (Conneau et al., 2018; Belinkov and Glass, 2019; Wang et al., 2019b) focus on an internal understanding of what is captured by the models (e.g. syntax, lexical content), rather than measuring performance on external tasks; this provides a complementary viewpoint, but it is outside the scope of this work.

3. PragmEval

3.1. Construction

Our goal is to compile a set of diverse pragmatics-related tasks. We restrict ourselves to classification either of sentences or sentence pairs, and only use publicly available datasets that are absent from other well-established benchmarks (such as SentEval, GLUE, and SuperGLUE), in order to have complementary benchmarks.

The scores in our tasks are not all meant to be compared to previous work, since we alter some datasets to yield more meaningful evaluations (we perform duplicate removal or class subsampling when mentioned). We found these operations necessary in order to leverage the rare classes and yield more meaningful scores. As an illustration, the GUM discourse corpus initially consists of more than 99% of *unattached* labels, and the dialog act annotations of the SwitchBoard conversation corpus contains 80% of *statements*. While disturbing the distributions of labels impacts the performance of models in real-world contexts, it seems reasonable when the goal is to indirectly evaluate the capacity of models to discriminate different semantic or pragmatic phenomena.

Section 3.2. presents the tasks we selected, while 3.3. proposes a rudimentary taxonomy of how they address different aspects of meaning. A summary of the tasks, together with some examples, is also given in table 1.

3.2. Task overview

PDTB The Penn Discourse Tree Bank (Prasad et al., 2014) contains a collection of fine-grained implicit (i.e. not signaled by a discourse marker) and explicit relations between sentences from the

news domain in the Penn TreeBank 2.0, which signal the purpose of an utterance given a context utterance. Explicit relations can be easily predicted from the discourse marker alone (Pitler et al., 2008) and are discarded. We select the level 2 relations, called types in PDTB terminology, as categories.

STAC (Asher et al., 2016) is a corpus of strategic chat conversations manually annotated with negotiation-related information, dialogue acts and discourse structures in the framework of Segmented Discourse Representation Theory (Asher and Lascarides, 2003). We only consider pairwise relations between all dialogue acts, following Badene et al. (2019). We remove duplicate pairs and dialogues that only have non-linguistic utterances (coming from the game server). We subsample dialogue act pairs with no relation so that they constitute 20% of each fold.

GUM (Zeldes, 2017) is a corpus of multilayer annotations for texts from various domains; it includes discourse structure annotations according to Rhetorical Structure Theory (Mann and Thompson, 1987). Once again, we only consider pairwise interactions between discourse units (e.g. sentences/clauses). We subsample discourse units with no relation so that they constitute 20% of each document. We split the examples in train/test/dev sets randomly according to the document they belong to.

Emergent (Ferreira and Vlachos, 2016) is composed of pairs of assertions and titles of news articles that are *against*, *for*, or *neutral* with respect to the opinion of the assertion.

SwitchBoard (Godfrey et al., 1992) contains textual transcriptions of dialogues about various topics with annotated speech acts. We remove duplicate examples and subsample *Statements* and *Non Statements* so that they constitute 20% of the examples. We use a custom train/validation split (90/10 ratio) since our preprocessing leads to a drastic size reduction of the original development set. The label of a speech act can be dependent on the context (previous utterances), but we discarded it in this work for the sake of simplicity, even though integration of context could improve the scores (Ribeiro et al., 2015).

MRDA (Shriberg et al., 2004) contains textual transcriptions of multi-party real meetings, with speech act annotations. We remove duplicate examples. We use a custom train/validation split (90/10 ratio) since this deduplication leads to a drastic size reduction of the original development set, and we subsample *Statement* examples so that they constitute 20% of the dataset. We also discarded the context.

Persuasion (Carlile et al., 2018) is a collection of arguments from student essays annotated with factors of persuasiveness with respect to a claim; considered factors are the following: Specificity, Eloquence, Relevance and Strength. For each graded target, we cast the ratings into three quantiles and discard the middle quantile.

SarcasmV2 (Oraby et al., 2016) consists of messages from online forums with responses that may or may not be sarcastic according to human annotations.

Squinky dataset (Lahiri, 2015) gathers annotations on Formality, Informativeness, and Implicature, where sentences were graded on a scale from 1 to 7. The Implicature score is defined as the amount of information that is not explicitly expressed in a sentence. For each target, we cast the ratings into three quantiles and discard the middle quantile.

Verifiability (Park and Cardie, 2014) is a collection of online user comments annotated as *Verifiable-Experiential* (verifiable and about writer's experience), *Verifiable-Non-Experiential*, or *Unverifiable*.

EmoBank (Buechel and Hahn, 2017) aggregates emotion annotations on texts from various domains using the VAD representation format. The authors define Valence as *corresponding to the concept of polarity*,[4] Arousal as *degree of calmness or excitement* and Dominance as *perceived degree of control over a situation*. For each target, we cast the ratings into three quantiles and discard the middle quantile.

3.3. Taxonomy

It has been argued by Halliday (1985) that linguistic phenomena fall into three metafunctions: *ideational* for semantics, *interpersonal* for appeals to the hearer/reader, and *textual* for form-related aspects. This forms the basis of discourse relation types by Hovy and Maier (1992), who call them semantic, interpersonal and presentational.

[4]This is the dimension that is widely used in sentiment analysis.

Dataset	Example	Class
PDTB	*it was censorship / it was outrageous*	conjunction
STAC	*what? / i literally lost*	question-answer-pair
GUM	*Do not drink / if underage in your country*	condition
Emergent	*a meteorite landed in nicaragua. / small meteorite hits managua*	for
SwitchBoard	*well, a little different, actually*	hedge
MRDA	*yeah that's that's that's what i meant .*	acknowledge-answer
Persuasion	*Co-operation is essential for team work / lions hunt in a team*	low specificity
SarcasmV2	*don't quit your day job / [...] i was going to sell this joke. [...]*	sarcasm
Squinky	*boo ya.*	uninformative, high implicature, informal
Verifiability	*I've been a physician for 20 years.*	verifiable-experiential
EmoBank	*I wanted to be there..*	low valence, high arousal, low dominance

Table 1: Example instances for each of the PragmEval tasks (these examples were selected for their conciseness and are not representative of the whole dataset)

PragmEval tasks cut across these categories, because some of the tasks integrate all aspects when they characterize the speech act or discourse relation category associated to a discourse unit (mostly sentences), an utterance, or a pair of these. However, most discourse relations involved focus on *ideational* aspects, which are thus complemented by tasks insisting on more interpersonal aspects (e.g. using appeal to emotions, or verifiable arguments) that help realize speech act intentions. Finally, intentions can achieve their goals with varying degrees of success. This leads us to a rudimentary grouping of our tasks:

A The speech act classification tasks (SwitchBoard, MRDA) deal with the detection of the intention of utterances. They use the same label set (Core and Allen, 1997) but different domains and annotation guidelines. Similarly, a discourse relation characterizes how an utterance contributes to the coherence of a document/conversation (e.g through *elaboration* or *contrast*), so this task requires a form of understanding of the use of a sentence, and how a sentence fits with another sentence in a broader discourse. A discourse relation can be seen as a speech act whose definition is tied to a structured context (Asher and Lascarides, 2003). Here, three tasks (PDTB, STAC, GUM) deal with discourse relation prediction with varying domains and formalisms.[5] The Stance detection task can be seen as a coarse-grained discourse relation classification.

B Detecting emotional content, verifiability, formality, informativeness or sarcasm is necessary in order to figure out in what realm communication is occurring. A statement can be persuasive, yet poorly informative and unverifiable. Emotions (Dolan, 2002) and power perception (Pfeffer, 1981) can have a strong influence on human behavior and text interpretation. Manipulating emotions can be the main purpose of a speech act as well. Sarcasm is another means of communication and sarcasm detection is in itself a suitable task for the evaluation of pragmatics, since sarcasm is a clear case of literal meaning being different from the intended meaning.

C Persuasiveness prediction is a useful tool to assess whether a model can measure how well a sentence can achieve its intended goal. This aspect is orthogonal to the determination of the goal itself, and is arguably equally important.

Note that the semantic tasks of GLUE can also be considered as a grouping of tasks, where the goal is to represent accurately the denotation of utterances (e.g. the identity of the objects and agents they involve, the relation between them, the temporal and spatial location). In contrast, solving PragmEval tasks requires knowledge of complementary aspects that characterize utterances in a different way. The A group characterizes the kind of frame into which semantic content fits; for instance, identical subjects, verbs, and objects can be used in a question, a claim, or an instruction. Semantic tasks (semantic similarity, NLI) usually compare

[5]These formalisms have different assumptions about the nature of discourse structure.

utterances within the same frame. Additionally, utterances with the same semantic content can differ according to aspects involved in group B and C, e.g. formality or persuasiveness. To ensure that these aspects are taken into account by NLU models, a pragmatic evaluation is required.

4. Evaluation

4.1. Models

Our goal is to assess the performance of popular NLU models and the influence of various training signals on PragmEval scores. We evaluate state of the art models and baselines on PragmEval using the Jiant framework (Wang et al., 2019c). Our baselines consist of an average of GloVe (Pennington et al., 2014) embeddings (CBoW), and a BiLSTM with both GloVe and ELMo (Peters et al., 2018) embeddings. We equally evaluate BERT (Devlin et al., 2019) base uncased models, and perform experiments with *Supplementary Training on Intermediate Labeled-data Tasks* (Phang et al., 2018). STILT is a further pretraining step on a data-rich task before the final fine-tuning evaluation on the target task. STILTs can be combined using multitask learning. We use Jiant's default parameters,[6] and uniform loss weighting when multitasking (a different task is optimized at each training batch). We finetune BERT with four of such training signals:

MNLI (Williams et al., 2018) is a collection of 433k sentence pairs manually annotated with *contradiction*, *entailment*, or *neutral* relations. Phang et al. (2018) showed that finetuning with this dataset leads to accuracy improvement on all GLUE tasks except CoLA (Warstadt et al., 2018).

DisSent (Nie et al., 2019) consists of $4.7M$ sentence pairs that are separated by a discourse marker (from a list of 15 markers). Prediction of discourse markers based on the context clauses/sentences with which they occur has been used as a training signal for sentence representation learning. The authors used handcrafted rules for each marker in order to ensure that the markers signal an actual relation. DisSent has underwhelming results on the GLUE tasks as a STILT (Wang et al., 2019a).

Discovery (Sileo et al., 2019b) is another dataset for discourse marker prediction, composed of 174 discourse markers with $10k$ usage examples for each marker. Sentence pairs were extracted from web data, and the markers come either from the PDTB or from a heuristic automatic extraction.

PragmEval refers to all PragmEval tasks used in a multitask setup; since we use a uniform loss weighting, we discard Persuasion classes other than Strength (note that the other classes can be considered subfactors for strength) in order to prevent the Persuasion task to overwhelm the others.

4.2. Human accuracy estimates

For a more insightful comparison, we propose derivations of human accuracy estimates for the datasets we used. The authors of SarcasmV2 (Oraby et al., 2016) dataset directly report 80% annotator accuracy compared to the gold standard. Prasad et al. (2014) report 84% annotator agreement for PDTB 2.0, which is a lower bound of accuracy. For GUM (Zeldes, 2017), an *attachment accuracy of* 87.22% *and labelling accuracy of* 86.58% *as compared to the 'gold standard' after instructor adjudication* is reported. We interleaved attachment and labelling in our task. Assuming human annotators never predict the non-attached relation, 69.3% is a lower bound for human accuracy. Authors of the Verifiability (Park and Cardie, 2014) dataset report an agreement $\kappa = 0.73$ which yields an agreement of 87% given the class distribution, which is a lower bound of human accuracy. We estimated human accuracy on EmoBank (Buechel and Hahn, 2017) with the intermediate datasets provided by the authors. For each target (V,A,D) we compute the average standard deviation, and compute the probability (under normality assumption) of each example rating of falling under the wrong category.

Unlike the GLUE benchmark (Nangia and Bowman, 2019), we do not yet provide human accuracy estimates obtained in a standardized way. The high number of classes would make that process rather more difficult. But our estimates are still useful even though they should be taken with a grain of salt.

4.3. Overall results

Task-wise results are presented in table 2. We report the average scores of 6 runs of STILT and finetuning phases.

PragmEval seems to be challenging even for the BERT base model, which has shown strong performance on GLUE (and vastly outperforms the

[6] https://github.com/nyu-mll/jiant/jiant/config/examples/stilts_example.conf

	PDTB	STAC	GUM	Emergent	SwitchB.	MRDA	Persuasion	Sarcasm	Squinky	Verif.	EmoBank
CBoW	27.4	32	20.5	59.7	3.8	0.7	70.6	61.1	75.5	74.0	64.0
BiLSTM	25.9	27.7	18.5	45.6	3.7	0.7	62.6	63.1	72.1	74.0	63.5
BiLSTM+ELMo	27.5	33.5	18.9	55.2	3.7	0.7	67.4	68.9	82.5	74.0	66.9
Previous work	48.2	-	-	73.1	-	-	-	-	-	81.1	-
BERT	48.8	48.2	40.9	79.2	38.8	22.3	74.8	**77.1**	87.5	86.7	76.2
BERT+MNLI	49.1	49.1	42.8	81.2	38.1	22.7	71.7	73.4	88.2	86.0	76.3
BERT+PragmEval	49.1	**57.1**	42.8	80.2	**40.3**	**23.1**	**76.2**	75.0	87.6	85.9	76.0
BERT+DisSent	49.4	49.0	43.9	79.8	39.2	22.0	74.7	74.9	87.5	85.9	76.2
B+DisSent+MNLI	49.6	49.2	**44.2**	80.9	39.8	22.1	74.0	74.1	87.6	85.6	76.4
BERT+Discovery	50.7	49.5	42.7	**81.7**	39.5	22.4	71.6	76.7	88.6	86.3	**76.6**
B+Discovery+MNLI	**51.3**	49.4	43.1	80.7	**40.3**	22.2	73.6	75.1	**88.9**	**86.8**	76.0
Human estimate	**84.0**	-	**69.3**	-	-	-	-	**80.0**	-	**87.0**	73.1

Table 2: Transfer test scores across PragmEval tasks; we report the average when the dataset has several classification tasks (as in Squinky, EmoBank and Persuasion); B(ERT)+$\mathcal{X}$ refers to BERT pretrained classification model after an auxiliary finetuning phase on task $\mathcal{X}$. All scores are accuracy scores except SwitchBoard/MRDA, which are macro-F1 scores. *Previous work* refers to the best scores from previous work that used a similar setup, where PDTB score is from (Bai and Zhao, 2018), Emergent score is from (Ferreira and Vlachos, 2016) and Verifiability score is derived from (Park and Cardie, 2014).

baselines on our tasks). For many tasks, there is a STILT that significantly improves the accuracy of BERT. The gap between human accuracy and BERT is particularly high on implicit discourse relation prediction (both for the PDTB corpus and the GUM RST corpus). This task is known to be difficult, and previous work has also shown that task dedicated models are not yet on par with human performance either on PDTB (Bai and Zhao, 2018) or RST data (Morey et al., 2017).

Pretraining on MNLI does not improve the PragmEval average score for the BERT base model. A lower sarcasm detection score could indicate that BERT+MNLI is more focused on the literal content of statements, even though no STILT improves sarcasm detection. All models score below human accuracies, with the exception of emotion classification (but it is only due to the valence prediction subtask).

Table 3 shows aggregate results alongside comparisons with GLUE scores. The best overall unsupervised result (GLUE+PragmEval average) is achieved with Discovery STILT. Combining Discovery and MNLI yields both a high PragmEval and GLUE score, and also yields a high GLUE diagnostics score. All discourse based STILTs improve GLUE score, while MNLI does not improve PragmEval average score. PragmEval tasks based on sentence pairs seem to account for the variance across STILTs.

MNLI has been suggested as a good default auxiliary training task based on evaluation with GLUE (Phang et al., 2018) and SentEval (Conneau et al., 2017). However, our evaluation suggests that finetuning a model with MNLI alone has significant drawbacks.

More detailed results for datasets with several subtasks are shown in table 4. We note that MNLI STILT significantly decreases relevance estimation performance (on BERT base and while multitasking with DisSent). Many models surpass the human estimate at valence prediction, a well studied task, but interestingly this is not the case for Arousal and Dominance prediction.

The categories of our benchmark tasks cover a broad range of pragmatic aspects. The overall accuracies only show a synthetic view of the tasks evaluated in PragmEval. Some datasets contain many subcategories that allow for a fine grained analysis through a wide array of classes (e.g. 51 categories for MRDA). Table 5 in appendix A shows a fine grained evaluation which yields some insights on the capabilities of BERT. We report the 5 most frequent classes per task. It is worth noting that the BERT models do not neglect rare classes. These detailed results reveal that BERT+MNLI scores for discourse relation prediction are inflated by good scores on predicting the absence of relation (pos-

	PragmEval$_{AVG}$	P.E.-Pairs$_{AVG}$	P.E.-Single$_{AVG}$	GLUE$_{AVG}$	GLUE$_{diagnostics}$
BERT	61.8±.4	57.9±.5	62.3±.3	74.7±.2	31.7±.3
BERT+MNLI	61.7±.5	57.2±.5	62.2±.4	**77.0**±.2	32.5±.6
BERT+PragmEval MTL	**63.0**±.4	**60.0**±.4	62.6±.2	75.3±.2	31.6±.3
BERT+DisSent	62.0±.4	58.4±.4	62.2±.3	75.1±.2	31.5±.3
B+DisSent+MNLI	62.1±.4	58.2±.4	62.3±.2	76.6±.1	32.4±.0
BERT+Discovery	62.4±.3	58.2±.4	62.7±.3	75.0±.2	31.3±.2
B+Discovery+MNLI	62.5±.4	58.5±.5	**62.8**±.3	76.6±.2	**33.3**±.2

Table 3: Aggregated transfer test accuracies across PragmEval and comparison with GLUE validation downstream and diagnostic tasks (GLUE diagnostic tasks evaluate NLI performance under presence of linguistic phenomena such as negation, quantification, use of common sense); BERT+$\mathcal{X}$ refers to BERT pretrained classification model after auxiliary finetuning phase on task $\mathcal{X}$; P.E.-Pairs$_{AVG}$ is the average of PragmEval sentence pair classification tasks.

	Persuasiveness				EmoBank			Squinky		
	Eloquence	Relevance	Specificity	Strength	Valence	Arousal	Dom.	Inf.	Implicature	Formality
BERT	75.6	63.5	81.6	78.3	87.1	72.0	69.5	92.2	72.1	98.3
BERT+MNLI	74.7	57.5	82.3	72.2	86.6	72.4	69.9	92.5	73.9	98.1
BERT+PragmEval	75.6	**64.0**	83.2	**82.0**	86.8	71.9	69.2	92.3	71.8	**98.6**
BERT+DisSent	73.8	63.0	82.6	79.5	87.1	71.4	70.1	92.6	72.0	97.7
B+DisSent+MNLI	**76.9**	61.5	**83.9**	73.9	**87.6**	72.1	69.4	91.5	73.4	97.9
BERT+Discovery	76.0	59.1	80.1	71.4	86.8	**72.6**	**70.5**	**93.2**	74.2	98.5
B+Discovery+MNLI	74.1	60.4	79.4	80.4	86.4	72.1	69.6	93.1	**75.3**	98.4
Human estimate	-	-	-	-	74.9	**73.8**	**70.5**	-	-	-

Table 4: Transfer test accuracies across PragmEval subtasks (Persuasiveness, EmoBank, Squinky) BERT+$\mathcal{X}$ refers to BERT pretrained classification model after auxiliary finetuning phase on task $\mathcal{X}$.

sibly close to the neutral class in NLI), which is useful but not sufficient for pragmatics understanding. The STILTs have complementary strengths even with given tasks, which can explain why combining them is helpful. However, we used a rather simplistic multitask setup, and efficient combination of the tasks remains an open problem.

5. Conclusion

We proposed PragmEval, a set of pragmatics related evaluation tasks, and used them to evaluate BERT finetuned on various auxiliary finetuning tasks. The results lead us to rethink the efficiency of mainly using NLI as an auxiliary training task. PragmEval can be used for training or evaluating NLU or pragmatics related work in general. Much effort has been devoted to NLI for training and evaluation for general purpose sentence understanding, but we just scratched the surface of the use of pragmatics oriented tasks. In further investigations, we plan to use more general tasks than classification on sentences or sentence pairs, such as longer and possibly structured sequences. Several of the datasets we used (MRDA, SwitchBoard, GUM, STAC) already contain such higher level structures. Of course defining a generic architecture for structured tasks in which to evaluate the contribution of trained representations is not straightforward. In addition, a more inclusive comparison with human annotators on pragmatics tasks could also help to pinpoint the weaknesses of current models dealing with pragmatics phenomena. Yet another step would be to study the correlations between performance metrics in deployed NLU systems and scores of the automated evaluation benchmarks (GLUE/PragmEval) in order to validate our claims about the centrality of pragmatics.

6. Acknowledgements

This work is part of the CALCULUS project, which is funded by the ERC Advanced Grant H2020-ERC-2017. ADG 788506[7].
Philippe Muller is partially supported by the ANR (ANR-19-PI3A-0004) through the AI Interdisci-

[7] https://calculus-project.eu/

plinary Institute, ANITI, as a part of France's "Investing for the Future - PIA" program. He is also part of the programme DesCartes and is supported by the National Research Foundation, Prime Minister's Office, Singapore under its Campus for Research Excellence and Technological Enterprise (CREATE) programme.

7. Bibliographical References

Agirre, E., Diab, M., Cer, D., and Gonzalez-Agirre, A. (2012). Semeval-2012 task 6: A pilot on semantic textual similarity. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics – Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, pages 385–393. Association for Computational Linguistics.

Asher, N. and Lascarides, A. (2003). *Logics of conversation*. Cambridge University Press.

Asher, N., Hunter, J., Morey, M., Farah, B., and Afantenos, S. (2016). Discourse structure and dialogue acts in multiparty dialogue: the STAC corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2721–2727, Portorož, Slovenia, May. European Language Resources Association (ELRA).

Austin, J. L. (1975). *How to do things with words*. Oxford university press.

Badene, S., Thompson, K., Lorré, J.-P., and Asher, N. (2019). Data programming for learning discourse structure. In *Proceedings of the 57th Conference of the Association for Computational Linguistics*, pages 640–645, Florence, Italy, July. Association for Computational Linguistics.

Bai, H. and Zhao, H. (2018). Deep enhanced representation for implicit discourse relation recognition. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 571–583, Santa Fe, New Mexico, USA, August. Association for Computational Linguistics.

Belinkov, Y. and Glass, J. (2019). Analysis methods in neural language processing: A survey. *Transactions of the Association for Computational Linguistics*, 7:49–72.

Bowman, S. R. (2016). *Modeling natural language semantics in learned representations*. Ph.D. thesis.

Buechel, S. and Hahn, U. (2017). EmoBank: Studying the impact of annotation perspective and representation format on dimensional emotion analysis. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 578–585, Valencia, Spain, April. Association for Computational Linguistics.

Carlile, W., Gurrapadi, N., Ke, Z., and Ng, V. (2018). Give me more feedback: Annotating argument persuasiveness and related attributes in student essays. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 621–631, Melbourne, Australia, July. Association for Computational Linguistics.

Cer, D., Yang, Y., yi Kong, S., Hua, N., Limtiaco, N., John, R. S., Constant, N., Guajardo-Cespedes, M., Yuan, S., Tar, C., Sung, Y.-H., Strope, B., and Kurzweil, R. (2018). Universal sentence encoder.

Chen, M., Chu, Z., and Gimpel, K. (2019). Evaluation benchmarks and learning criteria for discourse-aware sentence representations. 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 649–662, Hong Kong, China, November. Association for Computational Linguistics.

Conneau, A., Kiela, D., Schwenk, H., Barrault, L., and Bordes, A. (2017). Supervised Learning of Universal Sentence Representations from Natural Language Inference Data. *Emnlp*.

Conneau, A., Kruszewski, G., Lample, G., Barrault, L., and Baroni, M. (2018). What you can cram into a single vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2126–2136. Association for Computational Linguistics.

Core, M. G. and Allen, J. (1997). Coding dialogs with the DAMSL annotation scheme. In *AAAI fall symposium on communicative action in humans and machines*, volume 56, pages 28–35. Boston, MA.

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.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova,

K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Association for Computational Linguistics.

Dolan, R. J. (2002). Emotion, cognition, and behavior. *Science*, 298(5596):1191–1194.

Ferreira, W. and Vlachos, A. (2016). Emergent: a novel data-set for stance classification. In *HLT-NAACL*.

Godfrey, J. J., Holliman, E. C., and McDaniel, J. (1992). Switchboard: Telephone speech corpus for research and development. In *Proceedings of the 1992 IEEE International Conference on Acoustics, Speech and Signal Processing - Volume 1*, ICASSP'92, pages 517–520, Washington, DC, USA. IEEE Computer Society.

Green, M. S. (2000). Illocutionary force and semantic content. *Linguistics and Philosophy*, 23(5):435–473.

Grice, H. P. (1975). Logic and conversation. In Peter Cole et al., editors, *Syntax and Semantics: Vol. 3: Speech Acts*, pages 41–58. Academic Press, New York.

Halliday, M. (1985). *An Introduction to Functional Grammar*. Edward Arnold Press, Baltimore.

Hovy, E. and Maier, E. (1992). Parsimonious or profligate: How many and which discourse structure relations? Technical Report RR-93-373, USC Information Sciences Institute.

Kiros, J. and Chan, W. (2018). InferLite: Simple universal sentence representations from natural language inference data. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4868–4874, Brussels, Belgium, October-November. Association for Computational Linguistics.

Kiros, R., Zhu, Y., Salakhutdinov, R. R., Zemel, R., Urtasun, R., Torralba, A., and Fidler, S. (2015). Skip-thought vectors. In *Advances in neural information processing systems*, pages 3294–3302.

Lahiri, S. (2015). SQUINKY! A Corpus of Sentence-level Formality, Informativeness, and Implicature. *CoRR*, abs/1506.02306.

Läubli, S., Sennrich, R., and Volk, M. (2018). Has machine translation achieved human parity? a case for document-level evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4791–4796, Brussels, Belgium, October-November. Association for Computational Linguistics.

Liu, X., He, P., Chen, W., and Gao, J. (2019). Improving multi-task deep neural networks via knowledge distillation for natural language understanding. *arXiv preprint arXiv:1904.09482*.

Mann, W. and Thompson, S. (1987). Rhetorical structure theory : a theory of text organization. Technical report, Information Science Institute.

McCann, B., Keskar, N. S., Xiong, C., and Socher, R. (2018). The natural language decathlon: Multitask learning as question answering. *arXiv preprint arXiv:1806.08730*.

Morey, M., Muller, P., and Asher, N. (2017). How much progress have we made on RST discourse parsing? a replication study of recent results on the RST-DT. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1319–1324, Copenhagen, Denmark, September. Association for Computational Linguistics.

Nangia, N. and Bowman, S. R. (2019). Human vs. muppet: A conservative estimate of human performance on the GLUE benchmark. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4566–4575, Florence, Italy, July. Association for Computational Linguistics.

Nie, A., Bennett, E. D., and Goodman, N. D. (2019). DisSent: Sentence Representation Learning from Explicit Discourse Relations. pages 4497–4510, July.

Oraby, S., Harrison, V., Reed, L., Hernandez, E., Riloff, E., and Walker, M. (2016). Creating and characterizing a diverse corpus of sarcasm in dialogue. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 31–41. Association for Computational Linguistics.

Park, J. and Cardie, C. (2014). Identifying appropriate support for propositions in online user comments. In *Proceedings of the first workshop on argumentation mining*, pages 29–38.

Pennington, J., Socher, R., and Manning, C. D. (2014). GloVe: Global Vectors for Word Representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 1532–1543.

Peters, M., Neumann, M., Iyyer, M., Gardner, M.,

Clark, C., Lee, K., and Zettlemoyer, L. (2018). Deep contextualized word representations. 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 2227–2237, New Orleans, Louisiana, June. Association for Computational Linguistics.

Pfeffer, J. (1981). Understanding the role of power in decision making. *Jay M. Shafritz y J. Steven Ott, Classics of Organization Theory, Wadsworth*, pages 137–154.

Phang, J., Févry, T., and Bowman, S. R. (2018). Sentence encoders on stilts: Supplementary training on intermediate labeled-data tasks. *CoRR*, abs/1811.01088.

Pitler, E., Raghupathy, M., Mehta, H., Nenkova, A., Lee, A., and Joshi, A. (2008). Easily identifiable discourse relations. In *Coling 2008: Companion volume: Posters*, pages 87–90. Coling 2008 Organizing Committee.

Poliak, A., Haldar, A., Rudinger, R., Hu, J. E., Pavlick, E., White, A. S., and Van Durme, B. (2018). Collecting diverse natural language inference problems for sentence representation evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 67–81.

Prasad, R., Dinesh, N., Lee, A., Miltsakaki, E., Robaldo, L., Joshi, A., and Webber, B. (2008). The penn discourse treebank 2.0. In Nicoletta Calzolari, et al., editors, *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, may. European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2008/.

Prasad, R., Riley, K. F., and Lee, A. (2014). Towards Full Text Shallow Discourse Relation Annotation : Experiments with Cross-Paragraph Implicit Relations in the PDTB. (2009).

Ribeiro, E., Ribeiro, R., and de Matos, D. M. (2015). The influence of context on dialogue act recognition. *arXiv preprint arXiv:1506.00839*.

Searle, J. R., Kiefer, F., Bierwisch, M., et al. (1980). *Speech act theory and pragmatics*, volume 10. Springer.

Shriberg, E., Dhillon, R., Bhagat, S., Ang, J., and Carvey, H. (2004). The icsi meeting recorder dialog act (mrda) corpus. In *Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue at HLT-NAACL 2004*.

Sileo, D., Van De Cruys, T., Pradel, C., and Muller, P. (2019a). Composition of sentence embeddings: Lessons from statistical relational learning. In *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM 2019)*, pages 33–43, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Sileo, D., Van de Cruys, T., Pradel, C., and Muller, P. (2019b). Mining discourse markers for unsupervised sentence representation learning. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Association for Computational Linguistics.

Sileo, D., Van de Cruys, T., Pradel, C., and Muller, P. (2020). DiscSense: Automated semantic analysis of discourse markers. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 991–999, Marseille, France, May. European Language Resources Association.

Socher, R., Chen, D., Manning, C., Chen, D., and Ng, A. (2013). Reasoning With Neural Tensor Networks for Knowledge Base Completion. In *Neural Information Processing Systems (2003)*, pages 926–934.

Subramanian, S., Trischler, A., Bengio, Y., and Pal, C. J. (2018). Learning general purpose distributed sentence representations via large scale multi-task learning. *International Conference on Learning Representations*.

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. (2018). GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium, November. Association for Computational Linguistics.

Wang, A., Hula, J., Xia, P., Pappagari, R., McCoy, R. T., Patel, R., Kim, N., Tenney, I., Huang, Y., Yu, K., Jin, S., Chen, B., Durme, B. V., Grave, E., Pavlick, E., and Bowman, S. R. (2019a). Can you tell me how to get past sesame street? sentence-level pretraining beyond language modeling. In *ACL 2019*.

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. R. (2019b). GLUE: A multi-task benchmark and analysis platform for natural

language understanding. In *International Conference on Learning Representations*.

Wang, A., Tenney, I. F., Pruksachatkun, Y., Yu, K., Hula, J., Xia, P., Pappagari, R., Jin, S., McCoy, R. T., Patel, R., Huang, Y., Phang, J., Grave, E., Liu, H., Kim, N., Htut, P. M., F'evry, T., Chen, B., Nangia, N., Mohananey, A., Kann, K., Bordia, S., Patry, N., Benton, D., Pavlick, E., and Bowman, S. R. (2019c). `jiant` 1.2: A software toolkit for research on general-purpose text understanding models. `http://jiant.info/`.

Warstadt, A., Singh, A., and Bowman, S. R. (2018). Neural network acceptability judgments. *arXiv preprint arXiv:1805.12471*.

Wieting, J., Bansal, M., Gimpel, K., and Livescu, K. (2015). Towards universal paraphrastic sentence embeddings. *CoRR*, abs/1511.08198.

Williams, A., Nangia, N., and Bowman, S. (2018). A broad-coverage challenge corpus for sentence understanding through inference. 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 1112–1122. Association for Computational Linguistics.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., and Le, Q. V. (2019). Xlnet: Generalized autoregressive pretraining for language understanding.

Zeldes, A. (2017). The GUM corpus: Creating multilayer resources in the classroom. *Language Resources and Evaluation*, 51(3):581–612.

Appendix A

	BERT	B+MNLI	B+DisSent	B+Discovery	B+PragmEval	Support
GUM.no_relation	48.9	**51.0**	46.0	45.4	43.3	48
GUM.circumstance	77.1	**80.6**	73.2	77.8	74.6	35
GUM.elaboration	41.5	38.5	40.0	**46.1**	42.9	32
STAC.no_relation	59.9	**63.8**	55.4	61.3	46.9	117
STAC.Comment	77.8	76.1	74.9	**78.6**	54.4	115
STAC.Question_answer_pair	79.1	80.1	**83.3**	76.9	83.0	93
SwitchBoard.Uninterpretable	86.0	86.0	85.5	86.1	**86.3**	382
SwitchBoard.Statement-non-opinion	72.0	72.1	**72.4**	**72.4**	**72.4**	304
SwitchBoard.Yes-No-Question	**85.9**	85.2	85.5	**85.9**	85.8	303
PDTB.Cause	55.2	55.7	53.1	**57.2**	55.9	302
PDTB.Restatement	40.4	40.0	41.3	**43.9**	41.0	263
PDTB.Conjunction	52.8	**53.9**	52.1	53.3	52.5	262
MRDA.Statement	51.2	51.8	48.9	**53.4**	51.4	364
MRDA.Defending/Explanation	52.8	54.1	**55.3**	52.8	52.0	166
MRDA.Expansions of y/n Answers	**51.7**	48.7	50.3	49.6	49.4	139

Table 5: Transfer F1 scores across the categories of PragmEval tasks; B(ERT)+$\mathcal{X}$ denotes BERT pretrained classification model after auxiliary finetuning phase on task $\mathcal{X}$.