Pragmatic Knowledge in Transformer Models: The case of T5

Anonymous CogSci submission

Abstract

Here we offer a key test of this assumption that the acquisition of pragmatic knowledge requires social interactions. Recent developments in natural language processing have led to statistical learning systems capable of acquiring deep knowledge of syntax and semantics. Still to be determined is whether these systems can acquire knowledge of pragmatics. If such systems are capable of learning pragmatic rules, it would imply that the acquisition of pragmatics need not depend on social interactions given that such systems learn language in isolation. We examine the transformer network T5 with respect to three kinds of pragmatic phenomena: presupposition, manner implicatures, and scalar implicatures. The results suggest that transformer language models have significant amounts of pragmatic knowledge and that the larger the transformer, the more pragmatic knowledge the they are able to acquire. The results raises fruitful new questions with regards to the acquisition of pragmatics and the role of (non) linguistics context.

Keywords: pragmatics; transformers; T5; language models; scalar implicature; manner implicature; presupposition

Introduction

Linguistics is not merely a function of the rules grammar and the meaning we stored with words. Linguistic meaning must be understood as dependent on the surrounding context. Context refers to not just the words neighboring a sentence, but also information about what the listeners know and do not know, the referent, and other people in the environment (Fox 2007; Chierchia et al 2012). Among all these factors, linguistic meaning depends on mental models of what the speaker thinks the listener knows.

Given it's a key factor in communication, social interaction would appear to be necessary for the acquisition of rules of pragmatics (Watzlawick 1967). This potential assumption is implied in theories such as the Rational Speech Act framework (Franke 2009, 2011; Frank & Goodman 2012, Bergen et al 2016), which holds that knowledge of how to interpret an utterance is intrinsically embedded in social interactions.

Thus, we ask the question, is social interaction necessary for the acquisition of pragmatic knowledge? Recent developments in statistical models of language understanding from Natural Language Processing (NLP) now offer us a way of addressing this question (Devlin et al., 2019; Liu et al, 2019; Tenney et al., 2019). Such models have been shown to acquire significant amounts of knowledge about syntax and meaning. They also appear to have limited knowledge of word knowledge. Still unknown is whether these models have knowledge of pragmatics. If they are sensitive to pragmatic rules, this would have implications for our understanding of the acquisition of pragmatic knowledge, the key reason being that they learn language without social interactions.

Transformers

A transformer architecture is a type of statistical learning system that augments feed-forward networks with self-attention mechanisms that allow for the specification of dependencies between words both within and between sentences. In this research, we used the recently introduced Text-To-Text Transfer Transformer (T5) (Raffel, et al. 2020). T5 represents an update to the transformer model first proposed in Vaswani et al. (2017). Transformer models have two main parts: a sequence of encoders and decoders. Encoders, in effect, comprehend a section of text. The set of words is not limited to a sentence. Encoders can, in fact, read in entire paragraphs or longer. This ability to read in large sections of text makes transformer models potentially well suited for learning and applying pragmatic rules, because the context can span over multiple sentences. The main idea behind a decoder is that it is the part of the language model that "generates" text. When transformers are used in translation, the encoder reads in the first language, while the decoder translates it into a second language. T5 extends the basic idea of mapping a sequence of word onto another sequence of words to other language tasks, such as grammatical judgments, question answering, text summarization, and causal reasoning. The basic anatomy of a transformer is shown in Figure 1.

The first step in a transformer is to convert a set of words into word embeddings, that is strings of numbers that through training come to represent the meaning of a word. All of the word embeddings in a chunk of text are submitted simultaneously to the first encoder. Each encoder consists of two main parts: a self-attention mechanism followed by a feed-forward neural network. The self-attention mechanism is trained to form links between words within a sentence, such as the verb and its arguments or a preposition and its object. However, the self-attention mechanism can also learn to form connections between words and phrases across sentences, offering a potential mechanism for application of pragmatic information. Each encoder possesses, in fact, several self-

attention mechanisms, called "heads." Having multiple self-attention mechanisms allows the network to capture a wide range of dependencies across the words in a chunk of text.

Transformer networks typically link encoders into a stack of encoders. The stack allows for the word embeddings to be adjusted so that they best reflect the context of the surrounding words.

The encoders map an input sequence into an abstract continuous representation. The decoders then take that continuous representation and generate words in a step-by-step manner, using the output from the previous step as input on the current step. In addition, the decoders' output is constrained by attentional vectors formed from the output of the top encoder. The inclusion of both encoders and decoders is one of the ways T5 differs from several other recently transformer-based models, such as BERT (Devlin et al., 2018) and RoBERTa (Liu et al., 2019), which include only the encoder part of the transformer.

Table 1 Model size variants

Model	Parameters	# layers	# heads
T5-Small	60M	6	8
T5-Base	220M	12	12
RoBERTa-large	355M	24	12
T5-Large	770M	24	16
T5-3B	3B	24	32
T5-11B	11B	24	128

The T5 model comes in several sizes, as specified in Table 1, along with and RoBERTa-Large for comparison. T5-small is significantly smaller than RoBERTa-Large. T5-Large is approximately the same size as these two other networks. T5-11B is quite large, containing 11 billion parameters and requires approximately 40GB of memory on a GPU. The model can also be run on a CPU on a system having 120GB of RAM.

A second major innovation of T5 is the manner in which it is trained. In BERT and RoBERTa, the target for an individual mask is associated with a single word piece. In T5, the target for a mask can be several words. For example, given the original sentence is *An elephant is larger than a goat*, T5 might be presented with the string *An elephant is <X> a goat*, with the target being several words, namely *<X> larger than <Y>*. Multiword targets are made possible through the use of the stack of decoders.

All versions of T5 were trained on a cleaned version of the common crawl called the Colossal Cleaned Common Crawl (C4). The training side is over two times larger than Wikipedia. The largest version of the model, T5-11B, achieved state-of-the-art performance results on the GLUE,

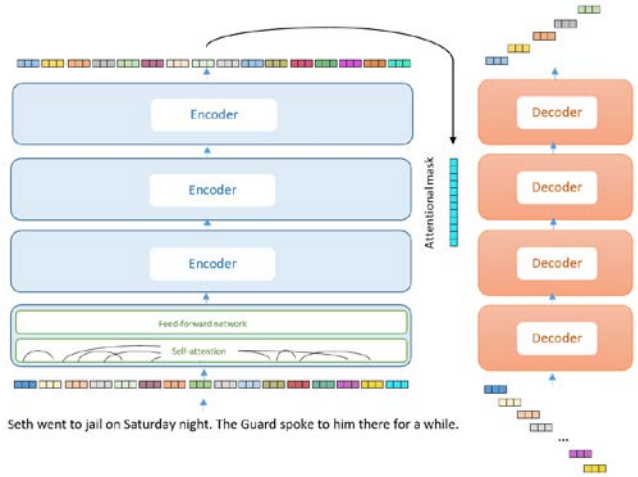

Figure 1: The basic anatomy of a T5 transformer

SuperGLUE, SQUAD, and benchmarks, which involve natural language processing tasks such as sentiment analysis, question answering, grammaticality judgments, paraphrase detection, selection of plausible causes and results, textual entailment detection, intended meaning detection, and reading comprehension with commonsense reasoning (Raffel, et al. 2020).

T5 represents a remarkable new class of language learning models that not only acquire a sophisticated understanding of language, such as syntax (Ganesh, Sagot & Saddah, 2019; Goldberg, 2019; Hewitt & Manning, 2019; Peters et al., 2018; Tenney, et al. 2019), but also seem to acquire general knowledge about the world (Petroni et al., 2019; Da & Kasai, 2019). The ability to capture at least certain aspects of world knowledge might suggest that T5 may be able to learn certain types of pragmatic rules. We investigated this possibility in a series of studies examining different kinds of pragmatic phenomena. It is certainly possible that T5 could show limited understanding of pragmatics. Under these circumstances, it would be good to know if the limit is a fundamental property of these statistical systems or else, possible, simply a function of the size of the system. To address this question, T5's knowledge was investigated for different model sizes. If the knowledge increases with model size, it would suggest that any limits in its knowledge might be a simple matter of the model's capacity rather than an inherent limitation of the architecture.

Study 1: Presupposition

In Study 1 we investigated whether T5 had knowledge of presupposition. Presupposition is the assumption of specific knowledge about a context. If the assumption is not warranted, the sentence will not sound normal or felicitous, even though it may be perfectly grammatical.. The phenomena of presupposition is shown in (1).

(1) a. Seth went to jail/ # a restaurant on Saturday night. The Guard spoke to him there for a while.
b. Kristen went to a restaurant/ # jail in the morning. The Waiter served her there quickly.

Notice that in (1a), it would be perfectly fine to say *Seth went to jail on Saturday* or *Seth went to a restaurant on Saturday*. However, only the first sentence is naturally followed by the sentence *The guard spoke to him there for a while*. Being at a jail implies one will likely meet guards. Going to a restaurant does not. The sentences in (1b) shows that the problem is not about going to restaurants, because it sounds perfectly fine to state that *The Waiter served her there quickly* after a sentences about going to a restaurant.

In this study, we investigated whether T5 and RoBERTa-Large would be sensitive to violations of presupposition. To examine this phenomenon, we used materials from Singh et al. (2016).

Methods

Materials The materials consisted of 41 minimal sentence pairs, as exemplified in (1). One member of the pair resulted in violations of presupposition, while the other did not. **Procedure** The sentences were presented to T5, and the cross-entropy loss was recorded. It was predicted that loss scores would be smaller for felicitous sentences than for sentences involving violations of presupposition. The judgments of RoBERTa were measured by looking at the relative rank of the target word. In (1), the target words would be *jail* and *restaurant*. Correct understanding was coded with a 1 and incorrect understanding with a 0. In this and following analyses, we used the HuggingFace implementation of T5 and RoBERTa (Wolf, et al., 2020).

Results and Discussion

The results indicate that T5 has pragmatic knowledge of the presupposition. The overall mean accuracies for the different versions of T5 are shown in Figure 2. All the models had mean accuracies that differed from chance (.5), as indicated by a binomial test, $p < .0001$. The mean accuracy T5-11b was the highest, and steadily decreased with smaller sized transformers. However, even for T5-small, accuracy was still relatively high at 80%. The results provide the first demonstration that transformer models are capable of encoding pragmatic information.

Study 2: Manner Implicature

In Study 2, we investigated whether transformer models have knowledge of manner implicature. Manner implicature is the phenomenon in which the subject of a comparison has the attribute of the object of the comparison. The inference is presumably drawn because the expression is more complex than is necessary for expressing the main idea. For example, in (2), a cake is compared to pudding, and described as as delicious as pudding. Manner implicature is the evaluative inference that the cake was delicious, not just the pudding.

(2) The cake was as delicious as the pudding.
manner implicature: the cake was delicious.

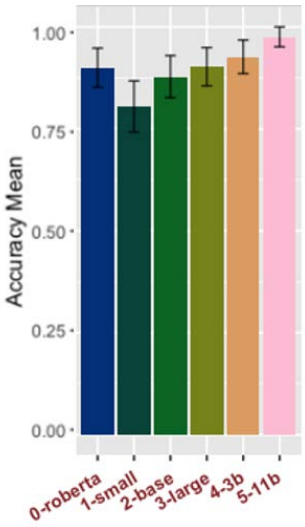

Figure 2: Mean accuracy levels for identifying violations of presupposition in different variants of transformers

In order to test this implicature, sentences can be constructed that either affirm or contradict the inference. For example, the sentence in (3a) affirms the inference that the cake was delicious, whereas (3b) denies the inference.

(3) a. The cake was as delicious as the pudding, which is to say the cake was the best.
b. # The cake was as delicious as the pudding, which is to say the cake was the worst.

Minimal pairs like the ones shown in (3) can be used to test whether transformer models have knowledge of manner implicature. In the case T5, the cross-entropy loss should be less for a felicitous sentence, like the one in (3a), than for one in (3b).

The phenomena of manner implicature appears to be associated with a sub-phenomenon regarding the nature of the adjective. Manner implicature sounds natural when the adjectives describe a subjective rather than an objective quantity. For example, the sentence in (2) uses an adjective that is subjective or evaluative, and the sentence sounds natural. However, the instance in (3) uses an adjective that is objective, the result is less natural.

(4) The cake was as heavy as the pudding.

The statement in (3) may not imply that the cake is heavy.

Methods

Materials The materials were drawn from a set of sentences developed by Rett (2014; 2019). They included 48 sentence items in total: 24 pragmatically good items, which consist of 12 sentences containing subjective adjectives (i.e., predicates of personal taste) and 12 objective adjectives (i.e.,

predicates that concern matter of fact), corresponding to 24 pragmatically incongruent sentences.

Procedure The procedure mirrored the one used in Study 1. Sentences were presented to T5 and RoBERTa and the cross-entropy loss was recorded. It was predicted that loss scores would be smaller sentences in which the inference was affirmed than for sentences in which the inference was contradicted. Correct understanding was coded with a 1 and incorrect understanding with a 0. In this and following analyses, we used the HuggingFace implementation of T5 and RoBERTa (Wolf, et al., 2020).

Results and Discussion

The results provided further evidence that T5 is capable of generating pragmatic inferences. The overall mean accuracies for the different versions of T5 are shown in Figure 3. All the models had mean accuracies that differed from chance (.5), as indicated by a binomial test, $p < .0001$. As in Study 1, the highest accuracy was obtained by T5-11b was the highest.

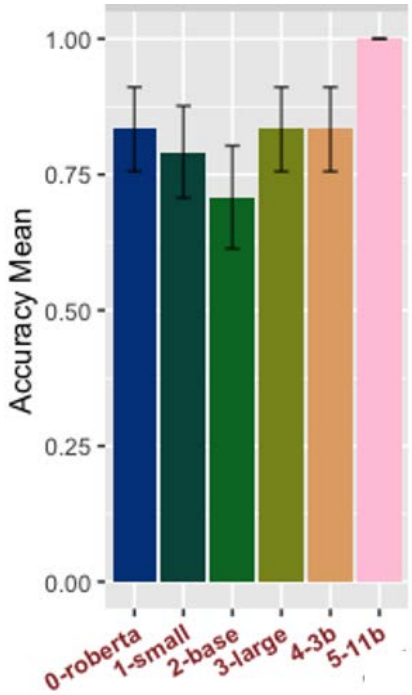

Figure 3: Mean accuracy levels for identifying Manner Implicature in different variants of transformers

Study 3: Scalar Implicature

In Study 3, we investigated utterances containing quantifier "some" and the Scalar Implicature it gives rise to, as shown in (5). Scalar implicatures is the phenomena in which *some* implies *not all*. In certain contexts, this implicature can give rise to a contradiction, as exemplified in (5a). The sentence in (5a) sounds bad because *some* implies *not all*, but all office buildings have desks. The sentence in (5b) sounds acceptable because not all office buildings have plants.

(5) a. # Some office buildings have desks and can become dusty. [SI noticed; bad]
b. Some office buildings have plants and can become dusty. [SI noticed; good]

Scalar implicature can be tested by presenting transformer models with sentence pairs like those in (5) and recording the cross-entropy loss.

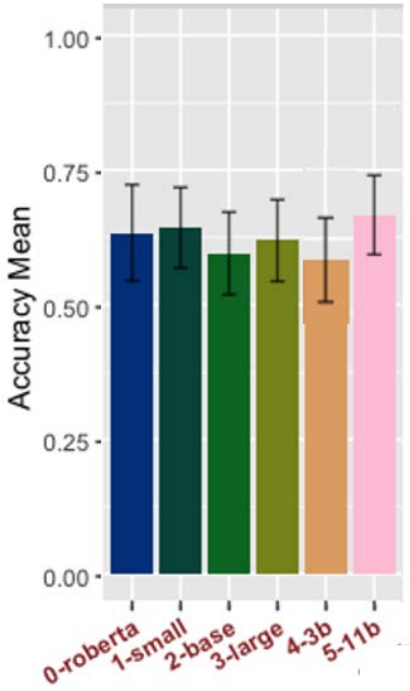

Figure 4: Mean accuracy levels for identifying Scalar Implicature in different variants of transformers

Methods

Materials The materials were drawn from a set of sentences developed by Nieuwland et al. (2010). They contained 42 pairs of sentences, with one member of each pair resulting in a contradiction.

Procedure The procedure was the same as the one used in Studies 1 and 2. Sentences were presented to T5 and RoBERTa and the cross-entropy loss was recorded.

Results and Discussion

The results indicated that provided further evidence that T5 is capable of generating pragmatic inferences. The overall mean accuracies for the different versions of T5 are shown in

Figure 3. Mean accuracies were not as high for scalar implicatures as they were for the other pragmatic phenomenon. Indeed, an analysis of the raw loss scores indicated that only T5-11b had a meaning accuracy that differed significantly from chance (.5), $p = .044$.

General Discussion

This paper provides a linguistic diagnosis of neural language models' understanding of pragmatically enriched meaning. In particular, we focus on RoBERTa-large and T5 models' accuracy as well as sensitivity to meanings that are not-at-issue, including Presupposition, Manner Implicatures, and Scalar Implicatures. We find that as the complexity of the models increases, their accuracy and sensitivity in handling nuanced meanings are getting better. Among all six models we examined, T5-11b repeatedly shows that it's capable of drawing inferences that are subtly implied or presupposed. This raises fruitful new questions. It appears that neural language models are not hard-wired with the iterative rationality models types of recursion, and they are not formally trained with the concept of "interlocutor". Despite all that setup, neural language models still understand conversational implicatures.

On the one hand, this sheds light on why the languages, particularly the pragmatic component of languages, are the way they are. Are humans cognitively hard-wired and the cognitive constraints shape the languages? Or is that more of a learning efficiency issue, in which languages are the way they are because that's most efficient for learning? Our findings provide evidence pointing towards the learning efficiency theory. On the other hand, the questions of exactly how neural language models achieved that goal of successfully drawing the desirable inferences are intriguing, and we leave it for future research.

Transformers can provide a new way to think about pragmatics. Linguists have made some observations about the meaning of sentences. The claim is that the meaning has both an entailed and implied component to it. One may wonder whether untrained, everyday language users recognize these inferences. If so, then these would be inferences that we would like a model of language to be able to draw. Recent pragmatic reasoning models such as the Rational Speech Act framework (Bergen et al 2016) give an initial attempt to simulate the derivation of pragmatically enriched meaning. Instead of explicitly modeling conversational participants as have been advocated by the rationality models, the transformer models achieve the same goal through a brand new setup. With three independent studies, we have shown that neural language models' performance increases as the number of parameters increases. Put otherwise, without learning Grice maxims, neural language models can still hit the target inference.

As an implication, we argue that pragmatic reasoning schemas are acquired implicitly and may not require human interaction, although the limitations of these models might point to the need for language samples that involve discourse. This will be examined in future research. We propose that Grice maxims might be a side-effect. Similarly for syntax, word meaning, reasoning, world knowledge, and pragmatics, which are not goals of studying language systems, instead they might turn out to be the side-effects of the goal. The goal is ambiguity resolution or predictive coding. The linguistic system is not trying to learn major problems like "syntax" or "pragmatics"; it's trying to simply predict what wasn't heard well because of the loud noise in the background. On top of that, the system discovers that if it learns syntax and pragmatics, it can do a better job of resolving the ambiguity or guessing the occluded words.

The way linguistics has couched things concerned how the child learns syntax. How does the child learn pragmatic inference? We maintain that these are not the problems the system is trying to solve. The system is focused on simpler issues, for example, simply "what was that you just said?" For future work, it would be interesting to know what layers in the neural network solve these problems. The findings could be used to prepare the network to acquire this knowledge with less data.

References

Bergen, L., Roger, L., & Goodman, Noah D. (2016). Pragmatic reasoning through semantic inference. *Semantics and Pragmatics*, 9(20):1-83.

Bumford, D., & Rett, J. (2020). Rationalizing evaluativity. In *Proceedings of the 25th Sinn und Bedeutung*.

Bylinina, L. (2017). Judge-dependence in degree constructions. *Journal of Semantics*, 34:291-331.

Devlin, J., Chang, M.W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language under- standing. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1:4171-4186.

Ettinger, A. (2020). What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics*, 8:34-48.

Frank, Michael C. & Goodman, Noah D. (2012). Predicting pragmatics reasoning in language games. *American Association for the Advancement of Science*, 336(6084):998.

Franke, M. (2011). Quantity implicatures, exhaustive interpretation, and rational conversation. *Semantics and Pragmatics*, 4(1):1-82.

Geurts, B. (2010). *Quantity Implicatures*. Cambridge University Press, Cambridge.

Horn, Laurence R. (1972). *On the Semantics of Logical Operators in English*. Ph.D. thesis, Yale University.

Jeretic, P., Warstadt, A., Bhooshan, S., & Williams, A. (2020). Are natural language inference models IMPPRESsive? Learning IMPlicature and PRESupposition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 8690-8705.

Nieuwland, Mante S., Ditman, T., & Kuperberg, Gina R. (2010). On the incrementality of pragmatic processing: An ERP investigation of informativeness and pragmatic abilities. *Journal of Memory and Language*, 63:324-346.

Rett, J. (2014a). Measure phrase equatives and modified numerals. *Journal of Semantics*, 32(10):425-475.

Rett, J. (2015). *The Semantics of Evaluativity*. Oxford: Oxford University Press.

Rett, J. (2019). Manner implicatures and how to spot them. *International Review of Pragmatics*, in press.

Sauerland, U. (2004). Scalar implicatures in complex sentences. *Linguistics and Philosophy*, 27:367-391.

Schuster, S., Chen, Y., & Degen, J. (2020). Harnessing the linguistic signal to predict scalar inferences. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5387-5403.

Singh, R., Fedorenko, E., Mahowald, K., & Gibson, E. (2016). Accommodating presuppositions is inappropriate in implausible contexts. *Cognitive Science*, 40:607-634.

Tenney, I., Das, D., & Pavlick, E. (2019). BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 4593-4601.

von Fintel, K. (2004). *Would you believe it? The king of France is back! Presuppositions and truth-value intuitions.* Oxford University Press.

Watzlawick, P., Bavelas, J. B., & Jackson, D. D. 1967. *Pragmatics of Human Communication: A Study of Interactional Patterns, Pathologies, and Paradoxes*. New York: W.W. Norton and Company.